Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), enero-febrero 2025,
Volumen 9, Número 1.

https://doi.org/10.37811/cl_rcm.v9i1

ALTERNATIVAS PARA EL ANÁLISIS DE CORRELACIÓN CUANDO NO SE CUMPLE EL SUPUESTO DE NORMALIDAD BIVARIANTE: SIMULACIÓN Y EJEMPLOS EN R

ALTERNATIVES FOR CORRELATION ANALYSIS WHEN BIVARIATE NORMALITY ASSUMPTION IS NOT MET: SIMULATION AND EXAMPLES IN R

Jhonier Rangel
Universidad ECCI

Liliana Andrea Ahumada Suárez
Universidad ECCI

Ulpiano Lara-Cristancho
Universidad ECCI

José Alazate
Universidad ECCI

Pitter Javier Cabezas-Chacón
Universidad ECCI

DOI: https://doi.org/10.37811/cl_rcm.v9i1.16701

Alternativas para el análisis de correlación cuando no se cumple el supuesto de normalidad bivariante: Simulación y ejemplos en R

Jhonier Rangel [1]
jrangelg@ecci.edu.co
https://orcid.org/0000-0002-6849-5551
Universidad ECCI
Colombia, Bogotá D.C.

Liliana Andrea Ahumada Suárez
lahumadas@ecci.edu.co
https://orcid.org/0000-0003-3061-5172
Universidad ECCI
Colombia, Bogotá D.C.

Ulpiano Lara-Cristancho
ularac@ecci.edu.co
https://orcid.org/0009-0008-4905-5041
Universidad ECCI
Colombia, Bogotá D.C.

José Alazate
jalzater@ecci.edu.co
https://orcid.org/
Universidad ECCI
Colombia, Bogotá D.C.

Pitter Javier Cabezas-Chacón
pcabezasc@ecci.edu.co
https://orcid.org/0009-0003-9639-4891
Universidad ECCI
Colombia, Bogotá D.C.

RESUMEN

En este trabajo se realiza un análisis crítico sobre el uso del coeficiente de correlación, ya que, en muchos casos, se emplea de manera inadecuada, lo que puede generar conclusiones e interpretaciones sesgadas, basadas en comportamientos diferentes a los supuestos considerados. En primer lugar, se realiza una estimación de muestras bivariantes, tanto de tendencia como de presencia de datos atípicos, y se argumenta por qué, bajo estos escenarios, es inapropiado utilizar el coeficiente de correlación lineal de Pearson. Posteriormente, se presentan alternativas para evaluar intervalos de confianza sobre los coeficientes de correlación de Spearman, Kendall y Kappa de Cohen. Finalmente, se presenta un ejemplo utilizando los datos de cobertura educativa en primaria y secundaria en Colombia, ya que estos no cumplen con el supuesto de normalidad bivariante.

Palabras clave: correlación, remuestreo, atipicidad, normalidad

[1] Autor principal
Correspondencia: jrangelg@ecci.edu.co

Alternatives for correlation analysis when bivariate normality assumption is not met: simulation and examples in r

ABSTRACT

This paper presents a critical analysis of the use of the correlation coefficient, as it is often applied inappropriately, which can lead to biased conclusions and interpretations based on behaviors that deviate from the assumptions considered. First, an assessment of bivariate samples is carried out, focusing on trends and the presence of outliers, and the reasons why, under these scenarios, it is inappropriate to use Pearson's linear correlation coefficient are discussed. Subsequently, alternatives are presented to evaluate confidence intervals for the correlation coefficients of Spearman, Kendall, and Cohen's Kappa. Finally, an example is provided using data on primary and secondary education coverage in Colombia, as this data does not meet the bivariate normality assumption.

***Keywords*:** correlation, resampling, atypicality, normality

Artículo recibido 18 enero 2025
Aceptado para publicación: 24 febrero 2025

INTRODUCCIÓN

Tanto en las ciencias exactas como en las ciencias humanas el análisis de la correlación entre variables ayuda a evaluar dependencias. Sin embargo, en algunos casos estas dependencias pueden ser no lineales y no cumplir los supuestos básicos para hacer un análisis, por ese motivo es necesario explorar otro tipo de coeficientes de correlación no lineal que permite flexibilizar el supuesto de normalidad bi-variante.

En (Martínez r., et al., 2009) se presenta una caracterización por rangos del coeficiente de correlación de Spearman, el cual permite evaluar la relación entre variables de manera no lineal, pero con presencia de monotonía, tanto creciente como decreciente. Por otro lado, (Bastías et al., 2013) destacan la importancia de las Buenas Prácticas de Manufactura, señalando que, en áreas como la medicina, muchas variables, aunque se representan numéricamente, no son necesariamente variables de razón. (Santabárbara, 2019) Explica que en muchos estudios solo se presenta el valor estimado del coeficiente de variación. Sin embargo, el autor argumenta que es una buena práctica registrar un intervalo de confianza que acompañe la estimación puntual, ya que en muchos casos el coeficiente aparente puede estar muy alejado de cero. No obstante, al calcular el intervalo, este puede incluir el cero o tener un límite inferior cercano a cero, lo que indicaría que no existe una correlación significativamente alta. El autor también destaca la ventaja de utilizar SPSS para calcular el coeficiente de correlación de manera precisa, y en este artículo se explora el uso del programa R (https://www.R-project.org/).

En Schober et al. (2018) se indica que el coeficiente de correlación lineal de Pearson refleja la asociación entre variables, donde un cambio en una variable está vinculado a un cambio en otra, ya sea en la misma (correlación positiva) o en la dirección opuesta (correlación negativa). Para datos continuos y normalmente distribuidos, se utiliza comúnmente este coeficiente. Sin embargo, para datos no distribuidos normalmente, datos ordinales o aquellos con valores atípicos relevantes, se emplea el coeficiente de correlación de rangos de Spearman, que mide la asociación monótona. Ambos coeficientes tienen un rango de -1 a +1, donde 0 indica ausencia de asociación. Además, se pueden realizar pruebas de hipótesis y calcular intervalos de confianza para evaluar la significancia y la fuerza de la relación en la población de origen de los datos. No obstante, en esta investigación también se examinan escenarios en los que no se cumple el supuesto de normalidad, así como situaciones en las que los datos siguen distribuidos con comportamientos distribucionales diferentes.

n este artículo también se exponen las técnicas de remuestreo Leave One Out (LOO), que consiste en dejar un individuo fuera en cada remuestreo, y el Bootstrap, que implica extraer muestras mediante muestreo aleatorio simple. Con estas muestras, se calcula un estimador; en este caso, se obtiene el coeficiente de correlación para los individuos que pertenecen a la submuestra seleccionada. (Bishara & Hittner, 2012) ya abordaron el tema, considerando distribuciones distintas a la normal. Sin embargo, como diferencia clave de esta investigación, en este trabajo se considerará la distribución T multivariante, contaminada con diferentes proporciones de datos que siguen otro parámetro de centralidad y una matriz de varianzas y covarianzas distinta.

En el trabajo de (Serna-Morales , J. K., etal.,2024) se expone que los datos de falla bivariados son comunes en estudios de confiabilidad y supervivencia, donde la estimación de la fuerza de dependencia es a menudo un paso importante en el análisis de los datos. En la literatura, se ha establecido que los coeficientes de correlación miden la relación lineal entre dos variables, pero también pueden existir relaciones no lineales fuertes entre ellas. El coeficiente de concordancia t de Kendall se ha convertido en una herramienta útil para el análisis de datos bivariados, la cual es usada en pruebas no paramétricas de independencia y como una medida complementaria de asociación. En el análisis de datos de confiabilidad, hay un fenómeno que ocurre cuando el valor de las observaciones se conoce parcialmente, lo cual se conoce como censura. En este trabajo, se comparan vía simulación dos métodos de estimación del t de Kendall, una de ellas suponiendo normalidad en las distribuciones marginales y ajustándose individualmente, y la otra basada en cópulas (Gaussiana y Clayton), donde los datos bivariados están censurados a intervalo.

Otra alternativa que se expone en este artículo es el *Bootstrap*, cuyo fundamento teórico se detalla en la literatura. En este caso, se aborda su aplicación en el contexto multivariante. Es importante señalar que las muestras bootstrap univariantes para cada variable no son equivalentes a las muestras multivariantes bootstrap, y de hecho, sería erróneo tratar de relacionarlas directamente. Este matiz es crucial para comprender cómo utilizar correctamente el método en escenarios multivariantes, evitando confusiones y errores en la interpretación de los resultados (Zientek, L. R., & Thompson, B. , 2007).

Por otro lado, se llevará a cabo una simulación para examinar el impacto de los datos atípicos en la inferencia del coeficiente clásico. El objetivo es investigar cómo la presencia de valores extremos puede

modificar los resultados obtenidos mediante métodos estadísticos tradicionales, como el cálculo del coeficiente de correlación. A través de esta simulación, se analizará cómo distintas proporciones de datos atípicos afectan la precisión y validez de las estimaciones, lo que permitirá obtener una comprensión más profunda de las limitaciones de los enfoques clásicos cuando se enfrentan a datos que no siguen distribuciones convencionales.

METODOLOGÍA

En este artículo se analizan los coeficientes de correlación más conocidos, tales como el coeficiente de correlación lineal de Pearson, el coeficiente de monotonía de Spearman, el coeficiente de concordancia de Kendall (Shi, X. et al., 2024) y el coeficiente de fiabilidad kappa de Cohen (Fleiss, J. L., et al., 1969)).

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Un teorema importante que ayuda a interpretar el coeficiente de correlación lineal de Pearson es el siguiente:

Teorema: El coeficiente de correlación lineal de Pearson de dos variables cuantitativas es el coseno euclidiano de los dos vectores de observaciones centrados por la media.

Demostración:

El coeficiente de correlación de Pearson se puede interpretar como el coseno del ángulo entre los vectores que representan las variables centradas. A continuación, se muestra cómo demostrar esta relación:

$.r = \frac{\widehat{cov}(X,Y)}{\widehat{\sigma_X}\,\widehat{\sigma_X}}$

$$cos\ cos\ (\theta) = \frac{\sum_{i=1}^{n} \left(x_i - \underline{x}\right)\left(y_i - \underline{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \underline{x}\right)^2 \sum_{i=1}^{n} \left(y_i - \underline{y}\right)^2}}$$

Muchos métodos estadísticos requieren el uso de supuestos de distribución (Gutiérrez, J. S. R. , 2024) en este caso este coeficiente tiene el supuesto de la distribución normal bivariada, cuya expresión es la siguiente:

$$X \sim N(\mu, \Sigma)$$

Es importante aclarar que el vector de medias no tiene restricciones en el espacio vectorial de dimensión *n*. La matriz de varianzas y covarianzas debe ser definida positiva (Smania, G., & Jonsson, E. N., 2021).

$$f(X) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} exp\ exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Donde el vector de medias es un objeto que organiza las medias de cada variable, mientras que la matriz de varianzas y covarianzas contiene las varianzas en su diagonal y las covarianzas fuera de ella.

Otra distribución multivariante importante es la distribución t de Student multivariante, que, al igual que la anterior, requiere un vector de medias y una matriz de varianzas y covarianzas.

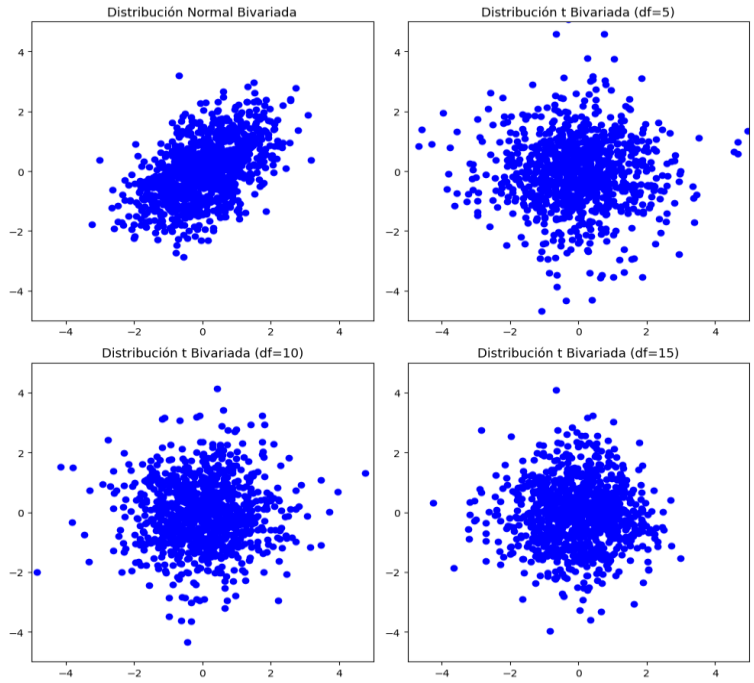
Un vector de medias, una matriz definida positiva y ν caracterizan la distribución t de Student multivariante. La distribución t multivariante es:

$$f_X(x) = \left(\frac{\Gamma\left(\frac{p+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)\left(\frac{1}{\pi}\right)^{\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\left(1+\frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+p}{2}}$$

Esta distribución multivariante tiene presencia de datos atípicos (Azzalini, A., & Capitanio, A. ,2003)).

Para visualizar esto se crea una grilla con cuatro paneles donde se simulan los valores de distribuciones normales bivariadas y t de Student bivariantes con 5, 10 y 15 grados de libertad, es posible usar Python junto con bibliotecas como Matplotlib para la visualización y NumPy para la generación de los datos en la extensión de Google Drive (https://colab.research.google.com/).

Figura 1. *Distribuciones bivariantes generadas con 100 puntos para cada caso. El primer panel muestra una distribución normal bivariada, mientras que los tres paneles siguientes muestran distribuciones t de Student bivariantes con 5, 10 y 15 grados de libertad, respectivamente. Todos los gráficos tienen la misma escala en los ejes X e Y para facilitar la comparación visual.*



En la figura 1, se ve que la distribución T multivariante se diferencia de la distribución normal bivariada principalmente en su manejo de los datos atípicos y en su forma analítica. La distribución normal bivariada se concentra alrededor de la media y, debido a su menor cola pesada, es menos propensa a generar valores extremos. En cambio, la distribución T multivariante, especialmente con grados de libertad bajos, tiene colas más pesadas, lo que aumenta la probabilidad de generar valores atípicos o extremos. Analíticamente, la distribución T depende de un parámetro adicional, los grados de libertad, que controla la forma de las colas. A medida que los grados de libertad disminuyen, las colas se hacen más gruesas y, por lo tanto, la probabilidad de observar datos atípicos aumenta significativamente. Este comportamiento contrasta con la distribución normal bivariada, en la cual los datos atípicos son menos frecuentes y la variabilidad se distribuye de manera más homogénea.

RESULTADOS Y DISCUSIÓN

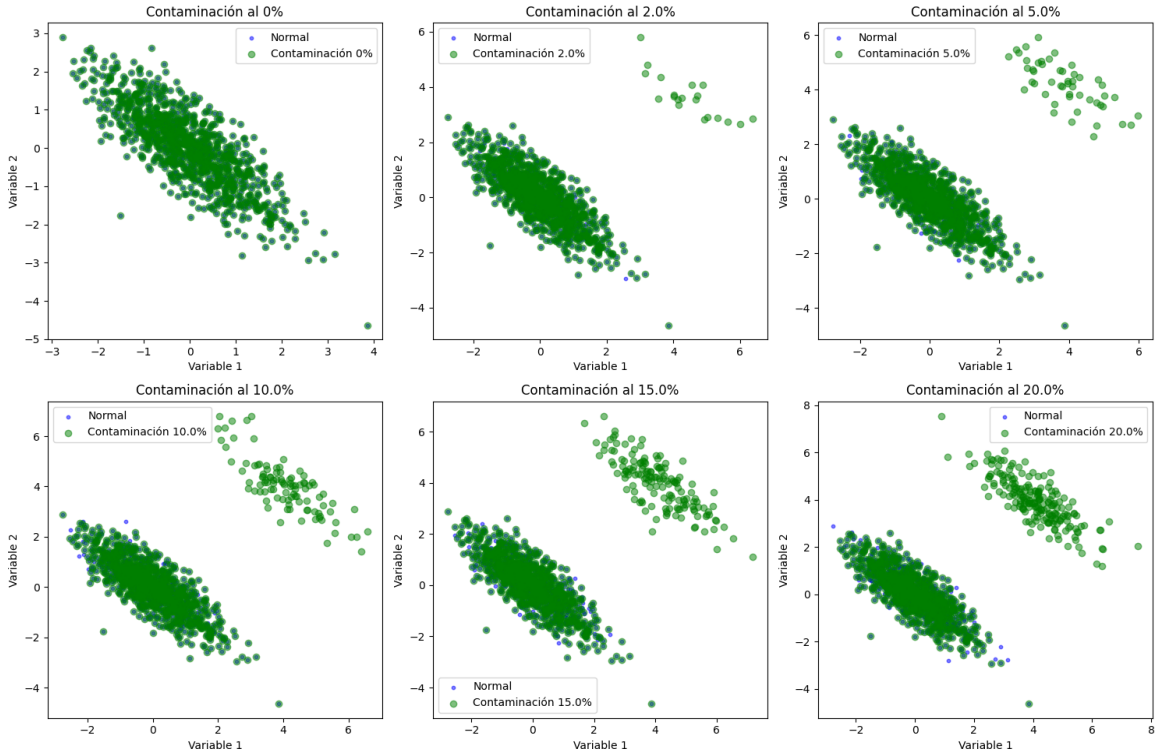
En este apartado se presentan las simulaciones y los resultados obtenidos a partir de datos normales bivariantes y de datos con distribución T bivariada, contaminados por diferentes proporciones de valores atípicos. Se analiza cómo estas contaminaciones afectan la inferencia basada en el supuesto de normalidad. Además, se realiza una aplicación utilizando los índices de cobertura primaria y secundaria en los municipios de Colombia durante el año 2023, con el fin de evaluar el impacto de estas

distribuciones en el análisis y los resultados obtenidos.

Se generó un análisis detallado de la relación entre dos variables bajo condiciones normales y contaminadas utilizando datos bivariantes con una distribución normal. Para ello, se generaron datos multivariantes con una media de [0,0] y una matriz de covarianzas que reflejaba un coeficiente de correlación de -0.8. Posteriormente, se crearon muestras contaminadas a diferentes niveles (5%, 10%, 15%, y 20%) mediante la modificación de una porción de las observaciones originales, lo que simuló la presencia de anomalías o perturbaciones en los datos. Además, se realizó una prueba de Mardia para evaluar la normalidad multivariante de cada conjunto de datos contaminado, obteniendo los valores p correspondientes para el sesgo y la curtosis, con el fin de examinar la influencia de la contaminación en la distribución de los datos.

En términos de correlación, se calcularon tres tipos de coeficientes: Kendall, Spearman y Pearson, para cada nivel de contaminación. El coeficiente de Pearson se utilizó para evaluar las relaciones lineales y normales entre las variables, mientras que los coeficientes de Kendall y Spearman fueron empleados para medir las relaciones no lineales o monotónicas. Los coeficientes de correlación, junto con los valores p obtenidos de la prueba de Mardia, fueron presentados en una tabla. Además, se generaron gráficos que ilustran la distribución de las observaciones originales y contaminadas, facilitando la interpretación visual del impacto de la contaminación en la relación entre las variables.

Figura 2. *Distribución de los datos originales y contaminados a diferentes niveles de contaminación (5%, 10%, 15%, 20%) en función de la variable 1 y variable 2. Los puntos azules representan los datos originales, mientras que los puntos verdes muestran los datos contaminados. Los gráficos ilustran cómo la contaminación afecta la relación entre las dos variables.*



Se evidencia que, aunque localmente hay presencia de correlación negativa, lo más probable es que estemos ante la presencia del efecto Simpson (Aiyou C., et al., 2009), ya que globalmente puede haber una correlación positiva. Este fenómeno ocurre cuando una tendencia observada en varios grupos desaparece o se invierte cuando los datos se combinan, lo que puede llevar a conclusiones erróneas si no se considera la estructura de los datos y la interacción entre las variables a nivel global y local.

Tabla 1. *Resultados de los valores p para los tests de sesgo y curtosis, junto con los coeficientes de correlación de Kendall, Spearman y Pearson, para distintos niveles de contaminación.*

Contaminación (%)	Coef. Kendall	Coef. Spearman	Coef. Pearson
0.00	-0.585578	-0.780344	-0.790705
0.02	-0.522967	-0.675797	-0.356312
0.05	-0.432012	-0.526342	-0.008416
0.10	-0.300268	-0.300594	0.269511
0.15	-0.177898	-0.100592	0.413265
0.20	-0.079512	0.067052	0.492009

En la tabla 1, se evidencia que el coeficiente de correlación de Pearson es especialmente sensible a las relaciones lineales entre dos variables, lo que lo hace más susceptible a los cambios en la magnitud de los datos. En la tabla proporcionada, se puede observar que a medida que aumenta el nivel de contaminación en los datos (de 0% a 20%), el valor de Pearson pasa de ser negativo (-0.79 en 0% de contaminación) a positivo (0.49 en 20% de contaminación), lo que indica una transición de una relación inversa a una directa a medida que las observaciones se desvían de la normalidad. Esta sensibilidad refleja cómo Pearson reacciona a las alteraciones en la relación lineal entre las variables, mientras que otros coeficientes como Kendall y Spearman, aunque también afectados, son menos sensibles a tales cambios y pueden captar relaciones no lineales o con más ruido.

En los datos contaminados, se puede observar cómo al aumentar el porcentaje de contaminación, los coeficientes de correlación cambian drásticamente. Lo interesante es que al contaminar los datos con un vector alejado del origen, los datos parecen mostrar una relación creciente entre las variables, que en realidad no existe en los datos originales. Este fenómeno es una manifestación clara del Efecto Simpson, que ocurre cuando, al combinar datos de diferentes grupos contaminados, se genera una ilusión de crecimiento en la relación entre las variables. Sin embargo, esta relación no es real, ya que al observar los datos por separado (sin contaminación o con poca contaminación), se evidencia que no hay una correlación positiva o incluso puede haber una relación negativa o inversa entre las variables. La contaminación de los datos al alejar el vector del origen produce un cambio significativo en la forma en que las variables parecen estar relacionadas, lo que puede generar conclusiones incorrectas si no se tiene cuidado al interpretar los resultados.

A continuación, se presentan las muestras generadas utilizando una distribución t multivariante con un vector de medias en el origen y una matriz de varianzas y covarianzas. Estas muestras se derivan de una mezcla de distribuciones normal y gamma, lo que da como resultado una distribución t multivariante. Para cada nivel de contaminación (0%, 2%, 5%, 10%, 15%, 20%), se generan datos contaminados a partir de esta mezcla, permitiendo observar los efectos de la contaminación en la estructura de los datos. El objetivo es estudiar cómo los datos contaminados afectan la distribución original y cómo los coeficientes de correlación, como los de Kendall, Spearman y Pearson, responden a la introducción de diferentes niveles de contaminación.

La metodología de mezcla entre distribuciones normal y gamma se utiliza para simular un escenario en el que las observaciones fuera de lo común son añadidas a los datos originales. Esta simulación ayuda a comprender el impacto de los valores atípicos en la distribución y permite evaluar cómo las correlaciones entre las variables cambian a medida que se aumenta la contaminación. En particular, la introducción de estos datos contaminados permite explorar los efectos del Efecto Simpson, el cual puede inducir a una interpretación incorrecta de las relaciones entre las variables, especialmente cuando los datos atípicos distorsionan la tendencia general.

Figura 3. *Visualización de los datos originales (en azul) y los datos contaminados (en rojo) a diferentes niveles de contaminación. Cada subgráfica corresponde a un nivel específico de contaminación, variando del 0% al 20%. Los puntos azules representan los datos generados sin contaminación, mientras que los puntos rojos muestran cómo los datos se ven alterados por la mezcla de distribuciones normal y gamma, evidenciando el efecto de la contaminación en la estructura de los datos.*

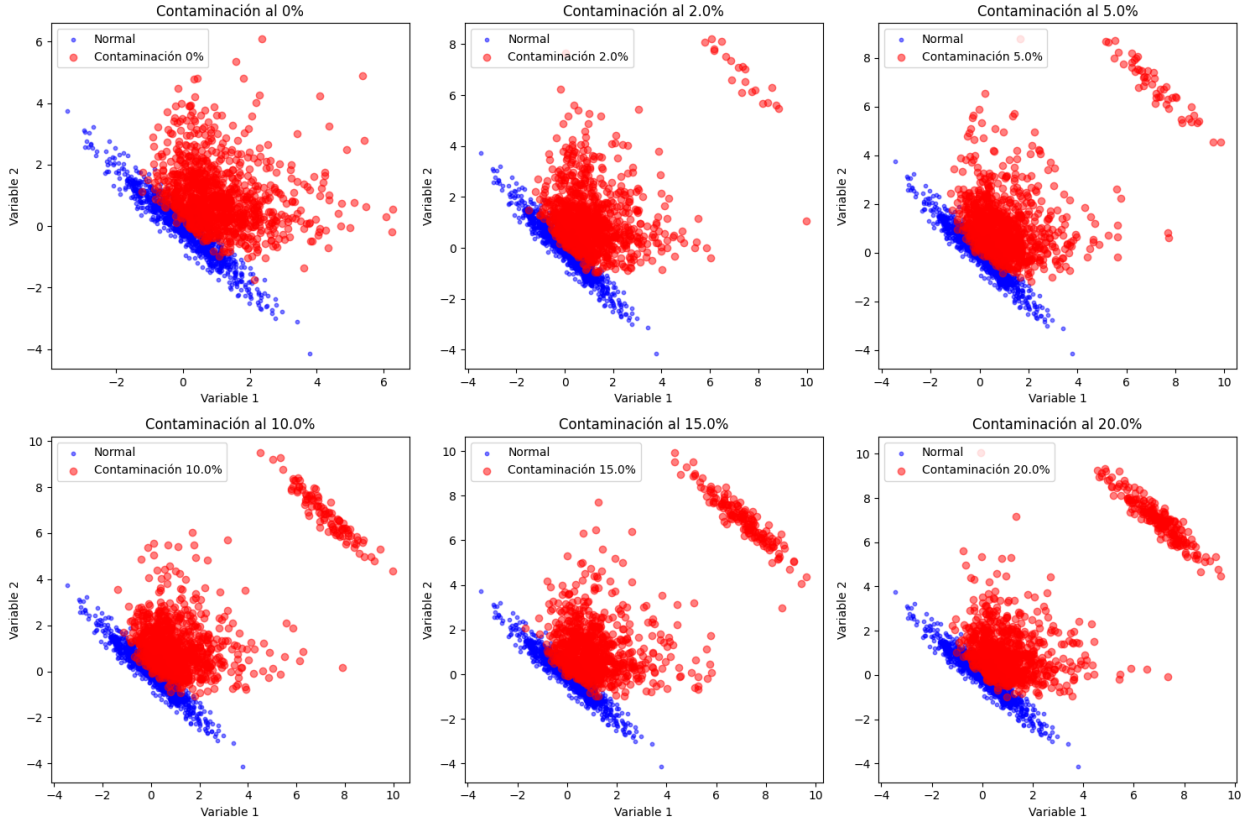


Tabla 2. Coeficientes de correlación de Kendall, Spearman y Pearson, junto con los valores de sesgo (Skewness) y curtosis (Kurtosis) para diferentes niveles de contaminación. Los valores de sesgo y curtosis están presentados con los límites inferior (Lim Inf) y superior (Lim Sup). Los coeficientes de correlación indican la relación entre las variables en función del porcentaje de contaminación en los datos.

%	Kendall	Spearman	Pearson	Sesgo (2.5%)	Sesgo (97.5%)	Curtosis (2.5%)	Curtosis (97.5%)
0.00	-0.011	-0.015	-0.024	1.058	1.235	1.724	2.271
0.02	0.011	0.017	0.034	1.963	2.260	5.499	7.692
0.05	-0.014	-0.022	-0.026	2.085	2.249	4.786	5.692
0.10	-0.048	-0.073	-0.064	1.836	1.867	2.741	2.754
0.15	0.031	0.047	0.049	1.426	1.451	0.916	0.919
0.20	0.005	0.009	0.014	1.185	1.209	0.065	-0.007

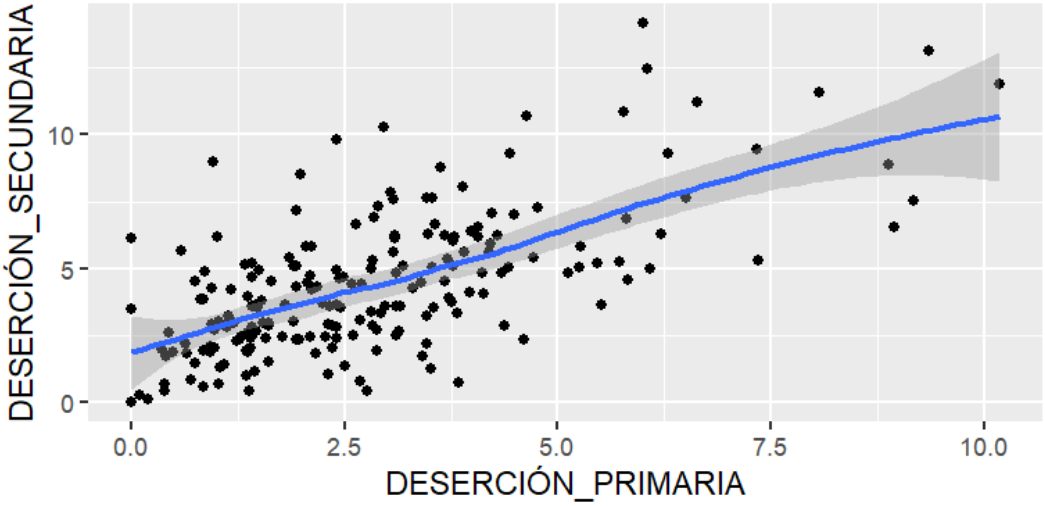
En la tabla 2, se observan variaciones en los coeficientes de correlación, sesgo y curtosis a medida que aumenta el nivel de contaminación de los datos. En cuanto a los coeficientes de correlación, el valor de Kendall, Spearman y Pearson muestra cambios moderados con el aumento de la contaminación. En particular, los coeficientes tienden a volverse más negativos o menos pronunciados cuando la contaminación es baja (como en el caso de 0% o 2%), sugiriendo que, a niveles bajos de contaminación, las relaciones entre las variables aún se mantienen relativamente fuertes, aunque con una ligera tendencia al decrecimiento. A medida que se incrementa la contaminación (especialmente a 10% y 15%), los coeficientes comienzan a estabilizarse o incluso a cambiar su signo, indicando una pérdida de la correlación entre las variables.

En cuanto a los valores de sesgo y curtosis, los datos muestran un comportamiento claro con la contaminación. Al principio, en niveles de contaminación bajos (0% a 5%), los valores de sesgo y curtosis se mantienen dentro de rangos razonables, pero a medida que la contaminación aumenta (especialmente a niveles más altos, como 10% y 15%), se observa un aumento significativo en la curtosis, lo que sugiere una mayor concentración de los datos alrededor de la media o la presencia de valores extremos más marcados. Este cambio en la curtosis es consistente con los efectos de la contaminación, que distorsionan la distribución original. En general, la contaminación tiene un impacto considerable sobre la normalidad de los datos, especialmente a medida que aumenta el nivel de contaminación, lo que se refleja en el cambio de los coeficientes y las medidas de dispersión como el sesgo y la curtosis.

Ahora, aplicaremos técnicas de remuestreo a los datos de deserción en primaria y secundaria en Colombia correspondientes al año 2023. Los datos utilizados están disponibles en el sitio web del Ministerio de Educación Nacional de Colombia y abarcan las tasas de deserción por cada departamento tanto en

primaria como en bachillerato. En la gráfica se presenta la estimación de la regresión kernel, destacando la posible tendencia creciente observada en las tasas de deserción.

Figura 4. Deserción en secundaria en relación con la deserción en primaria, representada mediante un gráfico de densidad kernel con intervalos de confianza del 95%.



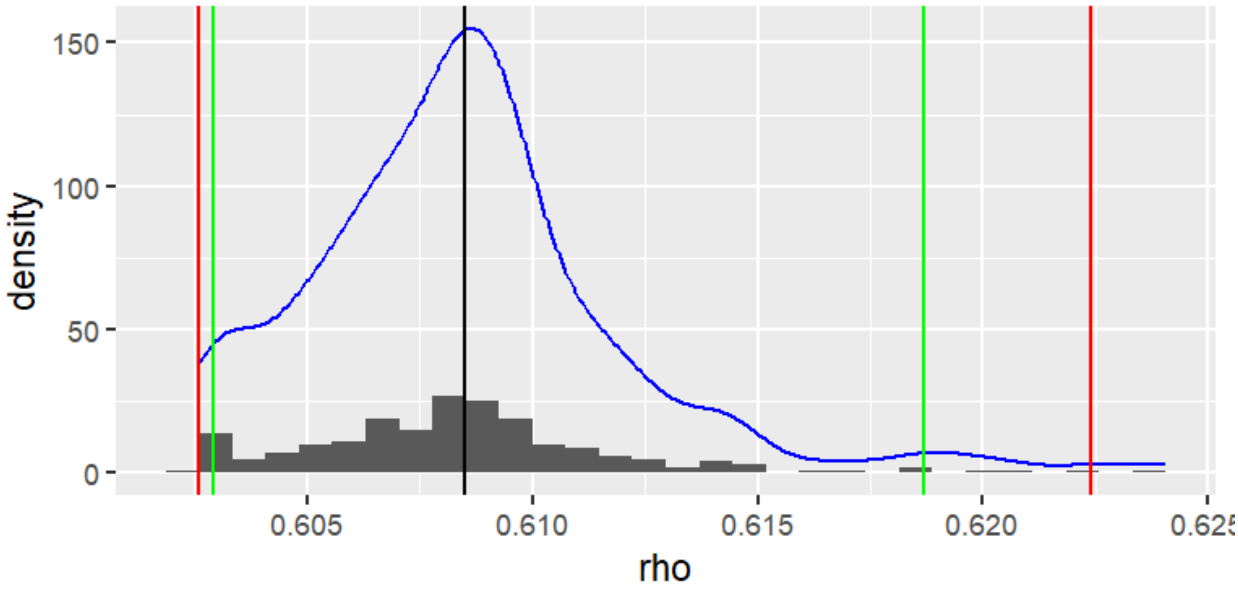
En la figura 4 se observa una relación positiva directa entre las variables. Sin embargo, para evaluar adecuadamente la correlación, es necesario recurrir a un índice de correlación y, en este caso, al coeficiente de correlación lineal de Pearson. No obstante, antes de utilizar este coeficiente, es fundamental someter los datos a una prueba de normalidad multivariante. En este estudio, se empleó el test de Mardia (Wulandari, D., et al., 2021), cuyo resultado arrojó un valor p de 0.012. Dado que este valor es menor al nivel de significancia del 5%, se rechaza la hipótesis nula, lo que indica que los datos no siguen una distribución normal multivariante. Por lo tanto, el uso del coeficiente de correlación lineal de Pearson no es adecuado en este caso.

En este artículo, se ha utilizado un enfoque basado en técnicas de remuestreo para analizar la relación entre la deserción en la educación en primaria y en secundaria en Colombia, a partir de datos proporcionados por el Ministerio de Educación Nacional en 2023. Las técnicas de remuestreo aplicadas incluyen "Leave One Out" (LOO), que permite evaluar la estabilidad de las estimaciones obtenidas al excluir sucesivamente una observación en cada iteración. Esta metodología es útil para obtener intervalos de confianza y evaluar el comportamiento de los estimadores sin depender de supuestos de normalidad, los cuales no se cumplen en este caso según lo indicado por los resultados de las pruebas de normalidad multivariante.

En lugar de utilizar un único conjunto de datos, la técnica Leave One Out permite crear muestras

repetidas al eliminar una observación de cada vez y calcular las estadísticas de interés, como la media y la correlación, para cada muestra. Este enfoque nos proporciona una visión más robusta de los resultados, ya que reduce el sesgo que podría generarse si se utilizara el conjunto de datos completo. De esta manera, se obtiene un intervalo de confianza del 95% para las estadísticas calculadas, como el coeficiente de correlación de Spearman entre la deserción primaria y secundaria. Además, se calcula el valor de la correlación en cada iteración para observar la variabilidad de los estimadores. Estas técnicas, combinadas con gráficos de distribución, permiten evaluar la confiabilidad de las conclusiones obtenidas sobre la relación entre las variables de deserción escolar.

Figura 5. Distribución de los coeficientes de correlación de Spearman para las diferentes muestras generadas mediante la técnica "Leave One Out". En la gráfica se muestra el histograma junto con la densidad estimada, así como los intervalos de confianza al 95% (en verde) y los percentiles del 0.5% y 99.5% (en rojo). La línea negra indica la media de los coeficientes de correlación.



En la Figura 5 se observa que los intervalos de confianza generados mediante las metodologías de remuestreo no son simétricos. Esto contrasta con los intervalos derivados bajo la suposición de un modelo distribucional previamente definido, los cuales sí presentan simetría. Esta asimetría sugiere que asumir un modelo de distribución específico podría llevar a una aproximación más precisa. Además, se puede concluir que existe una correlación directa significativa entre la deserción primaria y secundaria, ya que el intervalo de confianza no incluye el valor cero. De hecho, el intervalo obtenido es un subconjunto que supera el umbral de 0.5, lo que refuerza la presencia de una relación positiva entre ambas variables.

CONCLUSIONES

Esta investigación subraya la importancia de un uso adecuado del coeficiente de correlación,

especialmente cuando se consideran los supuestos subyacentes de los métodos estadísticos. El análisis reveló que, en situaciones donde los datos no cumplen con los supuestos de normalidad multivariante o presentan atípicos, el uso del coeficiente de correlación lineal de Pearson puede no ser apropiado, ya que podría generar interpretaciones erróneas. Ante esta situación, las alternativas de remuestreo, como el método Leave-One-Out y el bootstrap, se presentan como herramientas robustas, ya que permiten obtener estimaciones de intervalos de confianza sin depender de distribuciones paramétricas. Estas técnicas no solo ofrecen una mayor flexibilidad, sino que proporcionan estimaciones más confiables al evitar las restricciones asociadas con los métodos tradicionales. En el contexto colombiano, los resultados indican que existe una correlación significativa y directa entre la deserción escolar en primaria y secundaria. Este hallazgo es crucial para los responsables de las políticas educativas, ya que sugiere que las intervenciones en el nivel primario podrían tener un impacto directo en la reducción de la deserción en niveles educativos posteriores. En este sentido, es fundamental que las autoridades educativas utilicen estos análisis para diseñar políticas más eficaces que aborden de manera integral los factores que afectan la permanencia de los estudiantes en el sistema educativo.

REFERENCIAS BIBLIOGRÁFICAS

Martínez Ortega, R. M., Tuya Pendás, L. C., Martínez Ortega, M., Pérez Abreu, A., & Cánovas, A. M. (2009). El coeficiente de correlación de los rangos de Spearman caracterización. *Revista Habanera de Ciencias Médicas*, *8*(2), 0-0.

Bastías, J. M., Cuadra, M., Muñoz, O., & Quevedo, R. (2013). Correlación entre las buenas prácticas de manufactura y el cumplimiento de los criterios microbiológicos en la fabricación de helados en Chile. *Revista chilena de nutrición*, *40*(2), 161-168.

Santabárbara, J. (2019). Cálculo del intervalo de confianza para los coeficientes de correlación mediante sintaxis en SPSS. *REIRE Revista d'Innovació i Recerca en Educació*, *12*(2), 1-14.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763-1768. https://doi.org/10.1213/ane.0000000000002864

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological*

methods, *17*(3), 399.

Zientek, L. R., & Thompson, B. (2007). Applying the bootstrap to the multivariate case: Bootstrap component/factor analysis. *Behavior research methods*, *39*(2), 318-325. https://doi.org/10.3758/BF03193163

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*(5), 323–327. https://doi.org/10.1037/h0028106

Serna-Morales, J. K., Elorza, M. C. J., & Lopera-Gómez, C. M. (2024). Comparación de algunas estimaciones del t de Kendall para datos bivariados con censura a intervalo. *Ciencia En Desarrollo*, *15*(1), 130–140. https://doi.org/10.19053/01217488.v15.n1.2024.15586

Smania, G., & Jonsson, E. N. (2021). Conditional distribution modeling as an alternative method for covariates simulation: comparison with joint multivariate normal and bootstrap techniques. *CPT: pharmacometrics & systems pharmacology*, *10*(4), 330-339. https://doi.org/10.1002/psp4.12613

Gutiérrez, J. S. R. (2024). Ventajas del uso de las ecuaciones diferenciales estocásticas: puente browniano y movimiento browniano geométrico. *Multidisciplinary & Health Education Journal*, *6*(1), 830-838.

Gutierrez, J. S. R. (2024). Métodos inferenciales sobre cadenas de Markov en tiempo discreto con espacio de estados finito. In *Actas del II Congreso Internacional de Innovación, Ciencia y Tecnología INUDI–UH, 2024* (pp. 277-288). Instituto Universitario de Innovación Ciencia y Tecnología Inudi Perú. https://doi.org/10.35622/inudi.c.02.15

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(2), 367-389. https://doi.org/10.1111/1467-9868.00391

Chen, A., Bengtsson, T., & Ho, T. K. (2009). A regression paradox for linear models: Sufficient conditions and relation to Simpson's paradox. *The American Statistician*, *63*(3), 218-225. https://doi.org/10.1198/tast.2009.08220

Wulandari, D., Sutrisno, S., & Nirwana, M. B. (2021). Mardia's skewness and kurtosis for assessing

normality assumption in multivariate regression. *Enthusiastic: International Journal of Applied Statistics and Data Science*, 1-6. https://doi.org/10.20885/enthusiastic.vol1.iss1.art1

Shi, X., Jiang, Y., Du, J., & Miao, Z. (2024). An adaptive test based on Kendall's tau for independence in high dimensions. *Journal of Nonparametric Statistics*, *36*(4), 1064-1087. https://doi.org/10.1080/10485252.2023.2296521