Una medida de asimetría unidimensional para variables cualitativas

José Moral de la Rubia[1]
Universidad Autónoma de Nuevo León-México

Esta investigación metodológica tiene como objetivo definir un concepto de asimetría para variables cualitativas, cuantificarlo y mostrar su validez. Se usó un panel de cinco jueces expertos y simulaciones Monte Carlo. Se definió el estadístico diferencia promedio de frecuencia (*dpf*) entre pares de categorías ordenadas por homogeneidad de frecuencia. El estadístico *dpf* mostró un comportamiento ajustado a las expectativas con distintas variantes de la distribución binomial. La correlación entre el promedio de las puntuaciones de asimetría de los jueces y *dpf* fue muy alta. Para obtener normas orientativas de interpretación, se simularon 20,000 muestras de tamaños 20, 40, 100, 200, 500 y 1000 extraídas de una distribución binomial. Se concluye que *dpf* es válido para medir asimetría en variables cualitativas.
Palabras clave: asimetría, distribución discreta, escala nominal, variable cualitativa, simulación Monte Carlo.

A measure of one-dimensional asymmetry for qualitative variables
This methodological investigation aims to define a concept of asymmetry for qualitative variables, quantify it, and show its validity. A panel of five expert judges and Monte Carlo simulations were used. The statistic Mean Difference in Frequency (*MDF*) between pairs of categories ordered by frequency homogeneity was defined. The *MDF* statistic showed a behavior adjusted to expectations with different variants of the binomial distribution. The correlation between the mean skewness score of the judges and *MDF* was very high. To obtain interpretive guiding cutoffs, 20,000 samples of sizes 20, 40, 100, 200, 500, and 1000 were simulated, drawn from a binomial distribution. It is concluded that *MDF* is validity to measure asymmetry in qualitative variables.
Keywords: skewness, discrete distribution, nominal scale, qualitative variable, Monte Carlo simulation.

Uma medida de assimetria unidimensional para variáveis qualitativas
Esta pesquisa metodológica visa definir um conceito de assimetria para variáveis qualitativas, quantificá-lo e mostrar sua validade. Um painel de cinco juízes especialistas e simulações de Monte Carlo foram usados. Foi definida a diferença média de frequência (*dmf*) entre pares de categorias ordenadas por homogeneidade de frequência. A estatística *dmf* mostrou um

[1] Doctor en Filosofía y Ciencias de la Educación por la Universidad de Alcalá (Madrid, España). Psicólogo Especialista en Psicología Clínica (Madrid, España). Profesor-investigador de la Facultad de Psicología de la Universidad Autónoma de Nuevo León. Dirección postal: Calle/ Dr. Carlos Canseco 110. Col. Mitras Centro. CP. 64640 Monterrey, Nuevo León, México. Contacto: jose_moral@hotmail.com. http://orcid.org/0000-0003-1856-1458

https://doi.org/10.18800/psico.202201.017

comportamento ajustado às expectativas com diferentes variantes da distribuição binomial. A correlação entre os escores médios de assimetria dos juízes e *dmf* foi muito alta. Para obter diretrizes de interpretação, foram simuladas 20.000 amostras dos tamanhos 20, 40, 100, 200, 500 e 1000, extraídas de uma distribuição binomial. Conclui-se que *dmf* é válido para medir a assimetria em variáveis qualitativas.
Palavras-chave: assimetria, distribuição discreta, escala nominal, variável qualitativa, simulação de Monte Carlo.

Une mesure d'asymétrie unidimensionnelle pour les variables qualitatives
Cette recherche méthodologique vise à définir un concept d'asymétrie pour les variables qualitatives, à le quantifier et à montrer sa validité. Un panel de cinq juges experts et des simulations Monte Carlo ont été utilisés. La différence moyenne de fréquence (*dmf*) entre les paires de catégories classées par homogénéité de fréquence a été définie. La statistique *dmf* a montré un comportement ajusté aux attentes avec différentes variantes de la distribution binomiale. La corrélation entre les scores moyens d'asymétrie des juges et le *dmf* était très élevée. Pour obtenir des lignes directrices pour l'interprétation, 20 000 échantillons de tailles 20, 40, 100, 200, 500 et 1000 ont été simulés, tirés d'une distribution binomiale. Il est conclu que *dmf* est valide pour mesurer l'asymétrie dans les variables qualitatives.
Mots-clés: asymétrie, distribution discrète, échelle nominale, variable qualitative, simulation Monte Carlo.

El concepto de asimetría en estadística fue introducido por Karl Pearson en el desarrollo de su sistema de 12 distribuciones continuas a finales del siglo XIX. La asimetría y la curtosis como parámetros de forma junto con un parámetro de localización y otro de escala permiten definir las distintas distribuciones de Pearson (Provost et al., 2020). Más allá de las distribuciones continuas, el concepto se generalizó a distribuciones discretas y, de este modo, hoy en día se aplica tanto a variables cuantitativas continuas como cuantitativas discretas y ordinales (Weiss, 2020).

La asimetría se puede ver como una propiedad de la forma de la distribución al ser representada por medio de un diagrama de barras, en el caso de una variable ordinal o cuantitativa discreta con pocos valores (distribución discreta), o por medio de un histograma, en el caso de una variable cuantitativa continua (distribución continua) o discreta con muchos valores (distribución discreta). Como eje de simetría para dividir la distribución en dos partes, se toma una medida de tendencia central, como la media aritmética, mediana, moda o rango medio. Si ambas partes de la distribución son iguales, es decir, una es el reflejo de la otra, hay simetría. Si ambas partes son diferentes, hay asimetría (Mishra et al., 2019). Por ejemplo, si se toma como eje de simetría la media aritmética (μ), una distribución tendría simetría si $f(x-\mu)=f(x+\mu)$, donde $f(x)$ es la función de masa de probabilidad en una distribución discreta o la función de densidad en una distribución continua (Cole y Altman, 2017).

Las medidas de asimetría relativas (libres de unidad de medida o libres de los parámetros de localización y escala) se definen como cocientes, proporciones o promedios centrados en 0 (Altinay, 2016). El valor de 0 indica simetría y, en una distribución unimodal continua, refleja que los dos hombros y las dos colas a ambos lados del eje de simetría son idénticos, esto es, un lado es el reflejo del otro (Figura 1). Un valor

positivo en la medida de asimetría suele indicar que la cola derecha es más larga que la izquierda, lo que provoca que la media aritmética se vea desplazada hacia la derecha y haya menos valores por encima de la media aritmética que por debajo. En las distribuciones unimodales continuas con asimetría positiva, la moda (pico) queda por debajo de la mediana y la mediana por debajo de la media aritmética (Figura 1). No obstante, su generalización a otros tipos de distribución depende de cuan pesadas o ligeras sean las colas. Así, una cola acortada en un extremo puede compensar una cola alargada en el otro y surgir simetría (valor nulo) cuando ambas partes de la distribución son dispares.

En distribuciones unimodales continuas, un valor negativo en la medida de asimetría muestra que la cola izquierda es más larga que la derecha, lo que ocasiona que la media aritmética se vea desplazada hacia la izquierda y que haya menos casos por debajo de la media aritmética que por encima (Figura 1). En estas distribuciones con asimetría negativa, la moda (pico) queda por encima de la mediana y la mediana por encima de la media aritmética (Sarka, 2021). Esta regularidad se cumple bien en las distribuciones continuas unimodales, pero en las distribuciones discretas tiene muchos contraejemplos (Singh, Gewali, y Khatiwada, 2019), de ahí que es muy importante representar la distribución por medio de un diagrama de barras o un histograma cuando se evalúa la simetría.

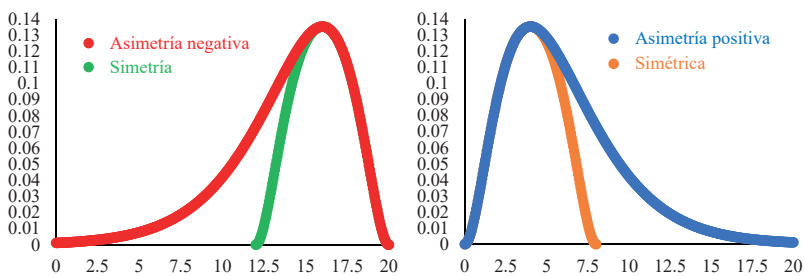


Figura 1. Funciones de densidad que muestran dos ejemplos de curvas asimétricas y la correspondiente curva simétrica de cuatro variables cuantitativas continuas

Existen varias medidas de asimetría relativas (Gupta & Kapoor, 2020; Versluis, 2017). Unas se basan en momentos centrales o cumulantes de tercer orden estandarizados, como el coeficiente $\sqrt{\beta_1}$ (poblacional) o $\sqrt{b_1}$ (muestral) de Pearson (1895) y el coeficiente γ_1 (poblacional) y g_1 (muestral) de Fisher (1930). Otras están basadas en cuantiles, como el coeficiente intercuartílico de Bowley (1901) y el percentílico de Kelley (1947). Además, existen unas terceras medidas mixtas que se basan en momentos y cuantiles, como la distancia orientada (con signo) y estandarizada de la media aritmética a la moda (Pearson, 1894) o a la mediana (Pearson, 1895), la distancia dirigida y estandarizada del semirrango a la moda, mediana o media aritmética (Altinay, 2016) o el área de asimetría de Singh et al. (2019). Todas ellas pueden tomar un valor negativo, nulo o positivo, y las medidas de Bowley (1901), Kelley (1947), Altinay (2016) y Singh et al. (2019) están limitadas al rango entre –1 y 1. También hay medidas de asimetría absolutas, como el índice de Kelley (1923).

La propiedad de la asimetría no se estudia en la descripción de una variable cualitativa, aunque sí en la relación entre dos variables cualitativas a partir de una tabla de contingencia cuadrada de datos correlacionados (Bowker, 1948; Fagerland, Lydersen y Laake, 2017). Cabría preguntarse si es posible definir un concepto de asimetría para describir una variable cualitativa y desarrollar una medida para su cuantificación. Retomando esta pregunta, el presente estudio tiene como objetivos: 1) definir un concepto de asimetría unidimensional para variables cualitativas y una medida para dicho concepto, 2) mostrar su validez y 3) generar unas normas orientativas de interpretación.

Método

Para el primer objetivo de definir asimetría cualitativa y una medición de la misma, se emplearon definiciones, argumentaciones y estas se expresaron en términos algebraicos. De aquí surge el estadístico *diferencia promedio de frecuencia* entre pares de categorías cualitativas

ordenadas por homogeneidad o proximidad de frecuencia que se denota por *dpf*.

Para el segundo objetivo de mostrar la validez de esta medida, se describe el estadístico con distintas variantes de la distribución binomial $\mathsf{B}(n, p)$ que permite emular el comportamiento de una variable cualitativa desde el estadístico propuesto. Se hace variar tanto el parámetro n de número de ensayos independientes de 1 (k = 2 categorías cualitativas) a 10 (k = 11 categorías cualitativas) como el parámetro p de probabilidad de éxito en cada ensayo independiente (p = .01, .1, .2, .3, .4 y .5). La función de masa de probabilidad $f_X(x)$ de las 60 distribuciones binomiales proporciona la frecuencia relativa de las k (= n + 1) categorías de la variable cualitativa. Se espera simetría con un valor p de .5 y asimetría en los demás casos. La expectativa es que el valor de *pdf* se aproxime a 1 cuanto más dispares sean las frecuencias o alturas de las barras en orden de simetría a ambos lados de la categoría central (número impar de categorías) o del eje imaginario entre las dos categorías centrales (número par).

La validez de *dpf* se comprobó por medio de correlaciones. A cinco jueces (profesores de estadística en facultades de psicología con grado de doctor), se les dio una escala para clasificar los 60 diagramas de barras en cinco categorías ordenadas. Se les pidió que evaluaran en cada gráfico la simetría o disparidad entre las barras a los lados de la moda (barra central cuando el número de categorías cualitativas es impar o línea imaginaria entre las dos barras centrales cuando el número de categorías es par). La escala tenía un rango de 1 (disposición de las barras a ambos lados de la moda totalmente simétrica) a 5 (disposición muy asimétrica). El instrumento proporcionado a los jueces vía correo electrónico en un archivo Excel, se puede ver en el Anexo. De este modo, cada diagrama de barras contaba con un orden promedio de asimetría desde los cinco jueces y un valor del estadístico *dpf*. La correlación entre ambos valores y la correlación entre los jueces se calcularon por el coeficiente de correlación por rangos de Spearman (r_S).

Para el tercer objetivo de tener puntos de corte de asimetría orientativos, se simularon 20 000 muestras de tamaños 20, 40, 100, 200,

500 y 1000 extraídas de una distribución binomial con parámetro n (n+1 categorías cualitativas) con valores de 1 (2 categorías) a 10 (11 categorías) y parámetro p (probabilidad de éxito) = ½ (simetría). Se calcularon los percentiles 50, 80, 90 y 95, así como el intervalo de la media con un nivel de confianza al 95% del estadístico *dpf* en las 20,000 simulaciones para cada tamaño muestral y número de categorías. Para describir las distribuciones en el muestreo de *pdf*, también se computó la asimetría (g_1 de Fisher) y exceso de curtosis (g_2 de Fisher). El análisis de datos se hizo con el Excel versión 2019 para Windows (Microsoft Corporation, 2019) y la simulación con el módulo *XLSTAT* para Excel (Addinsoft, 2021). Se dan más detalles sobre la simulación en la sección de Resultados.

Resultados

Definición de asimetría cualitativa

Todo concepto de asimetría implica un eje que permite dividir las distribuciones en dos partes. Si una parte con referencia al eje de simetría es reflejo de la otra, se considera que la distribución muestra simetría y la medida de asimetría debe arrojar un valor de 0. Por el contrario, si son distintas, se habla de asimetría, y el valor de la medida de asimetría debe ser distinto de 0. Cuanto más dispares sean los dos lados divididos por el eje de simetría, el valor de la medida de asimetría debe estar más alejado de 0.

En distribuciones aleatorias continuas, se toma como eje de simetría una medida de tendencia central, como la media aritmética, mediana, moda o rango medio. En el caso de las variables cualitativas, la única opción sería la moda, esto es, la categoría nominal con mayor frecuencia en la muestra. No obstante, la moda no siempre es única. Puede haber dos categorías modales (distribución bimodal), tres o más categorías modales (distribución multimodal) o ninguna (distribución uniforme). Consecuentemente, si se adopta la moda como eje de sime-

tría, el concepto solo podría aplicarse a las distribuciones unimodales y no ser universal.

Si las categorías de una variable en escala nominal se representan por números, estos carecen de cualquier propiedad algebraica. Perfectamente, las categorías se pueden identificar por letras, palabras o símbolos no numéricos para evidenciar el hecho de que representan las opciones de clasificación dentro de un sistema inclusivo (todo elemento de la población se puede clasificar) y exhaustivo (en una sola categoría) y no una medición en sentido estricto (determinación objetiva de cuantas veces la característica medida del objeto es la unidad de medida consensuada por expertos). La única cuantificación que admiten las variables cualitativas es el conteo de las veces en que aparece cada una de sus categorías en la muestra o la población, esto es, la frecuencia o probabilidad de cada categoría. Así, desde su frecuencia relativa o probabilidad, se abre una posibilidad de transformación. Las categorías cualitativas de la variable se pueden transformar en categorías ordenadas. De este modo, se puede crear una métrica ordinal de frecuencia.

Ante todo, la asimetría es una propiedad de la forma de una distribución. Al elaborar el diagrama de barras para estudiar la asimetría en variables cualitativas, no se procede a disponer la secuencia de categorías (ordenadas por frecuencia) en orden ascendente, como es lo estipulado, ya que aparecería una forma de escalera propia de una función monótona creciente, sino que se intenta crear una forma más o menos triangular o trapezoide. Si el número de categorías k es impar, se ubica la categoría de mayor frecuencia (categoría modal) o una de las categorías de frecuencia máxima, elegida al azar, en el centro. Tras emparejar las restantes categorías por homogeneidad o proximidad de frecuencia, estos pares se disponen en orden descendente a ambos lados de la moda. La categoría con frecuencia más alta de cada par se pone a la izquierda y la categoría con la frecuencia más baja del par se pone a la derecha. El par con las frecuencias más altas será el más próximo a la categoría central y el par con las frecuencias más bajas será el que se encuentre más alejado de la categoría central. Si el número de categorías es par, se ubican en el centro el par con las frecuencias más altas y

se procede del mismo modo. Si solo hay dos categorías, se puede ubicar en primer orden la más alta y en segundo orden la más baja.

La distribución se puede considerar simétrica, si las dos partes a ambos lados de la categoría de máxima frecuencia ubicada en el centro (número impar de categorías) o de la línea imaginaria perpendicular al eje de abscisas entre las dos categorías centrales de mayor frecuencia (número par de categorías) son iguales. Por el contrario, hay asimetría si son disimilares. A este concepto se le denomina *asimetría cualitativa*.

Propuesta de una medida de asimetría cualitativa

¿Cómo medir si hay o no similaridad o simetría entre las dos partes de la distribución? Se propone usar la diferencia promedio de frecuencia entre las categorías cualitativas emparejadas por homogeneidad o proximidad de frecuencia. Esta medida no requiere la existencia de una moda única, incluso aplica con una distribución uniforme, cuya forma en el diagrama de barras no es trapezoide, sino rectangular.

A continuación, se expresa la propuesta en términos algebraicos. Sea X una variable cualitativa con un número de k categorías nominales y cada una de ellas con frecuencia relativa f_i. Las frecuencias se ordenan en sentido descendente, esto es, de las más altas a las más bajas.

$$f_1 \geq f_2 \geq \cdots \geq f_{k-1} \geq f_k$$

Si k es impar, la categoría con la frecuencia modal o la categoría de frecuencia máxima que se ubicó en el centro del diagrama de barra (f_1), se excluye del emparejamiento de frecuencias. Se emparejan las restantes frecuencias: f_2 es igual o inmediatamente mayor que f_3, f_4 es igual o inmediatamente mayor que f_5, …, f_{k-1} es igual o inmediatamente mayor que f_k. Se restan los $(k-1)/2$ pares de frecuencias similares o más próximas, se suman las diferencias y se dividen por el número de diferencias sumadas: $(k-1)/2$. De este modo, se obtiene la *diferencia promedio de frecuencia* entre las categorías cualitativas emparejadas por homogeneidad o proximidad de frecuencia que se denota por *dpf*.

$$dpf = \frac{\sum_{i=1}^{(k-1)/2}\left(f_i - f_{i+1}\right)}{\left(k-1\right)/2} = \frac{2\times\sum_{i=1}^{(k-1)/2}\left(f_i - f_{i+1}\right)}{k-1}$$

Si k es par, se emparejan las frecuencias: f_1 es igual o inmediatamente mayor que f_2, f_3 es igual o inmediatamente mayor que f_4, …. f_{k-1} es igual o inmediatamente mayor que f_k. Se restan los $k/2$ pares de frecuencias similares o más próximas, se suman las diferencias y se dividen por el número de diferencias sumadas ($k/2$), con lo que se obtiene *dpf*.

$$dpf = \frac{\sum_{i=1}^{k/2}\left(f_i - f_{i+1}\right)}{k/2} = \frac{2\times\sum_{i=1}^{k/2}\left(f_i - f_{i+1}\right)}{k}$$

El estadístico *dpf* está acotado de 0 a 1. Un valor de 0 indica simetría, que puede corresponder a un perfil triangular (distribución unimodal), trapezoide (distribución bi o multimodal) o rectangular (distribución uniforme). Un valor de 1 representa la máxima asimetría y se alcanza con la distribución de una variable aleatoria discreta constante en la que un valor concentra toda la probabilidad o frecuencia (distribución Bernoulli de parámetro $p = 1$).

Comportamiento y validez del estadístico dpf

Para estudiar el comportamiento del estadístico *dpf*, se crearon 60 distribuciones cualitativas. El número de categorías nominales en estas variables osciló de 2 a 11. Las frecuencias de las categorías se generaron a partir de las funciones de masa de probabilidad de 60 distribuciones binomiales $\mathsf{B}\left(n,p\right)$. La probabilidad del número de éxitos ($x_i \in X \sim \mathsf{B}\left(n,p\right)$) pasó a ser la frecuencia relativa de la categoría nominal. Para lograr los grados variables de asimetría, se dieron seis valores diferentes a las probabilidades de éxito: .01 (probabilidad muy baja), .1, .2, .3, .4 y .5 (probabilidad media). Se obvió la otra mitad de probabilidades (.6, .7, .8, .9 y .99), ya que el índice *dpf* siempre es positivo y daría el mismo resultado, como se puede apreciar en la Tabla 1 y Figura 2.

Tabla 1

Generación de dos variables con 10 categorías nominales, una desde una distribución binomial de parámetros n = 9 y p = .3 y otra desde una distribución binomial de parámetros n = 9 y p = .7, y ordenamiento de sus categorías para calcular dpf y el diagrama de barras

x_i	A	$\mathrm{B}(n=9, p=0.3)$					B	$\mathrm{B}(n=9, p=0.7)$				
		$f_A = f_{x_i}$	↓	$A_{(\downarrow)}$	Δ	$A_{(\Delta)}$		$f_B = f_{x_i}$	↓	$B_{(\downarrow)}$	Δ	$B_{(\Delta)}$
0	*a*	.040	.267	*c*	4.13×10^{-4}	*i*	*k*	1.97×10^{-5}	.267	*q*	4.13×10^{-4}	*l*
1	*b*	.156	.267	*d*	.021	*g*	*l*	4.13×10^{-4}	.267	*r*	.021	*n*
2	*c*	.267	.172	*e*	.074	*f*	*m*	.004	.172	*p*	.074	*o*
3	*d*	.267	.156	*b*	.172	*e*	*n*	.021	.156	*s*	.172	*p*
4	*e*	.172	.074	*f*	.267	*c*	*o*	.074	.074	*o*	.267	*q*
5	*f*	.074	.040	*a*	.267	*d*	*p*	.172	.040	*t*	.267	*r*
6	*g*	.021	.021	*g*	.156	*b*	*q*	.267	.021	*n*	.156	*s*
7	*h*	.004	.004	*h*	.040	*a*	*r*	.267	.004	*m*	.040	*t*
8	*i*	4.13×10^{-4}	4.13×10^{-4}	*i*	.004	*h*	*s*	.156	4.13×10^{-4}	*l*	.004	*m*
9	*j*	1.97×10^{-5}	1.97×10^{-5}	*j*	1.97×10^{-5}	*j*	*t*	.040	1.97×10^{-5}	*k*	1.97×10^{-5}	*k*

Nota. x_i = valores discretos de las distribuciones binomiales $\mathrm{B}(n=9, p=.3)$ y $\mathrm{B}(n=9, p=.7)$ = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, A = variable cualitativa con dominio = {*a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, *j*}, B = variable cualitativa con dominio = {*k*, *l*, *m*, *n*, *o*, *p*, *q*, *r*, *s*, *t*} $f_A = f_{x_i}$ = frecuencia relativa simple de las categorías de A tomadas desde la función de masa de probabilidad de las distribuciones binomiales $\mathrm{B}(n=9, p=.3)$, $f_B = f_{x_i}$ = frecuencia relativa simple de las categorías de B tomadas desde la función de masa de probabilidad de la distribución binomial $\mathrm{B}(n=9, p=.7)$, ↓ = frecuencias ordenadas en sentido descendente, $A_{(\downarrow)}$ = valores o categorías nominales de A ordenadas en sentido descendente desde el valor de su frecuencia, Δ = frecuencias ordenadas para mostrar un perfil triangular, trapezoide o rectangular en el diagrama de barras, $B_{(\Delta)}$ = valores o categorías nominales de B en el orden que les corresponde en el diagrama de barras.

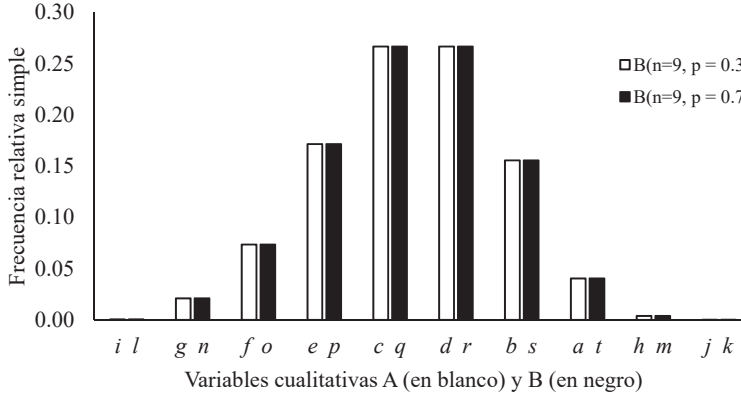


Figura 2. Diagrama de barras para valorar la asimetría cualitativa de las variables cualitativas A y B con 10 categorías, cuyas frecuencias corresponden a las probabilidades de una distribucional binomial $\mathsf{B}(\mathrm{n}=9,\mathrm{p}=.3)$ y $\mathsf{B}(\mathrm{n}=9,\mathrm{p}=.7)$, respectivamente.

$$dpf_A = \frac{\sum_{i=1}^{k/2}\left(f_i - f_{i+1}\right)}{k/2} = \frac{\left(f_c - f_d\right)+\left(f_e - f_b\right)+\left(f_f - f_a\right)+\left(f_g - f_h\right)+\left(f_i - f_j\right)}{10/2}$$

$$= \frac{\left(.27-.27\right)+\left(.17-.16\right)+\left(.07-.04\right)+\left(.02-3.86\times10^{-3}\right)+\left(4.13\times10^{-4}-1.97\times10^{-5}\right)}{5} = .07$$

$$dpf_B = \frac{\sum_{i=1}^{k/2}\left(f_i - f_{i+1}\right)}{k/2} = \frac{\left(f_c - f_d\right)+\left(f_e - f_b\right)+\left(f_f - f_a\right)+\left(f_g - f_h\right)+\left(f_i - f_j\right)}{10/2}$$

$$= \frac{\left(.27-.27\right)+\left(.17-.16\right)+\left(.07-.04\right)+\left(.02-3.86\times10^{-3}\right)+\left(4.13\times10^{-4}-1.97\times10^{-5}\right)}{5} = .07$$

De las 60 distribuciones (Tabla 2), 10 fueron simétricas (p = .5) y el resto fueron asimétricas en diferente grado ($p \neq .5$). El hecho de que el valor de p se aproxime a uno no implica necesariamente que la asimetría sea mayor, ya que depende de cuan dispares sean las frecuencias con respecto al eje de simetría (incluido con k par y excluido con k impar).

Tabla 2

Generación de distribuciones cualitativas de 2 a 11 categorías y asimetría variable desde la distribución binomial $\mathsf{B}(n,p)$.

Valores del parámetro de probabilidad de éxito p											
.01		**.1**		**.2**		**.3**		**.4**		**.5**	
$x_{(i)}$	$f_{x_{(i)}}$	$x_{(i)}$	$f_{x_{(i)}}$	$x_{(i)}$	$f_{x_{(i)}}$	$x_{(i)}$	$f_{x_{(i)}}$	$x_{(i)}$	$f_{x_{(i)}}$	$x_{(i)}$	$f_{x_{(i)}}$
2 categorías: a = 0 éxitos y b = 1 éxito; $\mathsf{B}(n=1,p)$											
a	.99	*a*	.90	*a*	.80	*a*	.70	*a*	.60	*a*	.50
b	.01	*b*	.10	*b*	.20	*b*	.30	*b*	.40	*b*	.50
3 categorías: a = 0, b = 1 y c = 2 éxitos; $\mathsf{B}(n=2,p)$											
b	.02	*b*	.18	*b*	.32	*b*	.42	*b*	.36	*a*	.25
a	.98	*a*	.81	*a*	.64	*a*	.49	*a*	.48	*b*	.50
c	1×10^{-4}	*c*	.01	*c*	.04	*c*	.09	*c*	.16	*c*	.25
4 categorías: a = 0, b = 1, c = 2 y d = 3 éxitos; $\mathsf{B}(n=3,p)$											
c	2.97×10^{-4}	*c*	.027	*c*	.096	*c*	.189	*a*	.216	*a*	.125
a	.970	*a*	.729	*a*	.512	*b*	.441	*b*	.432	*b*	.375
b	.029	*b*	.243	*b*	.384	*a*	.343	*c*	.288	*c*	.375
d	1×10^{-6}	*d*	.001	*d*	.008	*d*	.027	*d*	.064	*d*	.125
5 categorías: a = 0, b = 1, c = 2, d = 3 y e = 4 éxitos; $\mathsf{B}(n=4,p)$											
d	3.96×10^{-6}	*d*	.004	d	.026	*d*	.076	*a*	.130	*a*	.063
b	.039	*b*	.292	a	.410	*a*	.265	*c*	.346	*b*	.250
a	.961	*a*	.656	b	.410	*b*	.412	*b*	.346	*c*	.375
c	5.88×10^{-4}	*c*	.049	c	.154	*c*	.240	*d*	.154	*d*	.250
e	1×10^{-8}	*e*	1×10^{-4}	e	.002	*e*	.008	*e*	.026	*e*	.063
6 categorías: a = 0, b = 1, c = 2, d = 3, e = 4 y f = 5 éxitos; $\mathsf{B}(n=5,p)$											
e	4.95×10^{-8}	*e*	4.5×10^{-4}	*e*	.006	*e*	.028	*e*	.077	*a*	.031
c	.001	*c*	.073	*c*	.205	*a*	.168	*d*	.230	*b*	.156
a	.951	*a*	.590	*b*	.410	*b*	.360	*c*	.346	*c*	.313
b	.048	*b*	.328	*a*	.328	*c*	.309	*b*	.259	*d*	.313

Valores del parámetro de probabilidad de éxito p											
.01		**.1**		**.2**		**.3**		**.4**		**.5**	
d	9.8×10^{-6}	*d*	.008	*d*	.051	*d*	.132	*a*	.078	*e*	.156
f	1×10^{-10}	*f*	1×10^{-5}	*f*	3.2×10^{-4}	*f*	.002	*f*	.010	*f*	.031
7 categorías: $a = 0$, $b = 1$, $c = 2$, $d = 3$, $e = 4$, $f = 5$ y $g = 6$ éxitos; $\mathrm{B}(n = 6, p)$											
f	5.94×10^{-10}	*f*	5.40×10^{-5}	*f*	.002	*f*	.010	*f*	.037	*a*	.016
d	1.94×10^{-5}	*d*	.015	*d*	.082	*a*	.118	*e*	.138	*b*	.094
b	.057	*b*	.354	*a*	.262	*b*	.303	*d*	.276	*c*	.234
a	.941	*a*	.531	*b*	.393	*c*	.324	*c*	.311	*d*	.313
c	.001	*c*	.098	*c*	.246	*d*	.185	*b*	.187	*e*	.234
e	1.47×10^{-7}	*e*	.001	*e*	.015	*e*	.060	*a*	.047	*f*	.094
g	1×10^{-12}	*g*	1×10^{-6}	*g*	6.4×10^{-5}	*g*	.001	*g*	.004	*g*	.016
8 categorías: $a = 0$, $b = 1$, $c = 2$, $d = 3$, $e = 4$, $f = 5$, $g = 6$ y $h = 7$ éxitos; $\mathrm{B}(n = 7, p)$											
g	6.93×10^{-12}	*g*	6.3×10^{-6}	*g*	3.58×10^{-4}	*g*	.004	*g*	.017	*a*	.008
e	3.40×10^{-7}	*e*	.003	*e*	.029	*a*	.082	*f*	.077	*b*	.055
c	.002	*c*	.124	*a*	.210	*d*	.227	*e*	.194	*c*	.164
a	.932	*a*	.478	*b*	.367	*c*	.318	*d*	.290	*d*	.273
b	.066	*b*	.372	*c*	.275	*b*	.247	*c*	.261	*e*	.273
d	3.36×10^{-5}	*d*	.023	*d*	.115	*e*	.097	*b*	.131	*f*	.164
f	2.06×10^{-9}	*f*	1.70×10^{-4}	*f*	.004	*f*	.025	*a*	.028	*g*	.055
h	1×10^{-14}	*h*	1×10^{-7}	*h*	1.28×10^{-5}	*h*	2.19×10^{-4}	*h*	.002	*h*	.008
9 categorías: $a = 0$, $b = 1$, $c = 2$, $d = 3$, $e = 4$, $f = 5$, $g = 6$, $h = 7$ y $i = 8$ éxitos; $\mathrm{B}(n = 8, p)$											
h	7.92×10^{-14}	*h*	7.2×10^{-7}	*h*	8.19×10^{-5}	*h*	.001	*h*	.008	*a*	.004
f	5.43×10^{-9}	*f*	4.08×10^{-4}	*f*	.009	*f*	.047	*g*	.041	*b*	.031
d	5.33×10^{-5}	*d*	.033	*d*	.147	*e*	.136	*f*	.124	*c*	.109
b	.075	*b*	.383	*c*	.294	*d*	.254	*e*	.232	*d*	.219
a	.923	*a*	.430	*b*	.336	*c*	.296	*d*	.279	*e*	.273
c	.003	*c*	.149	*a*	.168	*b*	.198	*c*	.209	*f*	.219
e	6.72×10^{-7}	*e*	.005	*e*	.046	*a*	.058	*b*	.090	*g*	.109

Valores del parámetro de probabilidad de éxito p										
.01		**.1**		**.2**		**.3**		**.4**		**.5**	
g	2.74×10^{-11}	*g*	2.27×10^{-5}	*g*	.001	*g*	.010	*a*	.017	*h*	.031
i	1×10^{-16}	*i*	1×10^{-8}	*i*	2.56×10^{-6}	*i*	6.56×10^{-5}	*i*	.001	*i*	.004
10 categorías: $a = 0$, $b = 1$, $c = 2$, $d = 3$, $e = 4$, $f = 5$, $g = 6$, $h = 7$, $i = 8$ y $j = 9$ éxitos; $\mathrm{B}(n = 9, p)$											
i	8.91×10^{-16}	*i*	8.1×10^{-8}	*i*	1.84×10^{-5}	*i*	4.13×10^{-4}	*i*	.004	*a*	.002
g	8.15×10^{-11}	*g*	6.12×10^{-5}	*g*	.003	*g*	.021	*h*	.021	*b*	.018
e	1.20×10^{-6}	*e*	.007	*e*	.066	*f*	.074	*g*	.074	*c*	.070
c	.003	*c*	.172	*d*	.176	*e*	.172	*f*	.167	*d*	.164
a	.914	*a*	.387	*b*	.302	*c*	.267	*d*	.251	*e*	.246
b	.083	*b*	.387	*c*	.302	*d*	.267	*e*	.251	*f*	.246
d	7.91×10^{-5}	*d*	.045	*a*	.134	*b*	.156	*c*	.161	*g*	.164
f	1.21×10^{-8}	*f*	.001	*f*	.017	*a*	.040	*b*	.060	*h*	.070
h	3.53×10^{-13}	*h*	2.92×10^{-6}	*h*	2.95×10^{-4}	*h*	.004	*a*	.010	*i*	.018
j	1×10^{-18}	*j*	1×10^{-9}	*j*	5.12×10^{-7}	*j*	1.97×10^{-5}	*j*	2.62×10^{-4}	*j*	.002
11 categorías: $a = 0$, $b = 1$, $c = 2$, $d = 3$, $e = 4$, $f = 5$, $g = 6$, $h = 7$, $i = 8$, $j = 9$ y $k = 10$ éxitos; $\mathrm{B}(n = 10, p)$											
j	9.9×10^{-18}	*j*	9.0×10^{-9}	*j*	4.10×10^{-6}	*j*	1.38×10^{-4}	*j*	.002	*a*	.001
h	1.16×10^{-12}	*h*	8.75×10^{-6}	*h*	.001	*h*	.009	*i*	.011	*b*	.010
f	2.40×10^{-8}	*f*	.001	*f*	.026	*g*	.037	*h*	.042	*c*	.044
d	1.12×10^{-4}	*d*	.057	*a*	.107	*b*	.121	*c*	.121	*d*	.117
b	.091	*a*	.349	*b*	.268	*c*	.233	*d*	.215	*e*	.205
a	.904	*b*	.387	*c*	.302	*d*	.267	*e*	.251	*f*	.246
c	.004	*c*	.194	*d*	.201	*e*	.200	*f*	.201	*g*	.205
e	1.98×10^{-6}	*e*	.011	*e*	.088	*f*	.103	*g*	.111	*h*	.117
g	2.02×10^{-10}	*g*	1.38×10^{-4}	*g*	.006	*a*	.028	*b*	.040	*i*	.044
i	4.41×10^{-15}	*i*	3.65×10^{-7}	*i*	$7.37 \times .10^{-5}$	*i*	.001	*a*	.006	*j*	.010
k	1×10^{-20}	*k*	1×10^{-10}	*k*	1.02×10^{-7}	*k*	5.90×10^{-6}	*k*	1.05×10^{-4}	*k*	.001

Se observó que, cuando p = .5, *dfp* toma un valor nulo con cualquier número de categorías (k de 2 a 11), ya que las frecuencias son totalmente simétricas con respecto al eje de simetría. Con un valor de p extremo de .01, *dfp* alcanza los valores más altos cuando el número de categorías es par y, en esta condición, *dpf* es más alto cuanto menor es el número de categorías, ya que hay mucha diferencia entre las dos categorías centrales. No obstante, si k es impar, *dpf* presenta un valor próximo a 0, ya que la categoría modal excluida (eje de simetría) concentra casi toda la probabilidad y separa las restantes categorías (de muy baja frecuencia) en dos partes muy semejantes, resultando un perfil simétrico. Con una probabilidad de .1, este patrón es más tenue; con un número par de categorías el valor de *dpf* es más alto que con un número impar; no obstante, en ambos casos *dpf* disminuye a medida que el número de categorías aumenta. El hecho de que haya solo dos categorías marca el máximo de la asimetría para todos los niveles de probabilidad. Precisamente, con un valor de p = .01, la distribución se aproxima mucho a una variable aleatoria constante y *dpf* es igual a .98 (Tablas 2 y 3). Así, el comportamiento del estadístico se ajustó bien a las expectativas.

A cinco jueces se le dio una escala con cinco categorías ordenadas para clasificar los 60 diagramas de barras. El criterio de ordenación era si la disposición de las barras a ambos lados de la barra central (excluida) en el caso de un número impar de categorías o de la línea imaginaria entre las dos barras centrales (incluidas) en el caso de un número par de categorías se puede considerar: 1 = "disposición totalmente simétrica", 2 = "muy ligeramente asimétrica", 3 = "ligeramente asimétrica", 4 = "bastante asimétrica" y 5 = "muy asimétrica" (Anexo). Conforme al resultado previo, la correlación entre la puntuación promedio de asimetría de los cinco jueces y *dpf* fue muy alta, r_S = .87, *IC* al 95%: [.74, 1]. A su vez, la correlación entre los cinco jueces varió de .44 a .74 con un promedio entre las 10 correlaciones de $\bar{r}_S$ = .63, *IC* al 95% [.54, .65].

Tabla 3

Valor del estadístico dpf para distribuciones de 2 a 10 categorías cualitativas (parámetro n) y distintos grados de asimetría extrema (parámetro p = .01) a simetría (p = .5) creadas a partir de distribuciones binomiales $\mathrm{B}(n,p)$.

Número de categorías	Probabilidad de éxito (*p*)					
	.01	.1	.2	.3	.4	.5
2	.980	.800	.600	.400	.200	0
3	.020	.170	.280	.330	.200	0
4	.471	.256	.108	.130	.148	0
5	.019	.123	.140	.046	.148	0
6	.301	.109	.081	.038	.102	0
7	.019	.090	.028	.062	.071	0
8	.217	.052	.053	.065	.039	0
9	.018	.066	.059	.043	.022	0
10	.167	.027	.019	.013	.007	0
11	.017	.041	.022	.014	.006	0

Valores críticos orientativos de asimetría para dpf

Los valores críticos orientativos para rechazar la hipótesis nula de simetría (percentiles 90, 95 o 99) se obtuvieron por medio de simulación Monte Carlo. Se ejecutaron 20,000 extracciones. Las simulaciones se hicieron para muestras con seis tamaños distintos: 20, 40, 100, 200, 500 y 1000. Se partió de una distribución binomial $\mathrm{B}(n,p)$. La probabilidad de éxito (parámetro *p*) se fijó en .5 para tener simetría perfecta a nivel poblacional y, consecuentemente, un estadístico *dpf* nulo. El número de categorías (*k*) de la variable cualitativa A se hizo corresponder al número de éxitos (*x*) de la variable binomial X, con lo que el número de ensayos independientes (parámetro *n*) es $k - 1$. Se consideraron de 2 a 11 categorías nominales con incremento de 1.

Primero, se computó la función de masa de probabilidad de la distribución binomial $\mathrm{B}(k-1, p=.5)$. Segundo, se multiplicaron las *k*

probabilidades obtenidas por el tamaño muestral *n* y se redondearon al entero más próximo ($n_i = f_X(x) \times n$). La suma necesariamente debía ser *n* y conservarse la perfecta simetría desde la categoría modal excluida (*k* impar) o desde las dos categorías modales incluidas (*k* par). En caso de que faltara un valor al sumar las *k* frecuencias absolutas, se agregaba 1 a la frecuencia absoluta de la categoría modal. Por el contrario, si sobraba, se le restaba uno a la frecuencia absoluta de la categoría modal. Tercero, se dividió las frecuencias absolutas n_i por *n*. Las frecuencias relativas o probabilidades resultantes, en algunos casos, coincidían con las probabilidades de la distribución binomial y, en otros casos, diferían ligeramente. Cuarto, se definieron las distribuciones de los argumentos del estadístico *dpf* desde la aproximación de la proporción binomial a la distribución normal. Finalmente, se computó el estadístico *dpf* que en todos los casos era nulo (variable resultado). A continuación, se muestra un ejemplo de cómo se planteó la simulación: el caso de cuatro categorías nominales y una muestra de tamaño 20 (Tabla 4).

Se generaron un total de 60 simulaciones con reemplazamiento con 20,000 muestras simuladas (6 tamaños muestrales × 10 números distintos de categorías nominales). Las Tablas de la 5 a la 9 permiten apreciar que la distribución del estadístico *dpf* en todos los casos presentó asimetría positiva ($g_1 > 2EE_{g_1}$) y leptocurtosis ($g_2 > 2EE_{g_2}$), alejándose de una distribución normal. La cola derecha fue más larga que la izquierda y hubo un desplazamiento importante de la masa de probabilidad de la zona de los hombros hacia la cola derecha (Figura 3). Su variabilidad fue alta (.70 ≤ CV < 1) y alcanzó a ser muy alta con cuatro, cinco, diez y once categorías nominales ($CV > 1$).

Tabla 4

Simulación para cuatro categorías nominales y muestras de tamaño 20.

A	X	$f_X(x)=\binom{n}{i}0.5^n$	$n_i = f_X(x)\times n$	$p_i = \frac{n_i}{n} \sim N\left(p_i, \sqrt{\left(p_i(1-p_i)\right)/n}\right)$
a	0	$\binom{20}{0}.5^{20} = .125$	$.125\times 20 = 2$	$p_a = \frac{2}{20} = .1 \sim N(.1, .067)$
b	1	$\binom{20}{1}.5^{20} = .375$	$.375\times 20 = 8$	$p_b = \frac{8}{20} = .4 \sim N(.4, .110)$
c	2	$\binom{20}{2}.5^{20} = .375$	$.375\times 20 = 2$	$p_c = \frac{2}{20} = .4 \sim N(.4, .110)$
d	3	$\binom{20}{3}.5^{20} = .125$	$.125\times 20 = 2$	$p_d = \frac{2}{20} = .1 \sim N(.1, .067)$
$\sum$		1	20	1

Nota. A = variable cualitativa con dominio = {*a*, *b*, *c*, *d*}, *X* = variable discreta con distribución binomial $\mathsf{B}(n=3, p=0.5)$ y dominio = {0, 1, 2, 3}, $f_X(x)$ = función de masa de probabilidad de una distribución binomial $\mathsf{B}(n=3, p=0.5)$, n_i = frecuencia absoluta simple o valor esperado de las cuatro categorías nominales de variable cualitativa A con una muestra de 20 participantes, p_i = frecuencia relativa simple o probabilidad esperada de las cuatro categorías nominales de variable cualitativa A que, en su condición de proporciones binomiales, convergen a una distribución normal de parámetro de localización $\mu = p_i$ y parámetro de escala $\sigma = \sqrt{\left(p_i(1-p_i)\right)/n}$

El estadístico se definió del siguiente modo aprovechando la forma trapezoide y simétrica de la distribución: $dpf = \left(|p_b - p_c| + |p_a - p_d|\right)/2 = 0$

Tabla 5

Estadísticos descriptivos y normas para la distribución del estadístico dpf ante una distribución simétrica (p = .5) con dos y tres categorías para tamaños muestrales de 20, 40, 100, 200, 500 y 1000 participantes.

Est.	2 categorías nominales						3 categorías nominales					
	20	**40**	**100**	**200**	**500**	**1000**	**20**	**40**	**100**	**200**	**500**	**1000**
Mín	0	0	0	0	0	0	0	0	0	0	0	0
Máx	.627	.550	.319	.196	.127	.090	.545	.388	.242	.183	.110	.087
P50	.107	.075	.048	.033	.021	.015	.093	.065	.041	.030	.018	.013
P75	.183	.129	.082	.057	.036	.026	.158	.111	.070	.051	.031	.023
P80	.204	.143	.091	.064	.040	.029	.176	.124	.078	.056	.035	.026
P90	.262	.184	.117	.081	.051	.037	.228	.159	.100	.072	.044	.033
P95	.312	.220	.139	.096	.060	.044	.270	.188	.119	.085	.053	.039
$\bar{X}$	.127	.089	.056	.039	.025	.018	.110	.077	.048	.035	.022	.016
LI	.125	.088	.056	.039	.024	.018	.109	.076	.048	.035	.021	.016
LS	.128	.090	.057	.040	.025	.018	.111	.078	.049	.035	.022	.016
$\hat{S}_X$	.096	.067	.043	.030	.019	.014	.083	.058	.037	.026	.016	.012
CV	0.756	0.753	0.768	0.769	0.760	0.778	0.755	0.753	0.771	0.743	0.727	0.750
g_1	1.013	0.993	0.983	0.980	0.949	1.006	0.975	0.986	1.004	0.991	0.968	0.974
EE_{g_1}	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
g_2	0.961	0.794	0.825	0.805	0.723	0.936	0.769	0.829	0.907	0.879	0.762	0.811
EE_{g_2}	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035

Nota. Número de simulaciones = 20,000, valores perdidos = 0. *Min* = valor mínimo, *Max* = valor máximo, *P50* = mediana o percentil 50, *P75* = cuartil superior o percentil 75, *P80* = percentil 80, *P90* = percentil 90, *P95* = percentil 95, $\bar{X}$ = media aritmética, *LI* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, *LS* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, $\hat{S}_X$ = desviación estándar con la corrección de Bessel, *CV* = coeficiente de variación de Pearson, g_1 = coeficiente de asimetría basada en el tercer cumulante estandarizado de Fisher, EE_{g_1} = error estándar del coeficiente de asimetría de Fisher, g_2 = exceso de curtosis basado en el cuarto cumulante estandarizado de Fisher, EE_{g_2} = error estándar del exceso de curtosis de Fisher.

Tabla 6

Estadísticos descriptivos y normas para la distribución del estadístico dpf ante una distribución simétrica (p = .5) con cuatro y cinco categorías para tamaños muestrales de 20, 40, 100, 200, 500 y 1000 participantes.

Est.	4 categorías nominales						5 categorías nominales					
	20	**40**	**100**	**200**	**500**	**1000**	**20**	**40**	**100**	**200**	**500**	**1000**
Mín	0	0	0	0	0	0	0	0	0	0	0	0
Máx	.422	.269	.178	.118	.081	.054	.334	.200	.145	.110	.071	.044
P50	.093	.068	.043	.030	.019	.013	.075	.053	.035	.025	.016	.012
P75	.134	.097	.061	.043	.028	.019	.110	.077	.051	.036	.023	.017
P80	.145	.105	.066	.046	.030	.021	.120	.084	.055	.039	.025	.018
P90	.176	.127	.080	.056	.037	.025	.146	.102	.067	.048	.030	.022
P95	.203	.146	.091	.065	.042	.029	.170	.119	.077	.056	.034	.025
$\bar{X}$	.100	.073	.046	.032	.021	.014	.083	.058	.038	.027	.017	.013
LI	.100	.073	.046	.032	.021	.014	.082	.057	.038	.027	.017	.013
LS	.101	.074	.046	.033	.021	.014	.083	.058	.038	.027	.017	.013
$\hat{S}_X$	.110	.079	.049	.035	.023	.015	.093	.064	.042	.030	.019	.014
CV	1.100	1.082	1.065	1.094	1.095	1.071	1.120	1.103	1.105	1.111	1.118	1.077
g_1	0.745	0.750	0.740	0.743	0.769	0.747	0.813	0.786	0.770	0.820	0.762	0.769
EE_{g_1}	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
g_2	0.495	0.555	0.503	0.521	0.531	0.464	0.607	0.510	0.526	0.682	0.577	0.532
EE_{g_2}	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035

Nota. Número de simulaciones = 20,000, valores perdidos = 0. *Min* = valor mínimo, *Max* = valor máximo, *P50* = mediana o percentil 50, *P75* = cuartil superior o percentil 75, *P80* = percentil 80, *P90* = percentil 90, *P95* = percentil 95, $\bar{X}$ = media aritmética, *LI* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, *LS* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, $\hat{S}_X$ = desviación estándar con la corrección de Bessel, *CV* = coeficiente de variación de Pearson, g_1 = coeficiente de asimetría basada en el tercer cumulante estandarizado de Fisher, EE_{g_1} = error estándar del coeficiente de asimetría de Fisher, g_2 = exceso de curtosis basado en el cuarto cumulante estandarizado de Fisher, EE_{g_2} = error estándar del exceso de curtosis de Fisher.

Tabla 7

Estadísticos descriptivos y normas para la distribución del estadístico dpf ante una distribución simétrica (p = .5) con seis y siete categorías para tamaños muestrales de 20, 40, 100, 200, 500 y 1000 participantes

Est.	6 categorías nominales						7 categorías nominales					
	20	**40**	**100**	**200**	**500**	**1000**	**20**	**40**	**100**	**200**	**500**	**1000**
Mín	0	0	0	0	0	0	0	0	0	0	0	0
Máx	.431	.299	.186	.145	.079	.058	.375	.258	.170	.122	.079	.050
P50	.124	.083	.053	.038	.024	.017	.086	.074	.045	.032	.020	.014
P75	.167	.113	.073	.052	.033	.023	.123	.101	.062	.044	.028	.019
P80	.179	.121	.078	.055	.035	.025	.133	.108	.067	.047	.030	.021
P90	.211	.144	.092	.065	.041	.029	.161	.127	.079	.056	.035	.024
P95	.237	.162	.104	.074	.047	.033	.185	.144	.089	.064	.040	.028
$\bar{X}$	.130	.088	.057	.040	.025	.018	.093	.078	.048	.034	.022	.015
LI	.130	.087	.056	.040	.025	.018	.092	.078	.048	.034	.022	.015
LS	.131	.088	.057	.040	.026	.018	.093	.079	.049	.034	.022	.015
$\hat{S}_X$	.119	.082	.052	.037	.023	.017	.090	.073	.045	.032	.020	.014
CV	0.915	0.932	0.912	0.925	0.920	0.944	0.968	0.936	0.938	0.941	0.909	0.933
g_1	0.623	0.645	0.634	0.653	0.671	0.647	0.761	0.663	0.677	0.692	0.695	0.695
EE_{g_1}	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
g_2	0.297	0.288	0.300	0.344	0.566	0.407	0.624	0.420	0.417	0.449	0.487	0.552
EE_{g_2}	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035

Nota. Número de simulaciones = 20,000, valores perdidos = 0. *Min* = valor mínimo, *Max* = valor máximo, *P50* = mediana o percentil 50, *P75* = cuartil superior o percentil 75, *P80* = percentil 80, *P90* = percentil 90, *P95* = percentil 95, $\bar{X}$ = media aritmética, *LI* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, *LS* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, $\hat{S}_X$ = desviación estándar con la corrección de Bessel, *CV* = coeficiente de variación de Pearson, g_1 = coeficiente de asimetría basada en el tercer cumulante estandarizado de Fisher, EE_{g_1} = error estándar del coeficiente de asimetría de Fisher, g_2 = exceso de curtosis basado en el cuarto cumulante estandarizado de Fisher, EE_{g_2} = error estándar del exceso de curtosis de Fisher.

Tabla 8

Estadísticos descriptivos y normas para la distribución del estadístico dpf ante una distribución simétrica (p = .5) con ocho y nueve categorías para tamaños muestrales de 20, 40, 100, 200, 500 y 1000 participantes

Est.	8 categorías nominales						9 categorías nominales					
	20	40	100	200	500	1000	20	40	100	200	500	1000
Mín	0	0	0	0	0	0	0	0	0	0	0	0
Máx	.446	.342	.207	.136	.093	.068	.368	.264	.182	.171	.076	.064
P50	.124	.089	.062	.043	.028	.020	.110	.074	.048	.052	.023	.016
P75	.167	.118	.082	.057	.037	.026	.148	.100	.065	.067	.031	.022
P80	.179	.127	.087	.060	.039	.027	.159	.107	.069	.071	.033	.023
P90	.212	.150	.102	.070	.046	.032	.187	.127	.081	.082	.038	.027
P95	.238	.170	.114	.080	.051	.036	.210	.144	.092	.092	.043	.031
$\bar{X}$	.131	.093	.065	.045	.029	.021	.116	.078	.050	.054	.024	.017
LI	.130	.092	.065	.045	.029	.020	.115	.077	.050	.054	.024	.017
LS	.132	.093	.066	.045	.029	.021	.117	.078	.051	.054	.024	.017
$\hat{S}_X$	.118	.085	.054	.038	.024	.017	.104	.072	.046	.043	.021	.015
CV	0.901	0.914	0.831	0.844	0.828	0.810	0.897	0.923	0.920	0.796	0.875	0.882
g_1	0.589	0.652	0.620	0.600	0.613	0.600	0.613	0.659	0.651	0.560	0.626	0.675
EE_{g_1}	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
g_2	0.202	0.483	0.387	0.305	0.390	0.304	0.324	0.403	0.431	0.319	0.364	0.513
EE_{g_2}	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035

Nota. Número de simulaciones = 20,000, valores perdidos = 0. *Min* = valor mínimo, *Max* = valor máximo, *P50* = mediana o percentil 50, *P75* = cuartil superior o percentil 75, *P80* = percentil 80, *P90* = percentil 90, *P95* = percentil 95, $\bar{X}$ = media aritmética, *LI* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, *LS* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, $\hat{S}_X$ = desviación estándar con la corrección de Bessel, *CV* = coeficiente de variación de Pearson, g_1 = coeficiente de asimetría basada en el tercer cumulante estandarizado de Fisher, EE_{g_1} = error estándar del coeficiente de asimetría de Fisher, g_2 = exceso de curtosis basado en el cuarto cumulante estandarizado de Fisher, EE_{g_2} = error estándar del exceso de curtosis de Fisher.

Tabla 9

Estadísticos descriptivos y normas para la distribución del estadístico dpf ante una distribución simétrica (p = .5) con diez categorías para tamaños muestrales de 20, 40, 100, 200, 500 y 1000 participantes

Est	10 categorías nominales						11 categorías nominales					
	20	40	100	200	500	1000	20	40	100	200	500	1000
Mín	0	0	0	0	0	0	0	0	0	0	0	0
Máx	.956	.651	.384	.290	.176	.128	.748	.533	.360	.247	.168	.108
P50	.262	.207	.130	.091	.060	.042	.221	.159	.111	.080	.049	.036
P75	.352	.269	.170	.119	.077	.055	.296	.214	.146	.105	.064	.047
P80	.375	.285	.180	.126	.082	.058	.317	.230	.156	.112	.068	.050
P90	.439	.330	.211	.146	.094	.067	.372	.269	.182	.130	.080	.058
P95	.496	.369	.236	.163	.106	.075	.418	.304	.203	.145	.089	.065
$\bar{X}$	.275	.216	.135	.095	.062	.044	.232	.168	.116	.083	.051	.037
LI	.274	.215	.135	.094	.062	.044	.230	.167	.115	.083	.051	.037
LS	.277	.217	.137	.095	.062	.044	.233	.169	.117	.084	.051	.038
$\hat{S}_X$	.122	.085	.055	.038	.024	.017	.104	.075	.048	.034	.021	.015
CV	2.254	2.541	2.455	2.500	2.583	2.588	2.231	2.240	2.417	2.441	2.429	2.467
g_1	0.605	0.533	0.571	0.562	0.562	0.601	0.611	0.630	0.602	0.576	0.605	0.589
EE_{g_1}	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
g_2	0.311	0.191	0.331	0.265	0.314	0.467	0.328	0.361	0.331	0.239	0.331	0.270
EE_{g_2}	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035

Nota. Número de simulaciones = 20,000, valores perdidos = 0. *Min* = valor mínimo, *Max* = valor máximo, *P50* = mediana o percentil 50, *P75* = cuartil superior o percentil 75, *P80* = percentil 80, *P90* = percentil 90, *P95* = percentil 95, $\bar{X}$ = media aritmética, *LI* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, *LS* = límite inferior de la media aritmética en una estimación por intervalo con un nivel de confianza al 95%, $\hat{S}_X$ = desviación estándar con la corrección de Bessel, *CV* = coeficiente de variación de Pearson, g_1 = coeficiente de asimetría basada en el tercer cumulante estandarizado de Fisher, EE_{g_1} = error estándar del coeficiente de asimetría de Fisher, g_2 = exceso de curtosis basado en el cuarto cumulante estandarizado de Fisher, EE_{g_2} = error estándar del exceso de curtosis de Fisher.

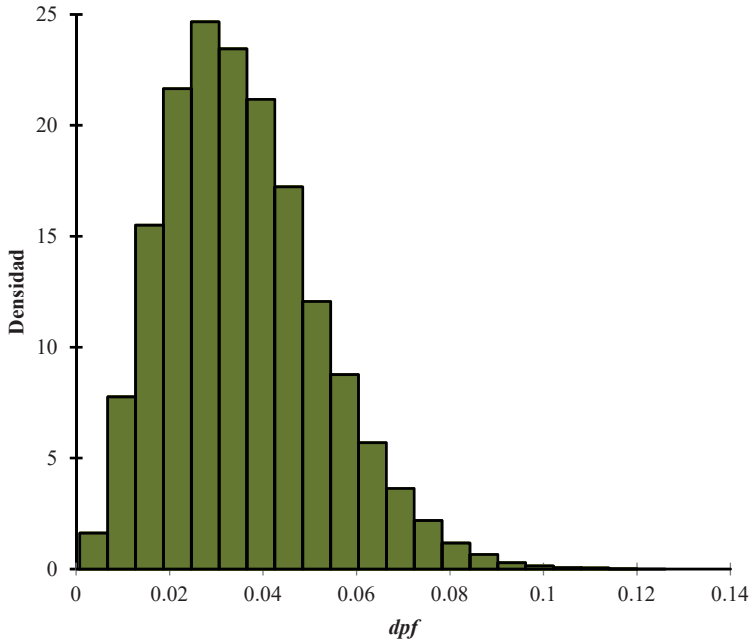


Figura 3. Histograma de la distribución de dpf para seis categorías nominales y un tamaño muestral de 1000

La mediana de los valores del percentil 95 (punto de corte más usual), generados para seis tamaños muestrales y 10 números distintos de categorías, fue de .09. Al estudiar el comportamiento del percentil 95 desde las Tablas 6 a 10, se observó que existe una correlación lineal inversa muy alta entre los valores del percentil 95 y el tamaño muestral (r_S = -.90). La relación entre los valores del percentil 95 y el número de categorías no fue lineal, sino que tuvo una forma de U (R^2 = .55 para un modelo cuadrático). A su vez, los valores del percentil 95 fueron ligeramente más altos cuando el número de categorías fue par (*Mdn* = .11) que cuando fue impar (*Mdn* = .09), aunque sin diferencia significativa (U = 398.5, Z = -.76, p = .446).

Discusión

Las variables cualitativas han recibido una amplia atención por parte de la estadística tanto descriptiva como inferencial. Se cuenta con tablas, gráficas, medidas de tendencia central, variación y asociación, pruebas de bondad de ajuste y técnicas de predicción, clasificación y reducción de dimensiones. A su vez, se ha estudiado la simetría en tablas de contingencia, tomando como eje de simetría la diagonal principal (frecuencias conjuntas de acuerdo), pero se ha ignorado el aspecto de la asimetría y el apuntamiento unidimensionales.

Para hablar de asimetría se requiere tener un eje a partir del cual se pueda evaluar si una parte de la distribución es igual a la otra parte (simetría) o es dispar (asimetría). En un principio, la opción natural parece la moda, pues es la medida de tendencia central de una variable cualitativa. No obstante, es una medida problemática, ya que puede ser múltiple, incluso no existir. Aparte, está el problema del orden de los datos. Las categorías nominales no poseen información intrínseca para su ordenación. Este problema se supera tomando una información extrínseca que sí tienen las categorías nominales, como es su frecuencia. La frecuencia permite un ordenamiento de las categorías. En este caso, no interesa simplemente un ordenamiento ascendente o descendente, pues no facilitaría observar gráficamente la simetría. Interesa un ordenamiento que lleve a un perfil triangular, si la moda es única; trapezoide, si hay dos o múltiples modas; o rectangular, si no hay moda. Se ubica en el centro una categoría de frecuencia máxima cuando k es impar o dos de las categorías con frecuencia máxima cuando es k par. Emparejadas las categorías por proximidad de frecuencia, se ubican los miembros de cada par a izquierda y a derecha de este centro, a la izquierda la categoría con más frecuencia del par y a la derecha la categoría con menos frecuencia del par. Los pares con frecuencias más altas quedan más cerca del centro y los pares con las frecuencias más bajas aparecen más alejados del centro. De este modo, el diagrama de barras permite visualizar si hay simetría o no.

Una cuantificación natural y muy sencilla de cálculo es promediar la diferencia entre las frecuencias de las categorías dispuestas en simetría. Cuando hay simetría, el valor de este promedio es cero. Cuando hay asimetría, el promedio se aleja de cero hasta alcanzar un máximo de 1. Esta medida nunca es negativa. Su máximo se encuentra con una variable aleatoria constante dicotómica. En esta distribución, una categoría concentra toda la frecuencia o probabilidad y la otra categoría tiene una frecuencia o probabilidad nula. El mínimo de 0 aparece con distribuciones como la uniforme discreta y aquellas que emulan a la Bernoulli o binomial con parámetro p = .5 o a la triangular discreta simétrica.

Los datos presentados muestran que el comportamiento del estadístico se ajusta al patrón esperado: bajar hacia cero cuando hay simetría con respecto al eje situado en la categoría central (k par) o el eje imaginario entre las dos categorías centrales (k impar) y subir hacia 1 cuanto más dispares son ambas partes. Este ajuste es constatado por la correlación entre el índice y la puntuación promedio de asimetría de los cinco jueces. La correlación es muy alta y, cuando se estima por intervalo con un nivel de confianza al 95%, incluye el 1. A su vez, la correlación entre los cinco jueces es alta, lo que indica que es una cualidad fácil de valorar visualmente por expertos. Para estudiar este comportamiento y obtener las gráficas evaluadas en grado de asimetría por los cinco jueces se acudió a la distribución binomial, al ser una variable discreta que puede variar de asimetría extrema negativa a asimetría extrema positiva, pasando por un punto medio de simetría.

También se usó la distribución binomial con parámetro p = .5 para estudiar la distribución del estadístico y obtener unos puntos de corte sugerentes de asimetría. La ventaja de esta distribución es que el dominio de la variable binomial (de 0 a n) permiten establecer un paralelismo con el número de categorías de una variable nominal (de 1 a k = n+1). Aparte esta distribución permite definir las distribuciones de los componentes del estadístico como proporciones binomiales y usar la aproximación a la distribución normal, lo que facilita la simulación

de datos. Finalmente, la probabilidad de éxito de un medio garantiza la simetría perfecta y el valor nulo del estadístico a nivel poblacional.

Se puede tomar como punto de corte para asimetría el percentil 95. Su mediana es .09 que constituiría la referencia más generalizada de punto de corte. No obstante, el valor del punto de corte es más alto cuanto menor es el tamaño muestral, ya que existe una relación lineal inversa muy alta entre el valor del percentil 95 y el tamaño muestral. A su vez, los valores más altos del percentil 95 aparecen con tamaños muestrales pequeños y grandes y los más bajos con las categorías centrales, alcanzándose el mínimo con cinco categorías, ya que existe una relación no lineal entre el percentil 95 y el número de categorías. También existe una tendencia a que el valor del percentil 95 sea más alto cuando el número de categorías es par que cuando es impar, aunque en última instancia no es significativa. Desde este patrón, se puede deducir que el punto de corte mínimo aparece con cinco categorías nominales y un tamaño de muestra de 1000 ($P95$ = .03) y el máximo con 10 categorías nominales y un tamaño muestral de 20 ($P95$ = .47).

Una limitación del estudio es haber usado exclusivamente la distribución binomial. Otra opción más compleja para simular datos y obtener los puntos de corte sería una distribución triangular discreta simétrica. Además, como la distribución binomial, la distribución triangular puede ser asimétrica en diversos grados, lo que posibilitaría constatar el adecuado comportamiento del estadístico *dpf* y corroborar los presentes resultados. Está distribución tiene tres parámetros: a (mínimo), b (moda), c (máximo) y estos tres parámetros deben cumplir la condición de que $a < b < c$. En caso de simetría, se pueden redefinir como: $Min = 2a < Mo = a + b < Max = 2b$. El dominio de una variable X con esta distribución es el conjunto finito= $\{2a, 2a+1, \ldots 2b-1, 2b\}$. Así, la cardinalidad del conjunto es: $n = b - a + 1$ y, como la distribución binomial, permite establecer un paralelismo con el número de categorías nominales.

Se concluye que sí se puede definir un concepto de asimetría para distribuciones de variables cualitativas. El promedio de las diferencias entre las frecuencias en disposición simétrica permite definir

un estadístico que oscila de 0 a 1, donde 0 indica simetría y 1 la asimetría máxima. Además, esta disposición permite una valoración visual confiable de la asimetría. Este estadístico es válido al mostrar un comportamiento ajustado a la definición de asimetría cualitativa. Se aproxima a 0 cuanto más simétricas o semejantes son las frecuencias o alturas de las barras equidistantes al eje de simetría y se aproxima a 1 cuanto más dispares son. A su vez, es preciso como indica su correlación muy alta con el promedio de las valoraciones de asimetría de los jueces expertos. La simulación Monte Carlo con base en la distribución binomial con parámetro $p = .5$, permite obtener unos puntos de corte (percentil 95) sugerentes de asimetría en función del número de categorías nominales y el tamaño muestral. El punto de corte es más alto cuanto menor es el tamaño muestral y con un número pequeño (2 o 3) y grande (< 6) de categorías nominales.

Se sugiere aplicar esta medida de asimetría en la descripción de las distribuciones de variables cualitativas, usar los puntos de corte generados y confirmar los presentes resultados, usando la distribución triangular simétrica. Por otra parte, se incita a desarrollar un concepto, una medida y unos puntos interpretativos de apuntamiento para describir distribuciones de variables cualitativas. La distancia vertical entre la frecuencia de la moda (pico) y el promedio de las frecuencias de las dos categorías inmediatamente adyacentes al centro (hombros) en la disposición triangular, trapezoide o rectangular descrita en este artículo podría ser de utilidad. En el caso de dos categorías, sería la diferencia entre las frecuencias de ambas categorías.

Agradecimientos

A la doctora Mónica González Ramírez, doctora Lucía del Carmen Quezada Berumen, doctor René Landero Hernández, doctor Leopoldo Daniel González y doctor Miguel Ángel Gutiérrez Barrón por su participación como jueces en la comprobación de la validez de *dpf* como medida de asimetría para variables cualitativas.

Referencias

Addinsoft (2021). *XLSTAT statistical and data analysis solution*. New York, USA. Disponible en https://www.xlstat.com/es

Altinay, G. (2016). A simple class of measures of skewness. *Munich Personal RePEc Archive,* Paper No. 72353, 1-13. Recuperado de https://mpra.ub.uni-muenchen.de/72353/

Bowker, A. H. (1948) A test for symmetry in contingency tables. *Journal of the American Statistical Association, 43*(244), 572–574. https://doi.org/10.1080/01621459.1948.10483284

Cole, T. J. y Altman, D. G. (2017). Statistics notes: percentage differences, symmetry, and natural logarithms. *British Medical Journal, 358*, 1-2. https://doi.org/10.1136/bmj.j3683

Fagerland, M. W., Lydersen, S. y Laake, P. (2017). *Statistical analysis of contingency tables*. Boca Raton, Fl: Chapman and Hall/CRC. https://doi.org/10.1201/9781315374116

Fisher, R. A. (1930). The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society of London, 130*(812), 16-28, https://doi.org/10.1098/rspa.1930.0185

Gupta, S. C. y Kapoor, V. K. (2020). Descriptive measures. In *Fundamentals of mathematical statistics*, *twelfth edition* (section 2, pp. 1-78). Sultan Chand & Sons.

Kelley, T. L. (1947). *Fundamentals of statistics*. Harvard University Press.

Microsoft Corporation. (2019). *Microsoft Excel 2019 for Windows*. Disponible en https://office.microsoft.com/excel

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C. & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia, 22*(1), 67-72. https://doi.org/10.4103/aca.ACA_157_18

Pearson, K. (1894). Contributions to the mathematical theory of evolution. I. On the dissection of asymmetrical frequency curves.

Philosophical Transactions of the Royal Society of London A, 185, 71-110. https://doi.org/10.1098/rsta.1894.0003

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London A, 186*, 343-414. https://doi.org/10.1098/rsta.1895.0010

Provost, S. B., Zareamoghaddam, H., Ahmed, S. E. & Ha, H. T. (2020). The generalized Pearson family of distributions and explicit representation of the associated density functions. *Communications in Statistics - Theory and Methods, 2020*, article 1843680, 1-14. https://doi.org/10.1080/03610926.2020.1843680

Sarka, D. (2021). Descriptive statistics. In *Advanced Analytics with Transact-SQL* (pp. 3-29). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-7173-5_1

Singh, A., Gewali, L. y Khatiwada, J. (2019). New measures of skewness of a probability distribution. *Open Journal of Statistics, 9*, 601-621. http://dx.doi.org/10.4236/ojs.2019.95039

Versluis, C. (2017). Skewness of continuous statistical distributions. *Social Science Research Network*, article 3088183. http://dx.doi.org/10.2139/ssrn.3088183

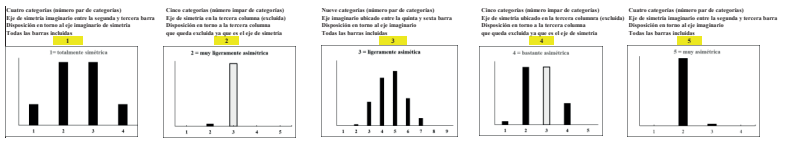
Weiss, C. H. (2020). Distance-based analysis of ordinal data and ordinal time series. *Journal of the American Statistical Association, 115*(531), 1189-1200. https://doi.org/10.1080/01621459.2019.1604370

Recibido: 2021-11-18
Revisado: 2021-11-20
Aceptado: 2021-11-26

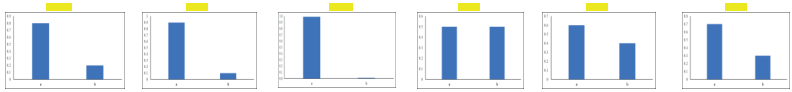
Anexo

Cuestionario entregado a los jueces para valorar la adecuación del estadístico dpf

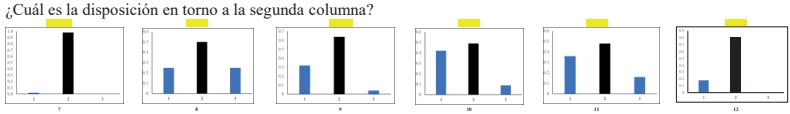
Instrucciones
Asigne a cada uno de los siguientes 60 diagramas de barras un grado de asimetría en el espacio marcado en amarillo. Valore si la disposición de las barras a ambos lados de la barra central (excluida) en el caso de un número impar de categorías o de la línea imaginaria entre las dos barras centrales (incluidas) en el caso de un número impar de categorías se puede considerar: 1 = totalmente simétrica, 2 = muy ligeramente asimétrica, 3 = ligeramente asimétrica, 4 = bastante asimétrica, o 5 = muy asimétrica. Use como referencia los siguientes ejemplos:



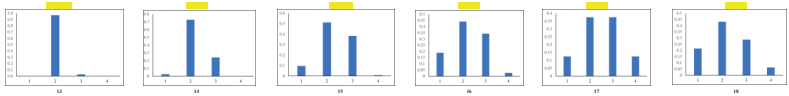
Dos categorías (número par de categorías)
Eje de simetría imaginario entre las dos barras
¿Cuál es la disposición en torno al eje imaginario de simetría?
Todas las barras incluidas



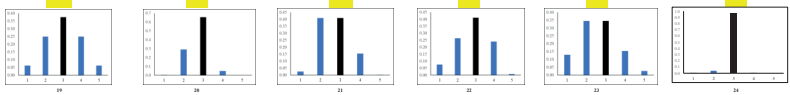
Tres categorías (número impar de categorías)
Eje imaginario de simetría ubicado en la segunda columna que queda excluida
¿Cuál es la disposición en torno a la segunda columna?



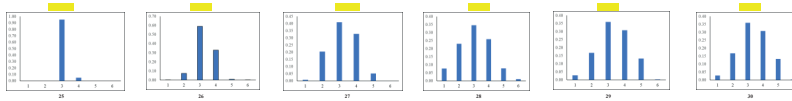
Cuatro categorías (número par de categorías)
Eje de simetría imaginario ubicado entre la segunda y tercera barra
¿Cuál es la disposición en torno al eje imaginario de simetría?
Todas las barras incluidas



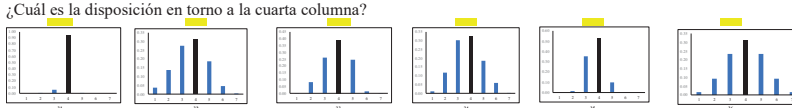
Cinco categorías (número impar de categorías)
Eje de simetría ubicado en la tercera columna que queda excluida
¿Cuál es la disposición en torno a la tercera columna?



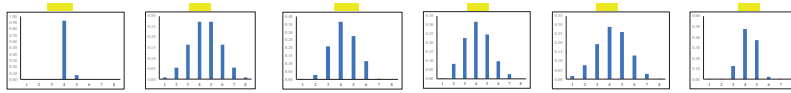
Seis categorías (número par de categorías)
Eje de simetría imaginario ubicado entre la tercera y cuarta barra
¿Cuál es la disposición en torno al eje imaginario de simetría?
Todas las barras incluidas



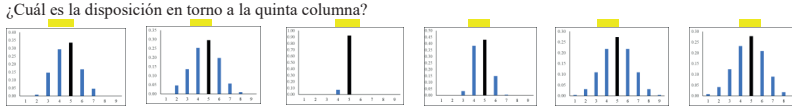
Siete categorías (número impar de categorías)
Eje de simetría ubicado en la cuarta columna que queda excluida
¿Cuál es la disposición en torno a la cuarta columna?



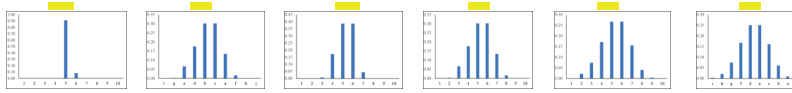
Ocho categorías (número par de categorías)
Eje de simetría imaginario ubicado entre la cuarta y quinta barra
¿Cuál es la disposición en torno al eje imaginario de simetría?
Todas las barras incluidas



Nueve categorías (número impar de categorías)
Eje de simetría ubicado en la quinta columna que queda excluida
¿Cuál es la disposición en torno a la quinta columna?



Diez categorías (número par de categorías)
Eje de simetría imaginario ubicado entre la quinta y sexta barra
¿Cuál es la disposición en torno al eje imaginario de simetría?
Todas las barras incluidas



Once categorías (número impar de categorías)
Eje de simetría ubicado en la sexta columna que queda excluida
¿Cuál es la disposición en torno a la sexta columna?