UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Febrero 2021, Vol. 14, No. 1, pp. 53–63

¿Pearson y Spearman, coeficientes intercambiables?

Pearson and Spearman, interchangeable coefficients?

Jorge Ortiz Pinilla[a]
jorgeortiz@usantotomas.edu.co

Andrés Felipe Ortiz Rico[b]
andresortiz@usantotomas.edu.co

Resumen

Se propone una discusión sobre la muy conocida forma de presentar los métodos no paramétricos como "alternativa" del estudio de parámetros cuando no se cumplen ciertos supuestos. Las consecuencias pueden ser el origen de muchas decepciones cuando ingenuamente se admite que un método resuelve el mismo problema que otro y al final ni siquiera se pregunta cuál se resolvió. Esta discusión se centra en los coeficientes de correlación de Pearson y de Spearman, pero bien puede llevarse a otras herramientas de análisis de datos. Adicionalmente, se incluyen discusiones sobre aspectos relacionados con la linealidad, la monotonía y el tamaño de muestra en relación con el uso y la interpretación de éstos coeficientes de correlación

***Palabras clave*:** Coeficiente de correlación de Pearson, Coeficiente de correlación de Spearman, No paramétrico, Asociación, Correlación lineal, Regresión lineal, Linealidad, Monotonía..

Abstract

We propose a discussion on the well-known way of presenting non-parametric methods as an "alternative" for the study of parameters when certain assumptions are not met. The consequences can be the source of many disappointments when we naively admit that one method solves the same problem as another and we end up without even wondering which one we have solved. Our discussion focuses on the Pearson and Spearman correlation coefficients, but it may well be carried over to other data analysis tools.

***Keywords*:** Pearson's correlation coefficient, Spearman's correlation coefficient, Nonparametric, Association, Linear correlation, Linear regression, Linearity, Monotoniciy. .

1. Introducción

Los coeficientes de correlación son herramientas extensamente utilizadas para analizar diferentes tipos de asociación entre variables en individuos pertenecientes a poblaciones bajo estudio. Se encuentran en prácticamente todas las áreas de conocimiento empírico, incluidas las ciencias sociales, agrarias, económicas, de la salud, ingenierías, física y muchas otras. Se destacan entre los más utilizados, el de Pearson (1895), definido para datos numéricos, y el de Spearman (1904), para datos ordinales. Es fácil encontrarlos con algunas pautas básicas de interpretación y uso en textos como Montgomery & Runger (2018), Navidi (2019) y Ortiz-Pinilla (2013), pero su estudio más teórico y profundo se encuentra en obras más especializadas como Kendall & Gibbons (1990). Rodgers & Nicewander (1988) describen una diversidad interesante de perspectivas de conceptualización e interpretación. Sin embargo, a pesar de tanta

[a]Docente de Maestría en Estadística Aplicada
[b]Docente de Maestría en Estadística Aplicada

información, es común encontrar publicaciones científicas en donde el uso y la interpretación de estos coeficientes son inadecuados. El trabajo de Porter (1999) ilustra esta problemática en artículos de tres revistas médicas. Se suma la gran cantidad de documentos en Internet con información imprecisa o errónea que induce fácilmente a decisiones equivocadas, por lo que se hace necesaria la lectura crítica y la selección de fuentes de confianza. Por otra parte, en la literatura se encuentran publicaciones que recomiendan el uso del coeficiente de Spearman como "alternativa no paramétrica" del de Pearson, que debe ser utilizada cuando los datos no provienen de distribuciones normales, cuando el tamaño de la muestra es inferior a 30, cuando las relaciones entre las variables no son lineales, o cuando se detectan datos atípicos. Con frecuencia la aplicación de estas recomendaciones no resuelve los problemas que los investigadores afrontan cuando deben decidir cuál escoger.

El propósito del artículo es analizar las recomendaciones anteriores para ayudar a los investigadores en su decisión de escoger uno u otro coeficiente. Se tratan los conceptos de monotonía, linealidad y tamaño de muestra, y la forma como se relacionan con la interpretación de los coeficientes. Se presentan y analizan ejemplos para ilustrar y dar soporte a las conclusiones. No se incluye el estudio de propiedades estadísticas que se encuentran fácilmente en otras publicaciones como las citadas en los párrafos iniciales.

La primera sección la dedicamos al concepto de asociación entre variables numéricas, mostrando la diferencia principal entre los estudios de correlación y los de regresión; en la segunda, presentamos los dos coeficientes y las pruebas de hipótesis más comunes; en la tercera, analizamos el concepto de monotonía y su relación con los coeficientes, y las implicaciones de los enfoques paramétrico y no paramétrico sobre la interpretabilidad de los resultados. En la sección 4 estudiamos algunas recomendaciones muy populares acerca del uso de uno de los coeficientes como sustituto del otro y en las conclusiones, hacemos un balance de los análisis anteriores.

2. Asociación entre variables numéricas

El concepto de asociación entre dos variables numéricas se ha desarrollado alrededor del comportamiento que se observa en las parejas de valores. El interés se centra en averiguar si tienden a agruparse según su magnitud. Si los valores grandes de las dos variables se encuentran juntos y lo mismo ocurre con los pequeños, la relación se define como *directa*. Si, por el contrario, los mayores valores de una de ellas tienden a presentarse con los más pequeños de la otra, se define como *inversa*. Las dos situaciones corresponden parcialmente a condiciones de monotonía y no se relacionan forzosamente con un modelo funcional específico entre las variables. En la sección 3.1.1 destacamos que sólo en condiciones muy extremas del coeficiente de correlación de *Pearson* se tiene una recta como modelo que describe de manera perfecta la relación. En los demás casos no se tiene esa garantía. En la actualidad se dispone de estudios e indicadores para relaciones más complejas que las de monotonía, pero no los consideramos en este artículo.

Cuando se utiliza una de las dos variables como base para pronosticar los valores de la otra mediante una recta, se construyen *modelos* de la forma $\hat{y} = B_0 + B_1 x$, en donde, a conveniencia del investigador, se denota x como la variable de base y y, la que se pronostica. Entonces el investigador puede *fijar* los valores de x y observar los que ocurren para y. De esta manera, las parejas tienen un componente (x) controlado y el otro (y) no. La pregunta central es cómo se comportan (cómo se distribuyen) los valores de y correspondientes a cada valor específico de x o, en otras palabras, cuál es la distribución condicional de y para cada x fijo, partiendo de suponer que una recta es adecuada como modelo. Se considera entonces que, *desde el punto de vista de los pronósticos*, la variable y es *dependiente* de la variable x que se llama *independiente*. Los desarrollos y las condiciones para obtener las respuestas hacen parte de los estudios de *regresión lineal simple*. Remitimos al lector a bibliografía sobre este tema como Draper & Smith (1998) para ver el papel que cumplen allí los coeficientes de correlación.

A diferencia de los estudios de regresión, en los de correlación no se ejerce control sobre ninguna de las variables y las dos cumplen papeles simétricos. Se trata de estudiar la tendencia en el comportamiento que tienen las dos variables como pareja, dejando la función de pronóstico en un segundo plano, como

complemento importante, pero no como objetivo principal. Se atribuye gran importancia a las siguientes preguntas:

1. ¿Qué tan fuertemente se relacionan, es decir, qué tan evidente es que el crecimiento de una variable sea concomitante con el de la otra? (monotonía de la relación)
2. ¿Qué tan evidente es que las dos variables cambien en el mismo sentido, o en sentidos contrarios? (sentido de la relación)
3. ¿Qué tan evidente es que exista una proporcionalidad en los cambios de las dos variables? (linealidad)

3. Los coeficientes

Con cualquiera de los enfoques de análisis de la sección 2 las evidencias mencionadas se hacen fuertes en la medida en que los puntos muestren una tendencia identificable *grosso modo* por alguna función monótona. Cuando así se percibe, se puede asumir que existe una función $y = f(x)$ y que los datos se agrupan alrededor de la curva definida por f. Además, se espera que para cada x, el *promedio* de los valores de y correspondientes coincida con el valor de la ordenada en el punto $(x, f(x))$. Así, la curva describe los valores esperados de Y para cada x. Como caso de interés especial, se asume que la relación entre las variables se modela por una recta, respondiendo también a la pregunta 3. A nivel individual, pueden encontrarse puntos cerca o lejos del modelo, pero lo ideal es que prevalezcan los cercanos. Los lejanos dan indicios de que la relación no es suficientemente fuerte.

3.1. El coeficiente de correlación de Pearson

Este coeficiente se calcula con los valores observados de parejas de datos numéricos

$$(x_1, y_1), (x_2, y_2), ..., (x_i, y_i), \ldots, (x_n, y_n)$$

de dos variables cuantitativas X, Y evaluadas en un conjunto de n individuos. Se define como la covarianza muestral entre los componentes tipificados de las parejas de datos:

$$r_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s(x)} \right) \left(\frac{y_i - \bar{y}}{s(y)} \right) \tag{1}$$

en donde $s()$ es la desviación estándar muestral.

3.1.1. El coeficiente de correlación de Pearson y la linealidad

En (1) resaltamos la dependencia de r_{XY} de las diferencias $x_i - \bar{x}$ y $y_i - \bar{y}$ entre los valores numéricos de las variables. Mediante la desigualdad de *Cauchy-Schwarz* se demuestra que el máximo valor, $|r_{XY}| = 1$, se obtiene cuando todos los puntos se encuentran en una línea recta. Cualquier alejamiento de esta condición impide alcanzar el máximo. Por esta razón, se precisa el nombre como el coeficiente de correlación *lineal* de Pearson.

3.1.2. Inferencia sobre el coeficiente de correlación de Pearson

Cuando (X, Y) sigue una distribución normal bivariada, es posible diseñar una prueba estadística para contrastar la hipótesis nula de que el coeficiente poblacional es igual a cero. Se utiliza la expresión:

$$T = \frac{r_{XY}}{\sqrt{\frac{1 - r_{XY}^2}{n-2}}} \tag{2}$$

Bajo la hipótesis nula que asigna el valor de cero al parámetro de correlación en la distribución normal bivariada, T sigue una distribución t de Student con $n-2$ grados de libertad, denotada como t_{n-2}.

3.2. El coeficiente de correlación de Spearman

Por otra parte, el coeficiente de correlación de *Spearman* se basa en los rangos $r(x_i)$, $r(y_i)$ de los valores x_i, y_i al ordenar cada muestra por separado.

Una asociación *directa* perfecta entre X y Y se pone en evidencia cuando los datos de las muestras aparecen ordenados exactamente de la misma forma. En este caso, las diferencias $d_i = r(x_i) - r(y_i) = 0$ para $i = 1, 2, 3, \ldots, n$, y $d = \sum d_i^2 = 0$. En la medida en que d aumente, se debil ita esta evidencia.

En el otro extremo, una asociación *inversa* perfecta se presenta cuando el ordenamiento de menor a mayor en una de las variables corresponde exactamente al ordenamiento de mayor a menor en la otra. Entonces, si un dato x está en la posición i, el dato y correspondiente estará en la posición $n-i+1$; por lo tanto, $d_i = i - (n-i+1) = 2i - (n+1)$ y $d = \sum d_i^2 = n(n^2-1)/3$.

La ecuación de la recta que pasa por los puntos $(d = 0,\ \rho = 1)$ (máxima asociación directa) y

$$\left(d = n(n^2-1)/3,\ \rho = -1\right)$$

(asociación inversa extrema) arroja la expresión del *coeficiente de correlación de Spearman*:

$$\rho = 1 - \frac{6\sum d_i^2}{n\,(n^2-1)} \tag{3}$$

Cuando no hay valores iguales, ni entre las x, ni entre las y, como d es un entero entre 0 y $n(n^2-1)/3$, el número máximo de valores diferentes que puede tomar ρ es $1 + n(n^2-1)/3$ y, además, coincide con el coeficiente de correlación de Pearson aplicado a los rangos de los datos.

El punto medio del segmento que une los extremos anteriores corresponde a $d = n(n^2-1)/6$. Allí $\rho = 0$, indicando que las evidencias no favorecen ni una relación directa ni una inversa entre X y Y.

3.2.1. Inferencia sobre el coeficiente de correlación de Spearman

En ausencia de relaciones directas o inversas (hipótesis nula) entre X y Y, se puede asumir que, en muestras aleatorias de pares $(X_i, Y_i), i = 1, 2, \ldots, n$, tanto los rangos de las X_i como los de las Y_i tienen distribuciones uniformes discretas entre 1 y n, y son independientes. Entonces, los $n!$ emparejamientos diferentes posibles de los rangos de X y de Y tienen la misma probabilidad. A partir de esta propiedad se obtiene la función de distribución del coeficiente de correlación de Spearman. El rápido crecimiento de $n!$ obliga a buscar opciones alternativas. El cálculo de ρ como un coeficiente de correlación de Pearson para los rangos lleva a considerar la misma expresión (2) como aproximación. La figura 1 ilustra su comportamiento para $n \leq 22$.

4. Linealidad, monotonía y parámetros

En la sección anterior hemos presentado los coeficientes de correlación y una forma de calcular, en términos de proporciones mediante las funciones de distribución de las estadísticas de prueba, qué tan cerca están de los extremos y, por lo tanto, qué tan alejados se encuentran de cero, que es el valor central indicador de ausencia de una tendencia de asociación dominante positiva o negativa. De esta manera, más que desarrollar el tema inferencial, se trata de evaluar la magnitud de los coeficientes considerando

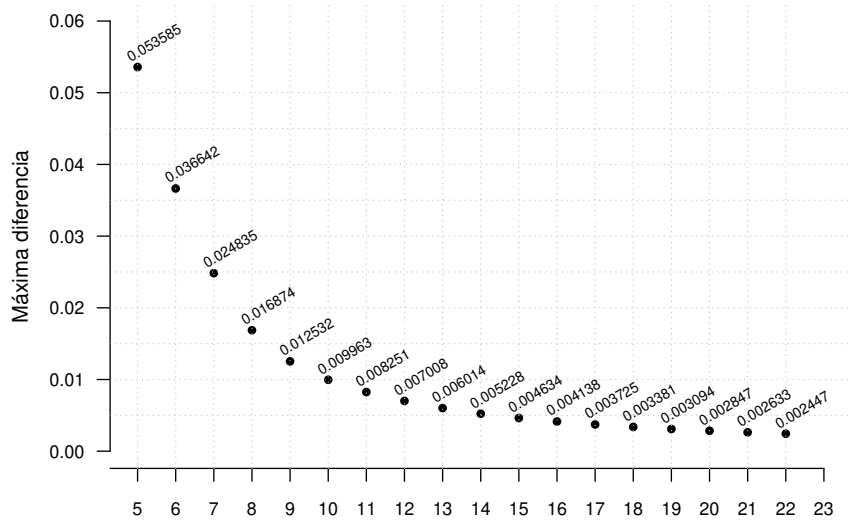

Figura 1: Máxima diferencia entre la función de distribución exacta de Spearman y la de Student con $n-2$ grados de libertad, en función del tamaño de muestra.

la cantidad de datos observados. Orientamos el artículo a su análisis como indicadores del grado de *asociación monótona, entendida como el comportamiento sostenido de crecimiento o decrecimiento de una de las variables en función de la otra.* Estudiaremos en detalle qué hacen bien y qué debilidades pueden presentar.

4.1. La monotonía y los coeficientes

Cuando los datos garantizan información de carácter ordinal, el concepto de asociación hace referencia al tipo de crecimiento de los datos de la variable Y con respecto a los de X ordenados. No está limitado a relaciones lineales. Cualquier función monótona creciente afecta las distancias entre los puntos correspondientes a los valores de Y, pero sin alterar su orden, por lo tanto, no afecta los rangos de Y, sin importar la variable afectada por la transformación. La función puede acelerar el crecimiento de Y o disminuirlo, es decir, cambiar la concavidad, sin perder su monotonía. Esto implica que cuando el coeficiente de correlación de Spearman vale 1, es posible ajustar una función monótona creciente, y cuando vale -1, es posible ajustar una función monótona decreciente, pero en ningún caso podemos especificar su curvatura. Recordemos que cuando el coeficiente de correlación de Pearson vale 1 o -1 el único tipo de función que se puede ajustar perfectamente a los datos es una recta. Como la recta representa a una función monótona, entonces $\rho_{XY} = r_{XY} = \pm 1$.

En la gráfica 2 observamos cuatro funciones estrictamente monótonas crecientes, por lo que, para datos seleccionados de una cualquiera de ellas, el coeficiente de Spearman vale exactamente 1, mientras que el de Pearson vale 1 sólo para la función 3 que es precisamente una recta. Para los demás casos, los valores podrían ser engañosos en cuanto al carácter lineal de las relaciones: $r_1 = 0.746$ indicaría una relación intermedia, mientras que $r_2 = 0.881$ es considerado por muchos como indicador de una relación fuerte. De manera especial, $r_4 = 0.974$ corresponde a una correlación extremadamente alta y si se considera ingenuamente que, como es un coeficiente de correlación lineal (de Pearson), entonces una recta se ajustaría casi perfectamente a los datos, lo cual es falso (ver la función 4 en la gráfica 2).

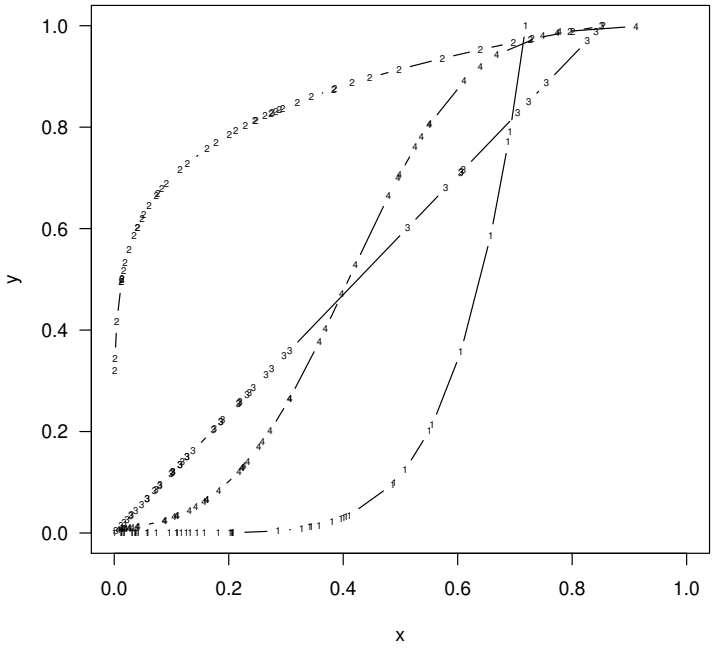

Figura 2: Funciones monótonas estrictamente crecientes, con coeficientes de correlación por rangos de Spearman $\rho = 1$, y con valores diferentes del coeficiente de correlación lineal de Pearson: $r_1 = 0.746,\ r_2 = 0.881,\ r_3 = 1,\ r_4 = 0.974$.

4.2. Paramétrico y no paramétrico

Hasta ahora hemos considerado los coeficientes r_{XY} y ρ_{XY} como indicadores del grado de asociación en conjuntos finitos de datos. A nivel de poblaciones infinitas, r_{XY} tiene un parámetro asociado, definido en términos de los momentos de primero y segundo orden cuando existen. En el caso de la distribución normal bivariada, este parámetro está incluido en la función de densidad que la define.

Para el coeficiente de correlación de Spearman, Kendall & Gibbons (1990, pag. 124) considera triplas de valores $(x_i, y_i), (x_j, y_j), (x_k, y_k)$ y define $\pi_2 = Pr(y_i < y_k \mid x_i < x_j)$ como la probabilidad de *concordancias de tipo 2*, y p_2 como la cantidad muestral de concordancias de tipo 2 dividida por el número total de posibilidades. Para n grande, toma $\rho_s = 6\Big(\pi_2 - \dfrac{1}{2}\Big)$ como la definición del coeficiente de correlación de Spearman para poblaciones continuas. Sin embargo, Kendall mismo muestra que el coeficiente muestral es un estimador sesgado del poblacional. No incluimos detalles inferenciales adicionales sobre él.

Cuando se trabaja con variables numéricas, el concepto de *relación monótona* es parametrizable sólo si se establece una función numérica específica cuyas constantes proporcionen los parámetros. Como comentamos en 4.1, esto es imposible con la sola información de los rangos. En este sentido, los análisis basados en el coeficiente de Pearson se conocen como *paramétricos*, y los basados en el de Spearman, como *no paramétricos*.

La consecuencia más importante consiste en que en una relación lineal la proporcionalidad de las diferencias permite identificar *tasas de cambio* que se calculan con los métodos de regresión –muy atractivas pero difícilmente reales, especialmente en las áreas sociales– mientras que en una relación monótona donde no se puede identificar la forma funcional, lo único que se puede establecer es una tendencia directa o indirecta de los componentes de las parejas. Obviamente esto ha generado un favoritismo por el coeficiente

de Pearson, pues los investigadores se sienten más satisfechos con algo concreto, un parámetro, con la ilusión de una interpretación, así sea falsa, que con un coeficiente más retador para describir, pero más claro en la información que ofrece. Basta con examinar cuidadosamente lo que se puede interpretar de un modelo de regresión para variables obtenidas con escalas tipo Likert en las que ni se tiene el significado de una unidad ni tiene cabida el concepto de proporcionalidad.

5. Interpretaciones y criterios de uso

El acceso masivo a los programas estadísticos facilita que se utilicen los coeficientes indistintamente, sin considerar lo que mide cada uno. Incluso resulta atractivo escoger uno de ellos con el criterio del valor más alto. Por lo general, esta práctica se toma como un indicio de que el investigador no tiene claridad sobre lo que está analizando. En esta sección abordaremos algunos aspectos del tema, con el fin de organizar las reflexiones que consideramos más importantes para el uso adecuado de los coeficientes.

5.1. Linealidad y monotonía

Es de conocimiento general que el coeficiente de Pearson indica qué tan ajustable es un *modelo lineal* a los datos. Sus valores extremos (± 1) están íntimamente asociados con los mejores de todos ellos que, según el criterio de mínimos cuadrados, son rectas. Por el contrario, el de Spearman no se asocia con un modelo numérico específico que relacione X y Y, y por lo tanto, no es posible establecer criterios de optimalidad de ajuste que permitan identificar el mejor. Su presentación como un coeficiente de Pearson para los rangos de las variables permite interpretarlo como un indicador de lo ajustable que pueda ser un línea recta *al conjunto de puntos cuyas coordenadas son los rangos*. Las situaciones óptimas se presentan cuando las parejas de datos se ordenan igual según la primera o la segunda componente (relación monótona creciente estricta), y cuando estos ordenamientos son contrarios (relación monótona decreciente estricta). Ya vimos en la sección 4.1 que la monotonía estricta no garantiza que se pueda precisar el tipo de función que relaciona las variables. De ahí que la información más directa que suministra el coeficiente de Spearman es acerca del grado de afinidad de los rangos, es decir de los ordenamientos de las dos variables.

Una relación lineal perfecta es también monótona perfecta. Por lo tanto, si el coeficiente de correlación de Pearson vale ± 1, entonces $\rho_{XY} = r_{XY} = \pm 1$. Sin embargo, una relación monótona perfecta ($\rho_{XY} = \pm 1$) no es necesariamente lineal y es posible que r_{XY} difiera considerablemente de ρ_{XY} (Sección 4.1).

5.1.1. El coeficiente de Spearman como reemplazo del de Pearson

Los comentarios en la sección 4 nos advierten sobre los riesgos de recomendar de manera sistemática el coeficiente de Spearman como "la alternativa no paramétrica" del de Pearson cuando los datos provienen de distribuciones no normales. Para que este uso sea válido, el investigador debe tener alguna garantía de la existencia de variables numéricas subyacentes a los rangos que puedan modelarse con una relación lineal. No tenerla implica el riesgo de obtener resultados o interpretaciones cuestionables. El coeficiente de Spearman no es entonces un reemplazo no paramétrico del de Pearson, sino un indicador de tendencia monótona de la relación, y corresponde a un enfoque diferente del análisis, que incluso permite considerarlos complementarios en algunos casos.

Con la figura 3 introducimos tres ejemplos para ilustrar algunos usos populares pero injustificados de los coeficientes.

5.1.2. Pearson si $n > 30$ y Spearman si $n < 30$

Ejemplo 5.1. En el conjunto de 80 puntos señalados con pequeñas circunferencias rojas, la tendencia de los primeros de la izquierda tiene una inclinación muy fuerte, mientras que la de los siguientes es muy plana.

Es un agregado de dos relaciones lineales, pero no lineal en su conjunto, aunque sí monótona creciente. Los valores $r_{XY} = 0.633$ para Pearson y $\rho_{XY} = 1$ para Spearman cumplen con indicar respectivamente la imperfección de la linealidad y la evidencia de la monotonía. Si pretendiéramos utilizar r de Pearson para evaluar la monotonía, tendríamos una herramienta inadecuada y un resultado confuso al respecto.

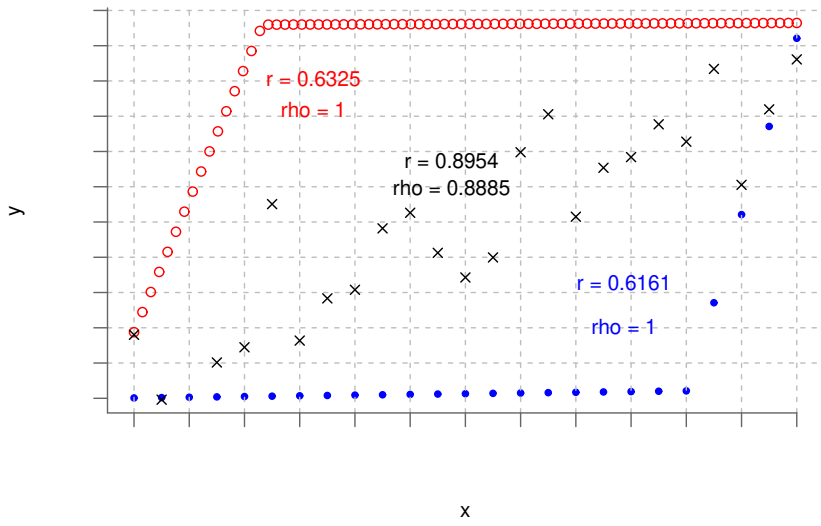

Figura 3: Pearson o Spearman: relación lineal o monótona

Ejemplo 5.2. Se trata de los 25 puntos sólidos azules en la parte inferior de la figura 3. Como en el ejemplo anterior, se observa un agregado de dos relaciones lineales con tendencias muy diferentes que hacen del conjunto una relación monótona creciente, pero no lineal. El valor $\rho = 1$ arroja la evidencia de esta propiedad, mientras que $r = 0.616$ indica que un ajuste lineal sería insuficiente para estos datos. Como antes, el uso de r como indicador de monotonía es inapropiado, pues 0.616 es demasiado bajo para valorar una relación estrictamente creciente.

Ejemplo 5.3. En la misma figura 3 agregamos 25 puntos identificados con $\times$, en una relación lineal dada por $y = 8x + e;\ x = 1, 2, \ldots, 25$ y $e \sim N(0, 25^2)$. Ambos coeficientes de correlación son cercanos a 0.89 y respaldan la idea de una *tendencia* lineal, monótona creciente que, sin ser estrictas, son ampliamente dominantes. Ya hemos visto que estos dos coeficientes no siempre dan valores cercanos, pero las condiciones de linealidad favorecen esta proximidad.

Los tres ejemplos nos permiten destacar las siguientes conclusiones:

1. Si el investigador está interesado en evaluar el grado de asociación lineal entre dos variables, la utilización del coeficiente de Spearman como "alternativa no paramétrica" del de Pearson es inadecuada con mucha frecuencia, y, más aún si se deriva del criterio de escoger el de mayor valor (ver ejemplos 5.1 y 5.2).

2. Cuando hay garantía de que la relación entre las variables sea lineal, considerando que ésta es monótona, se favorecen las condiciones para que los coeficientes de Pearson y de Spearman tomen valores similares. Esta es la única situación en donde se tiene la posibilidad de utilizar el coeficiente de Spearman como sustituto del de Pearson. El motivo para hacerlo puede ser la dificultad para obtener datos numéricos suficientemente confiables.

   La situación recíproca no es válida, como lo explicaremos más adelante.

3. Los mismos ejemplos 5.1 y 5.2 nos muestran que el coeficiente de correlación de Pearson puede cumplir un papel muy pobre como indicador de monotonía, es decir, como sustituto del coeficiente de Spearman.

4. El tamaño de muestra no sirve como criterio para escoger un coeficiente u otro. En el ejemplo 5.1, el tamaño de muestra es de 80, y en el ejemplo 5.2 es de 25. En los dos casos, el coeficiente de Pearson, inferior a 0.633, indica que una recta sería insuficiente para describir la relación entre las dos variables. El de Spearman, igual a 1, identifica relaciones estrictamente monótonas crecientes. El investigador debe estar atento al tipo de relación que quiere describir. Un mayor tamaño de muestra sirve para obtener información más completa y sacar conclusiones más estables, pero no elimina la necesidad de identificar el tipo relación que se estudia. Los coeficientes pueden ofrecer información complementaria: la relación es claramente monótona, pero no es lineal.

5. Puede ser conveniente, a partir de metadatos o de consideraciones teóricas, particionar los intervalos en subintervalos en donde se pueda garantizar el tipo de relación que se propone. En los ejemplos 5.1 y 5.2, es posible condicionar el estudio a los subintervalos en donde la relación sea lineal y el cálculo del coeficiente de Pearson tenga validez interpretativa.

5.1.3. Criterios de linealidad

A nivel conceptual, el tema más importante en una relación lineal es el de la proporcionalidad de los cambios. El investigador debe centrar su atención en este aspecto cuando trate de identificar este tipo de tendencia. Debe intentar explicar que cambios proporcionales en X corresponden a cambios igualmente proporcionales en Y, sin importar la ubicación de los intervalos donde haga las comparaciones. En la medida en que lo logre, tendrá la justificación de una relación lineal. Este análisis se hace con los cambios esperados en promedio y no a nivel individual, pues la linealidad que interesa es la del modelo.

5.1.4. La sola correlación alta no implica linealidad

En estudios exploratorios, es fácil pensar que, como una relación lineal fuerte se manifiesta con valores grandes de r, entonces que la recíproca es también válida: si encontramos valores grandes de r, entonces debe tratarse de una relación lineal fuerte. Veamos con un ejemplo que esto puede ser falso:

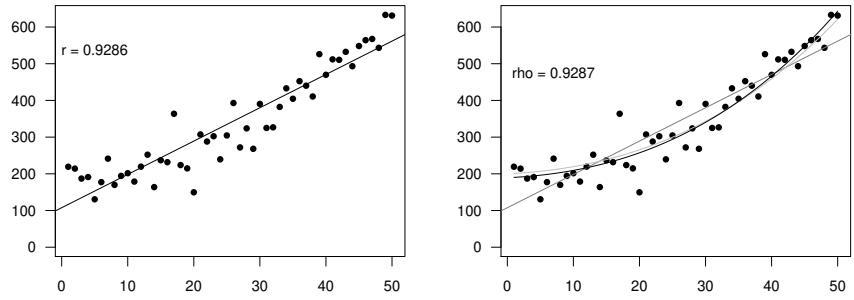

Figura 4: Un coeficiente alto de correlación lineal de Pearson no implica forzosamente que la relación sea lineal.

Ejemplo 5.4. El diagrama de dispersión de la izquierda en la figura 4 y el coeficiente de correlación lineal de Pearson con un valor de 0.9286 ofrecen cierta tranquilidad al investigador que quiera ver una recta como descriptor del comportamiento de las variables. En el diagrama de la derecha se presentan los mismos puntos, pero con la adición de una curva que resalta visualmente las deficiencias del modelo lineal: la recta subestima los extremos y sobrevalora el centro. En regresión, el análisis de residuos permite detectar esta anomalía del modelo lineal. Aquí sólo anotamos que los puntos fueron generados según un modelo de la forma $y - x^{x^{0.125}} = \varepsilon \sim N(200, 50^2)$. Aun con la ayuda de un diagrama de dispersión manejado ingenuamente, un investigador puede inclinarse por una linealidad engañosa.

6. Conclusiones y recomendaciones

Para conservar el contexto de las siguientes conclusiones, recordamos que no se abordó el aspecto inferencial y, por lo tanto, los comentarios no tienen nada que ver con este tema. El objetivo del estudio es presentar un análisis de algunos conceptos involucrados en la definición de los coeficientes para reforzar sus bases interpretativas y de aplicación.

1. La linealidad es un requisito conceptual para garantizar que el coeficiente de Pearson sea plenamente válido e interpretable, así como la monotonía lo es para el coeficiente de Spearman. La ausencia de argumentaciones para respaldar estos requisitos deja al investigador en condición frágil y limitada para interpretar a fondo sus resultados, especialmente si agrega estudios de regresión. La costumbre en el uso indiscriminado de herramientas correlacionales hace difícil aceptar este comentario, pero una reflexión objetiva debe llevar a adoptar estas técnicas con mayor cuidado.

2. Las teorías y los conocimientos disciplinares son la principal guía para proponer modelos que describan los patrones de comportamiento, y deben dar las pautas para interpretar los parámetros estudiados. Las mismas ayudas gráficas, fuertemente recomendadas, pueden resultar engañosas si se manejan de manera ingenua.

3. Es importante destacar que el principal objetivo de cualquier análisis estadístico de datos es facilitar la comprensión del comportamiento de un fenómeno en forma observacional (sin intervención humana) o experimental. Sin embargo, en muchas ocasiones, para el investigador es de suficiente importancia identificar o construir patrones que le sirvan de referencia para estudiar casos específicos o pronosticar resultados en algunas variables en función de otras aun con conocimiento limitado desde el punto de vista disciplinar.

4. El análisis basado en el cálculo de un coeficiente de correlación es solo un comienzo en un estudio correlacional, este análisis se complementa con modelos de regresión en donde una de las variables se considera como explicativa y la otra como explicada. El estudio de los residuos permite identificar posibles anomalías por tendencias, por debilidad de la relación, por la presencia de valores extremos y eventualmente de valores influyentes que puedan distorsionar la forma como se ven las variables relacionadas.

5. Cuando no se tiene una base teórica que oriente la búsqueda de modelos específicos, conviene tener presente que a mayor complejidad, la estabilidad de las estimaciones se debilita. Puede ser mas valioso proponer un modelo simple como una aproximación local inicial susceptible de perfeccionarse, que una expresión matemática mejor ajustada pero cuya complejidad se convierta en un obstáculo para entender y describir la relación.

Referencias

Draper, N. R. & Smith, H. (1998), *Applied Regression Analysis*, Wiley.

Kendall, M. & Gibbons, J. (1990), *Rank Correlation Methods*, A Charles Griffin title, Edward Arnold.

Montgomery, D. C. & Runger, G. C. (2018), *Applied Statistics and Probability for Engineers*, Wiley.

Navidi, W. (2019), *Statistics for Engineers and Scientists*, McGraw Hill.

Ortiz-Pinilla, J. (2013), *Principios de estadística aplicada*, Ediciones de la U.

Pearson, K. (1895), 'Notes on regression and inheritance in the case of two parents', *Proceedings of the Royal Society of London* **58**, 240–242.

Porter, A. M. W. (1999), 'Misuse of correlation and regression in three medical journals', *The Journal of the Royal Society of Medicine* **92**(1), 123–128.

Rodgers, J. L. & Nicewander, W. A. (1988), 'Thirteen ways to look at the correlation coefficient', *The American Statistician* **42**(1), 59–66.

Spearman, C. (1904), 'The proof and measurement of association between two things', *The American Journal of Psychology* **15**(1), 72–101.