CERES – *Health & Education Medical Journal*
ISSN: 2965-5722
2025, Vol. 3, N°. 2, p. 01-12
DOI:10.62234/ceresv3n2-003

El efecto relativo generalizado (ERG): su introducción y un tutorial en R

The generalized relative effect (GRE): its introduction and a tutorial in R

Jimmie Leppink [1]

[1] Leppink Analytics, Ondara – Alicante, Spain. Orcid: https://orcid.org/0000-0002-8713-1374

Autor Correspondiente:
Jimmie Leppink

Calle Hortetes, 5, Piso 2, Puerta 11, 03760 Ondara – Alicante, Spain.

Correo electrónico: leppinkanalytics@gmail.com

RESUMEN

Introducción: Cuando se comparan dos muestras en un resultado de interés, disponer de una estadística que pueda (1) aplicarse a todos los niveles de medición, (2) tener en cuenta la información de orden en variables ordinales y cuantitativas, (3) proporcionar resultados válidos para muestras de distintos tamaños y (4) tener en cuenta la información previa de la teoría o de investigaciones anteriores facilita la investigación cuantitativa y con métodos mixtos. Aunque existen varias estadísticas no paramétricas, todas se quedan cortas en al menos una de estas áreas. **Objetivo:** Este artículo tiene un doble objetivo: (1) introducir un método no paramétrico llamado 'efecto relativo generalizado' (ERG) que funciona bien en todas las áreas mencionadas y (2) proporcionar un tutorial de cómo calcular una estimación puntual del ERG y su correspondiente intervalo de credibilidad en R. **Marco teórico:** El ERG aúna los puntos fuertes de una variante bayesiana del porcentaje de todos los datos no solapados (PAND-B) y efecto relativo (ER) de Brunner-Munzel. Por un lado, el ER puede proporcionar resultados válidos para resultados dicotómicos, ordinales multicategoría y cuantitativos para tamaños de muestra iguales y desiguales, pero no puede aplicarse a resultados nominales multicategoría y no tiene en cuenta la información previa. Por otro lado, PAND-B es aplicable a todos los niveles de medición y tiene en cuenta la información previa, pero no tiene en cuenta adecuadamente la información de orden en los resultados ordinales multicategoría y cuantitativos, y tiende a exagerar las diferencias cuando los tamaños de las muestras son desiguales. **Método:** A través de un ejemplo simulado con un resultado ordinal de cuatro categorías y tamaños de muestra claramente desiguales, este trabajo compara PAND-B, el ER y el ERG. **Resultados y discusión:** El ERG reúne lo mejor de PAND-B y del ER y evita las debilidades de estas estadísticas. El ERG es fácil de implementar en código R básico, y se presentan directrices para la interpretación de los resultados del ERG en términos de 'pequeño', 'medio' y 'grande'. **Conclusión:** Dada su aplicabilidad a todos los niveles de medición y a tamaños de muestra iguales y desiguales, el ERG proporciona un enfoque coherente para la interpretación de la magnitud de las diferencias entre dos muestras en resultados cualitativos, cuantitativos o mixtos de resultados.

HISTÓRIA DO ARTIGO

Received 30 April 2025

Accepted 6 May 2025

PALABRAS CLAVE

Estadística; tamaño del efecto; métodos no paramétricos; métodos bayesianos; niveles de medición

ABSTRACT

KEYWORDS

Introduction: When comparing two samples on an outcome of interest, having a statistic that can (1) be applied to all levels of measurements, (2) account for order information in ordinal and quantitative variables, (3) provide valid outcomes for varying sample sizes, and (4) account for prior information from theory or previous research facilitates quantitative and mixed-methods research. Although several non-parametric statistics are available, they all fall short in at least one of these areas. **Objective:** The aim of this paper is twofold: (1) to introduce a non-parametric method called 'generalized relative effect' (GRE) that performs well in all of the aforementioned areas, and (2) to provide a tutorial of how to compute a GRE point estimate and its corresponding credible interval in R. **Theoretical Framework:** GRE unites the strengths of a Bayesian variant of the percentage of all non-overlapping data (PAND-B) and the Brunner-Munzel relative effect (RE). On the one hand, RE can provide valid outcomes for dichotomous, multicategory ordinal, and quantitative outcomes for equal and unequal sample sizes, but cannot be applied to multicategory nominal outcomes and does not account for prior information. On the other hand, PAND-B is applicable to all levels of measurement and accounts for prior information, but does not adequately account for order information in multicategory ordinal and quantitative outcomes, and tends to exaggerate differences when sample sizes are unequal. **Method:** Through a simulated example with a four-category ordinal outcome and clearly unequal sample sizes, this paper compares PAND-B, RE, and GRE. **Results and Discussion:** GRE unites the best of PAND-B and RE and avoids the weaknesses of these statistics. GRE is easy to implement in basic R-code, and guidelines for the interpretation of GRE outcomes in terms of 'small', 'medium', and 'large' are discussed. **Conclusion:** Given its applicability to all levels of measurement and to both equal and unequal sample sizes, GRE provides a consistent approach to the interpretation of the magnitude of differences between two samples in qualitative, quantitative, or mixed sets of outcomes.

Statistics; effect size; non-parametric methods; Bayesian methods; levels of measurement

INTRODUCCIÓN

Las comparaciones de dos muestras sobre uno o varios resultados de interés – cualitativos, cuantitativos o mixtos – constituyen una parte indispensable de la salud, la medicina, la educación y muchos otros campos de investigación. Aunque los investigadores disponen de muchas estadísticas paramétricas y no paramétricas para tales comparaciones, todos ellas se basan en suposiciones y tienen sus puntos fuertes y débiles. Se pueden distinguir al menos cuatro características deseables de estadísticas para comparaciones de dos muestras.

Características Principales de Estadísticas para Comparaciones de dos Muestras

Para empezar, los resultados de interés pueden ser nominales (ej.: zapatos azules, blancos o rojos), ordinales (ej.: rendimiento pobre, al límite, satisfactorio o excelente), de intervalo (ej.: puntuaciones en un test de inteligencia, asumiendo que un '0' absoluto no es un resultado posible) o de razón (ej.: precios de las acciones en la bolsa).[1] Casi todas las estadísticas para comparar dos muestras tienen en común que no se pueden aplicar a resultados de al menos uno de estos niveles de medición.

Una segunda característica deseable de una estadística para comparaciones de dos muestras, al menos en lo que se refiere a variables ordinales multicategoría y cuantitativas (es decir, nivel de medición de intervalo o razón), es un tratamiento adecuado de la información de orden. Por ejemplo,

un rendimiento excelente es superior a un rendimiento satisfactorio, un rendimiento satisfactorio es mejor que un rendimiento al límite, y este último sigue siendo algo mejor (aunque quizás no mucho) que un rendimiento pobre; una estadística que se supone útil para resultados ordinales multicategoría (y/o cuantitativos) debería tener en cuenta dicha información de orden. Al contrario, una estadística que trata el resultado de rendimiento mencionado como nominal pretende que las cuatro categorías de rendimiento son categorías distinguibles que no se pueden ordenar de menor a mayor, y este tratamiento inadecuado de un resultado ordinal tiende a producir conclusiones incorrectas. La única excepción se puede encontrar en los resultados dicotómicos; dada una codificación transparente y coherente (por ejemplo, codificación *dummy* 0/1), el orden de las categorías – al menos para el análisis en cuestión – está claro, y las diferencias entre dos muestras en el resultado dicotómico se deberían poder interpretar de forma inequívoca.

En adición a las características de nivel de medición e información de orden, una estadística para comparar dos muestras debe proporcionar resultados válidos independientemente de si las muestras son de tamaño igual o desigual. Aunque muchas de las estadísticas disponibles actualmente para comparaciones de dos muestras cumplen este requisito, como se analiza en la siguiente sección, no todas las estadísticas encontradas en la bibliografía lo hacen.

Por último, puesto que la investigación suele consistir en contribuir a la teoría y a la investigación disponible, y en el proceso suele estar influida por la teoría y la investigación previa, la capacidad de incorporar información procedente de la teoría y/o la investigación previa en el análisis de las diferencias entre grupos es una característica deseable de cualquier estadística.[2]

Estadísticas Clave para Comparaciones de dos Muestras

Aunque los investigadores tienen a su disposición una gran variedad de estadísticas para comparaciones de dos muestras, todas se quedan cortas en una o varias de las cuatro características deseables anteriormente comentadas. Sin embargo, algunas de las estadísticas disponibles son prometedoras, especialmente cuando se combinan con alternativas específicas.

Por un lado, Karch [3,4] proporciona argumentos convincentes de por qué se debería utilizar la prueba de Brunner-Munzel (BM) en lugar de la prueba de Mann-Whitney ampliamente utilizada, ya que varias suposiciones subyacentes a esta última no son realistas y su violación puede fácilmente dar lugar a conclusiones incorrectas. En la prueba de BM el efecto relativo (ER) es clave. En pocas palabras, cada observación de la muestra A se compara con cada observación de la muestra B y, para cada comparación, el resultado es uno de los siguientes: mayor en B, mayor en A o igual (en deportes como el fútbol, un empate). Por ejemplo, si se comparan dos muestras de n = 10 cada una, hay 10 x 10 = 100 comparaciones. Supongamos que 60 de estas comparaciones favorecen a la muestra A (es decir, son más altas en A), 30 favorecen a la muestra B (es decir, son más altas en B) y las 10 comparaciones restantes resultan en un empate. Los empates se dividen a partes iguales entre las dos muestras, lo que da como resultado 65 comparaciones a favor de A y 35 comparaciones a favor de B. Para esta comparación, el ER iguala 65 / 100 = 0,65 si se formula a favor de A y 35 / 100 = 0,35 si se formula a favor de B. En cualquier caso, hay un 65% de probabilidad de que una observación seleccionada al azar en la muestra A sea superior a una observación seleccionada al azar en la muestra B, y un 35% de probabilidad de que una observación seleccionada al azar en la muestra B sea superior a una observación seleccionada al azar en la muestra A. Como tal, el ER proporciona una estadística del tamaño del efecto bastante fácil de interpretar para comparaciones de dos muestras en resultados dicotómicos, ordinales multicategoría y cuantitativos. Sin embargo, dado que las categorías

nominales múltiples no se pueden ordenar de forma no arbitraria, la prueba de BM y su ER no se pueden aplicar a resultados nominales multicategoría, a menos que ese resultado nominal multicategoría se convierta en *dummies* 0/1 para cada categoría y se calcule el ER para cada categoría, lo que plantearía desafíos para la interpretación de los resultados. Por último, la incorporación de información previa no es actualmente una opción en la prueba de BM y su ER.

Por otro lado, una estadística de uso común en algunas áreas de investigación es el porcentaje de todos los datos no solapados (PAND).[5] Esta estadística responde a la pregunta de qué grado de 'no solapamiento' hay entre dos muestras en un resultado de interés. Por ejemplo, si en un grupo de 20 alumnos (muestra A) 19 llevan zapatos azules y uno lleva zapatos rojos, y en un segundo grupo de 20 alumnos (muestra B) 19 llevan zapatos rojos y uno lleva zapatos azules, estamos a sólo dos puntos de datos del no solapamiento perfecto: el alumno con zapatos rojos en la muestra A y el alumno con zapatos azules en la muestra B. En consecuencia, PAND es 95% o 0,95. Dada una codificación transparente y coherente, se puede demostrar que en el caso de resultados dicotómicos y tamaños de muestra iguales, el ER y PAND producen la misma estimación puntual y que el coeficiente de correlación de Pearson y Spearman son exactamente o aproximadamente iguales a: (2 x PAND) – 1. En otras palabras, las correlaciones de -0,5, 0, 0,1, 0,3 y 0,5 se corresponden, al menos aproximadamente, con valores de PAND y ER de 0,25, 0,5, 0,55, 0,65 y 0,75 respectivamente.

Sin embargo, también se puede demostrar fácilmente que cuando los tamaños de las muestras son desiguales, el coeficiente de correlación sigue siendo bastante cercano a '(2 x ER) – 1', aunque PAND exagera la diferencia. Consideremos el siguiente ejemplo: en la clase A, 21 alumnos suspenden y 7 aprueban el examen, mientras que en la clase B, 9 alumnos suspenden y 3 aprueban el examen. Para este ejemplo, encontramos una correlación de 0 y ER = 0,5, lo cual es correcto ya que la distribución de aprobados / suspendidos es 1:3 en ambas clases. Sin embargo, PAND iguala 0,6 en este caso, como si encontráramos ER = 0,6 y una correlación de aproximadamente 0,2 (¡!). Aunque una versión bayesiana de PAND – llamada PAND-B – que trata PAND como una variable binomial (es decir: cada observación puede ser un punto de 'éxito' no solapado o un punto 'fracaso' solapado) y añade una distribución Binomial a priori, ayuda a tener en cuenta la información previa (en ausencia de información previa: un éxito y un fracaso) [6,7], no elude el problema de PAND ante tamaños de muestra desiguales. Además, aunque puede proporcionar una estadística útil del tamaño del efecto para todos los niveles de medición, incluido el nominal, siempre que los tamaños muestrales sean iguales, no tiene en cuenta la información de orden como hace el ER. En otras palabras, aunque en el caso de tamaños de muestra iguales PAND y RE proporcionan la misma estimación puntual cuando los resultados son dicotómicos, se puede anticipar que el ER funcione de forma más adecuada cuando se comparan dos muestras en resultados ordinales multicategoría o cuantitativos.

En resumen, aunque las ventajas de PAND-B sobre el ER son la posibilidad de incorporar información previa y su aplicabilidad a los resultados de todos los niveles de medición, incluidos los resultados nominales multicategoría, el ER es preferible cuando los tamaños de las muestras son desiguales y/o cuando los resultados son ordinales multicategoría o cuantitativos. Si estamos de acuerdo en las características deseables de una estadística para comparaciones de dos muestras presentadas al principio de este artículo, una pregunta que surge es si existe una forma de unir las fortalezas relativas – y evitar las limitaciones – de PAND-B y del ER en una única estadística. Y aquí es donde se introduce el efecto relativo generalizado (ERG).

Uniendo las Fortalezas y Evitando las Limitaciones de las Estadísticas Disponibles Actualmente

Dada la división equitativa de los empates entre ambas muestras en el cálculo del ER, todas las comparaciones por pares de observaciones se reducen esencialmente a estar a favor de una muestra o de la otra. Sin embargo, el número de comparaciones por pares es mucho mayor que el tamaño total de la muestra; iguala al producto de los tamaños de las muestras. Como en un ejemplo anterior, cuando dos muestras tienen n = 10, hay 10 x 10 = 100 comparaciones. Alternativamente, si una muestra tiene n = 5 y la otra muestra n = 20, también tenemos 100 comparaciones, aunque por una vía diferente: 5 x 20. En consecuencia, para reescalar los números a favor de la muestra A y a favor de la muestra B (y los empates, que al final se dividen por igual entre A y B) al tamaño original de la muestra, tenemos que dividir los recuentos resultantes de todas las comparaciones por pares por el cociente del producto de los tamaños de las muestras (por ejemplo, 100 en el caso de dos muestras de tamaño 10) y el propio tamaño de la muestra (en este caso: 20). En el ejemplo anterior de 65 comparaciones a favor de A y 35 comparaciones a favor de B, eso significa dividir 65 y 35 entre (100 / 20), lo que da como resultado 13 y 7 respectivamente. Estas cifras se corresponden con proporciones de 0,65 y 0,35, aunque en una escala de un tamaño de muestra de 20 en lugar de en un recuento de 100 comparaciones. Añadiendo una distribución binomial a priori (en ausencia de información previa: un éxito y un fracaso) análoga a PAND-B [6], tenemos una alternativa bayesiana al ER de BM con una estimación puntual cercana a la del ER (la distribución a priori tiene cierta influencia, aunque menor con tamaños de muestra crecientes) y un intervalo de credibilidad del 90% o del 95% similar al intervalo de confianza del 90% o del 95% en torno a la estimación puntual del ER. En este caso, si se utiliza una distribución a prior de '1 éxito (A > B) y 1 fracaso (B > A)', la distribución posterior resultante sería:

B(13,7) + B(1,1) = B(14,8).

Se trata de una distribución posterior con una mediana (estimación puntual) de 0,641 y un intervalo de credibilidad del 95% de 0,430 a 0,819, lo cual es un intervalo muy amplio, ya que el tamaño de la muestra en este ejemplo es de sólo N = 20. El intervalo incluye 0,5, que representa la hipótesis nula por defecto (H_0) de que los grupos son comparables (es decir, un 50% de probabilidad de que 'B > A' y un 50% de probabilidad de que 'A > B', como en un partido de fútbol en el que, dados los rendimientos recientes de que se dispone, ambos equipos tienen la misma probabilidad de ganar).

Además, la estimación puntual no es exactamente 0,65, porque la distribución a priori no informativa de B(1,1) tiene una mediana de 0,5 y, por lo tanto, desplaza la estimación puntual ligeramente hacia 0,5. Sin embargo, como se ha explicado anteriormente, el peso de la distribución a prior en relación con los datos disminuye al aumentar el tamaño de la muestra, como también queda claro en el ejemplo más adelante en este artículo.

Como para el ER, el proceso de cálculo descrito anteriormente para el ERG funciona para resultados dicotómicos, ordinales multicategoría y cuantitativos, siempre y cuando la codificación sea transparente y coherente. Sin embargo, dado que para las variables nominales multicategoría no se puede establecer ningún sistema de codificación único y no arbitrario – que no sea, como ha comentado anteriormente, crear variables *dummy* para cada categoría y calcular para cada *dummy* – el ERG se debe calcular del siguiente modo. Supongamos que el ejemplo de 60 'A > B', 30 'B > A' y 10 'A = B' presentado anteriormente correspondiera a un resultado nominal multicategoría: en ese caso, no podríamos distinguir qué diferencias significan 'A > B' y cuáles deberían entenderse como 'B > A'. En este caso, en lugar de dividir por igual los empates entre 'B > A' y 'A > B', todas las

comparaciones que den como resultado 'A ≠ B' se deben interpretar como que A y B son *diferentes*, mientras que todas las comparaciones que den como resultado 'A = B' se codifican como que A y B son *iguales*. Para el ejemplo utilizado anteriormente, esto significaría 90 veces 'diferentes' y 10 veces 'iguales', lo que reescalado al tamaño original de la muestra (es decir, dividiendo por 100 / 20) nos da 18 y 2 respectivamente. En este caso, si se utiliza una distribución a prior de '1 éxito (A ≠ B) y un fracaso (B = A), la distribución posterior resultante sería:

$B(18,2) + B(1,1) = B(19,3)$.

Se trata de una distribución posterior con una mediana (estimación puntual) de 0,875 y un intervalo de credibilidad del 95% de 0,696 a 0,970. Sin embargo, hay que tener en cuenta que la interpretación de esta estimación puntual y del intervalo es diferente del anterior: en este caso, no se trata de la probabilidad de que una observación aleatoria de una muestra supere a una observación aleatoria de la otra muestra, sino de la probabilidad de que una observación aleatoria de una muestra sea *diferente* de una observación aleatoria de la otra muestra. Especialmente en el caso de resultados cuantitativos para los que es posible un rango de valores, cabe esperar altos niveles de diferencia. Además, este segundo intervalo no tiene en cuenta la información de orden en los resultados ordinales multicategoría y cuantitativos. Aunque especialmente cuando un resultado ordinal sólo tiene un número limitado de categorías (ej.: tres o cuatro), este segundo intervalo puede ayudar a proporcionar información útil sobre el grado de *empates* (es decir, igualdad) en el conjunto de comparaciones por pares, el principal interés cuando se trata de un resultado ordinal (como para un resultado cuantitativo y un resultado dicotómico codificado de forma coherente) suele residir en las interpretaciones de 'mejor-peor' más que en cualquier diferencia independientemente de su dirección.

En la sección de 'Métodos y Resultados' de este artículo, se explican ambos procesos computacionales con un ejemplo de datos simulados y un código R básico que se puede utilizar para obtener medianas posteriores e intervalos de credibilidad del X% (por ejemplo: 95%, 90% u 89%) para cada tipo de resultado. Además, para comparar PAND-B, el ER y el ERG en cuanto a las cuatro características deseadas de una estadística para comparaciones de dos muestras esbozadas al principio de este artículo, el ejemplo utilizado en el resto de este artículo es una comparación de dos muestras de tamaño claramente desigual en un resultado ordinal de cuatro categorías.

MÉTODOS Y RESULTADOS

Supongamos que una muestra de 130 estudiantes se divide aleatoriamente en una condición de control y una condición de tratamiento experimental. En la condición de control, los estudiantes se preparan para una estación de ECOE como lo hacen habitualmente, mientras que a los estudiantes de la condición de tratamiento experimental se les incita a utilizar un enfoque de razonamiento innovador. Tras el entrenamiento (es decir, la preparación), todos los estudiantes completan la misma estación de ECOE y un evaluador independiente (que no conoce las dos condiciones ni qué estudiante estaba en qué condición) les califica entre otros en la siguiente escala de rendimiento global: rendimiento pobre, al límite, satisfactorio o excelente. Aunque este tipo de escalas se trata con frecuencia como si tuviera un nivel de medición de intervalo, en realidad se trata de un resultado ordinal. Es decir, ¿qué garantía tenemos de que la diferencia entre cada dos categorías consecutivas de rendimiento sea siempre la misma? Además, para el estudio que nos ocupa, la investigación sólo dispone de recursos para 30 estudiantes en la condición de tratamiento experimental, por lo que la asignación

aleatoria a las condiciones se realiza de tal forma que 100 estudiantes terminan en la condición de control. Supongamos que los resultados son como presentados en la Tabla 1.

Tabla 1. Resultados del estudio hipotético ejemplar.

		Control	**Experimental**
		***n* = 100 (76,9%)**	***n* = 30 (23,1%)**
Rendimiento	Pobre	5 (5,0%)	1 (3,3%)
	Al límite	20 (20,0%)	4 (13,3%)
	Satisfactorio	65 (65,0%)	18 (60,0%)
	Excelente	10 (10,0%)	7 (23,3%)

Los porcentajes de las celdas indican números proporcionalmente ligeramente inferiores de calificaciones 'pobres' y 'al límite' y algo más de calificaciones 'excelentes' en la condición de tratamiento experimental. El coeficiente rho de Spearman y el coeficiente Tau-B de Kendall son 0,149 y 0,142 respectivamente, positivos porque hay una ligera tendencia hacia las diferencias a favor de la condición de tratamiento experimental y no al revés.

Para el ER, gracias a Karch[4] existe un paquete fácil de usar en *jamovi*[8], que para los datos del ejemplo actual proporciona una estimación puntual de 0,588, un intervalo de confianza del 95% de [0,480; 0,695] y un intervalo de confianza del 90% de [0,498; 0,677]. Aplicando la regla empírica de '(2 x PAND) – 1' introducida anteriormente, esto se traduciría en una correlación de 0,176, que no coincide ni con el coeficiente de Spearman ni con el de Kendall, pero se aproxima lo suficiente y daría lugar al mismo tipo de interpretación en cuanto a la fuerza de la asociación: relativamente débil.

Dado que PAND y su homólogo bayesiano PAND-B se tratan de la forma más sencilla de lograr un no solapamiento perfecto [5-7] y en el caso de resultados dicotómicos, ordinales multicategoría o cuantitativos en una dirección determinada (aquí: a favor de la condición de tratamiento experimental) [6], eliminar o intercambiar las observaciones 'pobres', 'al límite' y 'satisfactorias' en el grupo de tratamiento experimental (es decir, 1 + 4 + 18) y las observaciones 'excelentes' en el grupo de control (es decir, 10) daría como resultado un no solapamiento perfecto. En otras palabras, hay 1 + 4 + 18 + 10 = 33 observaciones que contribuyen al solapamiento, por lo que PAND iguala:

PAND = (130 – 33) / 130 = 0.746.

Aplicando la regla empírica de '(2 x PAND) – 1' introducida anteriormente, esto se traduciría en una correlación de 0,492, que por supuesto para los datos a mano está muy lejos de donde debería estar. Y para PAND-B encontraríamos:

B(97,33) + B(1,1) = B(98,34).

Se trata de una distribución posterior con una mediana (estimación puntual) de 0,744 y un intervalo de credibilidad del 95% de 0,665 y 0,813. Aplicando la regla empírica de '(2 x PAND) – 1' introducida anteriormente, esto se traduciría en una correlación de 0,488, que sigue muy lejos de donde debería estar una correlación calculada sobre este conjunto de datos.

Sin embargo, para el ERG procedemos de la siguiente manera. En primer lugar, en el **paso 1** tenemos que leer un archivo de datos, por ejemplo en formato '.csv', y definir nuestras variables:

```
# datos
data <- read.csv("file-location")
group <- data$group
```

```
value <- data$value
```

A continuación, en el **paso 2** tenemos que dividir nuestro conjunto de datos en dos grupos y determinar el tamaño de la muestra (en este caso: N = 130) para determinar el número de comparaciones por pares (en este caso: 100 x 30 = 3.000) y, posteriormente, volver a ajustar el tamaño de la muestra original como se ha explicado anteriormente:

```
# dividir los datos
c <- subset(data, group == "c")
e <- subset(data, group == "e")
# definir el tamaño de la muestra
k <- nrow(data)
```

En este punto, en el **paso 3** ejecutamos todas las comparaciones por pares, calculamos una estadística de diferencia y categorizamos cada diferencia en términos de 'Experimental > Control' (es decir, positivo, en el siguiente código 'pos'), 'Control > Experimental' (es decir, negativo, en el siguiente código 'neg') o empate:

```
# todas las comparaciones por pares
pairwise <- merge(c, e, by = NULL, suffixes = c("_c", "_e"))
# la diferencia por pares
difference <- pairwise$value_e - pairwise$value_c
# tres categorías
pos <- ifelse(difference > 0, 1, 0)
tie <- ifelse(difference == 0, 1, 0)
neg <- ifelse(difference < 0, 1, 0)
```

A continuación, en el **paso 4** tenemos que definir nuestra distribución a priori, B(1,1) en ausencia de cualquier información previa:

```
# los parámetros a priori
a <- 1
b <- 1
```

Penúltimo, en el **paso 5** dividimos por igual los empates entre las diferencias 'positivas' (a favor del tratamiento experimental) y 'negativas' (a favor de la condición control) y reescalamos al tamaño original de la muestra:

```
# suponiendo categorías ordinales, a favor del tratamiento experimental (e)
# datos de comparación por pares ajustados al tamaño original de la muestra
x <- k * (mean(pos) + 0.5 * mean(tie))
n <- k
re <- x / n
```

Y, por último, en el **paso 6** obtenemos la distribución posterior con su mediana (es decir, la estimación puntual del ERG) y el intervalo de credibilidad del X% deseado (aunque el siguiente ejemplo demuestra el 95%, también se puede calcular cualquier otro intervalo de credibilidad del X% deseado):

```
# parámetros posteriores
post_a <- a + x
post_b <- b + n - x
# mediana posterior
posterior_median <- qbeta(0.5, post_a, post_b)
posterior_median
# el intervalo de credibilidad del 95%
e_ci95 <- qbeta(c(0.025, 0.975), post_a, post_b)
e_ci95
```

Al hacerlo encontramos una estimación puntual del ERG (es decir, una mediana posterior) de 0,587, que casi coincide con la estimación puntual RE. Aplicando la regla empírica de '(2 x PAND) – 1' introducida anteriormente, esto se traduciría en una correlación de 0,174, que no coincide ni con el coeficiente de Spearman (0,149) ni con el de Kendall (0,142) pero se aproxima lo suficiente y daría lugar al mismo tipo de interpretación en cuanto a la fuerza de la asociación: relativamente débil. Además, en cuanto a la estimación del intervalo, encontramos un intervalo de credibilidad del 95% de [0,501; 0,668], un intervalo de credibilidad del 90% de [0,515; 0,656] y un intervalo de credibilidad del 89% de [0,517; 0,654]. Debido a la distribución a priori, un intervalo de credibilidad del X% es ligeramente más estrecho que su homólogo frecuentista, el intervalo de confianza del X%.

Por último, si estas cuatro categorías de rendimiento no fueran ordinales sino nominales, o además de una tendencia 'mejor-peor' nos interesara la distribución de valoraciones 'diferentes' frente a 'iguales', sólo *la primera línea de código* del **paso 5** es *ligeramente* diferente:

```
# datos de comparación por pares ajustados al tamaño original de la muestra
x <- k * (mean(pos) + mean(neg))
n <- k
re <- x / n
```

Es decir, en lugar de 'x <- k * (mean(pos) + 0.5 * mean(tie))' se escribe 'x <- k * (mean(pos) + mean(neg))' porque en esta segunda variante estamos interesados en la proporción de *diferencias* en vez de en una tendencia 'mejor-peor'. Con esta pequeña diferencia en una línea de codificación – nada más que sustituir '0.5 * mean(tie))' por 'mean(neg))' se obtiene una estimación puntual del ERG y el intervalo de credibilidad del X% deseado para diferente frente a igual, que es una variante del ERG que se puede utilizar para *todos tipos de resultados*, incluidos los nominales multicategoría. Para el ejemplo que nos ocupa, se obtiene una mediana posterior (es decir, una estimación puntual del ERG) de 0,558, un intervalo de credibilidad del 95% de [0,472; 0,641], un intervalo de credibilidad del 90% de [0,486; 0,628] y un intervalo de credibilidad del 89% de [0,488; 0,626].

En conclusión, aunque desde una perspectiva ordinal el intervalo de credibilidad del 95% para el ERG está completamente por encima de H_0: $p = 0.5$ (es decir, los dos grupos son comparables en

rendimiento), lo que indica un mejor rendimiento en la condición de tratamiento experimental, si desde una perspectiva nominal tuviéramos que probar H_0: $p = 0.5$ (es decir, un 50% de probabilidad de diferencia y un 50% de probabilidad de empate), concluiríamos que no hay pruebas suficientes de que haya más de un 50% de probabilidad de diferencia.

DISCUSIÓN

Este artículo partió de los argumentos de que (1) las comparaciones de dos muestras son muy comunes en muchos campos de investigación en estudios cualitativos, cuantitativos y de métodos mixtos, y (2) una estadística para comparaciones de dos muestras debería ser aplicable a todos los niveles de medición, tener en cuenta la información de orden en resultados ordinales y cuantitativos, proporcionar resultados válidos para tamaños de muestra variables y tener en cuenta la información previa. El ER proporciona resultados válidos para resultados dicotómicos, cuantitativos y ordinales multicategoría para tamaños de muestra iguales y desiguales, pero no es aplicable a resultados nominales multicategoría de forma directa. Y aunque PAND y PAND-B son aplicables a estos últimos y PAND-B permite la incorporación de información previa, ambas estadísticas de no solapamiento no tienen en cuenta adecuadamente la información de orden y se quedan claramente cortas ante tamaños muestrales desiguales. El ERG aúna los puntos fuertes del ER y de PAND-B y evita las limitaciones de ambas.

Aunque el ERG sigue en esencia el proceso computacional del ER de BM (es decir, una comparación por pares de cada observación en una muestra con cada observación en la otra muestra), añade una distribución a priori y, con sólo un pequeño ajuste (es decir, sustituyendo un término en una línea de código), se puede utilizar también para resultados nominales multicategoría. Dada su proximidad al ER, el ERG tiene en cuenta la información de orden como el ER y, ante tamaños de muestra desiguales, no produce estimaciones extrañas como las de PAND y PAND-B. Aplicando la regla empírica de '(2 x PAND) – 1' introducida anteriormente, las estimaciones del ERG de 0,25, 0,35, 0,45, 0,50, 0,55, 0,65 y 0,75 se corresponden aproximadamente con correlaciones de -0,5, -0,3, -0,1, 0, 0,1, 0,3 y 0,5 respectivamente. Por lo tanto, si en un contexto de investigación determinado se interpretan las correlaciones en torno a 0,1, 0,3 y 0,5 como 'pequeñas', 'medianas' y 'grandes' respectivamente, los valores del ERG en torno a 0,55 indican diferencias 'pequeñas', los valores del ERG en torno a 0,65 son indicativos de diferencias 'medianas' y los valores del ERG en torno a 0,75 se pueden interpretar como diferencias relativamente 'grandes'. Por otra parte, además de las estimaciones puntuales, se pueden utilizar intervalos de credibilidad para indicar rangos de hipótesis plausibles.

A diferencia de las pruebas *t* y otros enfoques paramétricos, las pruebas del ER y del ERG pueden tener en cuenta información previa sin los requisitos de una distribución Normal (es decir, en forma de campana) y homocedástica (es decir, de igual varianza). Dicho esto, cuando se aplica a resultados cuantitativos, la distribución de diferencias de comparación por pares se puede visualizar (por ejemplo con histogramas, gráficos de caja o nubes) para inspeccionar el grado de asimetría en esa distribución. Si se comparan dos muestras con una distribución aproximadamente Normal, cabe anticipar que la distribución de las diferencias de comparación por pares también se aproxime a una distribución Normal, mientras que la fuerte asimetría y/o los casos extremos en al menos una de las muestras también se reflejarán normalmente en una o ambas colas de la distribución de la las diferencias de comparación por pares.

Además, aunque el ejemplo de este artículo utiliza un diseño unidireccional – y se podrían dedicar futuras simulaciones y estudios con datos reales a cuestiones como el rendimiento y el poder estadístico del ERG en distintas circunstancias – el ERG también se puede aplicar a diseños 2 x 2 equilibrados. Es decir, en un diseño equilibrado de 2 x 2, las cuatro celdas tienen tamaños de muestra iguales, y el efecto principal de cada factor, así como su efecto de interacción, se pueden estimar de forma independiente. El efecto principal del factor A se calcula comparando las dos celdas en las que A = 1 frente a las dos celdas en las que A = 0, el efecto principal del factor B se calcula comparando las dos celdas en las que B = 1 frente a las dos celdas en las que B = 0, y una de las varias formas de calcular el efecto de interacción es contrastar '[A = 0 y B = 0] y [A = 1 y B = 1]' con '[A = 0 y B = 1] y [A = 1 y B = 0]'. De esta forma, el ERG puede – al menos para diseños 2 x 2 equilibrados – proporcionar estimaciones puntuales y de intervalo tanto para los efectos principales como para el efecto de interacción de A y B, independientemente de si el resultado es nominal, ordinal o cuantitativo. Esta generalización no es válida para diseños claramente desequilibrados, porque en estos últimos hay una correlación de diseño entre los factores y, por lo tanto, los efectos principales y de interacción ya no se pueden estimar de formar independiente.

Por último, el ERG también puede proporcionar un enfoque prometedor para las comparaciones multigrupo en diseños unidireccionales, ya sea aplicando el ERG a todos los pares de grupos posibles (en este caso, muchos estadísticos defenderán algún tipo de corrección para pruebas múltiples) o – si se dispone de una hipótesis alternativa dirigida específica – a comparaciones específicas. Esta extensión, así como la de los diseños 2 x 2 equilibrados, debería ser objeto de más investigación, entre otros porque comparaciones de múltiples grupos puede crear desafíos computacionales, especialmente en el caso de muestras más grandes.

CONCLUSIÓN

Dada su aplicabilidad a todos los niveles de medición y a tamaños de muestra iguales y desiguales, el ERG proporciona un enfoque coherente para la interpretación de la magnitud de las diferencias entre dos muestras en resultados cualitativos, cuantitativos o mixtos. Tanto si se trata de resultados cualitativos como cuantitativos, o si el interés radica en una integración de métodos mixtos de resultados cualitativos y cuantitativos, se puede incorporar información previa e interpretar las diferencias en términos de 'pequeñas', 'medianas' y 'grandes'. Además, se podría investigar más a fondo su generalización a comparaciones de más de dos grupos, así como a efectos principales y de interacción en diseños 2 x 2 equilibrados.

AGRADECIMIENTOS

El autor desea agradecer a la Dra. Patricia Pérez-Fuster sus perspicaces conversaciones sobre estadísticas para muestras pequeñas. Estas conversaciones contribuyeron al desarrollo de PAND-B hace años y crearon la base para la integración del ER y PAND-B en el ERG como descrito en este artículo.

FINANCIACIÓN

Este artículo y las investigaciones que han contribuido a él no han recibido ningún tipo de financiación en ningún momento.

CONFLICTO DE INTERESES

El autor declara no tener ningún conflicto de intereses.

REFERENCIAS

1. Stevens SS. On the theory of scales of measurement. Sci. 1946;26:677-680. http://www.jstor.org/stable/1671815

2. Lindley D. Bayesian statistics: A review. London: SIAM; 1972.

3. Karch JD. Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. Adv. Methods Pract. Psychol. Sci. 2021;4(2). https://doi.org/10.1177/2515245921999602

4. Karch JD. bmtest: A Jamovi Module for Brunner–Munzel's Test — A Robust Alternative to Wilcoxon–Mann–Whitney's Test. Psych. 2023;5:386-95. https://doi.org/10.3390/psych5020026

5. Parker RI, Hagan-Burke KJ, Vannest KJ. Percentage of all non-overlapping data (PAND): An alternative to PND. J Spec Educ. 2007;40(4):194-204. https://doi.org/10.1177/00224669070400040101

6. Leppink J. The art of modelling the learning process: Uniting educational research and practice. Cham: Springer; 2020. https://doi.org/10.1007/978-3-030-43082-5

7. Leppink J. Un modelo bayesiano para datos cualitativos en simulación. Rev Latam Sim Clin. 2021;3(3):117-9. https://doi.org/10.35366/103188

8. The jamovi Project. Jamovi (Version 2.6) [Computer Software]; 2024. Retrieved from https://www.jamovi.org