APLICACIÓN DE LOS COEFICIENTES CORRELACIÓN DE KENDALL Y SPEARMAN

(Application of the Kendall correlation and Spearman coefficients**)**

Pedro Morales[1], Luis Rodríguez[2]

[1] Departamento de Psicología. Decanato de Humanidades y Artes, Universidad Centroccidental "Lisandro Alvarado". (UCLA) Barquisimeto, Email: Pedromorales@ucla.edu.ve.

[2] Departamento Técnicas Cuantitativas. Decanato de Ciencias Económicas y Empresariales (DCEE). Universidad Centroccidental "Lisandro Alvarado". (UCLA,) Barquisimeto, Email: luisler@hotmail.com.

Recibido: 15/01/16 Aceptado: 28/03/16

RESUMEN

En este trabajo se presenta el marco conceptual y el desarrollo de los principios básicos que subyacen el análisis de correlación para su adecuada aplicación en la investigación científica. Se describen los indicadores más utilizados que permiten la cuantificación del grado de relación lineal entre variables del tipo cualitativa con escalas ordinal. Por otro lado, se quiere aclarar el significado de algunos términos relacionados con la correlación ya que comúnmente están siendo manejados de manera equívoca por el público en general. Es el caso de los términos correlación y relación causal donde muchos investigadores y estudiantes entienden que dos cosas están correlacionadas cuando creen que están relacionadas, y que las cosas no están correlacionadas cuando creen que no están relacionadas. Una relación causal entre dos variables existe si la ocurrencia de la primera causa la de la otra. La primera variable es llamada la causa y la segunda variable es llamada efecto. Una correlación entre dos variables no implica causalidad. Por tanto, si hay una relación causal entre dos variables, estas deben estar correlacionadas. También es importante acotar que dos variables al no estar correlacionadas es equivalente a decir que ambas variables son independiente, lo reciproco no se cumple en algunos casos. Los coeficientes de asociación lineal más aplicados en las investigaciones científicas, y que serán discutidos teóricamente corresponden al coeficiente de correlación de Kendall y de rangos de Spearman.

Palabras Clave: *Correlación lineal, correlación de Spearman, coeficiente de Kendall.*

SUMMARY

This paper presents the conceptual framework and the development of the basic principles underlying the correlation analysis for their proper application in scientific research. Described most commonly used indicators that allow the quantification of the degree of linear relationship between variables of the qualitative type scales ordinal. On the other hand, to clarify the meaning of some terms related to the correlation since they are commonly being handled misleadingly by the public in general. The case of terms is correlation and causation where many researchers and students understand that two things are correlated when they believe that they are related, and that things are not correlated when they believe that they are not related. There is a causal relationship between two variables if the occurrence of the first cause of the other. The first variable is called the cause and the second variable is called the effect. A correlation between two variables does not imply causality. Therefore, if there is a causal relationship between two variables, they must be correlated. It is also important to note that two variables not being correlated is equivalent to say that both variables are independent, not so reciprocal is true in some cases. Coefficients of linear Association applied in scientific research, and that they will be discussed theoretically corresponding to the ranges of Spearman and Kendall correlation coefficient.

Keywords: Linear correlation, Spearman's rank correlation coefficient of Kendall.

INTRODUCCIÓN

El análisis de correlación es una de las metodologías estadísticas descriptivas ampliamente utilizada en la mayoría de las investigaciones aplicadas, con la que se pretende comúnmente estimar el nivel de asociación lineal entre las variables objeto de estudio. Cuando las variables están correlacionadas, los investigadores se interesan en realizar predicciones para dichas variables, pero están limitados de realizar inferencias de la causa de la relación.

Existen ciertas recomendaciones básicas que deben de tomarse en cuenta antes de desarrollar un análisis de correlación, entre las que destacan: trazar el gráfico de dispersión entre las variables estudiadas, construir el histograma de frecuencia el cual dará cuenta de la asimetría y puntiagudez en la distribución de los datos, la presencia de valores atípicos y grado de dispersión; construir el gráfico de probabilidad normal; y determinar los estadístico de normalidad asociados con Kolgomorov-Smirnov ó el Wilk-Shapiro. Todos estos elementos descriptivos introductorios proporcionaran las características suficientes que rodean el comportamiento de las variables investigadas, para decidir cuál de los coeficientes de correlación es el más apropiado de acuerdo con estos elementos.

Al realizar estudios correlacionales los investigadores deben ser cautelosos ya que la relación entre variables puede ser explicada por una tercera variable, y en tal caso esa relación se le denomina correlación espuria o confundida. Bajo estas condiciones los investigadores pueden utilizar otras técnicas estadísticas como el análisis de variables mediadoras y moderadoras en el análisis de rutas (path analysis), o un enfoque multimetódico. Por tanto, el objetivo de este trabajo es presentar una serie de pautas teóricas y recomendaciones para la adecuada aplicación de los coeficientes de correlación ampliamente utilizados por la comunidad científica.

DESARROLLO

La asociación entre dos variables suele ser de interés en el análisis de datos y la investigación metodológica. Para Shong N. (2010, pag. 1) citando a Sheskin D.J.(2007), expresa que las medidas de asociación no son pruebas estadísticas inferenciales por el contrario son medidas estadísticas descriptivas que demuestran la dirección, fuerza o grado de relación entre dos o más variables. El análisis de correlación es una metodología estadística que trata de establecer la relación entre dos o más variables. Es decir, es un procedimiento que busca evaluar la relación entre las diferencias individuales (casos o sujetos) según dos o más variables aleatorias estudiadas, ya que lo fundamental de la técnica de comparación es evaluar la naturaleza de las

diferencias entre individuos y no entre los grupos o tratamientos. Es por ello, que este tipo de análisis estadístico se le denomina estudio de diferencias individuales.

Por tal razón, los datos obtenidos en este tipo de análisis consisten en pares de observaciones de una serie de individuos o casos (objetos, productos, etc.). En este sentido, la correlación mide el grado o la intensidad de la relación entre dos o más variables y refleja lo cerca que están los puntos (pares de coordenadas) de una línea recta con pendiente positiva o negativa. Si todos los puntos están muy cerca de la línea la relación es fuerte, y si muchos de los puntos se encuentran retirados de la línea se dice que la relación es débil. Cuando la relación es cero, se considera débil, ya que no existe relación entre las variables correlacionadas.

Al comparar los términos de correlación y relación causal se puede notar que existe diferencias desde el punto de vista estadístico, ya que cuando se estudia la correlación (magnitud y sentido) entre variables, se pretende constatar si existe o no una asociación lineal entre las mismas, en otras palabras, se realiza una comparación o descripción de dos o más variables diferentes, pero juntas, deduciéndose que las variables están o no correlacionadas. Mientras la causalidad se refiere a la causa y efecto de un fenómeno, en el que una cosa provoca directamente el cambio de otra.

Así, el hecho de que dos variables parezcan estar correlacionadas no necesariamente significa que una esté causando a la otra, por cuando pudiera estar ocurriendo una de dos cosas: a) la relación podría ser falsa o casual; ó b) la relación entre las variables también puede ser el resultado de una tercera variable que "causa" o explica las otras dos, provocando que las dos variables causadas por esta tercera parezcan estar relacionadas entre sí.

Un ejemplo ilustrativo que aclara perfectamente esta situación, es el caso cuando un investigador estudia las habilidades numéricas y la talla de los estudiantes en un colegio, llegando probablemente a la conclusión de que existe una correlación alta y positiva (r_{xy}>0,80) entre ambas variables, mientras mayor es la talla de los estudiantes, mayor son sus habilidades numéricas. Intuitivamente, se sabe que la talla no hace que los estudiantes aprendan matemática ni que el aprendizaje de la aritmética hace que los estudiantes sean más altos. En este ejemplo se evidencia la existencia de una tercera variable que explica la correlación entre las mejores habilidades numéricas y la talla de los alumnos, la cual sería la "edad" de los estudiantes. Así, no es que los estudiantes más altos sean mejores para los números, sino que los estudiantes con mayor edad, o los que se encuentran en grados superiores en el colegio, tienden a ser más altos y a tener mayores habilidades numéricas.

Existe una tercera variable que explica las dos variables, por este motivo parecen estar correlacionadas, pero en realidad no hay ninguna relación entre ellas. Kenny (1979, pág. 69) indica, si la correlación entre dos variables esta siendo explicada por una tercera se le denomina **relación espuria**.

Además, se hace necesario insistir que muchos de los métodos estadísticos inferenciales se basan en la pruebas de normalidad para las variables objeto de estudio. De hecho, si la falta de normalidad de la variable es suficientemente fuerte, muchos de los contrastes utilizados en los análisis estadístico-inferenciales no son válidos, incluyendo el análisis de correlación propuesto por Pearson (1896), y en muchas investigaciones cuando se presenta el análisis de correlación no se menciona la forma de distribución de los datos en las variables estudiadas. Incluso, aunque las muestras grandes tiendan a disminuir los efectos perjudiciales de la no normalidad, el investigador debería evaluar la normalidad de todas las variables incluidas en el estudio.

Existen diferentes técnicas para evaluar la normalidad de un conjunto de datos, que pueden dividirse en dos grupos: los ***métodos gráficos*** y los ***contrastes de hipótesis***. Entre los métodos gráficos destacan: el más simple de los gráficos univariantes es el **histograma de frecuencias** con la curva normal, por cuando da un percepción visual entre la distribución de los datos y la curva normal, teniendo presente que para su construcción se deben contar con muestras de al menos 30 datos, y el gráfico de **probabilidad normal**, que según Montgomery (2004, pag. 38) expresa que es una técnica útil para determinar si los datos de una muestra se ajustan a una distribución normal con base a un examen visual "subjetivo" de los datos. Mientras que los métodos numéricos basados en pruebas de hipótesis destacan: la prueba de **Kolgomorov-Smirnov** (versión modificada según Lillieffors, 1967) y la **Shapiro-Wilks** (1965). Pardo (2002, pag. 216) expone que la prueba de Kolgomorov-Smirnov es adecuada para muestras grandes y la de **Shapiro-Wilks** cuando los tamaños muestrales tienden a ser pequeños ($n \leq 50$).

En el análisis de datos existen tres métodos ampliamente utilizados para determinar la correlación monótona entre variables. Estos coeficientes son: el de Pearson, Spearman y Kendall; pero su estimación va a depender de la naturaleza de las variables utilizada, si es cualitativa (ordinal) o cuantitativa (razón). El coeficiente de correlación se denota como parámetro con la letra griega *rho* $\boldsymbol{\rho}$, cuyos valores presentan la característica de ser adimensionales y se mueven en un rango válido de -1 a +1 ($-1 \leq \boldsymbol{\rho} \leq +1$). Es este trabajo solamente se enfoca la atención en los coeficientes de correlación de Kendall y Spearman.

Coeficiente de Correlación por Rangos de Kendall (τ).

Cuando se estudia la relación entre variables cualitativas de tipo ordinal se debe utiliza el coeficiente de correlación de rangos de Kendall (1938), denominado τ (**tau**) de Kendall, del cual existen dos variantes **tau-b y tau-c**; además su aplicación tiene sentido si las variables objeto de estudio no poseen una distribución poblacional conjunta normal; es decir, si se requiere determinar el grado de asociación lineal entre dos variables cuantitativas pero las mismas no siguen un comportamiento normal, será preferible estimar este indicador mediante el coeficiente de Kendall. Como este indicador está basado en rangos y no en los datos originales, su estimación requiere que los valores de la variable ordinal sean transformados en rangos, este coeficiente se ve poco afectado ante la presencia de un número pequeño de valores atípicos (extremos) en la muestra estudiada, adaptándose bien en aquellas variables que reportan moderadas asimetrías en torno a la relación general.

Una característica notable del coeficiente de Kendall es que reporta valores más bajos con respecto a los coeficientes de Spearman y Pearson, en aquellas situaciones donde se analiza las asociaciones lineales con la misma intensidad (sin la presencia de valores atípicos); por ejemplo, se presentan casos donde fuertes correlaciones son reportadas por Spearman y Pearson, digamos al menos 0,90, mientras Kendall lo reportaría alrededor de 0,70. Este resultado no se puede traducir como si el coeficiente de Kendall es menos preciso que los otros dos, sino que **tau** se determina a partir de valores ranqueados, e inclusive reporta la misma tendencia en datos con distribución monótonas no lineales, en comparación con Spearman. Según Siegel y Castellan (1998, pag. 285, 287), para determinar el coeficiente de correlación de Kendall en una muestra de tamaño n pares de datos, tomados de dos variables aleatorias digamos X e Y, el cual se denota por $\mathbf{T_{xy}}$, se utiliza la expresión de la **Tabla 1**.

Tabla 1. Fórmulas para el cálculo de la correlación de Kendall.

Caso 1: Sin Observaciones Empatadas:

(1)

$$T_{XY} = \frac{2 \cdot S}{n(n-1)}$$

Donde:

S = P – M.

P= Nro. de valores positivos o "Acuerdos". Esto es el número de veces de incrementos de Y conforme se incrementa X, o el número de $y_i < y_j$ para todo i <j.

M = Nro. de valores Negativos o "Desacuerdos". Esto es el número de veces que disminuye Y cuando se incrementa X, o el número de $y_i > y_j$ para i <j.

Para todo i = 1, ..., (n - 1) y j = (i + 1), ..., n.

Caso 2: Con Observaciones Empatadas:

(2)

$$T_{XY} = \frac{2 \cdot S}{\sqrt{n(n-1) - T_X} \cdot \sqrt{n(n-1) - T_Y}}$$

Donde:

$T_X = \sum t\,(t-1)$, siendo t el número de observaciones empatadas en cada grupos de empates de la variable X.

$(\quad)$

$T_Y = \sum t \quad t-1$ = siendo t el número de observaciones empatadas en cada grupos de empates de la variable Y.

El coeficiente T_{xy} de Kendall varía de -1 a +1. Como se puede notar se realizan un total de n(n-1)/2 posibles comparaciones de n pares de datos, en consecuencia si todos los valores de Y se incrementan junto con los valores de X, entonces S= n(n-1)/2 de tal manera que T_{XY} = +1, por el contrario si todos los valores Y decrecen con el incremento de los valores de X, entonces S= -n(n-1)/2 y por lo tanto T_{XY} = -1. También Siegel y Castellan (1998), recomiendan que si se estudia la correlación en muestras con tamaños **n** mayor de 10 datos, se puede efectuar una aproximación de T_{XY} a una distribución normal, para las pruebas de significación.

Existen estudios donde dos o más casos estudiados dentro de una muestra reciben la misma puntuación en la misma variable, cuando esto sucede se dice que las puntuaciones están "empatadas", y por lo tanto a cada una de ellas se le debe asignar el promedio de los rangos que se le hubiese sido asignado si no estuvieran "empatados". Si ocurre que el número de valores empatados no es grande, su efecto en el sesgo generado sobre el estadístico T_{XY} no es tan importante: En la ecuación (2) se muestra su forma de cálculo (**ver Tabla 1**). Una de las ventajas de este coeficiente es que puede ser generalizado a un coeficiente de correlación parcial, también es apropiado para la evaluación de acuerdos entre múltiples jueces en pruebas de catación o juicios de expertos. La discusión detallada de este indicador será tema de otro trabajo.

Coeficiente de Correlación por Rangos de Spearman (r_s).

El coeficiente de correlación de rangos de Spearman (1904) es de gran utilidad en aquellos análisis de datos en donde se desea conocer la relación lineal entre variables cuyas escalas de medidas sean al menos ordinales, o que exista suficientes evidencias de que las variables en estudio a pesar de ser cuantitativas no siguen un comportamiento normal. Según Pardo *et al.* (2002, p. 345), el coeficiente de correlación *rho* de Spearman es el mismo coeficiente de correlación de Pearson pero aplicado luego de efectuar una transformación de los valores originales de las variables en rangos, tomando valores entre -1 y +1 y se interpreta de la misma manera que el de Pearson. Este coeficiente se denota comúnmente como r_s.

Pérez C. (2002, pag. 123), señala que cuando se determina a partir de variables cuantitativas, el grado de asociación lineal obtenido no es el de los valores de las variables, sino el de las clasificaciones por rangos de dichos valores.

Este indicador también es llamado coeficiente de correlación por rangos, entendiéndose por rango de un valor de una variable, el lugar que ocupa dicho valor en el conjunto total de valores de la variable, suponiendo de antemano que los valores se tienen ordenados de menor a mayor o viceversa. Este coeficiente se basa en la concordancia o discordancia de las clasificaciones por rangos de las modalidades de las variables cualitativas estudiadas.

Algebraicamente, el coeficiente de correlación de Spearman se obtiene como: Sean X e Y dos variables aleatorias cualitativas ordinales tomadas de una muestra de tamaño n, con categorías A_i y B_i, y sean x_i e y_i los rangos que les corresponden a A_i y B_i, entonces para Siegel y Castellan (1998, pag. 273) es:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad \textbf{(3)}$$

Donde: $d_i = x_i - y_i$ (i = 1,2,…n; n =Nro de parejas de rangos). Será $r_S = +1$ cuando la concordancia entre los rangos es perfecta por lo tanto $d_i = x_i - y_i = 0$, mientras que será $r_S = -1$ cuando la discordancia es perfecta y cuando no existe concordancia y discordancia entonces $r_S = 0$.

Como en el caso de visto para Kendall, puede ocurrir que dos o más casos (sujetos, objetos, etc.) presenten el mismo valor de rango en la misma variable, estando estos valores "empatados", en esta situación se debe calcular el promedio de los rangos que le correspondería a esos casos de no estar "empatados", y asignar este promedio a dichos casos, bajo esta situación se debe determinar el coeficiente de correlación de Spearman como se presenta en la **tabla 2**.

Tabla 2. Fórmula de cálculo del coeficiente de correlación de Spearman.

Coeficiente de Spearman	Donde:
$$r_s = \frac{(n^3 - n) - 6 \cdot \sum_{i=1}^{n} d_i^2 - (T_X + T_Y)/2}{\sqrt{(n^3 - n)^2 - (T_X + T_Y)(n^3 - n) + (T_X \cdot T_Y)}}$$ **(4)**	$$T_X = \sum_{i=1}^{g} (t_i^3 - t_i)$$ $$T_Y = \sum_{i=1}^{g} (t_i^3 - t_i)$$ T_X y T_Y : Factores de corrección para las Variables X e Y. g = Nro. de grupos de diferentes rangos. t = Nro. de rangos empatados en el i-ésimo grupo. X e Y : Variables Estudiadas.

(4)

Siegel y Castellan (1998, pag. 280), recomiendan que la corrección por "empates" en el coeficiente de Spearman es necesaria cuando se presenten un número importante de empates en los casos estudiados, así como también exponen que cuando el tamaño de la muestra es considerada grande (n>20), el estadístico r_s puede ser aproximado a una distribución normal. Tanto el coeficiente de correlación de Kendall como el de Spearman son denominados medidas de correlación no paramétrica.

CONCLUSIÓN

Es importante destacar que la correlación mide la intensidad de una relación lineal entre dos variables, que pudieran no tener una relación causal entre si, y sin embargo estar relacionadas. Los coeficientes de correlación que destacan en la aplicación de un análisis de datos corresponden al de Kendall y Spearman cuando las variables objeto de estudio son cualitativas con escalas de medidas de naturaleza ordinal. El coeficiente de correlación de Spearman es

pertinente si se presenta uno de los siguientes casos: el primero, supongamos que se estudia la asociación lineal entre variables cuantitativas con escalas de medidas al menos de intervalos, y bajo esta condición sería conveniente el uso del coeficiente de Pearson, pero si estas variables no siguen un comportamiento normal en sus datos, necesariamente se debe estimar Spearman; el segundo, cuando ambas variables originales presentan escalas de medidas ordinales y su determinación es directa. Por último, el coeficiente de Kendall es adecuado cuando ambas variables presentan escalas de medidas ordinales.

REFERENCIAS BILIOGRÁFICAS

Kenny D. (1979). *Correlation and Causality*. Recuperado de http://davidakenny.net/books.htm.
Download free: cc_v1.pdf.

Pagano R. (2009). *Estadística para las Ciencias del Comportamiento*. Séptima Edición, Editorial
CENGAGE Learning, México.

Pardo A., Ruiz, H. (2002) SPSS 13. *Guía para el Análisis de Datos*. Segunda Edición. McGraw
Hill, España.

Pérez C. (2002). *Estadística Aplicada a Través de Excel*. Segunda Edición, Editorial Prentice
Hall. España.

Montgomery D. (2004). Diseño y Análisis de Experimentos. Segunda Edición, Editorial Limusa
S.A., Mexico.

Shong N. (2010). Pearson's versus Spearman's and Kendall's Correlation Coefficients for Continuous Data. Tesis de grado de Maestria en Ciencias. Escuela Graduada de Salud Pública, Universidad de Pittsburgh, E.U.A.

Siegel S. y Castellan J. (1998). Estadística No Paramétrica: Aplicadas a las Ciencias de la Conducta.
Cuarta edición, editorial Trillas, México.

Sheskin D. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures.* Cuarta
Edición. Boca Raton, Florida: Chapman & Hall/CRC.