REVISTA CHILENA DE PEDIATRÍA
www.revistachilenadepediatria.cl
www.scielo.cl
Rev Chil Pediatr. 2020;91(4):642-643
DOI: 10.32641/rchped.v91i4.2300
CARTA AL EDITOR

Pearson Winsorizado: un coeficiente robusto para las correlaciones con muestras pequeñas

Pearson Winsorized: A robust coefficient for correlations with small samples

Sr. Editor

Recientemente se publicó un artículo de mucho interés en *La Revista Chilena de Pediatría*[1] el mismo que contiene una población pequeña por tratarse de un grupo con un diagnóstico específico, algo que resulta común en los estudios de pediatría. En ese sentido, la presente carta tiene como objetivo exponer al coeficiente Pearson Winzorizado (r_w) como una alternativa robusta para muestras pequeñas y ausencia de los supuestos de normalidad y presencia de valores atípicos en el ámbito de la pediatría.

Es conocido que el coeficiente por antonomasia en el estudio de relaciones bivariadas es el Producto Momento de Pearson (r_p), el mismo que requiere el cumplimiento de ciertos supuestos como normalidad bivariada, ausencia de valores atípicos, presencia de valores numéricos y dependencia lineal[2]. Pese a ello, es frecuente el incumplimiento de la normalidad y la presencia de valores atípicos en muestras pequeñas; en cuyo caso, se utiliza el coeficiente de Spearman (r_s). A pesar de eso, un estudio de simulación ha demostrado que el r_s tiene un mal desempeño en muestras pequeñas y diferentes formas de distribución, siendo el r_w un método más robusto ante estas situaciones[3].

Winzorizar, no es más que reemplazar valores extremos bajos o altos en el conjunto de datos con el fin de controlar los valores atípicos, los mismos que impactarán en la distribución de los datos. Lamentablemente, calcular el coeficiente r_w, no se encuentra en paquetes estadísticos tradicionales (v.e. SPSS) requiriendo explorar lenguajes de programación como R. En donde a través de la librería "WRS2"[4]. Se puede automáticamente obtener una correlación winzorizada (véase los códigos en la cuadricula).

```
library(WRS2)
rw1 = winall(data, tr = 0.20) #para correlacionar más de dos variables al mismo tiempo
rw1 = wincor(data, tr = 0.20) #para correlacionar solo dos variables.
rw1
```

A manera de ejemplo, se descarga una open data acerca del "Cuidado de la salud infantil en el Norte de Nueva Inglaterra" (https://data.world/dartmouthatlas/childrens-health-care-in-northern-new-england). Seleccionándose cuatro variables: "Visitas al consultorio por cada 1 000 niños" y "Visitas a la sala de emergencias por cada 1 000 niños", ambas variables siguen una distribución normal. Por otro lado, se selecciona "Número de niños en la población en estudio (2007-2010)" y "Médicos de familia por cada 100 000 niños" cuyas distribuciones son no-normales.

En la tabla 1, se observa la correlación de las variables previamente señaladas y se calcula los coeficientes Pearson, Spearman y Pearson Winzorizado. Los coeficientes de correlación en la condición de normalidad son casi similares (oscila entre 0,01 a 0,03) porque no existen valores atípicos y existe normalidad. Sin embargo, en la condición de no-normalidad y con 12 outliers, la diferencia entre los coeficientes se duplica (oscila entre 0,02 a 0,06). Si bien el coeficiente r_s presenta un valor más alto, recuerde que ahora dicho coeficiente representa rangos ordenados y que la función que utiliza es monotónica; es decir, se trata de una variable ordinal y se puede interpretar como "el grado en que las listas ordenadas se asocian". Pese a ello, r y r_w trabajan directamente con las variables continuas y al estar en la misma escala de medida sus valores son parecidos. Sin embargo, r_w resiste la no-normalidad y los valores atípicos, además, mantiene la escala continua y utiliza una función lineal, que son supuestos básicos de la relación bivariada.

Correspondencia:
José Ventura-León
jventuraleon@gmail.com

Cómo citar este artículo: Rev Chil Pediatr. 2020;91(4):642-643. DOI: 10.32641/rchped.v91i4.2300

En conclusión, la fórmula de r_p se basa en las medias aritméticas, que en muestras pequeñas son propensas a valores atípicos. Entonces, una medida robusta resulta necesaria. La tradición ha llevado a transformar los valores continuos a rangos ordenados y utilizar la r_s; sin embargo, la evidencia reciente demuestra que este coeficiente tiene altas tasas de error tipo I y baja potencia en diferentes formas de distribución y tamaños muestrales pequeños [3]. Además, el r_w es una mejor alternativa porque mantiene el supuesto de variable continua, atenúa el impacto de los valores atípicos; y trabaja mejor bajo condición de no-normalidad; algo que es común en los estudios reales.

Tabla 1. Comparación de los tres coeficientes en dos condiciones de normalidad

Valores	Normalidad	No-Normalidad
r_p	-0,65	-0,16
r_s	-0,64	-0,24
r_w	-0,62	-0,18
Outliers	0,00	12,00
n	69	69

Nota: r_p = Pearson; r_s = Spearman; r_w = Pearson Winzorizado

Referencias

1. Costa-Cordella S, Luyten P, Giraudo F, Mena F, Shmueli-Goetz Y, Fonagy P. Apego y estrés en niños con Diabetes tipo 1 y sus madres. Rev Chil Pediatr. 2020;91(1): 68-75. https://doi.org/10.32641/rchped.v91i1.1197
2. Lalinde JDH, Castro JFE, Tarazona MEP, Rodriguez JE, Rangel JGC, Sierra CAT, et al. Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones. Arch Venez de Farmacol Ter. 2018;37:587–95.
3. Tuğan E, Kocak M, Mirtagioğu H, Yiğt S, Mendes M. A simulation based comparison of correlation coefficients with regard to type I error rate and power. J Data Anal Inform Process. 2015;3:87. https://doi.org/ 10.4236/jdaip.2015.33010
4. Mair, P., Wilcox, R. Robust statistical methods in R using the WRS2 package. Behav Res. 2020;52:464-88. https://doi.org/10.3758/s13428-019-01246-w

José Ventura-León[a]

[a]*Universidad Privada del Norte, Lima, Perú.*

ORCID: 0000-0003-2996-4244