TAMAÑO DE EFECTO, POTENCIA DE LA PRUEBA, FACTOR DE BAYES Y META-ANÁLISIS EN EL MARCO DE LA CRISIS DE REPRODUCIBILIDAD DE LA CIENCIA. EL CASO DE LAS DIFERENCIAS DE PROPORCIONES Y TABLAS DE CONTINGENCIA CON VARIABLES NOMINALES Y MUESTRAS INDEPENDIENTES

(segunda parte)

Luis D' Angelo
Instituto Nacional de Parasitología 'Dr. Mario Fatala Chaben' (INP) - Administración Nacional de Laboratorios e Institutos de Salud "Dr. Carlos G. Malbrán"
Avenida Paseo Colón 568, CABA, Argentina (CP 1097)
luis11dangelo@gmail.com
https://orcid.org/0009-0008-5617-0750

Recibido 9 de abril de 2024, aceptado 27 de mayo de 2024

RESUMEN
Este trabajo es la continuación del artículo sobre tamaño de efecto para diferencia de medias independientes, centrándose ahora en las diferencias de proporciones y tablas de contingencia con variables nominales y muestras independientes.
La investigación estadística sobre proporciones desempeña un papel fundamental en una amplia gama de ámbitos científicos y sociales al proporcionar una valiosa comprensión sobre la distribución y la incidencia de fenómenos en una población específica. Más allá de evaluar la significación estadística de las pruebas es crucial considerar la relevancia práctica o clínica de los resultados, entendiendo la importancia del tamaño de efecto.
En el caso de las proporciones, nos encontramos con una amplia variedad de estadísticos de tamaño de efecto disponibles, como el coeficiente *Phi*, *V* de Cramer, *W* de Cohen, *Fei* de Ben-Shachar, *h* de Cohen, riesgo relativo, *odds ratio*, entre otros. Esta diversidad puede generar desafíos al momento de seleccionar el estadístico más adecuado para analizar los datos. Por lo tanto, este trabajo se propone abordar esta problemática en detalle, explorando las características y aplicaciones de cada estadístico, y ofreciendo orientación sobre su selección según el contexto y los objetivos de la investigación.

Palabras Clave: Tamaño del efecto, Proporciones, Potencia de la prueba, Factor de Bayes, Meta-análisis.
Códigos JEL: C12, C13, C14.

EFFECT SIZE, POWER ANALYSIS, BAYES FACTOR AND META-ANALYSIS IN THE FRAMEWORK OF SCIENCE'S REPLICATION CRISIS. THE CASE OF PROPORTIONS AND CONTINGENCY TABLES WITH NOMINAL VARIABLES AND INDEPENDENT SAMPLES

(second part)

Luis D' Angelo
Instituto Nacional de Parasitología 'Dr. Mario Fatala Chaben' (INP) - Administración Nacional de Laboratorios e Institutos de Salud "Dr. Carlos G. Malbrán"
Avenida Paseo Colón 568, CABA, Argentina (CP 1097)
luis11dangelo@gmail.com
https://orcid.org/0009-0008-5617-0750

Received April 9th 2024, accepted May 27nd 2024

ABSTRACT
This paper is a continuation of the article on effect size for difference of independent means, focusing now on differences of proportions and contingency tables with nominal variables and independent samples.
Statistical research on proportions plays a fundamental role in a wide range of scientific and social fields by providing valuable insights into the distribution and incidence of phenomena in a specific population. Beyond assessing the statistical significance of the evidence, it is crucial to consider the practical or clinical relevance of the results, understanding the importance of effect size.
In the case of proportions, we find a wide variety of effect size coefficients available, such as the Phi coefficient, Cramer's V, Cohen's W, Ben-Shachar's Fei, Cohen's h, relative risk, odds ratio, among others. This diversity can generate challenges when selecting the most appropriate statistic to analyze the data. Therefore, this paper aims to address this issue in detail, exploring the characteristics and applications of each statistic, and offering guidance on their selection according to the context and objetives of the research.

Keywords: effect size, proportions, power analysis, Bayes factor, meta-analysis.
JEL Codes: C12, C13, C14.

1. INTRODUCCIÓN

Este trabajo es una continuación del artículo sobre el tamaño de efecto para diferencias de medias independientes (D'Angelo, 2021). Al igual que en dicho artículo, nos enfocaremos en aspectos que trascienden la prueba de hipótesis, pero esta vez centrándonos en los tamaños de efecto y la potencia de la prueba relativas a las *diferencias de proporciones,* así como en su aplicación en el análisis de las *tablas de contingencia* con variables nominales y muestras independientes.

La investigación estadística sobre proporciones desempeña un papel fundamental en una amplia gama de ámbitos científicos y sociales, proporcionando información valiosa sobre la distribución y la incidencia de fenómenos clave dentro de una población. Estas medidas son críticas ya que ofrecen una comprensión detallada de cómo se distribuyen y ocurren eventos específicos en el contexto de una población determinada.

Existen diversos ámbitos en los que el uso de proporciones en términos de estadísticos resulta esencial. En salud pública y epidemiología, las proporciones son utilizadas para calcular la prevalencia de enfermedades en una población, así como para determinar tasas de mortalidad o de incidencia de enfermedades específicas. Estas medidas son esenciales para la planificación de políticas de salud pública y la asignación de recursos. En investigación social, estudios sociológicos, criminológicos y de ciencias sociales en general, las proporciones se utilizan para analizar fenómenos como la distribución de ingresos, tasas de desempleo, niveles de educación, entre otros. Estas proporciones ayudan a comprender la estructura social y a identificar desigualdades o necesidades específicas de la población. En economía, las proporciones son importantes para calcular indicadores clave como la tasa de inflación, la tasa de crecimiento económico, la relación deuda-PIB, entre otros. Estas medidas son fundamentales para comprender la salud económica de un país y para el diseño de políticas económicas efectivas. En el ámbito educativo, se utilizan para calcular tasas de graduación, tasas de deserción escolar, de estudiantes por docente, entre otros. En el campo del marketing y la investigación de mercado, se utilizan para analizar datos de encuestas, como la proporción de clientes satisfechos, la proporción de mercado de una empresa en particular, entre otros.

Ahora bien, en todos estos ámbitos los resultados obtenidos deben muchas veces ser referidos a una población habiendo partido de una muestra. Por tal motivo cobra importancia comprender cómo y de qué modo los errores de muestreo estadístico afectan la confiabilidad de las estimaciones. Identificar y cuantificar adecuadamente estos errores proporciona una medida de confianza en la precisión de las conclusiones obtenidas.

Para lograr este objetivo, es fundamental contar con un profundo entendimiento de las pruebas de hipótesis relacionadas con las proporciones. Garantizar la validez, la confiabilidad y una interpretación adecuada de los resultados obtenidos en los estudios

es de suma importancia. Es frecuente que el valor *p* (o *p-valor*) sea malinterpretado, lo que puede conducir a errores graves en la interpretación de los resultados. Las confusiones al respecto son diversas y, en ocasiones, ocurren simultáneamente. Es esencial comprender que el valor de *p* representa la probabilidad de obtener el resultado observado -o más extremo- si la hipótesis nula es cierta. Entre las interpretaciones erróneas más comunes del valor *p* se encuentran lo que la literatura designa como las falacias de la probabilidad inversa, de la replicación, del tamaño del efecto y de la significación clínica o práctica. Estas interpretaciones erróneas pueden llevar a conclusiones incorrectas y a una comprensión inadecuada de la significancia de los resultados obtenidos en un estudio. Por lo tanto, es crucial abordar estas confusiones y garantizar una interpretación precisa de los valores *p* en el contexto de las pruebas de hipótesis.

Pero más allá de la significación estadística de las pruebas es necesario dar cuenta de la relevancia práctica o clínica de los resultados. Es decir, cuando obtenemos evidencia sobre la existencia de un determinado efecto, buscaremos sumarle información en relación al tamaño de ese efecto.

En el campo de las proporciones, a diferencia del de las diferencias de medias, nos enfrentamos a una amplia gama de estadísticos de tamaño de efecto disponibles, que incluyen al coeficiente *Phi*, *V* de Cramer, *W* de Cohen, *Fei* de Ben-Shachar, *h* de Cohen, riesgo relativo, *odds ratio*, entre otros. Esta diversidad de opciones puede plantear un desafío significativo al momento de decidir qué estadístico utilizar para analizar adecuadamente los datos. En este trabajo, nos proponemos abordar esta problemática en profundidad, explorando las características y aplicaciones de cada estadístico y proporcionando orientación sobre cuándo y cómo seleccionar el estadístico más apropiado según el contexto y los objetivos de la investigación.

Comprender adecuadamente las pruebas de hipótesis relacionadas con las proporciones es esencial para garantizar la validez y la fiabilidad de los resultados obtenidos a partir de estudios basados en muestras independientes. Creemos que existe un problema importante en la aplicación de la prueba de hipótesis en diversos ámbitos relacionados a la interpretación del error de tipo *I* y *II*, al problema de la reproducibilidad y a las cuestiones relativas al tamaño del efecto o su significación práctica.

Cuando diseñemos nuestra investigación, o cuando estemos en presencia de un trabajo del cual necesitamos evaluar sus resultados, deberíamos tener en cuenta la capacidad del abordaje para conocer cuál es la probabilidad de obtener un resultado no significativo cuando la hipótesis nula sea falsa. Es decir, que exista un efecto, que además tenga evidentemente algún tamaño, y que no pudimos captar debido al deficiente poder de la prueba que hemos aplicado.

En último término, la cuestión de la reproducibilidad requiere ser atendida, preferiblemente mediante la realización de meta-análisis cuando sea factible. Esto cobra

especial relevancia en casos donde existe evidencia contradictoria entre diversos estudios, ya sea en términos de la significancia de las pruebas de hipótesis o en lo que respecta a los tamaños de efecto obtenidos. En tales circunstancias, se sugiere llevar a cabo estudios que tengan la capacidad de integrar y reconciliar las diferentes investigaciones, proporcionando así un análisis más completo y una visión más clara sobre el fenómeno en cuestión.

El tamaño del efecto ofrece una respuesta complementaria a una serie de problemas derivados de la utilización del test de hipótesis. No debemos olvidar que estos problemas no son cuestiones a subestimar. Las críticas durante estos últimos años han ido al fondo de la cuestión y no deben ser soslayadas. La denominada crisis de reproducibilidad ha sido puesta de manifiesto, lo que resalta la necesidad urgente de que los investigadores adopten y apliquen estas técnicas. Este artículo busca sensibilizar a los investigadores sobre la importancia de emplear el tamaño de efecto y la evaluación de la potencia en las pruebas de hipótesis. Redoblar los esfuerzos en la aplicación cuidadosa de estas técnicas es fundamental para garantizar que los resultados obtenidos en los distintos campos científicos sean más rigurosos y, por lo tanto, más creíbles.

En este trabajo, vamos a explorar diversos métodos y conceptos fundamentales en el análisis estadístico de proporciones y su aplicación en estudios con muestras independientes. En primer lugar, abordaremos la prueba de hipótesis para diferencias de proporciones, esencial cuando queremos comparar dos grupos diferentes. Examinaremos este tema desde dos perspectivas: primero, la aproximación de la distribución binomial a la normal y, segundo, desde la prueba exacta de Fisher, también conocida como test de permutaciones. Además, discutiremos en qué casos utilizar cada una de estas técnicas.

Luego, abordaremos los intervalos de confianza para las proporciones, comenzando con el método "exacto" binomial (Clopper-Pearson), seguido por el método asintótico (Wald), que se basa en una aproximación normal. También cubriremos otros métodos como el de Wilson, Jeffreys y el intervalo Agresti-Coull (una versión ajustada del método Wald). Además, profundizaremos en los intervalos de confianza para la estimación de la diferencia de proporciones y exploraremos cómo interpretar estos intervalos de manera gráfica. También examinaremos el intervalo de confianza *bootstrap*, que es una técnica más flexible en cuanto a los supuestos de aplicación de las técnicas tradicionales.

Luego nos adentraremos particularmente en los tamaños del efecto para los casos de la diferencia de proporciones para muestras independientes, como complemento a las pruebas de hipótesis e intervalos de confianza. En este análisis vamos a explorar varios coeficientes, como el coeficiente *Phi* (φ), *V* de Cramer, *T* de Tschuprow, *W* de Cohen, y la *lambda* de Goodman y Kruskal (λ). Cada uno de estos tiene su propia aplicación específica y su forma particular de interpretación. Para concluir este tema, también

trataremos el riesgo relativo y abordaremos conceptos clave como los *odds* y el *odds ratio*, incorporando su interpretación en términos del tamaño del efecto. En este contexto, discutiremos los errores estándar e intervalos de confianza, proporcionando una comprensión de cuando aplicar estas medidas.

Para dar una perspectiva más completa, incluiremos una sección sobre el Factor de Bayes, que mide la creencia en la hipótesis nula y alternativa, y exploraremos cómo realizar un meta-análisis para sintetizar resultados de múltiples estudios.

Finalmente, cerraremos con una síntesis de los conceptos abordados y su relevancia práctica en el ámbito de la investigación científica, destacando su aplicabilidad en la toma de decisiones y el diseño de futuros estudios.

2. PRUEBA DE HIPÓTESIS PARA DIFERENCIA DE PROPORCIONES CON MUESTRAS INDEPENDIENTES

2.1 Aproximación de la binomial a la normal

La prueba de hipótesis para la diferencia de proporciones con muestras independientes cuenta con similares virtudes y defectos respecto del resto de las pruebas de hipótesis. Por un lado, permite evaluar si hay una diferencia significativa entre las proporciones de dos grupos proporcionando una medida objetiva para la toma de decisiones, siendo una herramienta estándar ampliamente utilizada en el ámbito científico. Por el otro, no proporciona información sobre la magnitud o relevancia práctica de la diferencia observada, dependiendo del tamaño de la muestra, y requiriendo en muchos casos ciertas suposiciones, como la aleatorización de las muestras y la independencia de las observaciones.

Para determinar si dos porcentajes resultan diferentes, a partir de dos muestras, el modo clásico de obtener el valor de z viene dado por:

$$z_{dif.prop} = \frac{p_1 - p_2}{ES} \qquad (1)$$

Donde p_1 es la proporción en la muestra 1, p_2 la correspondiente a la muestra 2 y ES el error estándar para la diferencia de proporciones.

Una forma de calcular el error estándar ES viene dado por:

$$ES_{no\ combinado\ (unpooled)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \qquad (2)\ \text{(Agresti, 2007)}$$

Esta fórmula del error estándar es la raíz cuadrada de la suma de varianzas y definida en la literatura como *no combinado* o *unpooled*.

Cabe aclarar que esta forma de calcular el error estándar no es, para algunos autores, totalmente correcta si se la utiliza para poner a prueba la hipótesis nula $p_1 = p_2$, ya que implica justamente que las dos proporciones serían diferentes, y justamente cada una de las distribuciones debería aportar su propia e independiente varianza (Fleiss 2003). Esta forma del error estándar es la que se debería utilizar en los casos que estemos interesados en calcular el intervalo de confianza de la diferencia de proporciones como

veremos más adelante. Por esta vía, se podría sostener la hipótesis nula por la cual $p_1 = p_2$ en el caso que el intervalo de confianza incluya el cero.

Sin embargo, actualmente, para determinar si dos proporciones son diferentes se tiende a utilizar el siguiente error estándar:

$$ES_{combinado(pooled)} = \sqrt{\frac{p(1-p)}{n_1}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad (3)$$

Donde $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$,

aplicándole una corrección por continuidad $\left(-\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$ de modo que el valor de z calculado de este modo queda del siguiente modo:

$$z_{dif.prop} = \frac{p_1 - p_2 - \frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}{\sqrt{\frac{p(1-p)}{n_1}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (4), \text{(Fleiss, Levin y Paik, 2003, pág. 54)}$$

Si lo que se desea es probar la hipótesis alternativa de la diferencia de proporciones, la hipótesis nula debe dar cuenta de la igualdad de las proporciones. Por este motivo la versión *combinada* (*pooled*) es la más correcta. En general no hay grandes diferencias entre los dos z calculados con estas dos fórmulas, ambas basadas en el supuesto de normalidad y por lo tanto en el teorema del límite central, como no combinados (*unpooled*) o combinados (*pooled*). Es más, los programas estadísticos tampoco se han puesto de acuerdo por lo que es posible hallar pequeñísimas diferencias en el cálculo de z[1]. Esta confusión no zanjada también es posible hallarla en la bibliografía (Fleiss 2003).

Téngase en cuenta que estamos haciendo una aproximación de la distribución binomial a la normal. Para que el teorema del límite central pueda ser aplicado se tienen que cumplir las siguientes condiciones:

$p_1 * n_1 > 5, \quad (1-p_1) * n_1 > 5, \quad p_2 * n_2 > 5, \quad (1-p_2) * n_1 > 5$

Es decir, en todos los casos se debe chequear este supuesto.

Cabe consignar que este valor de z se relaciona con el chi^2 a partir de la siguiente expresión (z-test for independent proportions: Use & misuse, 2019):

$$chi^2 = z^2 \quad (5)$$

En este sentido, calcular el chi^2 o z^2 es completamente equivalente.

2.2 Prueba exacta de Fisher (test de permutaciones o *permutation test*)

Muchas veces los supuestos para la aplicación de la aproximación binomial a la normal no se cumplen, o se encuentran en el límite. Frente a esa situación o cualquier duda sobre su aplicación, es posible la utilización de la versión exacta conocida como test de Fisher, que se basa en el concepto de permutaciones. La idea es evaluar todas las

[1] Por ejemplo, el SPSS utiliza el *ES* combinado (*pooled*, es decir la correcta para nosotros) mientras que Minitab y las librerías prop.test o z.test de R, utilizan *ES* no combinado (*unpooled* es decir promedia dispersiones de proporciones que considera distintas).

posibles permutaciones de los datos para determinar la probabilidad de obtener esa configuración o configuraciones aún más extremas bajo la hipótesis nula. Este test puede ser realizado a partir del desarrollo de la distribución hipergeométrica manteniendo los marginales fijos (Edgington y Onghena, 2007; Good, 2005; Lock y otros, 2013).

Por ejemplo, en el caso de una de una tabla de 2x2 con una configuración como la siguiente:

Tabla 1. Tabla de contingencia 2x2 genérica.

	Evento (por ej. enfermedad)	No evento (ej. no enfermedad)
Grupo 1(expuestos)	a	b
Grupo 2 (no expuestos)	c	d

Fuente: adaptada de Fleiss, Levin y Paik, (2003, pág. 51).

podemos calcular la probabilidad de que las frecuencias provengan de una distribución de frecuencias independientes de acuerdo a la siguiente ecuación (Fleiss, Levin, & Paik, 2003, pág. 56):

$$p=\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \qquad (6)$$

Este test antes del uso de los programas informáticos era prácticamente inutilizable debido a la complejidad del cálculo. Entonces los métodos aproximados eran realmente necesarios. Hoy esa dificultad ha sido sorteada por lo cual en las tablas de 2x2 este test es posible utilizarlo siempre, y según el software en muchos casos en tablas mxn.

El test de Fisher es especialmente útil cuando se trabaja con muestras pequeñas o cuando las proporciones son muy extremas -es decir en términos prácticos cuando no se cumple $p_1 * n_1 > 5, \quad (1-p_1) * n_1 > 5, \quad p_2 * n_2 > 5, \quad (1-p_2) * n_1 > 5$-.

Esto es debido a que este test proporciona una p exacta y no depende de supuestos sobre la distribución de los datos.

3. INTERVALOS DE CONFIANZA PARA LAS PROPORCIONES

3.1 Método "exacto" binomial (Clopper-Pearson)

El *intervalo de confianza exacto de Clopper-Pearson* para una proporción p se basa en la inversión de una prueba binomial a dos colas iguales de la hipótesis H_0: $p=p_0$ (Agresti & Coull, 1998). Los límites inferior y superior son las soluciones en p de las ecuaciones:

$$\sum_{k=x}^{n} C_k^n \, p^k q^{n-k} = \frac{\alpha}{2} \qquad (7)$$

y

$$\sum_{k=0}^{x} C_k^c \, p^k q^{n-k} = \frac{\alpha}{2} \qquad (8)$$

Este intervalo suele ser considerado típicamente el *gold standard,* sin embargo, tiende a resultar muy conservador (Agresti & Coull, 1998) ya que se trata de una distribución

binomial, es decir, discreta. De modo que a menos que *n* sea muy grande este intervalo suele ser considerado inadecuado.

3.2 Método asintótico (Wald) basado en una aproximación normal

Cuando *n* es grande, en el sentido de que *n*p>5* y *n*q>5*, los procedimientos basados en una aproximación a la distribución normal proporcionan relativamente buenas aproximaciones en relación a los procedimientos binomiales exactos correspondientes. Basados en el teorema del límite central, los límites del intervalo de confianza para *p* pueden ser calculados como (Fleiss, Levin y Paik, 2003; Newcombre, 1998; Agresti y Coull, 1998):

$$IC_p = \mathrm{p} \pm z_c \sqrt{\frac{p(1-p)}{n}} \qquad (9)$$

Sin embargo, el intervalo de Wald a menudo tiene una cobertura inadecuada, particularmente para *n* pequeños y valores de *p* cercanos a 0 o 1. Por el contrario, el método exacto de Clopper-Pearson, como ya hemos señalado, es muy conservador y tiende a producir intervalos más amplios de lo necesario.

3.3 Método de "Wilson"

Brown y colaboradores (2001) recomiendan los métodos Wilson o Jeffreys para pequeñas *n*, y Agresti-Coull, Wilson o Jeffreys, para *n* más grandes, ya que proporcionan una cobertura más confiable que las alternativas. El método de Wilson, dentro de la lógica frecuentista, parece ser el que presenta mejores propiedades, tanto para *n* pequeños como para los grandes. Es que el comportamiento del intervalo de Agresti-Coull es ligeramente superior en el caso de n grande (*n*>40), pero como en estos casos todos los métodos de cálculo tienden a parecerse, entonces la ventaja se desvanece en contraste con la posibilidad de usar siempre el mismo intervalo.

El intervalo de Wilson sin corrección de continuidad viene dado por:

$$2np + z^2 \pm z\sqrt{\frac{(z^2+4npq)}{2(n+z^2)}} \qquad (10), \text{ (Newcombre, 1998).}$$

Y en el caso del intervalo con corrección de continuidad (también conocido como intervalo de score de Fleiss):

$$Inferior = \frac{2np+z^2-1-z\sqrt{z^2-2-\frac{1}{n}+4p(nq+1)}}{2(n+z^2)} \quad (11)$$

$$Superior = \frac{2np+z^2+1+z\sqrt{z^2+2-\frac{1}{n}+4p(nq+1)}}{2(n+z^2)} \quad (12), \text{ (Newcombre, 1998).}$$

3.4 Método de "Jeffreys"

También existe una versión del intervalo desde una perspectiva bayesiana, denominado intervalo de credibilidad de Jeffreys. Se trata de un abordaje bayesiano con distribución a priori no informativa, con buenas propiedades frecuentistas (Brown, Cai y Dasgupta, 2001).

Su cálculo depende de la obtención de los cuantiles de la función Beta (generalmente 0,025 y 0,975 para una credibilidad de 95%) con parámetros (0,5;0,5) y puede ser calculada a partir de:

$$\text{Límite inferior} = \frac{x+0{,}5}{n+1+(n-x+0{,}5)(e^{2\omega}-1)} \qquad (13)$$

donde

$$\omega = \frac{k\sqrt{\frac{4pq}{n}+\frac{(k^2-3)}{6n^2}}}{4pq} + \frac{(0{,}5-p)\left(pq(k^2+2)\right)-\frac{1}{n}}{6n(pq^2)} \qquad (14)$$

El límite superior puede ser aproximado por la misma expresión reemplazando k por $-k$ en ω (Brown, Cai y Dasgupta, 2001).

3.5 Intervalo "Agresti-Coull" (Wald ajustado)

También tenga en cuenta que existe una estimación puntual denominada de Agresti-Coull (1998) que presenta una amplitud ligeramente mayor que los otros métodos debido a la forma en que se calcula este intervalo.

Junto con la mayor parte de los autores que se han ocupado del tema no parece haber lugar para la duda. Los métodos de score de Wilson y el de Jeffreys son los de elección para la obtención de los intervalos de confianza. Para obtener los valores concretamente es necesario servirse de software estadístico específico o calculadores en la red[2].

3.6 Intervalos de confianza para la estimación de la diferencia de proporciones (Interpretación gráfica)

Uno de los problemas que suele aparecer en la interpretación de los intervalos de confianza para dos proporciones (también para otros estadísticos) es el de considerar que para que haya diferencia estadísticamente significativa no debe haber solapamiento entre los intervalos de confianza. Esto representa un grave error ya que un solapamiento de ¼ de los intervalos resulta en una significación del 0,05 y cuando los intervalos justo se tocan (solapamiento = 0) la significación aproximada es de 0,01 (Cumming y Finch, 2005).

Aquí vamos a ejemplificar con una tabla de contingencia de fantasía y analizada con SPSS (que genera intervalos de confianza de Jeffreys) en cuyo gráfico encontramos un solapamiento de ¼ aproximadamente con una significación de 0,047 (tanto para *chi²* como para la prueba exacta de Fisher):

Tabla 2. Tabla de contingencia 2x2 con significación al 0,05 y solapamiento de los intervalos de confianza.

	Control	Tratamiento	Total
Enfermos	31	20	51

[2] El programa estadístico SPSS adoptó el intervalo de Jeffreys. Otras opciones las ofrecen los programas Minitab, S-PLUS o Mathematica. También se ha desarrollado una página web que permite el cálculo de los intervalos por varios métodos: http://epitools.ausvet.com.au/content.php?page=CIProportion, o se puede consultar el excelente libro de Salvatore Mangiafico en el que se explica en detalle cómo realizar estos cálculos con *RStudio* (2023, pág. 93).

Sanos	20	31	51
Total	51	51	102

Fuente: elaboración propia.

Obtenemos el siguiente gráfico:

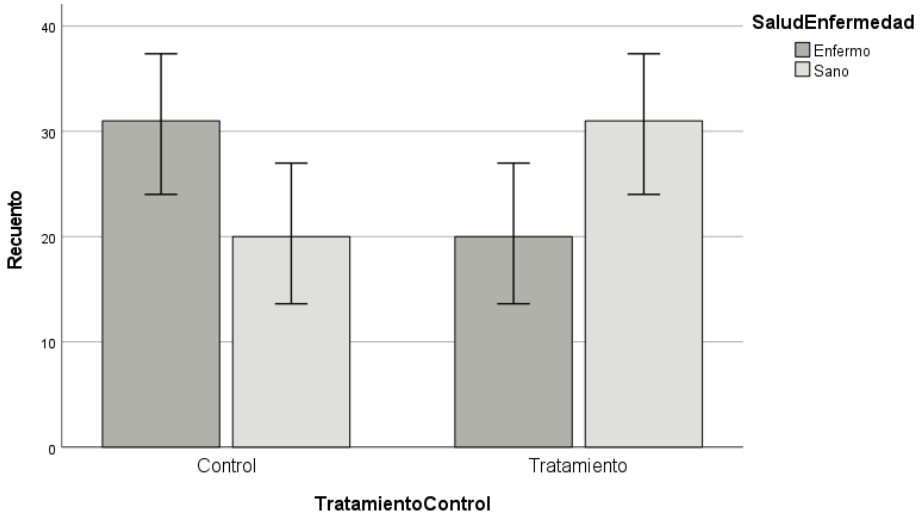

Gráfico 1. Tabla de contingencia 2x2 con significación al 0,05 y solapamiento de los intervalos de confianza.

Fuente: elaboración propia.

Según Cumming es necesario calcular el intervalo de confianza para la diferencia y analizar si el 0 es un valor probable para una determinada confianza (2014). Para ello es posible utilizar la fórmula (2) que ofrece Agresti (2007) con error estándar no combinado (*unpooled*). Sin embargo, se debe tener en cuenta que los softwares estadísticos pueden presentar alguna pequeña diferencia de acuerdo con la fórmula para el cálculo del intervalo de confianza, como fue señalado precedentemente.

3.7 Intervalo de confianza bootstrap

Otro método para obtener intervalos de confianza para la diferencia de proporciones es el método de remuestreo (*bootstrapping* o *bootstrap*) que fue propuesto por Bradley Efron (1979). Consiste básicamente en la obtención de una cantidad muy grande de muestras con reemplazamiento a partir de la muestra original en estudio. Para dichas muestras es posible obtener sus percentiles al nivel deseado y de ese modo obtener un intervalo de confianza. Este método es muy útil cuando se sospecha alguna dificultad para determinar la distribución de muestreo por lo cual es posible utilizarlo en infinidad de situaciones. La enorme potencia de cálculo de las computadoras actuales facilita considerablemente la aplicabilidad de este método generando un gran número de

muestras (generalmente 1000 o más), impracticable en el pasado[3].

3.8 Residuos estandarizados corregidos de Haberman

Finalmente, más allá de los resultados de la prueba de significación estadística para la diferencia de proporciones, existe otro estadístico que permite indicar en una tabla de contingencia si la frecuencia observada se aleja significativamente de la esperada de independencia. Este estadístico es denominado de *residuos estandarizados corregidos* (Haberman, 1973).

Su fórmula viene dada por:

$$REC = \frac{f_o - f_e}{\sqrt{f_e\left(1-\frac{f_i}{n}\right)\left(1-\frac{c_j}{n}\right)}} \quad (15)$$

Siendo f_o la frecuencia observada, f_e la frecuencia esperada de independencia, f_i el total marginal de la fila correspondiente, y c_j el total marginal de la columna de la intersección entre filas y columnas (celda[4] de aquí en adelante). Este estadístico se distribuye con una distribución normal estándar por lo que valores mayores a 1,96 estarían indicando que la frecuencia correspondiente es significativa al 0,05 (a dos colas). En este caso ya no se busca una diferencia entre las proporciones o porcentajes, sino que la comparación es con las proporciones marginales, de modo que la interpretación evidentemente es bastante diferente y complementaria a la diferencia de proporciones.

4. POTENCIA DE LA PRUEBA EN DIFERENCIA DE PROPORCIONES

Hemos comenzado por razones puramente didácticas con la prueba de hipótesis para proporciones, sin embargo, debiéramos haber comenzado por donde el investigador debería comenzar. El primer paso es el de determinar la cantidad de casos necesarios para que una vez obtenidos los resultados podamos dar cuenta de ellos en términos de los errores de tipo *I* y tipo *II*. Esta es una decisión importante de toda investigación ya que, evidentemente una vez decidida la significación de la prueba, digamos al 0,05 debemos conocer cuál es la probabilidad de que no rechacemos la hipótesis nula cuando ella no es cierta. Esto cobra especial relevancia en las pruebas de significación para las diferencias de proporciones en las que la exigencia en cantidad de casos suele ser muy superior a las relativas a diferencias de medias.

Este análisis debe ser realizado antes de efectuar el estudio. Como ya hemos señalado en otro lugar, tamaño del efecto, cantidad de casos, varianza y errores de tipo *I* y tipo *II*

[3] Este procedimiento puede ser utilizado desde programas estadísticos comerciales como SPSS, Stata o SAS, o por software libre como Jasp, Jamovi o en RStudio para lo cual volvemos a aconsejar el libro de Mangiafico (2023).

[4] Una celda (o casilla) en una matriz de datos es un punto de intersección entre una fila y una columna específica dentro de la matriz de datos. Cada celda contiene un único valor que representa una observación o una medida particular en el conjunto de datos, sea de carácter numérico, categórico o de otro tipo (binarios, geoespaciales, imagen o video, etc.)

están fuertemente ligados como es desarrollado en la siguiente expresión:

$$1-\beta = P\left\{\frac{p_2-p_1}{\sqrt{\frac{2\bar{p}\bar{q}}{n}}} > z_{\frac{\alpha}{2}}\right\} \qquad (16), \text{(Fleiss, Levin y Paik, 2003, pág. 70)}$$

Donde β representa el error de tipo *II*, $p_2 - p_1$ el tamaño del efecto, α error de tipo *I* (en este caso $\frac{\alpha}{2}$ para una prueba de dos colas), y $n = n_1 + n_2$.

La potencia estadística refleja la capacidad de un estudio para detectar efectos genuinos en la población de interés. Un bajo nivel de potencia aumenta la probabilidad de pasar por alto efectos significativos, lo que podría llevar a conclusiones erróneas de que no existe una relación entre las variables cuando en realidad sí podría existir. Es por eso que recomendamos encarecidamente el uso de herramientas especializadas como *G*Power*[5] u otras plataformas estadísticas disponibles en línea para calcular la potencia adecuadamente.

Al no prestar atención a este aspecto crítico, existe el riesgo de que estudios importantes se pasen por alto o se subestimen debido a limitaciones en el diseño o la ejecución, lo que podría impactar negativamente en el avance del conocimiento en el campo de estudio correspondiente. Por lo tanto, insistimos en la evaluación y consideración de la potencia estadística al diseñar, ejecutar y presentar sus investigaciones, garantizando así la integridad y la fiabilidad de los resultados obtenidos.

5. TAMAÑO DEL EFECTO EN DIFERENCIA DE PROPORCIONES PARA MUESTRAS INDEPENDIENTES

El tamaño del efecto para proporciones se puede considerar como una medida de la magnitud de la diferencia entre ellas. Esta medida puede ser evaluada utilizando valores obtenidos de una manera relativamente simple:

$$p_{dif} = p_2 - p_1 \quad (17)$$

Esta diferencia es en sí un tamaño de efecto. Pero en este caso muchas veces para no perder de vista la totalidad de la diferencia será necesario mantener los dos miembros de la igualdad. ¿Por qué? simplemente porque 50-45= 5, pero 20-15= 5 y evidentemente estas dos igualdades no significan lo mismo. El 5 en el segundo caso tiene una dimensión mayor que en el primero.

Es por este motivo que el tamaño del efecto para proporciones tiene particularidades que la diferencian sustancialmente del tamaño del efecto en el caso de la diferencia de medias, dónde en general, las cosas se presentan en principio de un modo bastante más simple.

[5] Disponible en la página de la universidad Heinrich-Heine-Universität Düsseldorfde modo libre y gratuito en https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

Para abordar este tema y comprender su complejidad, exploraremos brevemente los diferentes índices que pueden ser utilizados para medir el tamaño del efecto en tablas de contingencia.

Los coeficientes *Phi*, *V* de Cramer y w de Cohen son medidas comunes de tamaño del efecto que se utilizan en el contexto de análisis de tablas de contingencia o en el estudio de asociaciones entre variables categóricas. Todas ellas dan cuenta de la relación entre las variables, más que de la relación entre dos proporciones.

También se han utilizado los estadísticos *Phi* y *r* con el mismo objetivo. Sin embargo, estos índices de tamaño del efecto han sido criticados por Fleiss (2003) y por Haddock y otros (1998) ya que su utilización, basada en tomar las variables como si fueran cuantitativas, tienden algunas veces a subestimar el verdadero tamaño del efecto.

5.1 El coeficiente Phi (φ)

El coeficiente *Phi* (*φ*)se basa en el Chi-cuadrado:

$$Phi(\varphi) = \sqrt{\frac{chi^2}{n}} \quad (18)$$

Este coeficiente es muy utilizado como medida de asociación. Esta medida es similar al coeficiente de correlación de Pearson en su interpretación. De hecho, un coeficiente de correlación de Pearson estimado para dos variables binarias nos dará el coeficiente *Phi*. Los valores de *Phi*, al igual que los de *r* de Pearson, varían de -1 a 1, donde 0 indica ausencia de asociación, -1 indica asociación negativa perfecta y 1 indica asociación positiva perfecta. Muchos de los aspectos teóricos referidos a las tablas de 2x2 son aplicables a las tablas mxn. Obviamente la aplicabilidad de este indicador está limitada a la posibilidad de obtener un chi-cuadrado. Y su interpretación quedará sujeta a la posibilidad de interacción conjunta entre las variables. Es decir, se trata de un coeficiente simétrico.

La interpretación del tamaño de efecto sugerida por Cohen es que en la ausencia de una determinación en el campo concreto de aplicación se pueden considerar los valores contenidos en la siguiente tabla:

Tabla 3. Interpretación del coeficiente Phi como tamaño de efecto.

gl	Pequeño	Medio	Grande
1	0,1	0,3	0,5

Fuente: adaptada de Cohen (1988, pág. 222).

Un tamaño de efecto alrededor de *Phi* de 0,1 será un tamaño de efecto pequeño (aunque no insignificante), un valor de 0,3 sería medio, y en torno a 0,5 o superior será un tamaño de efecto grande.

5.2 El coeficiente *V* de Cramer

Vimos en un párrafo anterior el coeficiente *Phi*, asociado al *Chi-cuadrado* y al *n* de la muestra. La *V* de Cramer no es más que la extensión de *Phi* para el caso de tablas mxn:

$$V = \sqrt{\frac{chi^2}{n(k-1)}} \qquad (19)$$

Donde k es el menor del número de filas o de columnas.

Tanto *Phi* como V son utilizados como medidas de asociación, es decir, tamaños de efecto por lo que para la realización de cualquier meta-análisis se requiere el correspondiente intervalo de confianza. Para calcularlo, dado que la distribución de V no es normal, se suele utilizar la transformación *Z de Fisher* del siguiente modo:

Primero se calcula su transformación a valores Z:

$$Z_V = ln\left(\frac{1+V}{1-V}\right)/2 \qquad (20)$$

El error estándar de Z_V viene dado por:

$$ES_{Z_V} = \frac{1}{\sqrt{n-3}} \qquad (21)$$

Luego se pueden obtener los límites del intervalo en Z

$$Z_{rInf} = Z_V - Z_{conf} * ES_{Z_V} \qquad (22)$$

$$Z_{rSup} = Z_V + Z_{conf} * ES_{Z_V} \qquad (23)$$

Y finalmente es posible obtener el intervalo de confianza de V del siguiente modo:

$$V_{Inf} = \frac{e^{2*Z_{rInf}}-1}{e^{2*Z_{rInf}}+1} \qquad (24)$$

$$V_{Sup} = \frac{e^{2*Z_{rSup}}-1}{e^{2*Z_{rSup}}+1} \qquad (25)$$

El intervalo de confianza de *Phi* se calcula y se interpreta de la misma forma que el coeficiente V de Cramer. Siempre teniendo en cuenta la simetría de la relación entre las variables:

Tabla 4. Interpretación de la V de Cramer según grados de libertad.

gl	Pequeño	Medio	Grande
1	0,1	0,3	0,5
2	0,07	0,21	0,35
3	0,06	0,17	0,29
4	0,05	0,15	0,25
5	0,04	0,13	0,22

Fuente: adaptada de Cohen (1988, pág. 222).

Para *Phi* obviamente el grado de libertad es 1, por lo que un valor en torno a 0,1 indica un tamaño de efecto pequeño, un valor en torno a 0,3 medio, y en torno a 0,5 o superior, un tamaño de efecto grande. Otro ejemplo, en una tabla 3x2 con gl=2 un valor de V en torno a 0,07 será un valor pequeño, 0,21 medio, y de 0,35 o más será un valor grande.

5.3 *T* de Tschuprow

Alexander Tschuprow propuso otro indicador de asociación para tablas de contingencia en el año 1939. Es muy similar al coefieciente V de Cramer, sin embargo, resulta más estricto en cuanto a la obtención de su valor máximo (1), que solo puede alcanzarse en tablas de contingencia cuadradas:

$$T = \sqrt{\frac{chi^2}{n(c-1)(f-1)}} \qquad (26)$$

5.4 *W* de Cohen

El coeficiente W de Cohen es una medida de asociación análoga al coeficiente *phi* y V de Cramer de acuerdo con la siguiente fórmula Cohen (1988):

$$W = \sqrt{\sum_{i=1}^{m} \frac{(P_{1i}-P_{0i})^2}{P_{0i}}} \qquad (27)$$

Donde P_{0i} es la proporción en la celda i bajo la hipótesis nula,

P_{1i} es la proporción en la celda i a partir de la hipótesis alternativa, y refleja el efecto en dicha celda.

m es el número de celdas.

Para calcular el coeficiente W de Cohen, se toma la diferencia entre las dos proporciones hipotéticas en cada celda de la tabla de contingencia, se eleva al cuadrado y se divide por la proporción esperada bajo la hipótesis nula. Luego, estos valores se suman y se calcula la raíz cuadrada del resultado. Cohen nos alienta a observar la analogía con el cálculo de la chi-cuadrada y a considerar la simetría del coeficiente (Cohen, 1988, pág. 216).

Otra forma de calcular W es a partir de V de Cramer:

$$W = V * \sqrt{k-1} \qquad (28)$$

Donde k es el número de categorías de la variable con menor número de categorías.

Como W y V están relacionados linealmente podemos obtener el intervalo de confianza de W a partir del cálculo del intervalo de confianza de V y transformarlos en valores w de acuerdo a la fórmula correspondiente (28).

Otra forma aproximada es obtener el error estándar para los cálculos correspondientes del siguiente modo:

$$ES_W = \sqrt{Var_w} = \sqrt{\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(o_{ij}-e_{ij})^2/e_{ij}}{n-1}} \qquad (29)$$

que para valores grandes de n puede ser aproximada por

$$ES_w = \sqrt{w/n} \qquad (30)$$

La interpretación del tamaño del efecto es similar a la de *Phi* y V de Cramer. Sin embargo, un aspecto en contra de este coeficiente es que en tablas de dimensiones m×n con valores extremos, puede tomar valores superiores a 1.

Antes de dejar este tema, téngase en cuenta que Ben-Shachar y otros (2023) y Jané y otros (2024) han propuesto recientemente, para el caso de pruebas de bondad de ajuste, un coeficiente derivado de W de Cohen que tiene la siguiente forma:

$$\text{Ɒ (Fei)} = \sqrt{\frac{chi^2}{n*\left(\frac{1}{min(p_e)}-1\right)}} \qquad (31)$$

El coeficiente Fei, a diferencia de la W de Cohen, tiene la virtud de mantenerse entre los

valores 0 y 1.

5.5 *Lambda* de Goodman y Kruskal (λ)

Todos los coeficientes que hemos examinado hasta ahora como medidas de tamaño de efecto tienen tanto ventajas como desventajas al ser simétricos. En otras palabras, evalúan la fuerza de la asociación sin considerar una dirección específica. Esto implica que no reflejan el efecto que una variable pueda tener sobre la otra.

Goodman y Kruskal (1954) diseñaron un coeficiente que responde a la siguiente pregunta: ¿Qué proporción de errores que se cometerían al predecir sin información se reducirían al utilizar la variable considerada como base para la predicción?

Se la puede definir es "una medida de la reducción proporcional del error conociendo los valores de la variable independiente (factor) en tablas de contingencia" en base a las frecuencias no modales:

$$\lambda_{c\ dep} = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1} \qquad (32)$$

Donde,

$\lambda_{c\ dep}$ representa el coeficiente *lambda* considerando a las columnas como variable dependiente.

ε_1 es la suma de las frecuencias no modales marginales de las columnas. Representa la probabilidad de error.

ε_2 es la suma total de las frecuencias no modales de cada una de las filas de la tabla. Representa la probabilidad de error habiendo ocurrido la variable independiente.

Si realizamos el mismo procedimiento, pero ahora en la dirección de las filas vamos a obtener *lambda* en la otra dirección ($\lambda_{f\ dep}$) con las filas como dependiente.

El promedio entre ambos coeficientes es lambda simétrica:

$$\lambda_s = \frac{\lambda_{c\ dep} - \lambda_{f\ dep}}{2} \qquad (33)$$

Un valor de 0 estaría indicando ausencia de efecto, y un valor de 1 nos indica que conociendo el valor de la independiente podemos predecir sin error la variable dependiente o factor. De todos modos, se debe tener en cuenta que en algunos casos puede ocurrir que obtengamos un valor 0 cuando existe un efecto evidente, por lo que es necesario considerar esta eventual desventaja del coeficiente.

El error estándar de *lambda* puede ser calculado de acuerdo a:

$$ES_{\lambda_{c\ dep}} = \sqrt{\frac{n-s}{(n-r)^2} * (s + r - 2u)} \qquad (34)$$

Donde n es el total de casos, r representa el valor máximo marginal de las filas, s el valor resultante de sumar todas las casillas contingentes menos los valores máximos de cada columna, y u representa la suma de todos los valores de las casillas de la columna con marginal máximo que a su vez sean máximos de fila.

5.6 *h* de Cohen

Cohen (1988) propuso para la obtención del tamaño del efecto para la distancia entre

dos proporciones por el arcoseno de acuerdo a la siguiente expresión:

$$\varphi = 2 * \text{arcoseno}\left(\sqrt{p}\right) \quad (35)$$

de modo que

$$h = \varphi_1 - \varphi_2 \text{ , (Cohen, 1988)} \quad (36)$$

Este coeficiente ya no se limita a su aplicación en el marco de tablas de contingencia, sino que se extiende a la comparación entre dos proporciones, e incluso a la comparación entre una proporción y un valor teórico.

Para obtener el error estándar de h se procede del siguiente modo

$$ES_h = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (37)$$

De modo que el intervalo de confianza de h viene dado por las siguientes expresiones:

$$h_{Inf} = h - z_{conf} * ES_h \quad (38)$$

$$h_{Sup} = h + z_{conf} * ES_h \quad (39)$$

La interpretación de los valores de h de Cohen es similar a los de la d de Cohen:

Tabla 5. Interpretación de h como tamaño de efecto.

	Pequeño	Medio	Grande
h	0,2	0,5	0,8

Fuente: adaptada de Cohen (1988, pág. 184).

Antes de proseguir vamos a dar cuenta de algunas críticas que han recibido los coeficientes que miden tamaño de efecto en el caso de proporciones que hemos visto hasta ahora. Varios autores, incluyendo a Fleiss, Levin y Paik (2003), Agresti (2007), Haddock, Rindskopf y Shadish (1998), Sánchez-Meca, Marín-Martínez, y Salvador Chacón-Moscoso (2003), han señalado que *Phi* (*φ*), *V* de Cramer (*V*), *W* de Cohen y *h* de Cohen (*h*) tienden a subestimar el verdadero tamaño del efecto. Todos ellos apuntan, como posible solución, a la aplicación del *riesgo relativo* y de los *odds ratio* como solución al problema, temas que veremos a continuación.

El riesgo relativo es ampliamente utilizado y más fácil de comprender, pero presenta el problema que no es simétrico. En cambio, la propiedad de los *odds*, por la cual las inversas de los logaritmos con base *e* son simétricos en torno a 0 los hace de mayor utilidad (Agresti, 2007).

5.7 Riesgo relativo

Al final del apartado anterior, hemos destacado que los *odds ratio* son recomendados como medida de tamaño del efecto en casos de diferencias de proporciones. Sin embargo, existen algunas consideraciones importantes respecto a su utilización. En particular, la principal crítica se centra en la dificultad de interpretación. Se argumenta que el riesgo relativo es una medida más fácil de comprender, y estamos de acuerdo con esta afirmación.

¿Qué es un *riesgo relativo*? Bueno es simplemente el cociente entre dos riesgos, entre

dos proporciones y se puede expresar del siguiente modo:

$$RR = \frac{R_e}{R_o} = \frac{a/(a+b)}{c/(c+d)} \qquad (40) \text{ -ver Tabla 1-}$$

Donde R_e representa la prevalencia entre los expuestos y R_o la prevalencia entre los no expuestos (Molina Arias, 2014; Borenstein y Hedges, 2019, pág. 223).

Con una varianza aproximada dada por la siguiente ecuación:

$$V_{lnRR} = \frac{1}{a} + \frac{1}{a+b} + \frac{1}{c} + \frac{1}{c+d} \qquad (41)$$

Y entonces,

$$ES_{lnRR} = \sqrt{V_{lnRR}} \qquad (42),$$

(Borenstein, Hedges, Higgins y Rothstein, 2009, pág. 35).

Se suele afirmar que, en estudios sociales, económicos y de salud en los que se puedan calcular las prevalencias, como en investigaciones demográficas, epidemiológicas o de salud pública, el riesgo relativo puede ser una medida adecuada para determinar el tamaño del efecto correspondiente. Sin embargo, su utilidad se ve limitada en estudios en los que no se disponga de las prevalencias correspondientes (los marginales de las tablas deben ser representativos de la población de referencia), como en los estudios de casos y controles. Específicamente, se desaconseja su uso, especialmente cuando las prevalencias superan el 10% (Schmidt, 2008, pág. 165; Molina Arias, 2014, pág. 275). En resumen, mientras que el *riesgo relativo* puede calcularse cuando se tienen las prevalencias disponibles, se recomienda utilizar el *odds ratio* cuando no se dispone de esta información.

Odds ratio y *riesgo relativo* pueden ser relacionados de acuerdo a la siguiente ecuación, lógicamente si contamos con el valor de la prevalencia o riesgo general:

$$RR = \frac{OR}{(1-Prev)+(Prev*OR)}, \quad (43)$$

Donde, *Prev* está representando la prevalencia del grupo de control, o los no expuestos (verdadera prevalencia en la población) de acuerdo a la terminología que venimos usando (Molina Arias, 2014, pág. 278; Rita y Komonen, 2008, pág.71).

5.8 Definición de *odds*

Los *odds* dan cuenta de la razón de que un evento ocurra respecto de que no ocurra. Son las chances de que ocurra respecto de que no ocurra. Es conceptualmente distinto a la probabilidad. Su formalización matemática viene dada por:

$$odds = \frac{p}{(1-p)} = \frac{frecuencia_{evento}}{frecuencia_{no\ evento}} \quad (44)$$

Se subraya que no debe ser confundido con las probabilidades, aunque su interpretación sea similar cuando la frecuencia del evento sea relativamente pequeña en relación a la frecuencia del no evento. En esos casos *odds* y probabilidades se van a

parecer. Por lo demás, son conceptos diferentes y medidos en escalas completamente diferentes. De 0 a 1 las probabilidades, de 0 a infinito los *odds*.

5.9 Definición de *odds ratio*

Un *odds ratio* es esencialmente un cociente entre *odds*, lo que lo hace una medida poderosa al permitir la comparación de dos grupos en términos de sus respectivas chances. Puede ser interpretado como las chances de un grupo por unidad de chances del otro. Es de más fácil interpretación cuando el numerador es mayor que el denominador.

$$odds\ ratio = OR = \frac{p_1(1-p_1)}{p_2(1-p_2)} = \frac{ad}{bc} = \frac{odds_{grupo\ 1}}{odds_{grupo\ 2}} \qquad (45)$$ -ver tabla 1-

Obsérvese que en su cálculo no intervienen las frecuencias marginales.

El error estándar del *odds ratio* viene dado por:

$$EE_{\ln(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \qquad (46)$$

Como su distribución de muestreo es aproximadamente normal con media *OR* y desvío estándar $EE_{\ln(OR)}$ pueden ser calculados sus intervalos de confianza según:

$$\ln OR \ \pm z_{\frac{\alpha}{2}}(EE_{\ln(OR)}) \qquad (47)$$

Es posible convertir los *odds ratio* en unidades de *d* de Cohen de acuerdo con la siguiente expresión:

$$d_{HH} = Log_e OR * \frac{\sqrt{3}}{\pi} \qquad (48)$$

Y su varianza:

$$S^2{}_{dHH} = \frac{3}{\pi^2}\left(\frac{1}{O_{1e}} + \frac{1}{O_{2e}} + \frac{1}{O_{1c}} + \frac{1}{O_{2c}}\right) \qquad (49)$$

Varios autores sostienen que, además de que los *odds ratio* tienen un buen comportamiento en tanto tamaño de efecto, presentan una mejor performance en cuanto a su transformación a la métrica de la *d* de Cohen, en comparación con otros coeficientes (Borenstein, Hedges, Higgins, y Rothstein, 2009; Sanchez-Meca, Marín-Martínez, y Salvador Chacón-Moscoso, 2003).

5.10 La interpretación del tamaño del efecto (odds ratio)

Cohen (1988) nos ofrece una forma de interpretar los tamaños de efecto. Mientras que Sawilowsky (2009) amplía la descripción de Cohen del siguiente modo:

Tabla 6. Interpretación de la *OR* bajo la métrica de la *d* de Cohen y otras medidas de tamaño de efecto.

Tamaño del efecto	*d, h*	*r, Phi,* *V* de Cramer, *W* de Cohen	*OR* (aprox d_{HH})
Muy pequeña	0,01	0,005	1,083
Pequeña (pero no trivial)	0,2	0,1	1,4
Media o moderada	0,5	0,3	2,5

Grande	0,8	0,5	4,3
Muy grande	1,2	0,8	8,8
Enorme	2	0,9	37,6

Fuente: adaptada de Cohen (1988) y Sawilowsky (2009).

Nótese que la aproximación la hemos hecho basándonos en la aproximación d_{HH} que hemos indicado en el apartado precedente de acuerdo con la ecuación (48). Si bien es posible utilizar directamente el *odds ratio* como medida del tamaño del efecto, es común recurrir a la métrica de *d* de Cohen. Esto se debe a la conveniencia de estandarizar los resultados, especialmente en meta-análisis que integran estudios con diferentes métricas. Es importante muchas veces considerar la relación entre estos dos estadísticos para una interpretación completa y precisa de los hallazgos.

Sin embargo, como hemos señalado en otra oportunidad (D'Angelo, 2021) es necesario ser cuidadosos al trabajar con los tamaños de efecto. Cohen afirmó que estos valores de interpretación deben ser considerados en cada escenario de investigación (1988). Ya que en cada ámbito esto puede cambiar de modo sustancial. Por ejemplo, un muy pequeño efecto en un tratamiento capaz de salvar vidas, puede ser considerado un efecto importante de por sí, y porque puede evidentemente abrir las puertas a nuevas investigaciones. Consideremos otro ejemplo, un estudio que evalúa la efectividad de un nuevo sistema de alerta temprana de fallas en los motores de aviones. Después de realizar un análisis, se encuentra que este sistema tiene un efecto relativamente pequeño en la reducción de accidentes aéreos. A primera vista, este efecto puede parecer insignificante. Sin embargo, considerando el contexto de la seguridad de los aviones, incluso una pequeña reducción en la tasa de accidentes podría tener un impacto significativo en la seguridad y la confianza de los pasajeros en la industria de la aviación. Por lo tanto, a pesar de ser un efecto pequeño en términos absolutos, en el contexto de la seguridad aérea, este resultado podría ser considerado de gran importancia y justificaría una mayor investigación y desarrollo del sistema de alerta temprana. Por el lado contrario, podemos hallar tamaños de efecto medianos o incluso grandes según estas categorías de interpretación, pero que pueden carecer de importancia teórica, práctica o clínica.

Uno de los problemas que se esgrime contra los *odds ratio* es justamente el relativo a la dificultad de uso e interpretación. En particular se señala que muchas veces se interpretan los *odds ratio* como *riesgos relativos*. Los dos estadísticos son indicadores del tamaño del efecto. Por nuestra parte, dados a elegir entre *odds ratio* y *riesgo relativo*, vamos a optar, en general, por el primero. Por cuatro razones, la primera es relativa a sus propiedades matemáticas a partir del hecho que los *odds* son simétricos respecto al 0 como ya hemos señalado, y su ratio comparable con respecto a 1. La segunda es que es aplicable en mayor cantidad de situaciones. En particular, cuando tenemos

sospechas en relación a la prevalencia del fenómeno en cuestión, podemos fiarnos de este estadístico que no lo presupone. Tercero, resulta más sencillo a la hora de transformar sus valores a otras métricas de tamaño de efecto. Finalmente, los *odds ratio* pueden ser calculados a partir de una regresión logística, con el importante agregado de que en esos casos podemos controlar estadísticamente algunas variables. En contra se le puede esgrimir que para prevalencias mayores el valor se hace muy superior al *riesgo relativo*. Pero aquí solamente volvemos al problema de la interpretación. Esto se resuelve trabajando para mejorar el nivel conceptual de los investigadores en cuanto a la compresión y aplicación del estadístico.

Para concluir este debate vamos a citar a Rita y Komonen (2008, pág.71), quienes sostienen que la mayor parte del debate en relación al *riesgo relativo* y el *odds ratio* no se refiere a sus propiedades estadísticas *per se*. Los *odds ratio* poseen la mayoría de las propiedades estadísticas que debería tener una medida de tamaño de efecto ideal, y, por lo tanto, es más adecuado para una variedad de diseños de investigación y análisis de datos.

5.11 Sobre los errores estándar e intervalos de confianza en este trabajo

En este trabajo se ha intentado complementar cada uno de los estadísticos de tamaño de efecto con alguna forma de cálculo a partir de su error estándar y sus respectivos cálculos.

En los últimos años, ha habido importantes avances en la aplicación de procedimientos para calcular intervalos de confianza utilizando una variedad de métodos alternativos.

- Método *Bootstrap*: Es una técnica de *remuestreo* utilizada para estimar la distribución de un estadístico a partir de datos observados. Este método genera múltiples muestras *bootstrap* seleccionando aleatoriamente observaciones con reemplazamiento. A partir de estas muestras se pueden estimar intervalos de confianza. Este procedimiento puede llevarse a cabo en casos de estudios con muestras representativas. Existen varias librerías gratuitas y programas estadísticos comerciales para realizar este procedimiento[6].

- Prueba de permutación: Es otra técnica de *remuestreo* utilizada para evaluar la significación de las diferencias observadas entre grupos sin supuestos sobre la distribución de los datos. Consiste en permutar aleatoriamente las asignaciones de los grupos para generar una distribución nula y compararla con las estadísticas observadas (Edgington & Onghena, Randomization Test, 2007). Esta técnica es especialmente útil en casos de diseños experimentales[7].

- Método del parámetro chi-cuadrado no central *(ncp)*: Este método consiste en

[6] Recomendamos utilizar la librería *rcompanion*, desarrollada por Salvatore Mangiafico (2023).

[7] Los cálculos pueden obtenerse a partir de librerías en *R* como *coin* o *perm*, o en el sitio web de Lock y otros (2021) https://www.lock5stat.com/StatKey.

determinar los valores del parámetro no central de una distribución chi-cuadrado para establecer los límites del intervalo de confianza de los coeficientes en estudio. El enfoque *ncp* puede aplicarse tanto a estudios experimentales, en los que la asignación de grupos es aleatoria, como a estudios con muestras representativas[8].

6. MEDIDA DE CREENCIA EN LA HIPÓTESIS NULA Y ALTERNATIVA: FACTOR DE BAYES

Hasta ahora nos hemos enfocado en la existencia o no de un efecto entre variables, y luego en el tamaño de este efecto. Ahora nos toca avanzar sobre la probabilidad de que la hipótesis alternativa (o la nula) sean ciertas dados unos ciertos datos. El factor de Bayes (FB) cuantifica la evidencia proporcionada por los datos observados a favor de una hipótesis sobre la otra. El factor de Bayes nos informa cuán probable es, de acuerdo a nuestros datos, tanto la hipótesis nula como la alternativa (Jeffreys, 1961).

Hemos desarrollado este punto en otro lugar, por lo que remitimos al lector a dicho trabajo (D'Angelo, 2021). Sin embargo, en el caso de las proporciones el trabajo de Erdonan Gunel y James Dickey es central para la aplicación de esta técnica (1974).

Ellos definen cuatro esquemas de muestreo que van a determinar el modo de calcular el factor de Bayes de acuerdo con los grados de libertad implicados en el plan:

- El esquema de Poisson
- Esquema de muestreo multinomial conjunto
- Esquema de muestreo multinomial independiente
- Esquema de muestreo hipergeométrico

Entre el primer esquema muestral (Poisson) y el último (hipergeométrico) se observa una graduación hacia menores grados de libertad, de modo que en el esquema de muestreo de Poisson cada recuento de celdas es aleatorio, por lo cual el total general también lo es. El siguiente esquema, de muestreo multinomial conjunto, es similar al de Poisson, excepto que el gran total (suma total de filas o de columnas en tablas de contingencia) queda determinado. En cambio, en el esquema de muestreo multinomial independiente hay dos restricciones, ya sea en los totales de fila o en los totales de la columna. En otras palabras, todos los márgenes de fila o todos los márgenes de columna son fijos. En consecuencia, los recuentos de frecuencias se distribuyen multinomialmente dentro de cada fila o columna. En la mayor parte de los estudios, este esquema es el más común.

Finalmente, el esquema de muestreo hipergeométrico tanto los totales marginales de

[8] Para llevar a cabo estos procedimientos, recomendamos utilizar la librería en *R effectsize* de Mattan Ben-Shachar (Ben-Shachar, Patil, Thériault, Wiernik, & Lüdecke, 2023). Otra opción para realizar estos cálculos es el complemento de Excel *Real Statistics* de Charles Zaiontz (2023, https://real-statistics.com).

fila y columna quedan establecidos de antemano de modo fijo. La aplicación práctica del esquema de muestreo hipergeométrico es rara (Jamil, Ly, Morey, Marsman y Wagenmakers, 2016).

7. META-ANÁLISIS

Hemos recorrido diversos aspectos relativos al tamaño del efecto, es decir, a un valor que refleja la magnitud del efecto del tratamiento o la fuerza de una relación entre dos variables. Sin embargo, también representa la unidad de análisis en un meta-análisis. El principal objetivo de la síntesis es la de calcular un efecto de resumen de los diversos estudios, así como una medida de precisión y un valor *p* relativo al conjunto de los estudios incluidos (Borenstein, Hedges, Higgins y Rothstein, 2009, págs. 3-6).

El propósito del meta-análisis "es la integración de los resultados dispersos en múltiples estudios en los que se ponen a prueba una misma hipótesis, es decir, se trata de un análisis conjunto de otros análisis realizados anteriormente" (Macbeth, Cortada de Kohan y Razumiejczyk, 2007).

Un aspecto a tener en cuenta para el cálculo es la naturaleza de la variabilidad entre los estudios. Es decir, si podemos considerar que la variabilidad de los resultados queda definida exclusivamente por la selección de las muestras, entonces esa variabilidad es intrínsecamente dependiente del proceso de muestreo, en el caso que el tamaño sea idéntico para todos los estudios, entonces decimos que estamos frente a un meta-análisis de efectos fijos. Cuando por el contrario además de la variabilidad dependiente exclusivamente del proceso de selección de los casos podemos sospechar diferencias en los procedimientos de análisis, es decir, que vamos a encontrar variabilidad entre las unidades y entre los estudios, decimos que estamos frente a un meta-análisis de efectos variables (Borenstein, Hedges, Higgins, y Rothstein, 2009, págs. 78-79).

Así, en el caso de los meta-análisis de efectos fijos, la variabilidad tiende a ser siempre igual o menor que en los meta-análisis de efectos variables ya que:

$$SE_{efectos\ fijos} = \sqrt{\frac{\sigma^2}{k*n}} \quad (50)$$

$$SE_{efectos\ variables} = \sqrt{\frac{\sigma^2}{k*n} + \frac{\tau^2}{k}} \quad (51), \text{ (Borenstein, Hedges, Higgins, y Rothstein, 2009, pág. 82).}$$

Donde σ^2 representa la varianza del tamaño del efecto, k la cantidad de estudios, n el total de casos y τ^2 (T^2 para un meta-análisis concreto) la variabilidad entre los estudios, que como se ha definido en los modelos de efectos variables es mayor que cero, es decir algún valor que debe incrementar consiguientemente el *SE*.

El resultado obtenido debe ser validado. En primer lugar, se debe dar cuenta de la heterogeneidad de las distribuciones. Para ello se ha propuesto un indicador basado en Chi^2 con k-1 grados de libertad:

$$Q = \sum w_i(T_i - \bar{T}), \qquad (52)$$

donde T_i representa cada uno de los diversos tamaños de efectos hallados en los diversos estudios, y $\bar{T}$ representa el tamaño de efecto de resumen.

Diversos autores han señalado que este indicador presenta problemas, especialmente, cuando el número de estudios es pequeño (SEH-LELHA, 2003).

Por este motivo se ha propuesto otro índice denominado I^2, en términos de la proporción de la variación entre estudios respecto de la variación total, es decir la proporción de la variación total que es atribuible a la heterogeneidad (Borenstein, Hedges, Higgins y Rothstein, 2009, pág. 117):

$$I^2 = \left(\frac{Q-gl}{Q}\right)100\%, \qquad (53)$$

donde $gl=k-1$

Los umbrales para la interpretación de I^2 dependen de varios factores. Una guía aproximada de interpretación es la siguiente:

0% a 40%: puede no ser importante;

30% a 60%: puede representar heterogeneidad moderada;

50% a 90%: puede representar heterogeneidad sustancial;

75% a 100%: heterogeneidad considerable.

(Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0, 2011)

Al igual que respecto al apartado anterior reenviamos al lector al artículo sobre tamaño del efecto, potencia de la prueba, factor de Bayes y meta-análisis en el marco de la crisis de reproducibilidad de la ciencia, pero referido a la diferencia de medias que puede resultar una buena introducción también para el caso de la diferencia de proporciones (D'Angelo, 2021). Sin embargo, quienes deseen profundizar sobre estos temas les sugerimos consultar los excelentes trabajos de Borenstein, Hedges, Higgins, y Rothstein, (2009), Cochrane Handbook (2011), Cumming (2012) y Cooper y otros (2019), entre otros.

8. CONCLUSIONES

Este trabajo ha intentado poner de manifiesto la necesidad de profundizar en los conocimientos en el caso de las proporciones y sus diferencias. Se trata de uno de los estadísticos más básicos y ampliamente utilizados. Sin embargo, en diversos ámbitos se puede observar un uso deficiente del mismo. Quizá debido a su aparente simplicidad, que no siempre conlleva una comprensión profunda y, por ende, implicando una mayor propensión a cometer errores en su aplicación.

Al igual que en el caso de las diferencias de medias aritméticas no alcanza con la obtención de una significación estadística de las diferencias de proporciones. Es necesario abordar la cuestión del tamaño del efecto. Aquí es dónde la tarea se torna más desafiante. Es que en el caso de las diferencias de medias contamos con pocos

estadísticos establecidos, la *d* de Cohen, *g* de Hedges y *delta* de Glass. En el caso de las proporciones es necesario definir el estadístico más apropiado para este fin.

A lo largo de este trabajo, hemos explorado una serie de estadísticos con el objetivo de proporcionar fundamentos para su elección, aunque sin llegar a una determinación definitiva.

Creemos junto a una gran parte de los investigadores que los *odds ratio* representan el mejor indicador del tamaño del efecto en el caso de las diferencias de proporciones. Esto se debe a las propiedades matemáticas de los *odds*, que los hacen simétricos en relación al 1, y a su facilidad para convertirse en la d de Cohen (d_{HH}) y otros aspectos, como se ha expuesto en el cuerpo del artículo. Esta mayor facilidad para ser convertidos en otras métricas lo hace sobresalir respecto del riesgo relativo, además de su consabida mayor capacidad para ser utilizado en diversos escenarios de investigación.

En resumen, además de obtener los intervalos de confianza, indicando claramente qué ecuación se ha utilizado, y realizar la prueba de hipótesis para las diferencias de proporciones, señalando el valor p obtenido y la potencia de la prueba (incluyendo el cálculo de los residuos estandarizados corregidos cuando se desea examinar el comportamiento de una casilla en particular), es imprescindible incorporar una medida de tamaño del efecto y sus errores estándar correspondientes. En este contexto, preferimos utilizar los *odds ratio*, como se ha mencionado anteriormente.

Sin embargo, no debemos dejar de señalar que, en muchos casos vamos a querer dar cuenta de la relación entre dos variables, y no entre dos proporciones. En esos casos nos inclinamos, a pesar del gran desconocimiento que la acompaña, por la utilización de la W de Cohen o bien por la V de Cramer ya que se encuentran linealmente relacionadas. Ahora bien, si se trata de dar cuenta de una relación asimétrica *Lambda* de Goodman y Kruskal (o Tau-y), a pesar de sus debilidades, es el estadístico que se debiera utilizar.

Finalmente, analizar la probabilidad de rechazar tanto la hipótesis nula como la alternativa es otra práctica recomendable. Aunque este trabajo no aborda cuestiones prácticas, en varias situaciones específicas se hace referencia a la dificultad de obtener estudios con un mayor número de casos justificándose la realización del estudio de todas maneras. Nuestra postura al respecto es que, si nuestra prueba de hipótesis cuenta con un diseño demasiado limitado, es evidente que tendremos pocas probabilidades de obtener una significación estadística que respalde la hipótesis alternativa. Aunque podemos tener suerte y obtener una muestra con diferencias significativas, esta suerte generalmente alcanza un límite. Es muy probable que los intervalos de confianza de nuestro tamaño de efecto cuestionen nuestros resultados y, además, que obtengamos evidencia insuficiente en relación con nuestra hipótesis alternativa. Estas reflexiones no buscan desalentar la realización de este tipo de estudios con un número reducido de casos, sino fomentar la conciencia sobre sus

limitaciones y la necesidad de abordar el problema del *p-hacking* con rigor y transparencia.

Como se ha enfatizado previamente, es imperativo que el investigador se esfuerce por divulgar los resultados de sus estudios, independientemente de los resultados obtenidos en la prueba de hipótesis. Todos los estudios son valiosos y pueden contribuir al avance del conocimiento científico. Los metaanálisis bien ejecutados pueden facilitar la síntesis de información y desempeñar un papel crucial en la consolidación del conocimiento científico, es decir, en la búsqueda de la verdad.

CONFLICTO DE INTERESES

El autor declara no presentar conflictos de intereses en relación con la preparación y publicación de este artículo.

REFERENCIAS

AERA. (2006). Standards for reporting on empirical social science research in AERA publications. Educational Researcher, 36(6), 33-40.

Agresti, A. (2007). An Introduction to Categorical Data Analysis (Second ed.). Florida: Willey-Interscience.

Agresti, A., & Coull, B. A. (1998). Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician, 52(2).

Algina, J., & Keselman, H. J. (2006). Confidence Interval Coverage for Cohen's Effect Size Statistic. (S. Publications, Ed.) Educational and Psychological Measurement.

APA. (2007). Manual of the American Psychological Association (APA) (Sixth ed.). Washington. DC.

Aromataris, E., & Munn Z. (Editors). (2017). Joanna Briggs Institute Reviewer's Manual. Retrieved from The Joanna Briggs Institute: https://reviewersmanual.joannabriggs.org

Ben-Shachar, M. S., Patil, I., Thériault, R., Wiernik, B. M., & Lüdecke, D. (2023). Phi, Fei, Fo, Fum: Effect Sizes for Categorical Data That Use the Chi-Squared Statistic. Mathematics, 11(9).

Bolstad, W. M. (2007). Introduction to Bayesian Statistics. New Jersey: John Wiley & Sons, Inc.

Borenstein, M., Hedges, L. V., Higgins , J. P., & Rothstein, H. R. (2009). Introduction to Meta-Analysis. Chichester: John Wiley & Sons, Ltd.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). Introduction to Meta-Analysis. West Sussex: John Wiley & Sons, Ltd.

Borenstein, M., & Hedges, L. V. (2019). Effect Sizes for Meta-Analysis. In H. Cooper, L. V. Hedges, & V. Jeffrey, The Handbook of Research Synthesis and Meta-Analysis (pp. 208-241). Russell Sage Foundation.

Brand, A., Bradley, M. T., Best , L. A., & Stoica, G. (2008). Accuracy of Effect Size

Estimates from Published Psychological Research. Perceptual and Motor Skills, 106, 645-649.

Breaugh, J. A. (2003). Effect Size Estimation: Factors to Consider and Mistakes to Avoid. Journal of Management, 29(1), 79-97.

Brown, L. D., Cai, T. T., & Dasgupta, A. (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 16(2).

Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. (2011). Retrieved from https://handbook-5-1.cochrane.org/chapter_9/9_5_2_identifying_and_measuring_heterogeneity.htm

Coe, R., & Merino Soto, C. (2003). Magnitud del Efecto: Una guía para investigadores y usuarios. Revista de Psicología de la PUCP. Vol. XXI,, XXI(1).

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (Second ed.). New York: Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1992). A Power Primer. Psychological Bulletin, 112(1).

Cousineau, D., & Laurencelle, L. (2011). Non-central t distribution and the power of the t test: A rejoinder. Tutorials in Quantitative Methods for Psychology, 7(1).

Cumming, G. (2014). The New Statistics: Why and How. Psychological Science (Sage), 25(1).

Cumming, G., & Finch, S. (2005). Inference by Eye: Confidence Intervals and How to Read Pictures of Data . American Psychologist.

D'Angelo, L. A. (2021). Tamaño de efecto, potencia de la prueba, factor de Bayes y meta-análisis en el marco de la crisis de reproducibilidad de la ciencia. El caso de la diferencia de medias -con muestras independientes- (primera parte). Cuadernos del CIMBAGE, 1(23).

Edgington, E. S., & Onghena, P. (2007). Randomization Test. Boca Raton: Chapman & Hall/CRC.

Edgington, E. S., & Onghena, P. (2007). Randomization Tests. Boca Raton: Chapman & Hall/CRC.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1).

Ellis, P. D. (2010). The Essencial guide to Effect Size. Cambridge: University Press.

Epitat . (2020, marzo). Retrieved from Servizo Galego Saùde - Conselería de Sanidede: https://www.sergas.es/Saude-publica/Epidat-3-1-descargar-Epidat-3-1-(espanol)

Faulkenberry, T. J. (2018). Computing Bayes factors to measure evidence from experiments: An extension of the BIC approximation. Biometrical Letters.

Field, A. (2013). Discovering Statistics Using IBM SPSS Statistics. London: Sage Publications Ltd.

Fleiss, J., Levin, B., & Paik, M. C. (2003). Statistical Methods for Rates and Proportions. New Jersey: WileyIntercience.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Newbury Park: Sage Publications.

Good, P. (2005). Permutation, Parametric, and Bootstrap Tests of Hypotheses. Huntington Beach: Springer.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. Journal of the American Statistical Association, 49(268), 732-764.

Grissom, R. J., & Kim, J. J. (2014). Effect Sizes for Research: Univariate and Multivariate Application. Routledge: Psychology Press.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2018). Informed Bayesian T-Tests. Retrieved from https://arxiv.org/abs/1704.02479

Gunel, E., & Dickey, J. (1974). Bayes factors for independence in contingence tables. Biometrika, 61(3), 545.

Haberman, S. J. (1973). The analysis of residuals in cross-classification tables. Biometrics(29).

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using Odds Ratios as Effect Sizes for Meta-Analysis of Dichotomous Data: A Primer on Methods and Issues. Psychological Methods, 3(3).

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. Journal of Educational Statistics, 6.

Higgins J. P. T., G. S. (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. Retrieved from The Cochrane Collaboration: www.handbook.cochrane.org

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. PLoS ONE , 13(4), https://doi.org/10.1371/journal.pone.0195474.

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. Plos One.

Introduction to New Statistics. (2019, 10). Retrieved from https://thenewstatistics.com/itns/

Ioannidis, J. P. (2005). Why Most Published Research Findings are False. PLoS Medicine, 2(8).

Ioannidis, J. P. (2014). How to make more published research true. PLoS Med, 11(10).

Ioannidis, J. P. (2016). Why Most Clinical Research Is Not Useful. PLoS Med, 13(6).

Iraurgi, I. (2009). Evaluación de resultados clínicos (II): Las medidas de la significación clínica o los tamaños del efecto. NORTE DE SALUD MENTAL(34), 94–110.

Iyengar, S., & Greenhouse, J. B. (1988). Selection Models and de File Drawer Problem. Statisticasl Science, 3(1).

Jamil , T., Ly , A., Morey, R. D., Marsman, M., & Wagenmakers, E.-J. (2016). Default “Gunel and Dickey” Bayes factors for contingency tables. Retrieved from Springerlink.com: https://link.springer.com/article/10.3758/s13428-016-0739-8

Jané, M. B., Ben-Shachar, M. S., Moreau, D., Steele, J., Qinyu, X., Caldwell, A. R., . . . Zloteanu, M. (2024). Guide to E□ect Sizes and Con□dence Intervals. Retrieved from https://www.researchgate.net/publication/367462417_Guide_to_Effect_Sizes_and_Confidence_Intervals

Jeffreys, H. (1961). Theory of probability (3rd. ed.). New York, NY: Oxford University Press.

Lock, R. H., Frazer Lock, P., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2021). Statistics - Unlocking the Power of Data. Wiley.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2013). Statistics. Unlocking the Power of Data. Willey.

Macbeth, G., Cortada de Kohan, N., & Razumiejczyk, E. (2007). El Meta-Análisis: La Integración de los Resultados Científicos. Evaluar, 7.

Mangiafico, S. (2023). An R Companion for the Handbook of Biological Statistics, version 1.3.8.

Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian benefits with JASP. European Journal of Developmental Psycholog, 14(5), 545-555.

Meng-Yun , L. (2013). Bayesian Statistics. https://www.bu.edu/sph/files/2014/05/Bayesian-Statistics_final_20140416.pdf. Boston University School of Public Health.

Molina Arias, M. (2014). La odds ratio puede ser engañosa. Revista de Pediatría, 16, 275-279.

Morales Vallejo, P. (2012, Octubre 3). El tamaño del efecto (effect size):análisis complementarios al contraste de medias. Retrieved 2019, from https://web.upcomillas.es/personal/peter/investigacion/Tama%f1oDelEfecto.pdf

Morey, R. D., & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. Psychological Methods.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews, 82, 591-605.

Newcombre, R. G. (1998). Two-sided Confidence Intervals for the single Proportion: Comparison of seven Methods. Statistics in Medicine.

Nunnaly , J. (1960). The Place of Statistics in Psychology. Educational and Psychological Measurement, XX(4).

Pardo, A., & San Martín, R. (1994). Análisis de datos en Psicología II. Madrid: Pirámide.

Pértegas Díaz, S., & Pita Fernández, S. (2003). Cálculo del poder estadístico de un estudio. Cad Aten Primaria, https://www.fisterra.com/mbe/investiga/poder_estadistico/poder_estadistico.asp,

59-63.

Quintana, D. S., & Williams, D. R. (2018). Quintana,Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using JASP. BMC Psychiatry, https://doi.org/10.1186/s12888-018-1761-4.

Rita, H., & Komonen, A. (2008). Odds ratio: an ecologically sound tool to compare proportions. Ann. Zool. Fennici, 45, 66-72.

Rouder, J. N., Speckman , P. L., Dongchu , S., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review, 16(2), Psychonomic Bulletin & Review.

Sanchez-Meca, J., Marín-Martínez, F., & Salvador Chacón-Moscoso, S. (2003). Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis. Psychological Methods, 8(4).

Sawilowsky, S. S. (2009). New effect size rules of thumb. Journal of Modern Applied Statistical Methods, 8(2), 597 – 599.

Schmidt, C. (2008). When to use the odds ratio or the relative risk. International Journal of Public Health, 165-167.

SEH-LELHA. (2003). Heterogeneidad entre los estudios incluidos en un meta-análisis. Retrieved from Liga española para la lucha contra la hipertensión arterial: https://www.seh-lelha.org/heterogeneidad-los-estudios-incluidos-meta-analisis/

Tortosa Ybáñez, M. T., Alvarez Teruel, J. D., & Pellín Buades, N. (2015). Necesidad de re-educación estadística en profesores universitarios: errores de interpretación valores p y tamaño del efecto. (X. J. cambio, Ed.) Retrieved from Repositorio Institucional de la Universidad de Alicante: http://rua.ua.es/dspace/handle/10045/49978

Wuensch's SPSS Links Page. (2019, 10). Retrieved from http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Programs.htm

Zaiontz, C. (2023, 12 12). Real Statistics Using Excel. Retrieved from https://real-statistics.com/

z-test for independent proportions: Use & misuse. (2019, 11). Retrieved from InfluentialPoints:
https://influentialpoints.com/Training/Statistics_bibliography_3_stats.htm#fli