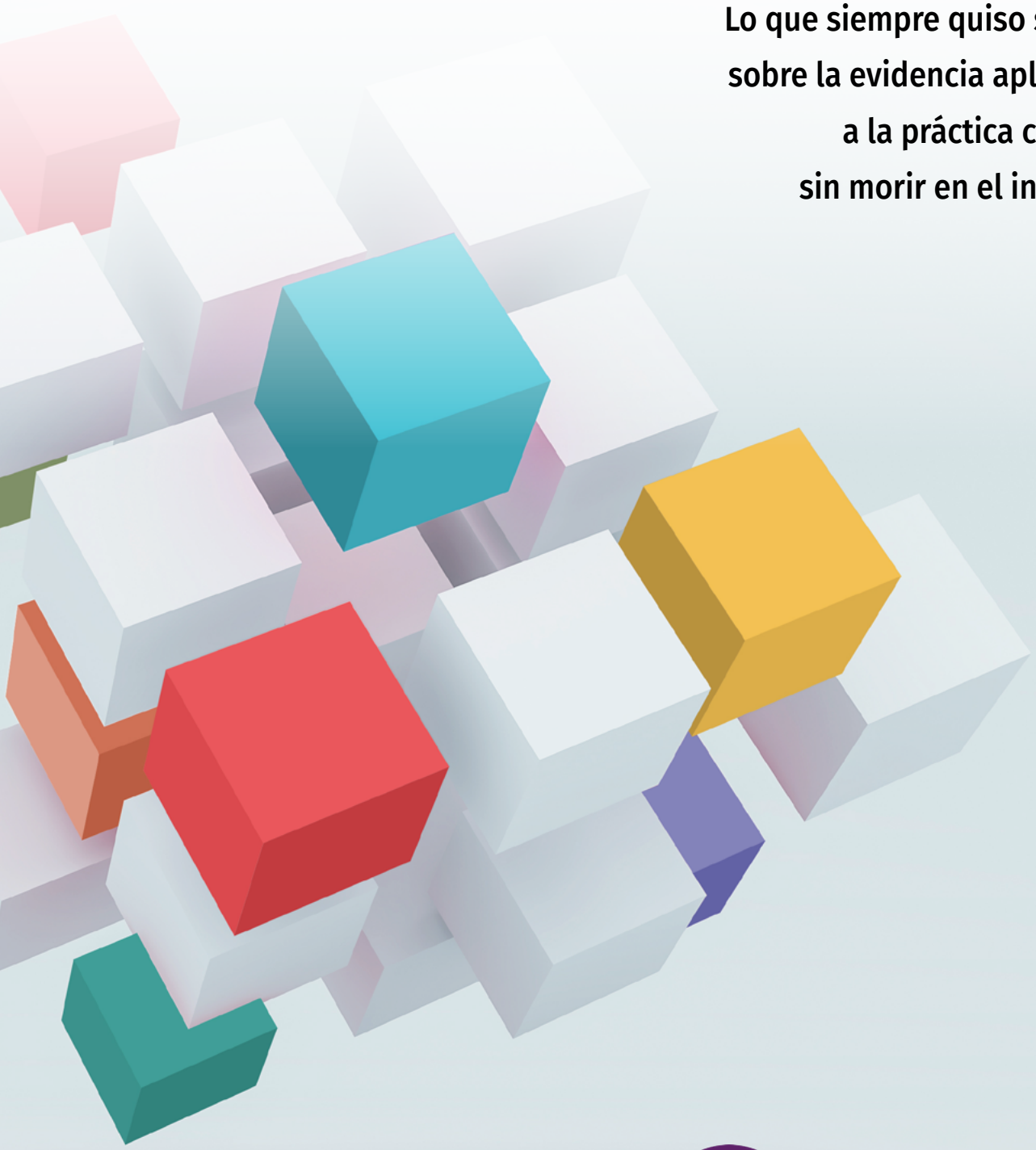
Medicina Basada en la Evidencia

Lo que siempre quiso saber
sobre la evidencia aplicada
a la práctica clínica
sin morir en el intento



Grupo Pediatría Basada en Evidencias
AEPap

aepCPBE
Comité de Pediatría Basada En la Evidencia

Cómo citar este documento:
Comité de Pediatría Basada en la Evidencia de la AEP; Grupo de Trabajo de Pediatría Basada en la Evidencia de la AEPap (coords.). Medicina Basada en la Evidencia. Lo que siempre quiso saber la evidencia aplicada a la práctica clínica sin morir en el intento. Madrid: Lúa Ediciones; 2024.

Han participado como coordinadores del libro:

Comité de Pediatría Basada en la Evidencia de la AEP:
Paz González Rodríguez, María Aparicio Rodrigo, Pilar Aizpurua Galdeano, M.ª Jesús Esparza Olcina, Eduardo Ortega Páez

Grupo de Trabajo de Pediatría Basada en la Evidencia de la AEPap:
M.ª Salomé Albi Rodríguez, Javier González de Dios, Rafael Martín Masot, Carlos Ochoa Sangrador, Manolo Molina Arias

Realización y edición:
Lúa Ediciones 3.0, S.L.
www.luaediciones.com

ISBN: 978-84-128758-1-2

6.10.

Correlación. Modelos de regresión

Eduardo Cuestas Montañés
Servicio de Pediatría y Neonatología. Hospital Privado Universitario de Córdoba. Argentina

Manuel Molina Arias
Servicio de Gastroenterología Pediátrica. Hospital Infantil Universitario La Paz. Madrid. España

Carlos Ochoa Sangrador
Servicio de Pediatría. Complejo Asistencial de Zamora. Zamora. España

"La regresión lineal es como el GPS de los datos:
te muestra la dirección a seguir en un mundo lleno de incertidumbre estadística"

William S. Cleveland

OBJETIVOS:

- Comprender los conceptos de correlación y regresión, y sus diferencias
- Saber valorar un coeficiente de correlación
- Saber cuándo utilizar el coeficiente de correlación de Pearson
- Saber cuándo utilizar el coeficiente de correlación de Spearman
- Comprender los modelos básicos de regresión

Muchas veces nos interesa determinar si dos variables cuantitativas están asociadas y, si es posible, a partir de una de ellas predecir el valor de la otra. Para ello se utilizan dos análisis estadísticos diferentes denominados análisis de correlación y de regresión.

Si el objetivo es establecer una asociación entre las dos variables, recurriremos a la correlación, que es una medida de la relación lineal entre dos variables numéricas. En cambio, si la meta es la predicción, recurriremos a un modelo de regresión, en el que una de las variables se considera independiente o predictora y la otra variable dependiente o resultado.

CORRELACIÓN

Si dos parámetros tienen una relación lineal, existe correlación entre ellos. Por ejemplo, la glucemia y la insulina están correlacionadas entre sí, lo que significa que cuando cambia una, también cambia la otra. La relación puede ser positiva o negativa. En la correlación positiva, si el valor de un parámetro aumenta o disminuye, el valor del otro parámetro varía también en el mismo sentido. La correlación también puede ser negativa y, en este caso, si el valor de un parámetro disminuye, el valor del otro parámetro aumenta, y viceversa. La correlación solo indica una relación, pero no implica necesariamente relación causa-efecto.

La correlación es un término general para la asociación entre pares de variables. Esta puede ser positiva, cuando las dos variables cambian en el mismo sentido, o negativa, cuando lo hacen en sentido contrario

COEFICIENTE DE CORRELACIÓN

El coeficiente de correlación mide la dirección y la intensidad de la relación, es decir, lo fuerte o débil que es, pudiendo tener valores entre -1 y 1. Si existe una relación perfecta, el coeficiente será de 1 si la correlación es positiva, o de -1 si es negativa. Si no existe ninguna relación lineal, el coeficiente será 0. De esta forma, si el valor del coeficiente de correlación se aproxima a 0, la relación es débil, mientras que será más fuerte cuanto más se aproxime a 1 o -1.

Existen varios coeficientes de correlación. A continuación, describiremos los utilizados con más frecuencia.

El coeficiente de correlación cuantifica la fuerza de la asociación entre dos variables, con valores entre -1 y 1, siendo los valores extremos una relación lineal perfecta, y el signo, la dirección de la relación. Un valor de cero indica que no existe una relación lineal entre las variables

1. Coeficiente de correlación de Pearson

El coeficiente de correlación lineal producto-momento, más conocido como **coeficiente de correlación de Pearson** (*r*), es el más utilizado y se obtiene al dividir la covarianza entre el producto de la varianza de las dos variables:

$$r = \frac{S_{xy}}{S_x S_y}$$

donde s_{xy} representa la covarianza y s_x y s_y las varianzas de las variables "x" e "y".

Para utilizar el coeficiente de correlación de Pearson entre dos variables continuas deben cumplirse los siguientes supuestos:

a) La relación entre las dos variables debe ser lineal. Esto puede comprobarse de forma sencilla con un diagrama de dispersión, observando que la forma de la nube de puntos sigue la dirección de una recta.

b) Ambas variables deben seguir una distribución normal. Podemos comprobarlo con una prueba de hipótesis, como la de Shapiro-Wilk o la de Kolmogorov-Smirnov, o con un método gráfico, como el histograma o el gráfico de cuantiles teóricos.

c) Debe existir homocedasticidad, es decir, que la varianza de la variable "y" debe ser constante a lo largo de los valores de la variable "x". Podemos confirmar si se cumple este supuesto de forma sencilla comprobando que la nube de puntos se dispersa de forma similar a lo largo de los valores de la variable "x".

Para utilizar el coeficiente de correlación de Pearson la relación entre las variables debe ser lineal, deben distribuirse de forma normal en la población y deben cumplir el criterio de homocedasticidad

Por último, debemos saber que el valor de este coeficiente es sensible a la presencia de valores extremos en la distribución, que pueden sesgar la magnitud del efecto estimado. En estos casos, nos plantearemos si lo más idóneo es utilizar alguna alternativa al coeficiente de correlación de Pearson.

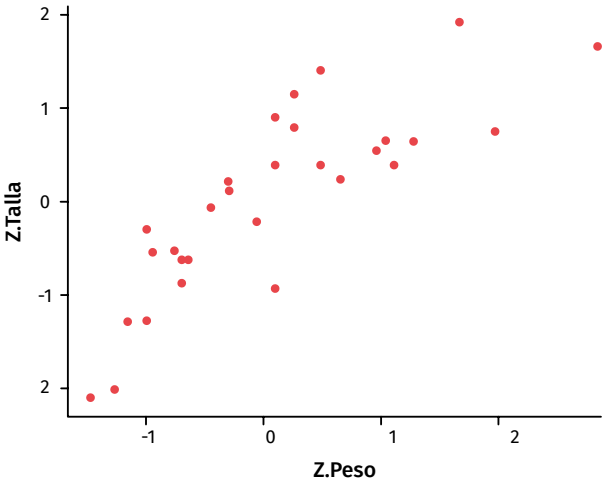
Veamos un ejemplo:

Podemos calcular el coeficiente de correlación de Pearson utilizando un programa de acceso libre, el *software* estadístico R (https://www.r-project.org/) con el *plugin* RCommander y esta base de datos.

En la base de datos se recogen una serie de registros con información sobre niños asmáticos. Vamos a determinar si existe correlación entre los valores de peso y talla estandarizados (Z.Peso, Z.Talla). En el **Anexo 1** de este capítulo se muestran las instrucciones para realizar este ejercicio. Si lo necesita, puede revisar el capítulo 6.18. correspondiente a la instalación de R y RCommander.

Una vez cargados los datos, representamos el diagrama de dispersión (**Figura 1**), con el que podemos asumir que ambas variables se relacionan de forma lineal.

Figura 1. Diagrama de dispersión entre las variables peso estandarizado (Z.Peso) y talla estandarizada (Z.Talla)



Seguidamente, comprobamos la asunción de normalidad de la variable Z.Peso, mediante una prueba de Shapiro-Wilk. El programa nos ofrece el resultado, con un estadístico W = 0,948 y un valor de significación de p = 0,158. No podemos rechazar la hipótesis nula, por lo que asumimos que la variable peso estandarizado sigue una distribución normal. Podemos repetir el proceso para la variable Z.Talla, llegando a la misma conclusión (W = 0,982, p = 0,882).

Por último, vamos a comprobar el supuesto de homocedasticidad. Si observamos el diagrama de puntos de la **Figura 1**, podemos asumir que la nube se dispersa de forma similar en todo el rango de valores de la variable representada en el eje x.

Una vez comprobado que se cumplen los tres requisitos necesarios, calculamos el coeficiente de correlación de Pearson.

El programa nos ofrece un valor de r = 0,82, con un valor de significación estadística p <0,05. Por lo tanto, podemos concluir que existe una asociación alta entre las dos variables.

El programa R nos ofrece también el intervalo de confianza del 95% del coeficiente, que es de 0,66 a 0,91. El intervalo no incluye el valor nulo (0), por lo que ya sabemos que alcanza significación estadística sin necesidad de conocer el valor de p.

2. Coeficiente de correlación de Spearman

El coeficiente de correlación por rangos, más conocido como **coeficiente de correlación de Spearman** (ρ) es el equivalente no paramétrico del coeficiente de Pearson.

Como ocurre con el resto de las técnicas no paramétricas, no se emplean los datos directos para el cálculo del coeficiente, sino su transformación en rangos. El coeficiente de Spearman no precisa asumir el supuesto de normalidad de las variables, por lo que puede utilizarse cuando no se cumplen los supuestos necesarios para el coeficiente de correlación de Pearson.

Aunque la potencia del coeficiente de Spearman es menor que la del coeficiente de Pearson, tiene una serie de ventajas sobre este último. En primer lugar, no exige supuesto de linealidad, por lo que puede utilizarse en casos de relación logística y exponencial. Solo debe cumplirse que la relación entre las dos variables sea monótona, lo cual quiere decir que cuando una de las variables cambia, la otra lo hace con una tendencia constante.

En segundo lugar, como ya se ha dicho, al ser una prueba no paramétrica, no precisa asumir el supuesto de normalidad de las variables. Por último, al calcularse con los rangos en lugar de con los datos directos, es mucho más robusto a la presencia de valores extremos que el coeficiente de Pearson.

Volviendo al ejemplo anterior:

Partiendo de la base de datos sobre niños asmáticos, vamos a suponer que no se cumple alguno de los tres requisitos necesarios para poder emplear el coeficiente de Pearson. En ese caso, podríamos calcular su alternativa no paramétrica, el coeficiente de correlación de Spearman.

En el **Anexo 2** de este capítulo se muestran las instrucciones para realizar este ejercicio.

El programa nos ofrece un valor del coeficiente ρ = 0,85, con un valor de p <0,05. En el caso del coeficiente de Spearman, RCommander no calcula de forma directa su intervalo de confianza, para lo cual habría que recurrir a paquetes adicionales.

Al igual que en el ejemplo anterior, podemos concluir que existe una asociación alta entre las dos variables.

3. Otros coeficientes de correlación

Además, existen otros coeficientes, el más utilizado de los cuales es el coeficiente tau de Kendall (τ).

La **tau de Kendall** es otra alternativa no paramétrica, cuyo uso puede preferirse al de Spearman en aquellos casos de muestras pequeñas y en las que exista una alta ligadura de rangos (al ordenar los datos por rangos, existen múltiples coincidencias en la misma posición).

Otros coeficientes menos utilizados son el **coeficiente de correlación parcial**, que estudia la relación entre dos variables, pero teniendo en cuenta y eliminando la influencia de otras variables existentes; el **coeficiente de correlación semiparcial**, similar al anterior, pero que discrimina el efecto de terceras variables sobre las dos correlacionadas de forma independiente (no sobre las dos de forma simultánea, como el coeficiente parcial); y el **coeficiente de correlación múltiple**, que permite conocer la correlación entre una variable y un conjunto de variables, todas ellas cuantitativas.

REGRESIÓN

La idea es similar a la correlación y a veces se confunde con ella. Es importante aclarar la diferencia entre correlación y regresión. La correlación solo indica la fuerza de la relación entre dos variables. La regresión permite estimar o predecir el valor de una variable dependiente o explicativa en función del valor que tome la otra variable, la independiente o de respuesta.

Mientras que la correlación mide únicamente la fuerza y dirección de la asociación entre dos variables, la regresión permite estimar los valores de una de ellas (dependiente) a partir de los valores de la otra (independiente)

La regresión es un instrumento potente porque puede demostrar la asociación entre muchas variables explicativas y la variable respuesta, y puede ponderar el efecto independiente de cada una de ellas. También permite mostrar relaciones no lineales entre las variables explicativas y la de respuesta.

En general, describimos la regresión como simple o univariable cuando en el modelo solo se incluye una variable independiente; esto contrasta con la regresión múltiple o multivariable, en la que intervienen dos o más variables independientes o predictoras.

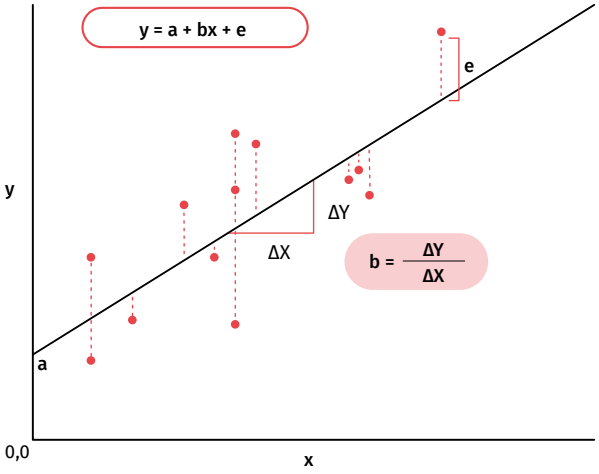
MODELOS DE REGRESIÓN

Si tomamos una variable independiente "x" y una variable dependiente "y", todos los modelos de regresión simple se ajustan a la siguiente ecuación:

$$\text{Función}(y) = a + bx + e$$

El componente "Función(y)" dependerá del tipo de variable dependiente del modelo, lo que nos condicionará el modelo de regresión concreto que tendremos que utilizar. En la **Figura 2** se muestra un ejemplo de diagrama de dispersión de dos variables con la línea de regresión del modelo, en este caso lineal, así como el significado de los diferentes coeficientes de la ecuación de regresión: "a" y "b" son los denominados coeficientes de regresión.

Figura 2. Representación de los distintos componentes de un modelo de regresión lineal simple. El punto "a" muestra el intercepto con el eje de ordenadas; "b" representa la pendiente de la recta de regresión, en cuántas unidades aumenta la variable y por cada unidad de aumento de la variable x; "e" representa el error entre el valor real y la predicción de la recta, llamados residuos y representado por las líneas de puntos



El componente "a" representa el valor de "y" cuando "x" vale 0. Suele denominarse interceptor, ya que es el punto donde la representación gráfica de la línea de regresión cruza el eje de ordenadas (eje y).

El componente "b" representa la pendiente de la línea y nos informa de en cuántas unidades aumenta la variable "y" por cada unidad que aumenta la variable "x".

Por último, el cuarto componente, "e", representa la variabilidad aleatoria del modelo. Esta variabilidad será la responsable de la diferencia que se produzca entre la predicción del modelo de regresión y el valor real observado en el estudio.

Según los componentes descritos en la ecuación anterior, podemos definir los cuatro modelos de regresión simple utilizados con más frecuencia:

1. Regresión lineal simple

Es el caso más sencillo y se aplica a dos variables cuantitativas. En este caso, la función del modelo es la media aritmética de la variable dependiente.

2. Regresión logística

La regresión logística es muy similar a la regresión lineal; lo usamos cuando la variable dependiente es cualitativa dicotómica (por ej.: la presencia o ausencia de una enfermedad, que se codifica como 1 o 0) y una serie de variables explicativas (independientes) discretas o continuas.

La función del modelo será el logaritmo neperiano (natural) de la *odds* de que la variable dependiente tenga el valor 1. El coeficiente "b" representa el logaritmo neperiano de la *odds ratio* de que ocurra un fenómeno por unidad de cambio de la variable independiente, por lo que podremos estimar la *odds ratio* calculando su antilogaritmo (e^b).

3. Regresión de Poisson

La distribución de Poisson es la distribución de probabilidad del recuento de sucesos raros que ocurren aleatoriamente en un intervalo de tiempo o espacio a una tasa media constante. Constituye la base de la regresión de Poisson, que se utiliza fundamentalmente para analizar la tasa de una enfermedad cuando los individuos tienen diferentes tiempos de seguimiento.

La distribución de Poisson es una distribución discreta, por lo que los valores de la variable dependiente son enteros positivos. Esto la convierte en la técnica ideal para situaciones de recuento, como número de ingresos, número de hijos, etc. La función del modelo de regresión de Poisson es el logaritmo neperiano de lambda (λ), que representa la probabilidad de que ocurra un evento en un intervalo determinado, lo que suele corresponder a la densidad de incidencia en los estudios longitudinales.

4. Regresión de riesgos proporcionales de Cox

Se utiliza en estudios de supervivencia, cuando la variable dependiente es de tipo tiempo a suceso. El modelo es similar al de la regresión logística, con la diferencia de que la función representa el logaritmo neperiano de la tasa de riesgos instantáneos (*hazard ratio*).

La interpretación de todos estos modelos se verá de forma más clara cuando se desarrollen en próximos capítulos, donde se describirán sus peculiaridades, sus requisitos de aplicación y su modo de llevar a cabo con ejemplos prácticos.

Anexo 1

1. Lanzar RCommander

2. Cargar esta base de datos

3. Comprobar el supuesto de linealidad. Representar el diagrama de dispersión:
   Gráficas\Diagrama de dispersión...
   Seleccionar Z.Peso como variable "x" y Z.Talla como variable "y" → Aceptar

4. Comprobar el supuesto de normalidad:
   Estadísticos\Resúmenes\Test de normalidad...
   Marcamos la variable Z.Peso y la prueba elegida, en este caso, Shapiro-Wilk → Aceptar
   Repetir los mismos pasos para la variable Z.Talla

5. Comprobar el supuesto de homocedasticidad. Utilizar el mismo gráfico de dispersión del paso

6. Calcular el coeficiente de correlación de Pearson:
   Estadísticos\Resúmenes\Test de correlación...
   Seleccionar las dos variables, el coeficiente de correlación elegido (el de Pearson) y marcar la opción para un contraste bilateral (salvo que se conozca el sentido de la asociación, en cuyo caso se podría seleccionar una de las dos opciones de contraste unilateral) → Aceptar

Anexo 2

Los pasos son similares a los del ejemplo del **Anexo 1**, pero en el paso 6 seleccionaríamos el coeficiente de Spearman.

PREGUNTAS DE AUTOEVALUACIÓN

1. ¿Cuál de los siguientes análisis permite predecir una variable a partir de otra?:
a) Error estándar.
b) Correlación.
c) Regresión.
d) Análisis de la varianza.

2. Si existe una correlación muy fuerte entre dos variables, entonces el coeficiente de correlación debe ser:
a) Mayor que 1 si la correlación es positiva.
b) Cercana a 0, pero negativa, si la correlación es negativa.
c) Lo más cercano posible a 1 (+1 o -1).
d) Lo más cercano posible a 0.

3. ¿Cuál de los siguientes supuestos no es necesario cumplir para utilizar correctamente el coeficiente de correlación de Pearson?:
a) Homocedasticidad.
b) Tamaño muestral grande.
c) Normalidad de ambas variables.
d) Relación lineal entre las dos variables.

Ver respuestas

BIBLIOGRAFÍA

De Maris A, Selman SH. Converting Data into Evidence. New York: Springer New York; 2013.

Everitt BS. Medical Statistics from A to Z: A Guide for Clinicians and Medical Students. (3ª ed.). Cambridge University Press; 2021.

Hossain M. Making Sense of Medical Statistics: A Bite Sized Visual Guide. (1ª ed.). Cambridge University Press; 2021.

Maxwell AE. Basic Statistics. Dordrecht: Springer Netherlands; 1978.

Michael Harris, GT. Medical Statistics Made Easy. Scion Publishing; 2020.

Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Correlación. Modelos de regresión. Evid Pediatr. 2021;17:25.

Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Regresión lineal simple. Evid Pediatr. 2021;17:46.

Smith DJ, Samuel S. Basic Statistical Techniques for Medical and Other Professionals: A Course in Statistics to Assist in Interpreting Numerical Data. (1ª ed.). New York: Productivity Press; 2021.