

# Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones

*On the proper use of the Pearson correlation coefficient: definitions, properties and assumptions*

Juan Diego Hernández Lalinde, Mg.<sup>1\*</sup>; <https://orcid.org/0000-0001-6768-1873>, Jhon Franklin Espinosa Castro, Mg.<sup>1</sup>; <https://orcid.org/0000-0003-2186-3000>, Mariana Elena Peñalosa Tarazona, Dr.<sup>1</sup>; <https://orcid.org/0000-0002-3863-0580>, Johel E Rodríguez, Mg.<sup>2</sup>; <https://orcid.org/0000-0002-8353-2736>, José Gerardo Chacón Rangel, Dr.<sup>2</sup>; <https://orcid.org/0000-0003-3615-578>, Cristian Andrés Toloza Sierra, Ing.<sup>3</sup>; <https://orcid.org/0000-0002-9908-0076>, Marly Karina Arenas Torrado, Mg.<sup>4</sup>; <https://orcid.org/0000-0001-9084-2460>, Sandra Milena Carrillo Sierra, Mg.<sup>5</sup>; <https://orcid.org/0000-0001-9848-2367>, Valmore José Bermúdez Pirela, Dr.<sup>6</sup>; <https://orcid.org/0000-0003-1880-8887>

<sup>1</sup>Universidad Simón Bolívar, Departamento de Ciencias Sociales y Humanas, Cúcuta, Colombia.

<sup>2</sup>Universidad Simón Bolívar, Facultad de Ingenierías, Cúcuta, Colombia.

<sup>3</sup>Universidad Simón Bolívar, Departamento de Sistemas, Cúcuta, Colombia.

<sup>4</sup>Universidad Simón Bolívar, Facultad de Administración y Negocios, Cúcuta, Colombia.

<sup>5</sup>Universidad Simón Bolívar, Facultad de Ciencias Jurídicas y Sociales, Cúcuta, Colombia.

<sup>6</sup>Universidad Simón Bolívar, Facultad de Ciencias de la Salud, Cúcuta, Colombia.

\*Autor de correspondencia: Juan Diego Hernández Lalinde. Universidad Simón Bolívar, Departamento de Ciencias Sociales y Humanas. Calle 14 entre avenidas 4 y 5, Barrio La Playa. C. P.: 540006. Cúcuta, Colombia. Correo electrónico: [hernandezjuandiego@gmail.com](mailto:hernandezjuandiego@gmail.com)

## Resumen

El coeficiente de correlación de *Pearson* es una medida considerablemente utilizada en diversas áreas del quehacer científico, desde estudios técnicos, econométricos o de ingeniería; hasta investigaciones relacionadas con las ciencias sociales, del comportamiento o de la salud. Es precisamente esta extensa y profusa divulgación una de las razones que explicaría el uso indebido que se le da a esta herramienta estadística, especialmente en aquellos escenarios en los que debe ser interpretada correctamente o en los que se tienen que comprobar las suposiciones matemáticas que la sustentan. Un ejemplo de esto se halla cuando se asume que la *correlación* implica *causalidad*, confusión en la que se incurre con frecuencia y en la que se ven involucrados, tanto investigadores noveles, como algunos más experimentados. Pero tal vez el foco de mayores errores se encuentre al momento de comprobar premisas como la de la *normalidad*, siendo que esta se verifica únicamente a nivel univariado y se omite su revisión bivariada, quizás por desconocimiento o porque exige la utilización de técnicas más complejas. Situación similar se observa cuando se intentan detectar *datos atípicos*. En este caso, lo común es que se empleen diagramas de caja y bigotes para identificar valores extremos en cada variable, cuando lo apropiado sería abordar esta tarea con procedimientos que cuantifiquen la distancia que separa a dicha observación del centro de gravedad de los datos, pero de manera simultánea y tomando en cuenta todos los componentes del espacio vectorial en que se encuentra. En tal sentido, se propone la presente revisión como aporte para esclarecer estas dudas y como guía metodológica para orientar en la verificación de tales supuestos, abordando el aspecto matemático de manera general, pero enfatizando en las alternativas de que dispone el investigador para acometer debidamente este tipo de análisis.

**Palabras Clave:** coeficiente, correlación, *Pearson*, supuestos, normalidad bivariada, datos atípicos multivariados.

## Abstract

The Pearson correlation coefficient is a measure widely used in several areas of scientific work, from technical, econometric or engineering studies; to social, behavioral or health sciences researches. It's precisely this extensive and profuse disclosure one of the reasons that would explain the misuse of this statistical tool, especially in those cases in which it must be correctly interpreted, or in those situations in which the mathematical assumptions that support it have to be checked. An example of this arise when it's assumed that correlation implies causation, confusion that occurs frequently and involves both, novice and experienced researches. But perhaps the mayor focus of errors is found when checking assumptions such as normality, since it's verified only at univariate level omitting its bivariate verification, possibly due to lack of knowledge or because it requires more complex techniques. Similar situation is observed when trying to detect outliers. In this case, it's common to use box and whisker plots to identify extreme values in each variable, when the appropriate would be to approach this task using procedures that calculate the distances that separates this observation of the center of the data, taking into account all its vector space components. In this regard, this review is proposed as a contribution to clarify these doubts and as a methodological guide to help in the verification of such assumptions, addressing the mathematical aspect in a general manner, but emphasizing the alternatives available to undertake this type of analysis.

**Keywords:** coefficient, correlation, *Pearson*, assumptions, bivariate normality, multivariate outliers.

Los términos *relación* o *asociación* son equivalentes y se usan para designar aquella área de la estadística en la que se evalúa la covariación entre al menos dos variables. Dentro de este grupo, la correlación lineal es un caso particular en el que tal correspondencia tiene características bien definidas y que suele medirse mediante el coeficiente  $R$  de *Pearson*. Originalmente desarrollado por *Karl Pearson* en 1895 —quien se basó en las investigaciones de *Sir Francis Galton* publicadas diez años antes<sup>1,2</sup>— es ampliamente utilizado en diversos campos del quehacer humano. En la ingeniería; por ejemplo, se ha empleado recientemente para evaluar el efecto de la deformación del viento en modelos de validación aerodinámica<sup>3</sup> o para estimar la rapidez con que este fluido pasa a través de turbinas generadoras de energía<sup>4</sup>. También ha tenido una aplicación importante en la industria petrolera, desde las fases de exploración, perforación y completación, hasta aspectos relacionados con la proyección de nuevos yacimientos en suelos marinos<sup>5</sup>.

Otra área del saber que se ha valido enormemente de esta herramienta estadística es la de la psicología. Es común que este coeficiente se use al momento de definir las propiedades psicométricas de escalas, inventarios o pruebas, tal y como refleja la validación del cuestionario de estigmatización percibida de pacientes que han sufrido quemaduras en Brasil<sup>6</sup>. Pero su alcance va más allá: investigaciones en las que se explora la dirección de la relación entre el agotamiento laboral y el estrés postraumático secundario<sup>7</sup> o en las que se indaga sobre la asociación entre la cantidad de estacionamientos públicos y el bienestar ciudadano<sup>8</sup>, reflejan la variedad de usos de este coeficiente.

La ciencia de datos, una disciplina relativamente nueva, ha basado gran parte de sus técnicas en la correlación de *Pearson*, sobre todo cuando estas son puestas a disposición de campos como la medicina. Tal cosa se aprecia en artículos en los que se proponen marcos computacionales para estimar el riesgo de obesidad<sup>9</sup>, métodos de segmentación no lineales para separar hematomas subdurales<sup>10</sup> o aislamiento de secciones epidurales mediante tomografía computarizada<sup>11</sup>. También la economía y finanzas se han servido de esta herramienta. Ejemplos de esto se hallan en publicaciones en las que se establecen las principales causas de la volatilidad de los negocios brasileños<sup>12</sup>, así como en investigaciones destinadas a medir el impacto y la interacción entre la inversión extranjera y el mercado bursátil de la India<sup>13</sup>.

La medicina, por supuesto, no escapa a esta realidad; de hecho, es tal vez una de las ramas que más se ha beneficiado de los análisis de correlación y regresión. Casos notables se encuentran al explorar tratamientos de intervención en pacientes que han experimentado un ataque cerebral, en los que se ha tanteado —hasta ahora sin mucho éxito— la asociación entre la proteína C reactiva y la disminución de la fatiga<sup>14</sup>. Un estudio interesante publicado por *Tarabichi* y *Detours*<sup>15</sup> discrepa sobre los hallazgos de *Tomasetti* y *Vogelstein*<sup>16</sup>, quienes proponen que el riesgo de padecer distintos tipos de cáncer está altamente correlacionado con el número

de divisiones de las células madre y que, por ende, mucha de su aparición se debe simplemente a mala suerte. Los primeros sostienen que el esquema de clasificación empleado por *Tomasetti* y *Vogelstein* falla al basarse principalmente en la incidencia de cáncer reportada en poblaciones norteamericanas. Otros tratados se han dedicado a establecer la correspondencia entre diferentes condiciones, como las vinculadas a enfermedades renales<sup>17</sup>, a los perfiles de masa grasa en el tejido adiposo<sup>18</sup> y a la función visual cuando se presenta macularidad precoz<sup>19</sup>.

Lo interesante es que, a pesar de que el uso del coeficiente de correlación de *Pearson* es extensísimo, también lo son las incorrecciones y omisiones al momento de emplearlo, las imprecisiones cuando de verificar sus supuestos se trata, o incluso, la confusión y desinformación a la hora de interpretar los resultados<sup>20</sup>. Es habitual encontrar en la literatura académica y científica ausencia de claridad al esbozar las nociones de regresión y correlación, por ejemplo; siendo que se hace más énfasis en las semejanzas que en las puntuales e importantes diferencias<sup>21</sup>. Este desconcierto aun alcanza connotaciones semánticas: el vocablo *correlación* suele aplicarse a cualquier tipo de asociación entre variables, cuando; en realidad, este se refiere solo a un caso particular de dicha relación<sup>22</sup>. Tal vez el mayor problema surge cuando se asume erróneamente que la correlación implica causalidad. Cuando dos variables  $X$  y  $Y$  exhiben una fuerte correlación, varias situaciones podrían explicarla: (a), que  $X$  origine  $Y$ ; (b), que  $Y$  origine  $X$ ; (c), que una variable externa  $Z$  origine tanto a  $X$  como a  $Y$ ; o (d), que la relación observada se deba meramente al azar<sup>22-24</sup>. Este hecho, si bien ha sido discutido ampliamente, sigue reportándose en un número no despreciable de textos, artículos, seminarios o ponencias<sup>22,23</sup>.

Esta revisión constituye un aporte para dar solución a esta problemática. La intención es la de ofrecer una fuente de consulta bibliográfica que promueva el uso apropiado de esta herramienta y sirva como guía a investigadores, profesores o estudiantes vinculados a la medicina y áreas afines.

### Covarianza y Correlación

Cada vez que se analicen al menos dos variables de forma simultánea, surgirá una pregunta natural encaminada a conocer el grado de coherencia que estas exhiben. En tales situaciones, el interés del investigador recaerá en determinar si los valores de una característica cambian consistentemente conforme a los de la otra, o si; por el contrario, no hay ningún patrón que las asocie. Estadísticamente este fenómeno se cuantifica a través de la covarianza<sup>20</sup>. En líneas generales, la covarianza entre dos variables aleatorias  $X$  y  $Y$  se define mediante<sup>25-27</sup>:

$$\text{Cov}(X, Y) \equiv \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)], \quad (1)$$

donde  $E$  es el operador de la esperanza matemática,  $\mu_X$  y  $\mu_Y$  son las medias poblacionales de  $X$  y  $Y$ , respectivamente; y las expresiones  $\text{Cov}(X, Y)$  y  $\sigma_{XY}$  son equivalentes. Nótese que si se reemplaza  $Y$  por  $X$  en la ecuación (1) se obtiene:

$$\text{Cov}(X, X) \equiv \sigma_{XX} = \sigma_X^2 = E[(X - \mu_X)(X - \mu_X)] = E(X - \mu_X)^2. \quad (2)$$

De esto se deduce que la covarianza de una variable con ella misma es, simplemente, la varianza. De manera particular, la fórmula (1) se transforma en las siguientes ecuaciones dependiendo de si  $X$  y  $Y$  son continuas o discretas; a saber<sup>28</sup>:

$$\text{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X - \mu_X)(Y - \mu_Y) f_{XY} dx dy, \quad (3)$$

$$\text{Cov}(X, Y) = \sum_x \sum_y (X - \mu_X)(Y - \mu_Y) f_{XY}, \quad (4)$$

donde el término  $f_{XY}$  es la función de probabilidad conjunta. Lo anterior es importante puesto que revela lo siguiente: si en el producto  $(X - \mu_X)(Y - \mu_Y)$  los valores de  $X$  y  $Y$  tienden a ser mayores o menores que los de las medias poblacionales, la covarianza tendrá signo positivo; por el contrario, si los valores de  $X$  tienden a ser mayores o menores que los de su media, mientras que los valores de  $Y$  varían de forma inversa, la covarianza tendrá signo negativo<sup>29,30</sup>. A su vez, la magnitud del resultado dependerá de cuán grandes sean las diferencias en  $(X - \mu_X)(Y - \mu_Y)$ .

Otro aspecto a resaltar tiene que ver con las medidas de las variables. En vista de que  $X$  y  $Y$  pueden estar expresadas en distintas unidades, el resultado de la covarianza bien podría ser no interpretable, desventaja que condujo a desarrollar el coeficiente de correlación, el cual no es más que la estandarización de este término<sup>31</sup>. Defínase a  $\rho_{XY}$  como la correlación entre dos variables aleatorias  $X$  y  $Y$ , tal que<sup>25</sup>:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \quad (5)$$

donde  $\sigma_X$  y  $\sigma_Y$  son las desviaciones estándares de  $X$  y  $Y$ , respectivamente. Como puede apreciarse en la ecuación (5), las unidades del numerador se cancelarán con las del denominador, lo que conduce a que  $\rho_{XY}$  sea adimensional y ofrezca ventajas importantes al momento de su comprensión y manejo.

### Coeficiente De Correlación Muestral De Pearson

Pearson presentó por primera vez en 1895 esta importante medida en términos matemáticos como<sup>31</sup>:

$$r_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{[\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2]^{1/2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}. \quad (6)$$

En la expresión (6), las barras indican las medias muestrales de  $X$  y  $Y$ , en tanto que  $S_{XX}$ ,  $S_{YY}$  y  $S_{XY}$  corresponden a las sumas de cuadrados corregidas para  $X$ ,  $Y$  y el producto cruzado  $XY$ . Tal y como comentan *Rodgers* y *Nicewander*<sup>1</sup>, existen otras formas de calcular el coeficiente de correlación, aunque la ecuación (6) es la más común. Obsérvese que en el numerador las puntuaciones de las variables se centran al restarlas del promedio, para luego acumular tales valores en el producto cruzado. Por otra parte, el denominador sirve para ajustar la escala de las características involucradas y que estas se anulen.

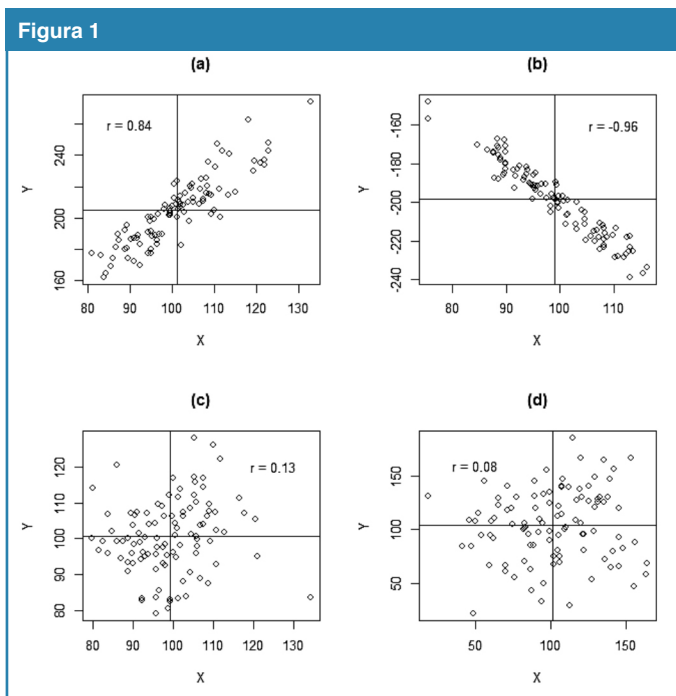
Para ahondar sobre lo que se planteó anteriormente acerca del sentido de la correlación, supóngase ahora que se construye un plano cartesiano cuyo origen es  $(\bar{x}, \bar{y})$ . Considérese también que se dispone de un conjunto de  $n$  pares ordenados de la forma  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  y que se examina el comportamiento de estos en torno al centro del diagrama. Los cuadrantes en los que se localizaría la nube de puntos y el signo del coeficiente se muestran en el cuadro 1, en tanto que en la figura 1 se presentan algunos ejemplos. Nótese que en la medida en que los valores de  $x_i$  y  $y_i$  sean mayores o menores que los de  $\bar{x}$  y  $\bar{y}$ , la correlación será positiva y el gráfico se asemejará a una recta que se desplaza de forma creciente desde el cuadrante III al I (casos 1 y 2, **figura 1a**). Por el contrario, cuando los valores de  $x_i$  sean mayores que los de  $\bar{x}$  y los de  $y_i$  sean menores que los de  $\bar{y}$ ; o viceversa, la correlación será negativa y el diagrama evocará una línea decreciente que se mueve desde el cuadrante II al IV (casos 2 y 3, **figura 1b**). Por último, si las desviaciones de  $X$  y  $Y$  con relación a sus medias tienden a compensarse, la correlación será débil o nula, podrá asumir cualquier signo y los puntos se extenderán en todos los cuadrantes (caso 5, figuras 1c y 1d).

Además de señalar la dirección de la asociación lineal entre las variables, el coeficiente de correlación de *Pearson* es un indicador de la fuerza con que estas se vinculan<sup>20,32</sup>. Se ha establecido con suficiencia en la teoría estadística, que factores como el tamaño muestral, la falta de linealidad o la variabilidad en las características de interés pueden alterar considerablemente la magnitud de esta medida. *Goodwin* y *Leech* hacen una revisión cuidadosa de este tema y explican cuáles son los seis aspectos que mayor impacto tienen en el coeficiente; a saber<sup>33</sup>: (a), la variabilidad en los datos correspondientes a cada variable aleatoria; (b), las diferencias en las formas de las distribuciones marginales de  $X$  y  $Y$ ; (c), la falta de linealidad en la relación entre  $X$  y  $Y$ ; (d), la presencia importante de datos atípicos a nivel biviariado; (e), el tamaño de la muestra; y (f), los posibles errores de medición. Aunque este asunto escape al alcance de la presente revisión, sí se puntualiza acá que el valor de  $R$ ,  $r$  o  $r_{XY}$  —cualquiera sea su notación— será un número real que oscile desde -1 hasta 1. Así pues, la fuerza de la relación entre  $X$  y  $Y$  será mayor, conforme este valor tienda a los extremos antes citados<sup>34</sup>. En la próxima sección se establecerá esto con mayor nivel de detalle.

**Cuadro 1.** Interpretación geométrica del coeficiente de correlación de *Pearson*. Cuando la covariación de  $X$  y  $Y$  es directa, la correlación es positiva y los puntos ocupan los cuadrantes I y III. Cuando la covariación es inversa, la correlación es negativa y los puntos ocupan los cuadrantes II y IV. Cuando las desviaciones se compensan entre sí, la correlación es débil o nula y los puntos ocupan todos los cuadrantes.

Caso	Covariación de X y Y	Valor de la suma del producto cruzado	Signo corr.	Cuadrantes
Caso 1	$x_i > \bar{x} \quad y_i > \bar{y}$	$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) > 0$	+	I y III
Caso 2	$x_i < \bar{x} \quad y_i < \bar{y}$	$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) > 0$	+	I y III
Caso 3	$x_i > \bar{x} \quad y_i < \bar{y}$	$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) < 0$	-	II y IV
Caso 4	$x_i < \bar{x} \quad y_i > \bar{y}$	$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) < 0$	-	II y IV
Caso 5	Desv. compensadas	Cualquiera	Cualquiera	Todos

**Figura 1.** (a), coeficiente de correlación positivo y de magnitud elevada, nube de puntos con pendiente positiva ocupando los cuadrantes I y III; (b), coeficiente de correlación negativo y de magnitud elevada, nube de puntos con pendiente negativa ocupando los cuadrantes II y IV; (c) y (d), coeficientes de correlación con magnitud débil o nula, nube de puntos esparcida en todos los cuadrantes. El eje cartesiano se ha construido a partir de  $(\bar{x}, \bar{y})$  en cada caso.



### Propiedades Del Coeficiente De R De Pearson

Son muchas las propiedades que podrían citarse acerca de este coeficiente; sin embargo, y para efectos de practicidad, se considerarán únicamente las más relevantes.

**1. Adimensionalidad:** tal y como se estableció en la ecuación (6), al dividir la suma de cuadrados del producto XY entre las raíces individuales de las sumas de cuadrados de X y Y, se obtiene un índice sin dimensiones que se origina cuando las unidades del numerador se cancelan con las del denominador. Esta propiedad representa una ventaja esencial que hace de este coeficiente una medida versátil y fácilmente interpretable.

**2. Rango definido entre -1 y 1:** la demostración formal de esta característica es intrincada y va más allá del objetivo de

este artículo; no obstante, puede abordarse intuitivamente si se visualiza al R de Pearson desde una perspectiva geométrica. Tal y como especifican Rodgers y Nicewander<sup>1</sup>, este coeficiente puede entenderse como el coseno del ángulo formado por los vectores asociados a X y Y. Cuando dicho ángulo es cercano a 0, el coseno tiende a 1, lo que implicaría que las variables poseen una alta proximidad en el espacio. Situación similar se presenta cuando el ángulo es aproximadamente igual a 180, en cuyo caso, el coseno será igual a -1 e indicará que las variables siguen exhibiendo una elevada cercanía, pero en direcciones opuestas. Ahora bien, cuando el ángulo entre los vectores es de 90 grados o similar, el coseno tenderá a 0 y esto supondrá que las variables son ortogonales, y que, en consecuencia, no están relacionadas linealmente.

En cuanto a esta propiedad, mucho se ha escrito y discutido acerca de cómo interpretar cualitativamente la magnitud de esta medida. En tal sentido —y para efectos de la presente revisión— se asumirán las sugerencias de Cohen<sup>35</sup> por considerarlas de las más extendidas y respetadas a nivel de la comunidad científica; sin embargo, vale la pena mencionar que estas referencias solo deben ser empleadas como elemento de ayuda al momento de reflexionar sobre la importancia de los hallazgos. El **cuadro 2** ofrece esta clasificación.

**Cuadro 2.** Interpretación de la magnitud del coeficiente de correlación de Pearson según las sugerencias de Cohen. Se supone en este cuadro que la relación se da entre X y Y, pero aplica a cualquier par de variables. Se plantea el valor absoluto del coeficiente, de modo que la magnitud es independiente del signo.

Rango de valores de $r_{XY}$	Interpretación
$0.00 \leq  r_{XY}  < 0.10$	Correlación nula
$0.10 \leq  r_{XY}  < 0.30$	Correlación débil
$0.30 \leq  r_{XY}  < 0.50$	Correlación moderada
$0.50 \leq  r_{XY}  < 1.00$	Correlación fuerte

**3. Relación lineal:** esta propiedad es una de las más importantes y tal vez sea en la que más errores de interpretación se comenten. Es fundamental distinguir que lo que mide el coeficiente de Pearson es la fuerza y la dirección de la relación lineal entre las variables. Así, un  $r_{XY} = \pm 1 r_{XY} = \pm 1$  indicará que hay una perfecta asociación lineal positiva o negativa —según sea el signo— entre X y Y. En este orden de ideas, un  $r_{XY} = 0 r_{XY} = 0$  será evidencia de que no existe relación lineal entre las características de interés, pero no constituye prueba alguna de independencia. En otras palabras, si el coeficiente de correlación es nulo, únicamente se podría establecer que no hay asociación lineal entre las variables aunque pueda existir otro tipo de vinculación<sup>34,36</sup>.

**4. Simetría:** la simetría en este caso establece que, sin importar si se intercambian las posiciones de X y Y, el resultado del coeficiente será el mismo. En términos matemáticos, esto podría expresarse a través de la igualdad  $r_{XY} = r_{YX}$ , lo que a su vez implica que al realizar un análisis de correlación, ninguna de las características de interés

debe asumirse como explicativa de la otra. Referencias que apoyan esta afirmación pueden encontrarse en *Wiedermann* y *Hagmann*<sup>37</sup>, quienes arguyen que el coeficiente de correlación no puede ser usado para determinar la dirección de la relación lineal entre las variables.

**5. Independencia con respecto al origen y a la escala:** el valor del *R* de *Pearson* una vez ha sido calculado, no cambiará a pesar de que se modifique el origen o la escala de los datos; es decir, el coeficiente no se ve afectado por aquellas transformaciones lineales que se apliquen a las variables. En palabras simples, sumar o restar constantes uniformemente a cada variable no alterará el resultado, lo mismo si se multiplica o divide.

**Suposiciones Vinculadas  
Al Uso Del Coeficiente *R* De *Pearson***

El uso adecuado de este coeficiente debe sustentarse en el cumplimiento de las siguientes premisas:

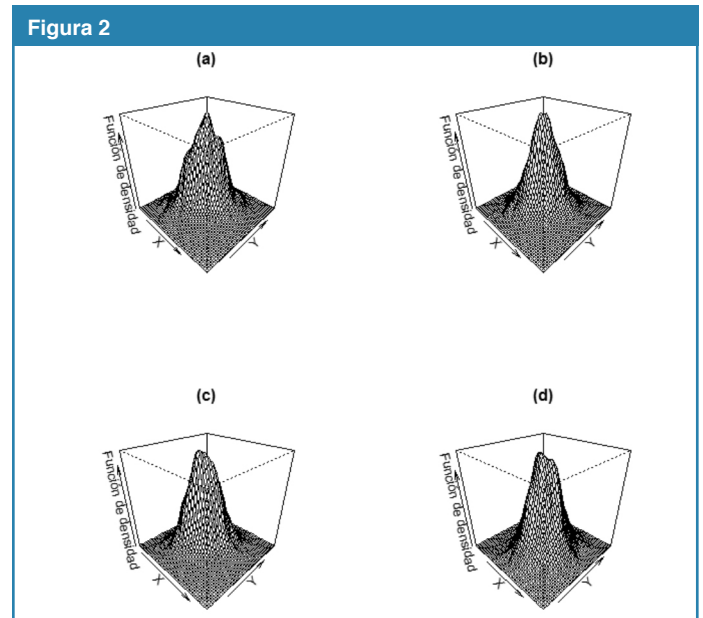
**1. Nivel de medición de las variables:** las dos variables deben ser de intervalo o de razón, aunque no es necesario que ambas tengan el mismo nivel de medición<sup>38-41</sup>. Ejemplos de estas características pueden ser: en ingeniería, la velocidad de un fluido (en m/s) y el factor de fricción de una tubería (adimensional); en psicología, el nivel de ansiedad de un sujeto (en puntos) y la frecuencia cardíaca (en ppm); en medicina, la presión arterial media (en mm de Hg) y la concentración de glicemia en la sangre (en mg/dL); en economía, el índice de precios al consumidor (en porcentaje) y el producto interno bruto (en dólares americanos); entre otras.

Si bien este tópico no se aborda en este artículo, sí se mencionará que existen algunos casos especiales de correlación conocidos como *correlación biserial-puntual* y *coeficiente phi*. En el primero, se utiliza una derivación del *R* de *Pearson* cuando una variable es cuantitativa y la otra es dicotómica<sup>42</sup>; en el segundo, se emplea tal ecuación cuando ambas características son binarias<sup>50,51</sup>.

**2. Datos pareados:** para que el cálculo de esta medida pueda realizarse, se necesitará que los casos en cuestión tengan datos en cada variable<sup>38</sup>. Si hay valores perdidos, estos registros se descartarán por completo del análisis.

**3. Normalidad bivariada:** contrario a lo que sucede en muchas investigaciones en las que se comprueba únicamente la normalidad marginal de *X* y *Y*, el uso apropiado del coeficiente de correlación de *Pearson* exige que se satisfaga el supuesto de normalidad bivariada; esto es, que la distribución de probabilidad conjunta de *X* y *Y* sea normal. La **figura 2** muestra cuatro distribuciones normales bivariadas obtenidas mediante simulación con *R-Studio*. Nótese que, a pesar de que la curva de tales poblaciones no se ajusta de manera perfecta a la densidad esperada, sí se asemeja de forma importante a una campana de *Gauss* construida a partir de ternas ordenadas  $(xx, yy, f_{XY}f_{XY})$ .

**Figura 2.** (a), distribución normal bivariada con medias (0, 0) y matriz de varianzas-covarianzas (1, 0.5, 0.5, 1); (b), distribución normal bivariada con medias (10, 5) y matriz de varianzas-covarianzas (1, 0.2, 0.2, 1); (c), distribución normal bivariada con medias (5, 10) y matriz de varianzas-covarianzas (1, 0.5, 0.5, 1); y (d), distribución normal bivariada con medias (0, 0) y matriz de varianzas-covarianzas (1, 0.2, 0.2, 1).



Retomando el comentario inicial sobre esta premisa y con la intención de enfatizar en la importancia de verificar apropiadamente esta hipótesis, se refuerza acá la idea de no emplear la comprobación de la normalidad de cada variable de manera individualizada. *Rencher*, en su libro titulado *Methods of Multivariate Analysis*, sostiene que cuando se trabaja con varias características, escrutar la normalidad univariada no debe ser el único enfoque ya que existe correlación entre las variables y debido a que la normalidad marginal no garantiza la normalidad conjunta<sup>29</sup>. Esto significa que la normalidad bivariada de los datos podría ser rechazada, aunque haya sido corroborada en términos univariados. También podría ocurrir lo contrario; a saber: autores como *Timm* hacen hincapié que, si una de las variables tiene una distribución que se aleja de la normalidad, entonces el vector multivariado será no normal<sup>45</sup>.

En tal sentido, la recomendación pasa por utilizar pruebas de bondad de ajuste que se adapten al tipo de datos que se manejan en estas situaciones; pruebas que, en la mayoría de los casos, son una generalización de sus similares univariadas. Uno de estos métodos es el sugerido por *Mardia* que se basa en analizar la asimetría y la curtosis de los datos a nivel multivariado<sup>46,47</sup>, siendo tal vez el que mayor estabilidad y fiabilidad ofrece de entre todas las alternativas disponibles<sup>29,45,48</sup>. En la actualidad, este test puede ser fácilmente ejecutado en programas como *R*, *R-Studio*, *Stata* o *Matlab*.

**4. Ausencia de datos atípicos a nivel bivariado:** tal y como sucede con el supuesto anterior, la conjetura relacionada con la presencia de datos atípicos a nivel multivariado suele ser malinterpretada y verificada erróneamente. La causa de esto

puede deberse a que, en estas situaciones, una observación no tiene que ser necesariamente grande o pequeña en relación con el resto de valores de la base de datos para que tenga el efecto de un *outlier*<sup>49</sup>. Adicionalmente, *Rencher* plantea cuatro escenarios que explican el porqué es más complejo el análisis de datos atípicos multivariados que univariados<sup>29</sup>: (a), cuando hay más de dos variables, los datos no pueden ser graficados con facilidad para identificar los *outliers*; (b), las bases de datos multivariadas no pueden ser ordenadas como sí puede hacerse en el caso de una sola variable, en el que es sencillo identificar los extremos de la distribución; (c), el vector de una observación en particular, bien podría tener un error de medición considerable en uno de sus componentes, y errores de medición insignificantes en los otros; y (d), los *outliers* multivariados pueden tener diferentes efectos sobre estadísticos como la media, varianza o la correlación.

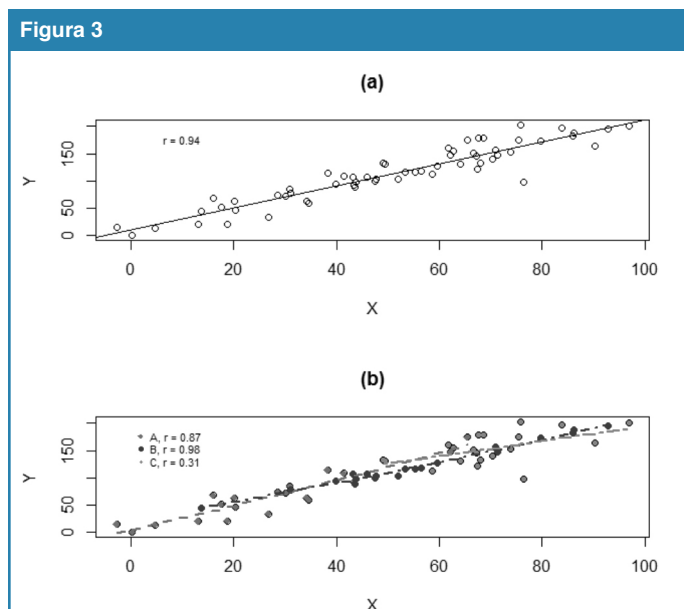
La alternativa clásica para investigar la presencia de datos atípicos multivariados es mediante el cálculo de las distancias de *Mahalanobis*, las cuales usan estimadores de ubicación y dispersión para detectar aquellos puntos que se alejan significativamente del centroide<sup>49-51</sup>. Sin embargo, es necesario enfatizar en el hecho de que, para proporcionar información fiable, el investigador debe cerciorarse de que no se presenten los efectos de enmascaramiento (*masking effect*) o empantanamiento (*swamping effect*). El enmascaramiento sobreviene cuando un dato aberrante no es descubierto debido a la presencia de otros valores atípicos adyacentes. Por otro lado, el empantanamiento ocurre cuando una observación no extrema es clasificada como *outlier* producto de la existencia de otros datos normales<sup>52,53</sup>. Autores como Muñoz-García y Amón-Urbe advierten acerca de las consecuencias potenciales de emplear las distancias de *Mahalanobis* en presencia de estos efectos: el enmascaramiento podría disminuir esta medida cuando en realidad el valor es atípico; el empantanamiento, al contrario, podría aumentarla cuando la observación no es verdaderamente un *outlier*. Estos problemas pueden resolverse si se usan estimaciones robustas para las medias y la matriz de varianzas-covarianzas, tales como el estimador M multivariado, el S bicuadrático o el estimador de covarianza de mínimo determinante<sup>54</sup>.

Así pues, la recomendación que se propone en esta publicación es la de utilizar las distancias de *Mahalanobis* como estrategia principal para la detección de datos atípicos multivariados, siempre y cuando se tenga la precaución de calcular tales medidas con base en estimaciones que no se vean afectadas por la presencia de *outliers*. Esto puede abordarse con facilidad en paquetes como *SPSS*, *R*, *R-Studio* o *Stata*.

**5. Linealidad:** sobre esta propiedad ya se han suministrado suficientes elementos teóricos en esta revisión, de manera que solo se recordará en este punto que lo único que mide el coeficiente *R* de *Pearson* es la fuerza y la dirección de la relación lineal entre dos variables. Ahora bien, si se mencionará que la forma más idónea y más empleada de constatar si esta suposición se cumple o no, es a través de los diagramas de dispersión. Para algunos investigadores, estos gráficos ofrecen incluso más información que el coeficiente en sí; no obstante, conviene adoptarlos cuidadosamente, en especial

cuando la base de datos contiene factores que sirven como variables de agrupación. Así pues, siempre que sea posible, el investigador deberá decidir si lo más provechoso es construir diagramas de dispersión para el conjunto de observaciones como un todo, o generar gráficos parciales<sup>20,31</sup>. La **figura 3** exhibe estas dos situaciones.

**Figura 3.** (a), gráfico de dispersión con línea de ajuste en un conjunto de datos no desagregado, coeficiente de correlación bruto de 0.94; (b), gráfico de dispersión con líneas de ajuste desagregadas según grupos A, B y C, coeficientes de correlación individuales.



**6. Independencia de observaciones:** el postulado de independencia de observaciones es crucial al momento de realizar pruebas de hipótesis o construir intervalos de confianza para el coeficiente de correlación o cualquier otro parámetro. Este supuesto básicamente supone la observancia de dos premisas: en primer lugar, *independencia entre grupos*, lo que implica que los participantes de cada lote deben estar integrados por diferentes sujetos y que cada individuo aparece solo una vez y pertenece solo a un grupo. En segundo lugar, *independencia dentro del grupo*; es decir, el valor obtenido para un sujeto cualquiera en una de las variables, no estará relacionado con los valores del resto de sujetos en esa misma variable.

La consecuencia de violar esta suposición es la de obtener resultados erróneos debido al sesgo que potencialmente se introduce en los datos, de manera que su consecución supone una labor decisiva para el equipo de investigación. Desafortunadamente, no hay una manera sencilla o eficaz de comprobar este supuesto *una vez* que la muestra ha sido recogida; al contrario, la recomendación pasa por velar escrupulosamente que la escogencia y asignación de los sujetos a los diferentes grupos sea al azar, *mientras* se toma la muestra<sup>40</sup>. La planeación cuidadosa de las etapas del experimento es la alternativa ideal para garantizar esto, aunque no siempre se pueda desarrollar este tipo de investigación en áreas vinculadas a las ciencias sociales.

**7. Condiciones del muestreo:** las técnicas inferenciales que se enseñan en los cursos o textos básicos de estadística suponen que la muestra constituye un conjunto de variables aleatorias independientes e idénticamente distribuidas (vaid). En términos matemáticos, esto significa que cada variable aleatoria tiene la misma distribución de probabilidad y que todas son mutuamente independientes<sup>25-27</sup>. Un ejemplo sencillo podría ilustrar esta noción: supóngase que se lanza una moneda equilibrada  $n$  veces. Puesto que el peso de la moneda está distribuido equitativamente, la probabilidad teórica de obtener *cara* o *sello* es la misma en cada repetición; esto es,  $P(C) = P(S) = 1/2$ . Además, el resultado de cada ensayo no tendrá efecto en los lanzamientos subsiguientes, en consecuencia, este ejemplo es un caso de *vaid* y se ajusta al modelo de la distribución binomial, el cual es una sucesión de  $n$  ensayos de *Bernoulli* descritos como sigue: (a), el experimento se repite un número  $n$  finito de veces; (b), cada repetición tiene solo dos resultados posibles; (c), la probabilidad de éxito, denotada por  $pp$ , es constante en cada ensayo; (d), la probabilidad de fracaso, definida como  $q = 1 - p$ , también permanece constante en cada repetición; y (e), los ensayos son independientes.

Ahora bien, el problema se origina cuando se asume erróneamente que cualquier tipo de muestreo aleatorio garantiza lo anterior. Esto no es cierto y puede deducirse con facilidad al contemplar situaciones reales en las que los sujetos son escogidos de una población finita y descartados para las próximas extracciones. Considérese una población de  $N = 100$  adultos de ambos sexos de la cual se seleccionan al azar un total de  $n = 30$  personas. Suponga además que el muestreo se realiza sin reposición; es decir, una vez que un individuo ha sido seleccionado, se elimina su nombre de la población correspondiente y se evitan así los registros duplicados. Bajo estas circunstancias, la probabilidad de inclusión de cada sujeto no es constante y, aunque las variables aleatorias que son analizadas en los individuos tengan la misma distribución de probabilidad, no serán mutuamente independientes. Sin embargo, si el muestreo se realizase con reposición y el nombre de cada persona se repusiese luego de haber sido extraído, se cumpliría la condición de independencia y la muestra sería, en efecto, un conjunto de *vaid*.

En tal sentido, resulta esencial detallar las condiciones que debe reunir el muestreo para que las ecuaciones mostradas en esta revisión sean válidas, así como también las técnicas de inferencia estadística obtenidas a partir de ellas. El método de muestreo que asegura la noción de *vaid* es el muestreo aleatorio simple (MAS) con reposición. *Lohr*, en su libro titulado *Sampling: design and analysis*, establece que esta técnica es la más elemental y la que sirve como fundamento para métodos más sofisticados. El requisito matemático principal de este procedimiento es la equiprobabilidad; esto es, cada sujeto en la población tiene la misma probabilidad de ser escogido para pertenecer a la muestra<sup>55-57</sup>. Otra característica distintiva de esta técnica es la homogeneidad de las unidades de información; a saber: cuando se implementa un MAS, se asume que los sujetos que conforman la población poseen características similares en las variables de interés

para el estudio. Lo anterior, gracias al uso de algoritmos<sup>58</sup> y test estadísticos entre algunos ya mencionados<sup>59</sup>.

Como puede desprenderse del párrafo anterior, la implementación del MAS en casos reales es sumamente limitada. Por una parte, sería ilógico seleccionar a la misma persona más de una vez para que participe en la investigación; por otra, aquellos escenarios ideales en los que la población es homogénea constituyen eventos muy poco probables. Así pues, la sugerencia que se esboza en el presente artículo es la de diseñar las etapas iniciales del estudio de modo que garanticen, no solo que se está seleccionando una muestra acorde a las características de un MAS, sino también que asegure el supuesto de independencia descrito en el numeral anterior. Si no es posible lograr estos objetivos, el equipo de investigación podría recurrir a técnicas como el muestreo aleatorio estratificado, aleatorio por conglomerados o al de muestras complejas, teniendo la previsión de incorporar las *ponderaciones muestrales* al momento de obtener las estimaciones puntuales y sus correspondientes errores estándares. Programas estadísticos como *Stata*, *SAS*, *R* o *R-Studio* poseen módulos que realizan estas labores con relativa simplicidad.

## Referencias

1. Rodgers JL, Nicewander WA. Thirteen Ways to Look at the Correlation Coefficient. *Am Stat.* febrero de 1988;42(1):59.
2. J. J. Strossmayer University of Osijek, Faculty of Agriculture in Osijek, Kralja Petra Svačića 1d, 31000 Osijek, Croatia, Rebekić A, Lončarić Z, Petrović S, Marić S. Pearson's or Spearman's correlation coefficient - Which one to use? *Poljoprivreda.* 18 de diciembre de 2015;21(2):47-54.
3. Zhong M, Zheng S, Wang G, Hua J, Gebbink R. Correlation analysis of combined and separated effects of wing deformation and support system in the CAE-AVM study. *Chin J Aeronaut.* marzo de 2018;31(3):429-38.
4. Kusiak A, Li W. Estimation of wind speed: A data-driven approach. *J Wind Eng Ind Aerodyn.* octubre de 2010;98(10-11):559-67.
5. Fu X, Wang J, Tan F, Feng X, Wang D. Occurrence and enrichment of trace elements in marine oil shale (china) and their behavior during combustion. *Oil Shale.* 2015;32(1):42.
6. Freitas N de O, Forero CG, Caltran MP, Alonso J, Dantas RAS, Piccolo MS, et al. Validation of the Perceived Stigmatization Questionnaire for Brazilian adult burn patients. Lin C-Y, editor. *PLOS ONE.* 30 de enero de 2018;13(1):e0190747.
7. Shoji K, Lesniewska M, Smoktunowicz E, Bock J, Luszczynska A, Benight CC, et al. What Comes First, Job Burnout or Secondary Traumatic Stress? Findings from Two Longitudinal Studies from the U.S. and Poland. Elhai JD, editor. *PLOS ONE.* 25 de agosto de 2015;10(8):e0136730.
8. Larson LR, Jennings V, Cloutier SA. Public Parks and Wellbeing in Urban Areas of the United States. Lepczyk CA, editor. *PLOS ONE.* 7 de abril de 2016;11(4):e0153211.
9. Joseph PV, Wang Y, Fourie NH, Henderson WA. A computational framework for predicting obesity risk based on optimizing and integrating genetic risk score and gene expression profiles. Devaney J, editor. *PLOS ONE.* 24 de mayo de 2018;13(5):e0197843.

10. Vera M, Huérfano Y, Contreras J, Vera M, Salazar W, Vargas S, et al. Desarrollo de una técnica computacional no lineal para la segmentación de hematomas subdurales, presentes en imágenes de tomografía computarizada cerebral. *Arch Venez Farmacol Ter.* 2017;36(6):168-73.
11. Vera M, Huérfano Y, Contreras J, Vera M, Salazar W, Vargas S, et al. Técnica computacional no lineal para la detección de hemoventrículo, en imágenes de tomografía computarizada cerebral. *Lati-noam Hipertens* [Internet]. 29 de enero de 2018 [citado 11 de enero de 2019];12(5). Disponible en: [http://saber.ucv.ve/ojs/index.php/rev\\_lh/article/view/14695](http://saber.ucv.ve/ojs/index.php/rev_lh/article/view/14695)
12. Bernardo SM, Anholon R, Novaski O, Silva D, Quelhas OLG. Main causes that lead strategies to decline at execution phase: an analysis of Brazilian companies. *Int J Product Perform Manag.* 6 de marzo de 2017;66(3):424-40.
13. Rani K. Analysis of Dynamic Interaction between Foreign Investments and Indian Stock Market. *J Int Econ.* 2015;6(2):10.
14. Wu S, Duncan F, Anderson NH, Kuppaswamy A, Macloed MR, Mead GE. Exploratory Cohort Study of Associations between Serum C - Reactive Protein and Fatigue after Stroke. Hashimoto K, editor. *PLOS ONE.* 24 de noviembre de 2015;10(11):e0143784.
15. Tarabichi M, Detours V. A research note regarding «Variation in cancer risk among tissues can be explained by the number of stem cell divisions». *F1000Research.* 22 de agosto de 2016;5:2044.
16. Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science.* 2 de enero de 2015;347(6217):78-81.
17. Fenton A, Jesky MD, Webster R, Stringer SJ, Yadav P, Chapple I, et al. Association between urinary free light chains and progression to end stage renal disease in chronic kidney disease. Bjornstad P, editor. *PLOS ONE.* 9 de mayo de 2018;13(5):e0197043.
18. Dinas PC, Nintou E, Psychou D, Granzotto M, Rossato M, Vettor R, et al. Association of fat mass profile with natriuretic peptide receptor alpha in subcutaneous adipose tissue of medication-free healthy men: A cross-sectional study. *F1000Research.* 15 de marzo de 2018;7:327.
19. Akuffo KO, Nolan JM, Peto T, Stack J, Leung I, Corcoran L, et al. Relationship between macular pigment and visual function in subjects with early age-related macular degeneration. *Br J Ophthalmol.* febrero de 2017;101(2):190-7.
20. Kozak M, Krzanowski W, Tartanus M. Use of the correlation coefficient in agricultural sciences: problems, pitfalls and how to deal with them. *An Acad Bras Ciênc.* diciembre de 2012;84(4):1147-56.
21. Warren WG. Correlation or Regression: Bias or Precision. *Appl Stat.* 1971;20(2):148.
22. Mukaka M. A guide to appropriate use of Correlation coefficient in medical research. *Malawi Med J J Med Assoc Malawi.* septiembre de 2012;24(3):69-71.
23. Porter AM. Misuse of correlation and regression in three medical journals. *J R Soc Med.* marzo de 1999;92(3):123-8.
24. Use and Misuse of Correlation Coefficients [Internet]. *STAT 509.* [citado 15 de enero de 2019]. Disponible en: <https://newonlinecourses.science.psu.edu/stat509/node/160/>
25. Montgomery DC, Runger GC. *Applied statistics and probability for engineers.* 3rd ed. New York: Wiley; 2003. 706 p.
26. Walpole RE, Myers RH, Myers SL, Ye K. *Probability & statistics for engineers & scientists: MyStatLab update* [Internet]. 2017 [citado 24 de junio de 2018]. Disponible en: <http://www.mylibrary.com?id=947904>
27. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences,* 5th Edition. [Internet]. Vol. 47. 1991 [citado 24 de junio de 2018]. Disponible en: <https://www.jstor.org/stable/2532686?origin=crossref>
28. Weisberg S. *Applied linear regression.* 3rd ed. Hoboken, N.J.: Wiley-Interscience; 2005. 310 p. (Wiley series in probability and statistics).
29. Rencher AC. *Methods of multivariate analysis.* 2nd ed. New York: J. Wiley; 2002. 708 p. (Wiley series in probability and mathematical statistics).
30. Covariance and Correlation. En: *Hedge Funds* [Internet]. Oxford, UK: John Wiley & Sons Ltd; 2013 [citado 26 de junio de 2018]. p. 121-46. Disponible en: <http://doi.wiley.com/10.1002/9781118673546.ch6>
31. Asuero AG, Sayago A, González AG. The Correlation Coefficient: An Overview. *Crit Rev Anal Chem.* enero de 2006;36(1):41-59.
32. Rigby AS. *Statistical methods in epidemiology.* VI. Correlation and regression: the same or different? *Disabil Rehabil.* enero de 2000;22(18):813-9.
33. Goodwin LD, Leech NL. Understanding Correlation: Factors That Affect the Size of r. *J Exp Educ.* abril de 2006;74(3):249-66.
34. Ratner B. The correlation coefficient: Its values range between +1/-1, or do they? *J Target Meas Anal Mark.* junio de 2009;17(2):139-42.
35. Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988. 567 p.
36. Sedgwick P. Pearson's correlation coefficient. *BMJ.* 4 de julio de 2012;345(jul04 1):e4483-e4483.
37. Wiedermann W, Hagmann M. Asymmetric properties of the Pearson correlation coefficient: Correlation as the negative association between linear regression residuals. *Commun Stat - Theory Methods.* noviembre de 2016;45(21):6263-83.
38. Yeager K. *LibGuides: SPSS Tutorials: Pearson Correlation* [Internet]. [citado 18 de diciembre de 2018]. Disponible en: <https://libguides.library.kent.edu/SPSS/PearsonCorr>
39. Measure of association I statistics [Internet]. *Encyclopedia Britannica.* [citado 17 de diciembre de 2018]. Disponible en: <https://www.britannica.com/topic/measure-of-association>
40. Pearson Product-Moment Correlation - When you should run this test, the range of values the coefficient can take and how to measure strength of association. [Internet]. [citado 18 de diciembre de 2018]. Disponible en: <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>
41. Pearson's Product-Moment Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example. [Internet]. [citado 18 de diciembre de 2018]. Disponible en: <https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php>
42. Point-Biserial Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example. [Internet]. [citado 18 de diciembre de 2018]. Disponible en: <https://statistics.laerd.com/spss-tutorials/point-biserial-correlation-using-spss-statistics.php>
43. Chi-Square Test for Association using SPSS Statistics - Procedure, assumptions and reporting the output [Internet]. [citado 18 de diciembre de 2018]. Disponible en: <https://statistics.laerd.com/spss-tutorials/chi-square-test-for-association-using-spss-statistics.php>



44. Cui Y, Wang S, Yan S. Evaluating the Applicability of Phi Coefficient in Indicating Habitat Preferences of Forest Soil Fauna Based on a Single Field Study in Subtropical China. PLoS ONE [Internet]. 1 de marzo de 2016 [citado 18 de diciembre de 2018];11(3). Disponible en: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4773244/>
45. Timm NH. Applied multivariate analysis. New York: Springer; 2002. 693 p. (Springer texts in statistics).
46. Mardia KV. Measures of Multivariate Skewness and Kurtosis with Applications. Biometrika. diciembre de 1970;57(3):519.
47. Kankainen A, Taskinen S, Oja H. On Mardia's Tests of Multinormality. En: Hubert M, Pison G, Struyf A, Van Aelst S, editores. Theory and Applications of Recent Robust Methods [Internet]. Basel: Birkhäuser Basel; 2004 [citado 30 de diciembre de 2018]. p. 153-64. Disponible en: [http://link.springer.com/10.1007/978-3-0348-7958-3\\_14](http://link.springer.com/10.1007/978-3-0348-7958-3_14)
48. Romeu JL, Ozturk A. A Comparative Study of Goodness-of-Fit Tests for Multivariate Normality. J Multivar Anal. agosto de 1993;46(2):309-34.
49. Cabana E, Laniado H, Lillo RE. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. :80.
50. De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemom Intell Lab Syst. enero de 2000;50(1):1-18.
51. Franklin S, Thomas S, Franklin S. Robust multivariate outlier detection using Mahalanobis' distance and modified Stahel-Donoho estimators. Semantic Sch. 2001;35.
52. Bendre SM. Masking and swamping effects on tests for multiple outliers in normal sample. Commun Stat - Theory Methods. enero de 1989;18(2):697-710.
53. Chiang J-T. The masking and swamping effects using the planted mean-shift outliers models. Int J Contemp Math Sci. 2007;2:297-307.
54. García JAM, Uribe IA. Técnicas para detección de outliers multivariantes. Rev En Telecomunicaciones E Informática. 2013;3(5):11-25.
55. Lohr SL. Sampling: Design and Analysis. 3.ª ed. Brooks/Cole Cengage Learning; 2010. 596 p.
56. Alf C, Lohr S. Sampling Assumptions in Introductory Statistics Classes. Am Stat. febrero de 2007;61(1):71-7.
57. Pineda-Roa C, Navarro-Segura M. Validación de una prueba para medir eventos vitales estresantes en adultos gay, lesbianas y bisexuales colombianos. Psicogente [Internet]. 1ene.2019 [citado 1mar.2019];22(41):1-9. Available from: <http://revistas.unisimon.edu.co/index.php/psicogente/article/view/3305>
58. Londoño González B. & Sánchez, P. Algoritmo Novedoso Para la Detección de Tareas Repetitivas en el Teclado, Investigación e Innovación en Ingenierías, vol. 3, no. 2, 2015. DOI: <https://doi.org/10.17081/invinno.3.2.2031>
59. Insignares Movilla, J. & Orozco E. The t Distribution: A Transformation of the Employee of the Brewery, Investigación e Innovación en Ingenierías, vol. 2, no. 2, 2014. DOI: <https://doi.org/10.17081/invinno.2.2.2049>