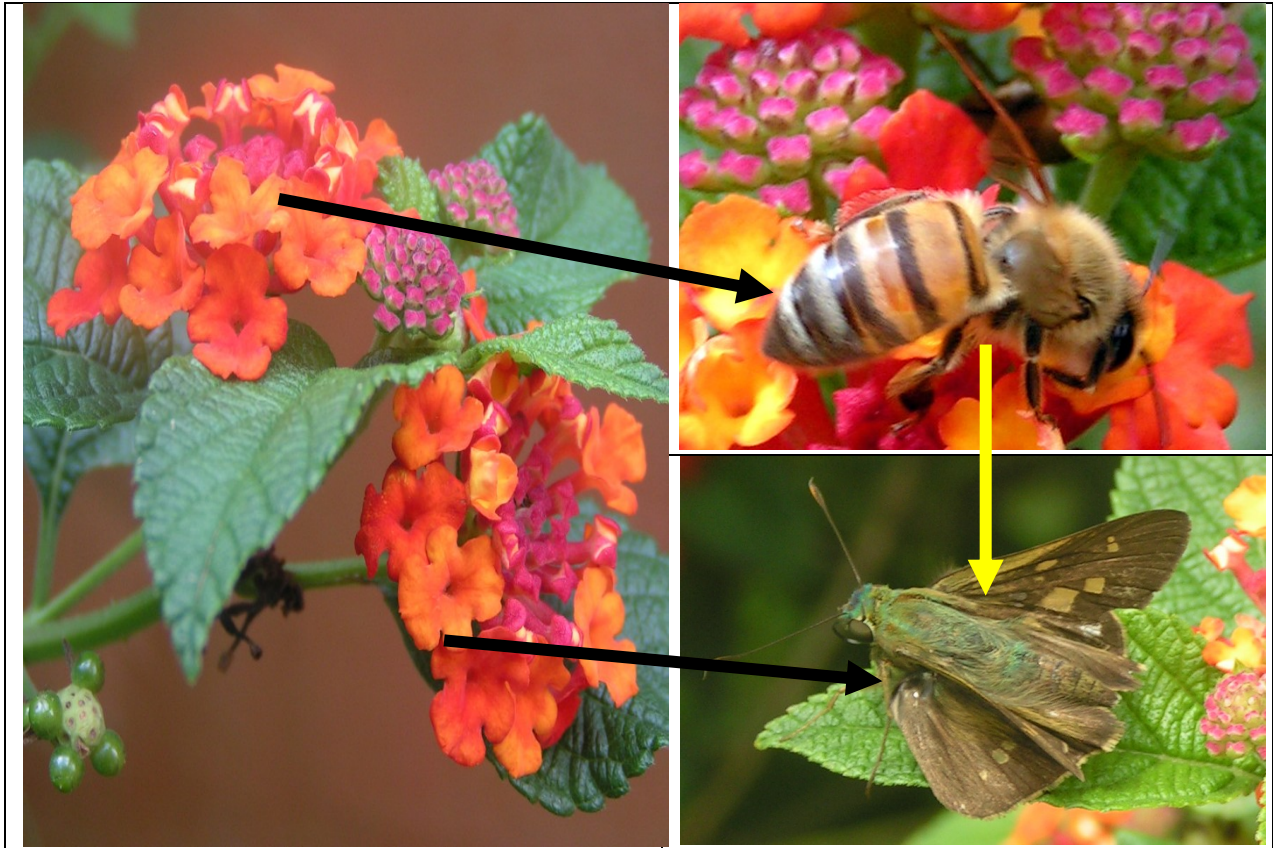


CORRELACIÓN LINEAL

Midiendo la relación entre dos variables



Correlación no es sinónimo de causa-efecto

JORGE FALLAS

2012

Índice

1. Introducción	1
2. Diagrama de dispersión	1
3. Coeficiente de correlación lineal de Pearson	2
6. Interpretación de r	6
4. Correlación no es sinónimo de causa-efecto	8
5. ¿De qué depende la intensidad de la correlación?.....	9
7. Significancia de r : Prueba de hipótesis	11
8. Coeficiente de correlación de órdenes de Spearman.....	16
8.1. Prueba de hipótesis.....	18
9. Coeficiente de correlación de órdenes de Kendall's (Tau-a y Tau-b)	21
9. Bibliografía.....	23
10. Ejercicios.....	25

El presente documento se distribuye bajo licencia “reconocimiento-No comercial-Compartir bajo la misma licencia” (CC BY-NC-SA de “Creative Commons); la cual permite entremezclar, ajustar y construir con base en su trabajo para fines no comerciales, siempre y cuando se de crédito y licencia de sus nuevas creaciones, en términos idénticos.

1. Introducción

En los capítulos anteriores hemos trabajado con variables tales como altura (h), diámetro (d), precipitación y peso en forma independiente; sin embargo ¿son en el mundo real realmente independientes dichas variables? Si observamos un bosque, podemos apreciar que usualmente los árboles de mayor altura también tienen el mayor diámetro y viceversa. Sin embargo, es evidente que esta relación no es perfecta, pues el árbol más alto no necesariamente tiene el mayor diámetro, ni tampoco el árbol más pequeño tiene el menor diámetro. Esto implica que las dos variables están correlacionadas y que el grado o intensidad de la correlación no es perfecto. Usted puede pensar en situaciones particulares de su quehacer profesional donde se presentan situaciones similares. Esto nos lleva a plantearnos las siguientes preguntas: dadas dos variables ¿Existe algún tipo de correlación entre ellas? y en caso de existir, ¿Cuál es la dirección e intensidad de la misma?

En el presente capítulo aprenderemos a cuantificar la intensidad y dirección de la correlación utilizando los siguientes coeficientes de correlación lineal:

R de Pearson	Variables cuantitativas
Rho de Spearman	Variables ordinales
Tau de Kendall	Variables ordinales

2. Diagrama de dispersión

El diagrama de dispersión expresa gráficamente la relación entre dos variables cuantitativas utilizando un sistema de ejes cartesianos. Su uso, como medio para representar gráficamente una distribución bivariable, se atribuye a Sir Francis Galton. Por ejemplo, si estamos estudiando la relación entre el diámetro y altura de un bosque, cada par de valores se gráfica en un sistema de coordenadas X, Y. La figura 1 ilustra este concepto para 46 valores de diámetro (cm) y altura total (m) de Jaúl. La distribución conjunta de diámetro y altura se denomina bivariable porque está formada por la interacción tanto del diámetro como de la altura. Esta gráfica nos permite hacer las siguientes observaciones:

1. Existe un alto grado de correlación lineal entre las dos variables. La naturaleza de la asociación puede describirse adecuadamente por una recta como la trazada en la figura 1.
2. La relación es imperfecta; ya que no siempre los árboles de mayor diámetro corresponden a árboles de mayor altura. Existe una covariación entre los dos sets de datos; sin embargo no es perfecta.

El análisis anterior permite apreciar la importancia de graficar y determinar, en forma preliminar, la presencia o ausencia de correlación entre las variables en estudio. El próximo paso es cuantificar el grado y dirección de la correlación. En 1896, Karl Pearson, colega de Galton, publicó un artículo titulado "*Contribuciones matemáticas a la teoría de la evolución, III. Regresión, herencia y panmixia*"; en el cual analizó las características del coeficiente de correlación. El índice, aun cuando no fue desarrollado por Pearson, lleva su nombre debido a esta publicación. El coeficiente de

correlación de Pearson es apropiado únicamente para aquellos casos en que la asociación es lineal y la escala de medición cuantitativa (intervalo o razón).

Con frecuencia, es necesario determinar la ausencia o presencia de correlación entre observaciones a un nivel de medición nominal u ordinal; por ejemplo la apariencia de un producto y su grado de aceptación en el mercado o el color de las hojas y el grado de enraizamiento en estacas. Para estos casos debemos utilizar índices de asociación no paramétricos tales como Tau de Kendall, Rho de Spearman o Chi-cuadrado (X^2) y sus variaciones.

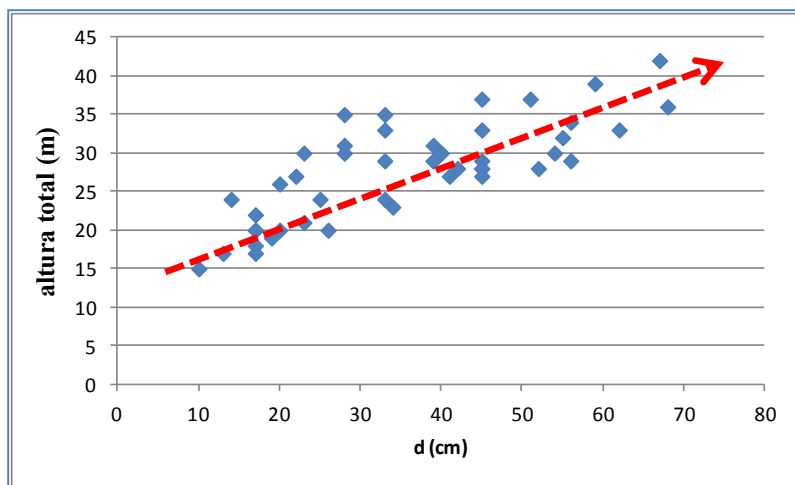


Figura 1: Diagrama de dispersión para diámetro (cm) y altura total (m) de jaúl. La recta indica la tendencia general de los datos.

3. Coeficiente de correlación lineal de Pearson

El coeficiente de correlación lineal de Pearson se define en términos de la covarianza de las variables aleatorias X y Y. La covarianza es una medida que indica la forma en que X y Y varían conjuntamente. Dadas dos variables aleatorias X y Y, con medias μ_x y μ_y y varianzas σ_x^2 y σ_y^2 , respectivamente; la correlación entre X y Y está dada por:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

El estimador de ρ es “r” y su fórmula es:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2)$$

El estadístico “r” posee las siguientes características (ver Fig. 2):

1. El coeficiente mide la intensidad y dirección de la correlación lineal y no excluye la posibilidad de que exista otra forma de correlación no lineal.

2. Su valor fluctúa entre 1 y -1. Cuando la asociación entre X y Y es perfecta “r” es igual a 1 (todos los valores se ubican en una recta); cuando no existe correlación lineal su valor es 0.
3. El signo del coeficiente de correlación indica la dirección de la asociación. Cuando el valor de “r” es positivo se dice que existe una correlación positiva o directa; o sea que valores grandes de X están asociados con valores grandes de Y y viceversa. Por otra parte, si el signo es negativo la correlación es negativa; valores grandes de X se asocian con valores pequeños de Y y viceversa.

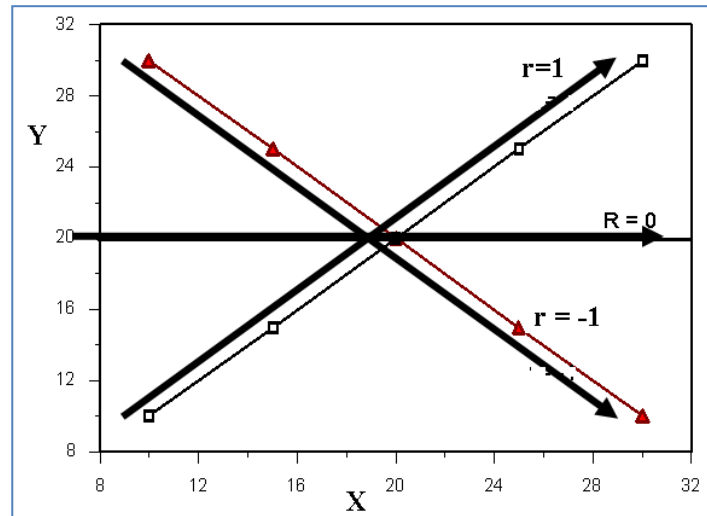


Figura 2: Relación entre el valor y signo del coeficiente de correlación lineal de Pearson y la tendencia de valores.

La formula 2 facilita el cálculo de “r” sin embargo no nos permite apreciar como los valores estandarizados de X y Y afectan su valor. Una expresión más simple y que nos permite apreciar lo anterior es:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \quad (3)$$

Donde:

$\frac{X_i - \bar{X}}{s_X}$ son valores estandarizados de X
 $\left(\frac{Y_i - \bar{Y}}{s_Y} \right)$ son valores estandarizados de Y

$\bar{X}$ la media de X

$\bar{Y}$ la media de Y

s_X la desviación estándar de X

s_Y la desviación estándar de Y

La suma de productos determina la magnitud y signo del coeficiente. La figura 3 muestra el diagrama de dispersión para cuatro observaciones de diámetro y altura total. En los ejes X y Y se indican los valores originales así como las desviaciones con respecto a la media de cada observación. El diagrama se ha dividido en cuatro cuadrantes: I, II, III, y IV. Las líneas que dividen los cuadrantes se originan en la media de cada una de las variables. Para cualquier punto, el producto $X*Y$ será positivo cuando las desviaciones de la media sean ambos positivos o ambos negativos; por otra parte el producto será negativo cuando cualquiera de las desviaciones tenga un signo opuesto. Los productos $X*Y$ serán positivos en los cuadrantes I y III y negativos en los cuadrantes II y IV.

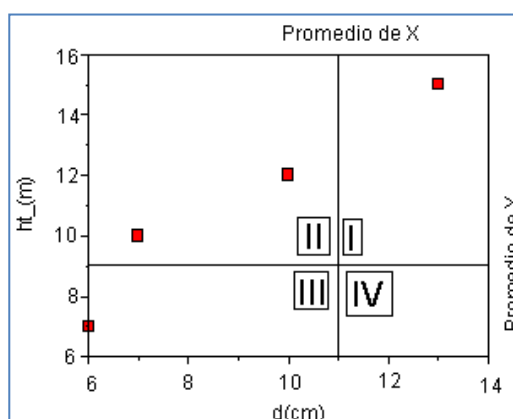


Figura 3: Diagrama de dispersión para cuatro pares de valores de diámetro (cm) y altura total (m). Las letras romanas indican los cuadrantes.

El signo de “r” está dado por el signo de los valores estandarizados en la formula tres; ya que n , S_x y S_y son positivos. Además cuanto mayor sea el número de casos que se ubiquen en los cuadrantes I y III o II y IV mayor será la magnitud de la suma de productos XY y por ende de r . Para los puntos de la figura tres, el valor del “r” es 0,97.

A continuación se ilustra el uso de la fórmula tres. El número de observaciones es muy reducido para un uso adecuado de “r”; sin embargo la simplicidad de la serie estadística permite ilustrar el cálculo de los diferentes componentes de la ecuación.

d (cm)	Alt. tot (m)
6	7
7	10
10	12
13	15

Sustituyendo los valores respectivos en la ecuación tres tenemos:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

Cuadro 1: Cálculo del coeficiente de correlación lineal de Pearson.

Estadísticos (1)	d (cm)(X) (2)	Alt. Tot. (m) (Y) (3)	X estandarizado (4)	Y estandarizado (5)	Producto (4)*(5)
	6	7	-0,9487	-1,1882	1,1272
	7	10	-0,6324	-0,2970	0,1879
	10	12	0,3162	0,2970	0,0939
	13	15	1,2649	1,1882	1,5029
Media	9	11		Suma	2,9119
S	3,162	3,367		n-1	3
Covarianza población	7,75			1/n-1	0,333
Covarianza muestra	10,33			R	0.97

Nota: Excel calcula la covarianza de la población y por lo tanto utiliza n en lugar de n-1.

Cuando se conoce la desviación estándar de X (S_x), de Y (S_y) y la covarianza entre X y Y (Cov_{xy}), el coeficiente de correlación puede calcularse utilizando la siguiente fórmula:

$$r = Syx / Sx * Sy \quad (4)$$

Para los datos del presente ejemplo $S_x=3,16$ cm; $S_y=3,37$ m y $S_{yx}=10,33$. Sustituyendo los valores en 4 tenemos que r es igual a:

$$r = 10,33 / (3,16 * 3,37) = \mathbf{0,97}$$

A continuación se muestran los cálculos realizados con XLStatics:

Two-variable		Single-variable				Correlation Coeff	
Number	4	Alt. tot (m)		d (cm)		Correlation	0.970648
Covariance	10.333	Mean	11	Mean	9		
Correlation	0.9706	St Dev	3.3665	St Dev	3.1623		
R ²	0.9422	Min	7	Min	6		
		Max	15	Max	13		
		Median	11	Median	8.5		

El coeficiente de correlación obtenido para una muestra (r) es un estimador puntual que mide la dirección e intensidad de la covarianza de los datos analizados y no implica necesariamente que dicho valor sea la correlación de la población (ρ). Por ejemplo, si para cuatro árboles no se obtiene una correlación entre el diámetro y la altura, no se puede afirmar que tal correlación no existe. Como con cualquier otro estimador, es esencial que el valor de "r" se interprete considerando las circunstancias y condiciones bajo las cuales fue calculado.

6. Interpretación de r

Al observar el valor del coeficiente de correlación podemos preguntarnos, por ejemplo, ¿qué nos indica un $r = 0,5$? ó ¿Cuánto más intensa es la correlación cuando $r = 0,8$? El coeficiente de correlación puede interpretarse tomando como referencia el intervalo $[-1 \leq \rho \leq 1]$; así como el concepto de variabilidad. A continuación se discute cada enfoque.

6.1. Valor de r con respecto a 0, 1 y -1

Esta es una de las formas más simples de interpretar el valor de “r”. Valores cercanos a cero indican poca o ninguna de correlación lineal, en tanto que valores cercanos a uno indican un alto grado de correlación lineal. Muy bien, pero ¿qué tan cerca? y además, ¿cómo interpretar valores intermedios? La respuesta no es simple, porque desafortunadamente el valor de “r” no puede interpretarse como directamente proporcional a su magnitud o tamaño. Un valor de $r = 0,50$ indica menos que la mitad de la intensidad de asociación con respecto a un valor de $r = 1$. Desde esta perspectiva un aumento en r de 0,15 tiene un mayor impacto en la intensidad de la asociación cuanto mayor sea el valor de “r”; ya que la ganancia marginal en precisión al estimar un valor dado de “Y” será mayor. Esto se apreciará mejor en cuando se estudie el tema de regresión.

EL cuadro 2 y la figura 4 muestran el error de predicción en porcentaje para diferentes valores de “r”. Puede observarse claramente que la relación entre “r” y el error de predicción (S_{yx}) es no lineal. Una reducción en el error de predicción de 50% se obtiene cuando el valor de “r” es 0,87 ($\Delta r = 1 - 0,87 = 0,13$). El siguiente incremento de 0,75 a 0,87 ($\Delta r = 0,12$) sólo proporciona un 16% adicional en la reducción del error de predicción. Por otra parte si observamos la reducción en el error de predicción para valores de “r” cercanos a cero podemos apreciar que la ganancia es aún menor. Un incremento de “r” de 0 a 0,10 resulta en una reducción en el error de predicción de tan solo 1%.

Cuadro 2: Relación entre el coeficiente de correlación lineal y el error de predicción*

coeficiente de correlación %	Error de predicción %
100	0
87	50
75	66
50	87
40	92
25	97
10	99
0	100

* $S_{yx} = S_y (1 - r^2)^{1/2}$

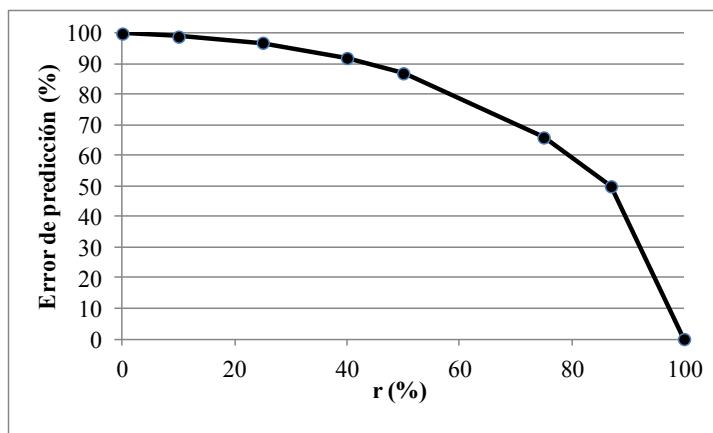


Figura 4: Relación entre el coeficiente de correlación y el error de predicción.

6.2. Proporción de la variabilidad explicada

Supongamos que para una especie determinada la relación peso (gr) y tamaño de la huella tiene un coeficiente de correlación de 0,80. Esto indica que parte de la variación en peso de los animales está asociada con la variación en tamaño de la huella y que además la relación es positiva. Sin embargo la covariación o correlación no es perfecta; ya que la variabilidad en peso no puede ser explicada totalmente por la variabilidad en tamaño de la huella. Alguna proporción de la variación en peso está asociada a otros factores tales como estado fisiológico del animal, sexo o edad.

El coeficiente de correlación nos permite determinar la proporción de la variabilidad total en peso asociada con la variación en tamaño de la huella. La proporción no asociada con la variabilidad en tamaño de la huella se denomina variabilidad no explicada o error. El error representa la variabilidad asociada a otras variables explicativas no consideradas en el análisis de correlación simple. Cuando dos o más variables se asocian con una variable dependiente la correlación se denomina múltiple.

Para proseguir con nuestro análisis la variabilidad, la variabilidad en peso debe expresarse en términos de la varianza (S^2). Si los valores difieren drásticamente entre sí (datos muy heterogéneos) el valor de S^2 será grande; por otra parte, si son relativamente homogéneos, S^2 será pequeña; y si todos los datos tienen el mismo peso, S^2 será cero (no existe variabilidad).

La variación total puede dividirse en dos componentes: variación asociada al factor tamaño de la huella y variación asociada a otros factores o error (cuadro 3). Para hacer esta división, el valor de “r” debe expresarse en términos de R^2 (coeficiente de determinación). Si el peso de la muestra tiene una varianza de $S^2 = 2,60 \text{ kg}^2$ y el coeficiente de correlación lineal es 0,80; la variación total en peso se puede dividir de la siguiente manera:

Cuadro 3: Partición de la variabilidad total en peso en variabilidad asociada al tamaño de la huella y a otras factores o variabilidad no explicada.

Variación total en peso	=	Variación asociada con tamaño de huella	+	Variación asociada con otros factores
S^2	=	R^2	+	$1 - R^2$
2,6 Kgr ²	=	1,664 Kgr ²	+	0,936 Kgr ²
100 %	=	64%	+	36%

Para nuestro ejemplo, el 64 % de la variabilidad en peso está explicada por la variación en la dimensión de las huellas y el remanente 36 % (error) está asociada a otros factores no medidos. Los cálculos anteriores elaborados bajo el supuesto de linealidad o aditividad indican que R^2 expresa la proporción de la variabilidad en Y asociada con la variabilidad en X; o sea la covarianza al cuadrado como se muestra a continuación.

$$R^2 = (S_{yx})^2 / S^2_x * S^2_y \quad (5)$$

Donde, S_x es la desviación estándar de X, (S_y) es la desviación estándar de Y y Cov_{xy} la covarianza entre X y Y

El concepto de variabilidad explicada permite entender mejor porque un coeficiente de correlación de 0,50 indica menos que la mitad de la intensidad de la correlación mostrada por una asociación perfecta. Para explicar el 50 % de la variabilidad se requiere un coeficiente de correlación de 0,71 ($0,71 * 0,741 = 0,50$).

4. Correlación no es sinónimo de causa-efecto

Al evaluar la correlación entre dos variables no se debe incurrir en el error de utilizar el valor de “r” como argumento para indicar que existe una relación de causa-efecto. Es sumamente importante no confundir ambos términos. No siempre un alto grado de correlación entre dos variables implica necesariamente que exista una relación de causa-efecto. Si deseamos argumentar sobre una relación de causa-efecto, debemos considerar elementos biológicos, físicos, químicos, económicos o de otra índole y no solo el valor del coeficiente de correlación.

Dos variables pueden presentar un alto grado de correlación por diferentes motivos como se ilustra en la figura 5. Por ejemplo, la variación en X puede ser la responsable de la variación en Y o viceversa. También pueden existir otros factores que afecten tanto a X como a Y (ver efecto de confusión). Por ejemplo, la respuesta de una planta a un determinado fertilizante puede estar afectada por el pH, la fertilidad inicial del suelo y época de aplicación del fertilizante. En otras ocasiones dos cosas o eventos ocurren al mismo tiempo pero sin embargo no están correlacionados y mucho menos uno es causa del otro.

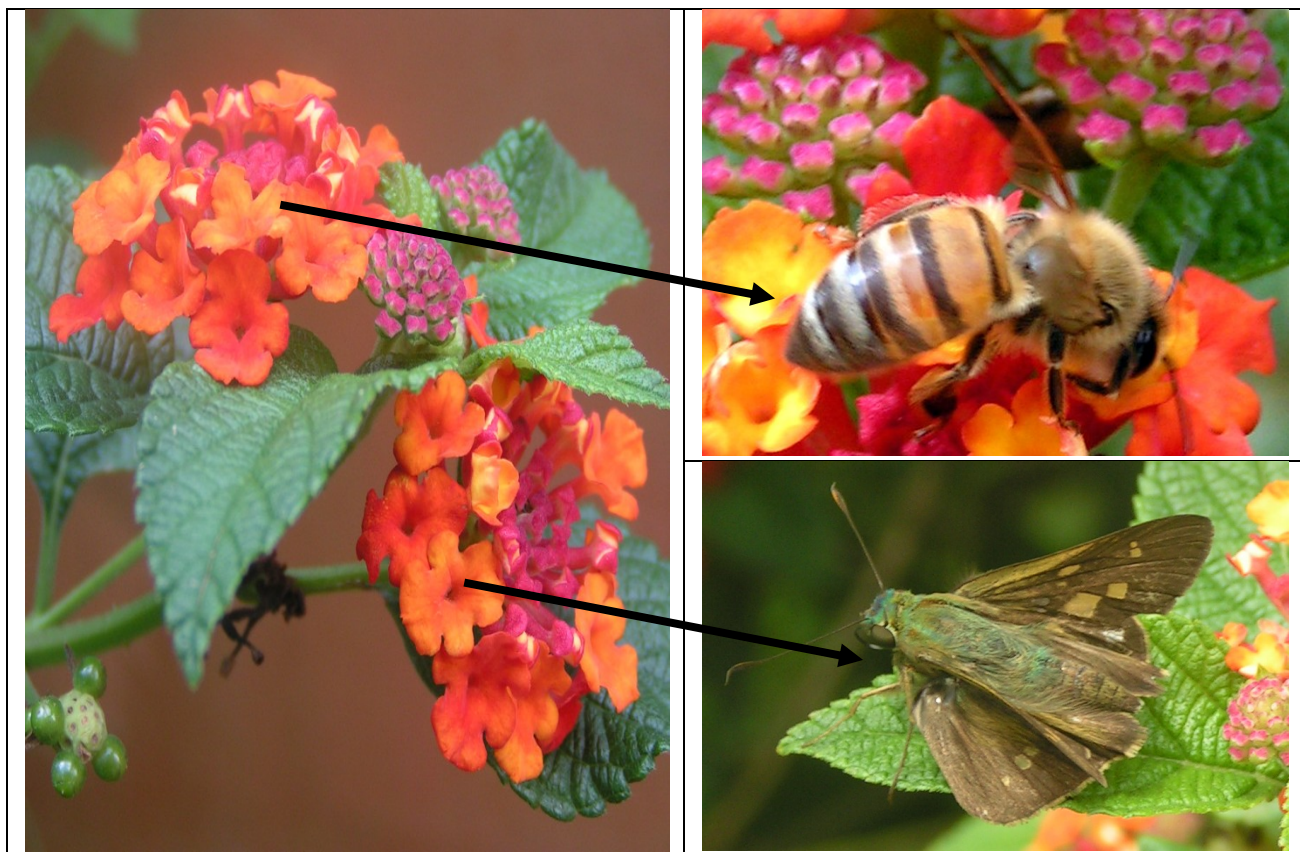


Figura 5: La presencia de la mariposa y la abeja está correlacionada con la presencia de la planta (existe una relación de causa efecto). Sin embargo, la presencia de la abeja, aunque está correlacionada con la presencia de la mariposa y viceversa, no implica una relación de causa-efecto. En este caso la presencia de la planta es lo que explica la presencia de la abeja y la mariposa.

5. ¿De qué depende la intensidad de la correlación?

El grado o intensidad de la correlación es afectada por los siguientes factores:

1. Principio de linealidad

La figura 6 muestra tres diagramas de dispersión con sus respectivas curvas de mejor ajuste. En el diagrama A el grado de correlación es moderado, en tanto que en el B es alto. La figura indica que cuanto mayor sea el número de observaciones que se ubiquen alrededor de la línea de mejor ajuste mayor será el valor de “r”.

Por otra parte cuando una recta no describe adecuadamente la relación entre X y Y, como es el caso de la figura 6C, se dice que las observaciones no satisfacen el principio de linealidad de regresión. Esto no implica que los datos no estén correlacionados, simplemente significa que la relación es no lineal. La relación del diagrama 6C es curvilínea y si calculamos el valor de “r” encontraríamos que es muy pequeño y posiblemente cercano a cero. Esta gráfica ilustra claramente que si la relación es curvilínea, el coeficiente de correlación de Pearson subestimarán en gran medida la intensidad de la asociación. El coeficiente de correlación de Pearson solo debe usarse con distribuciones bivariadas que no se aparten sensiblemente del principio de linealidad.

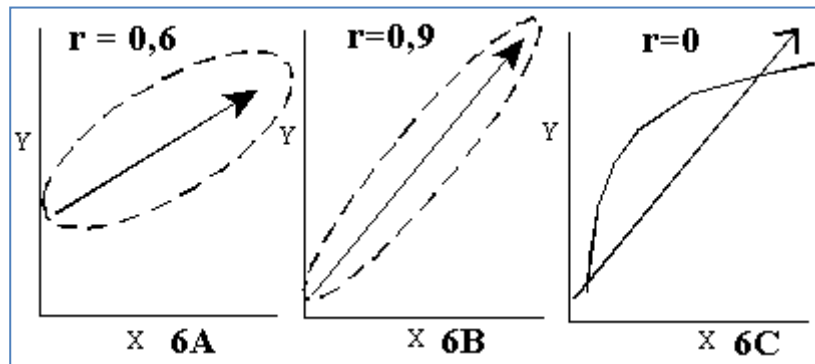


Figura 6: Diagramas de dispersión mostrando la tendencia de las observaciones para diferentes intensidades de correlación. Observe que en los datos de la gráfica C no existe correlación lineal.

2. Ámbito de las observaciones

El ámbito o dispersión de los valores de X y Y afectan sustancialmente la intensidad de la correlación como puede apreciarse en la figura 7. La figura de la izquierda contiene la totalidad de las observaciones de diámetro y altura; en tanto que la figura de la derecha sólo contiene aquellas observaciones cuyo diámetro es inferior a 19 cm.

El coeficiente de correlación para la totalidad de las observaciones es 0,73; en tanto que para el subset es 0,13. Esto nos indica que, dada una relación lineal, la magnitud del coeficiente de correlación depende del grado de variabilidad de los datos. Si mantenemos constantes los otros factores que afectan el valor de “r”, restringiendo el ámbito de X y Y reduciremos el valor del coeficiente de correlación.

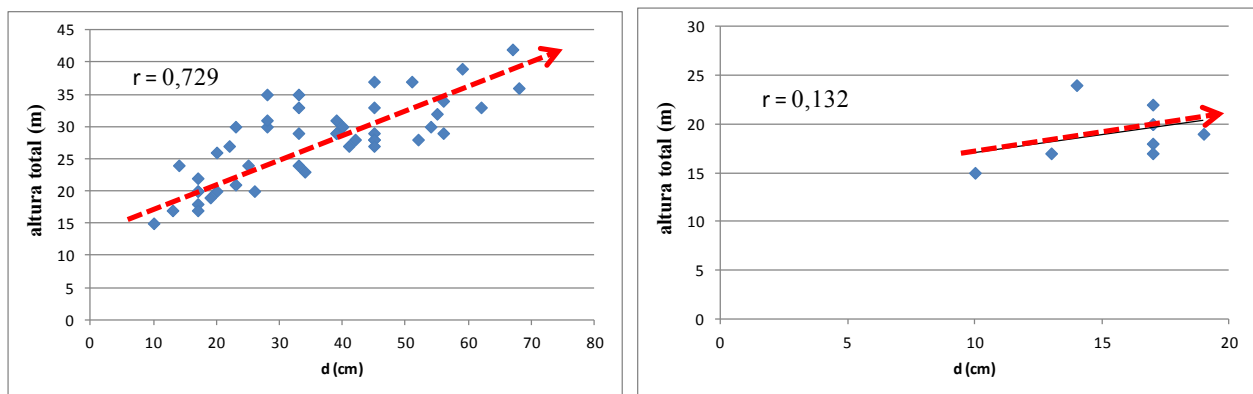


Figura 7: Efecto del ámbito de las observaciones en el valor del coeficiente de correlación lineal de Pearson.

3. Unidad de medición y codificación de los datos

El valor de “r” refleja el grado de correspondencia entre las posiciones relativas de X y Y. Por esta razón, ni la intensidad ni la dirección del coeficiente de correlación son afectadas por la unidad de medición ni por transformaciones que impliquen el uso de una constante y de operaciones algebraicas tales como suma, resta, multiplicación y división.

Las posiciones relativas de X y Y no son afectadas por las transformaciones indicadas previamente y por lo tanto no afectan el valor de “r”. Por ejemplo, la posición relativa del peso de un animal en una serie estadística será la misma indistintamente de si su peso se mide en gramos ó kilogramos.

4. Variación muestral

Al igual que la media o la desviación estándar, el coeficiente de correlación se obtiene a partir de una muestra y por lo tanto está sujeto a las leyes del azar; las cuales se expresan en variabilidad. El valor de “r” para una muestra es diferente al valor de la población (ρ). Si la muestra es representativa y lo suficientemente grande se esperaría que la discrepancia o error sea mínimo. El coeficiente de correlación para una muestra **no es el** coeficiente de correlación poblacional, sino simplemente un valor que responde a las características de la muestra seleccionada.

7. Significancia de r: Prueba de hipótesis

El coeficiente de correlación para la población se designa con la letra griega (ρ) y al igual que otros parámetros estadísticos es una constante desconocida. Cuando estimamos su valor a partir de una muestra podemos hacernos las siguientes preguntas:

1. ¿Es el valor de “r” estadísticamente diferente de cero?; o en otras palabras ¿ existe correlación entre X y Y ?
2. ¿Es el valor de r igual a un valor dado, por ejemplo 0,90?

La prueba de hipótesis permite discernir si el valor de “r” se debe a la correlación entre X y Y o si es el resultado del azar. A continuación se ilustra el uso de la distribución “t” de Estudiante para someter a prueba esta hipótesis.

En la sección tres se calculó un coeficiente de correlación lineal de 0,97 para cuatro valores de diámetro y altura. A simple vista pareciera que la relación es lo suficientemente fuerte y positiva (lo cual tiene sentido biológico) como para suponer que efectivamente existe una relación entre ambas variables; sin embargo ¿es el valor de “r” significativamente diferente de cero? Para responder a esta pregunta debemos realizar una prueba de hipótesis de dos colas.

Dado un conjunto de pares de observaciones no correlacionadas de una distribución normal bivariada, la distribución muestral de “r” sigue una distribución t de Estudiante con n-2 grados de libertad. Esta afirmación es válida incluso si los valores observados no son normales, siempre y cuando el tamaño de muestra no sea muy pequeño. Los percentiles de la distribución “t” de Estudiante con n-2 grados de libertad se utilizan para someter a prueba la hipótesis $H_0: \rho = 0$; o sea X y Y no están linealmente correlacionados. La fórmula del estadístico de prueba es:

$$t = r \sqrt{\frac{n-2}{1-r^2}} \quad (6)$$

Los valores de t calculados se comparan con valores críticos obtenidos de la distribución “t” de Estudiante con n-2 grados de libertad.

Para la determinación de los valores críticos de “r” se utiliza la transformación inversa:

$$r = \frac{t}{\sqrt{n-2+t^2}}. \quad (7)$$

El intervalo de confianza para “ρ” de Pearson se basa en la transformación “r” a “z” de Fisher. Si los pares ordenados (X,Y) tienen una distribución bivariable normal y son independientes, entonces Z es aproximadamente normal con media igual a:

$$z_r = 0.5 \ln \frac{1+r}{1-r} \quad (8)$$

Y su error estándar igual a:

$$\frac{1}{\sqrt{N-3}}, \text{ donde } N \text{ es el tamaño de la muestra.}$$

Una vez determinado el valor de “z”, el intervalo de confianza para ρ es igual a:

$$z \pm (\text{valor de } z \text{ para nivel de confianza deseado}) \times (\text{error estándar})$$

El valor de “z” puede convertirse a “r” mediante la siguiente ecuación:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \quad (9)$$

EJEMPLO

1. Calcular r:

$$r = 0,97$$

2. Hipótesis nula y alternativa

$$H_0: \rho = 0$$

$$H_a: \rho \neq 0$$

3. Elegir el valor de alfa crítico. Recuerde que debe elegir dicho valor antes de realizar la prueba de hipótesis. Para el presente ejercicio se eligió del valor de **0,05** (*recuerde que usted puede elegir cualquier valor inferior o igual a 0,1*)

4. Estadístico de prueba y P calculado

$$t = 0,97 * \left[\frac{4-2}{1 - (0,97*0,97)} \right]^{0,5}$$

$$t = 5,708$$

Si usted busca el valor 5,708 en una tabla “t” de Estudiante con n-2 grados de libertad (en este caso 4-2=2) encontrará que corresponde a una probabilidad de 0,0293 (este es el valor de P calculado por los paquetes estadísticos).

5. Decisión

Ho se rechaza si el valor de P calculado es menor que el alfa elegido (P critico).

En este caso el valor de P calculado (0,0293) es menor que el valor de P critico (0,05) y por tanto se rechaza Ho a un nivel de significancia de 5%.

6. Conclusión

A un nivel de significancia de 5% la muestra provee suficiente evidencia estadística para suponer que las variables provienen de una población con un coeficiente de correlación lineal diferente de 0. Por tanto para la muestra en estudio se concluye que el diámetro y la altura están linealmente correlacionados ($P < 0,05$).

Es posible que la conclusión no le sorprenda ya que biológicamente existe una relación entre el diámetro y altura; además, el valor del coeficiente de correlación es lo suficientemente alto como para suponer que tal correlación existe. Sin embargo debemos estar conscientes de que el tamaño de muestra determina los *grados de libertad* y este a su vez juega un papel importante en la determinación de la significancia de “r” como puede observarse en el siguiente cuadro.

Cuadro 4: Efecto del tamaño de la muestra en la significancia de r.

Pares (X,Y) (n)	g.l. (n-2)	r	Ho: R = 0 Ha: R >0 (2 colas alfa=0,05)
12	10	0,55	no rechazar
18	16	- 0,45	no rechazar
22	20	0,45	rechazar Ho
102	100	0,21	rechazar Ho
500	498	- 0,15	rechazar Ho
1002	1000	0,10	rechazar Ho

Observe como para un “r” de 0,10 con $n=1.002$ se rechaza la hipótesis nula, en tanto que para un “r” de 0,55 con $n=12$ no se rechaza. El factor que controla la decisión de rechazar o no rechazar H_0 es el número de observaciones (grados de libertad). Para muestras muy grandes (e.g. 500 o 1.000 pares de observaciones), coeficientes de correlación relativamente pequeños serán significativos en tanto que con muestras pequeñas sucederá lo opuesto.

Al evaluar los resultados de la prueba de hipótesis debemos diferenciar entre significancia estadística e importancia o aplicación de los resultados y los argumentos lógicos que sustentan la correlación. El no rechazar la hipótesis nula no significa que no exista una correlación entre las variables en estudio; así como el rechazarla tampoco nos asegura que exista tal relación. Los ejemplos del cuadro cuatro ilustran este punto. La conclusión final debe basarse tanto en el análisis de la evidencia estadística como en el conocimiento del tema en estudio.

El intervalo de confianza para ρ es igual a:

1. Convertir r a z

$$z = 0,5 \ln [(1+0,97)/(1-0,9)] = 2.09229572$$

2. Calcular error estándar

$$EE = 1/(4-2)^{0,5} = 0,7071$$

3. Calcular IC para ρ

Para un nivel de confianza de 95% el valor de Z es igual a 1,96.

Límite superior: $z = 2.09229572 + 1,96 * 0,7071 = 3,4782$

Límite inferior: $z = 2.09229572 - 1,96 * 0,7071 = 0,706379$

4. Convertir valores de z a r :

$$\text{Límite superior: } r = e^{2*3,4782} - 1 / e^{2*3,4782} + 1 = 0,998$$

$$\text{Límite inferior: } r = e^{2*0,706379} - 1 / e^{2*0,706379} + 1 = 0,608$$

Y el intervalo de confianza es: $0,608 < \rho < 0,998$. Dado el set de datos analizado, tenemos una confianza de 95% de que el parámetro de la población se encuentre entre 0,608 y 0,998.

EJEMPLO 2

¿Cuál es el coeficiente de correlación para los datos de diámetro y altura del archivo d_h_jaul.xlsx? Es el coeficiente diferente de cero a un nivel de confianza del 99%? El procedimiento a seguir es:

1. Diagrama de dispersión
2. Calcular r
3. Plantear hipótesis nula y alternativa
4. Elegir el valor de alfa crítico (“ p ” crítico)

5. Estadístico de prueba y “p” calculado
6. Decisión
7. Conclusión

A continuación se muestra cómo utilizar XLStaticis para realizar la prueba de hipótesis.

1. Diagrama de dispersión

El grafico muestra que existe una correlación lineal entre las variables diámetro (cm) y altura total (m).

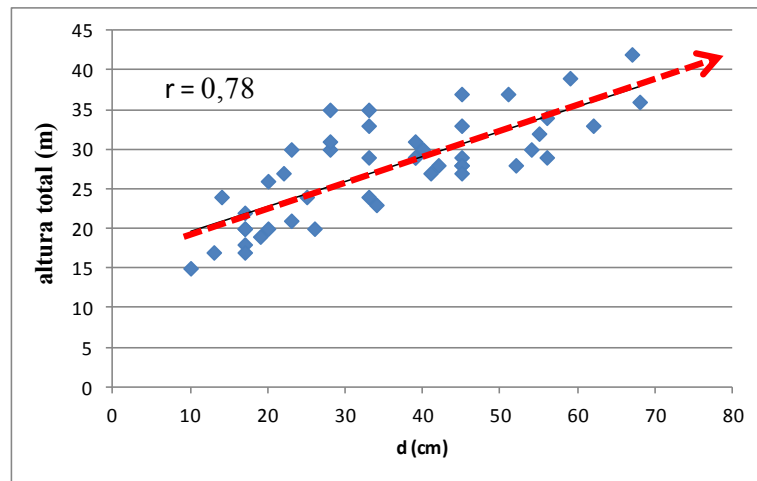


Figura 8: Diagrama de dispersión para diámetro (cm) y altura total (m) de Jaúl.

2. Calcular r

Two-variable		Single-variable			
		altura total (m)		diámetro (cm)	
Number	46	Number	46	Number	46
Covariance	77.752	Mean	27.739	Mean	35.652
Correlation	0.7818	St Dev	6.3541	St Dev	15.651
R ²	0.6113	Min	15	Min	10
		Max	42	Max	68
		Median	28.5	Median	33.5
Correlation Coeff					
Correlation	0.781847				

El coeficiente de correlación lineal de Pearson para las variables diámetro (cm) y altura total (m) es 0,78.

3. Plantear las hipótesis nula y alternativa

$H_0: \rho = 0$

$H_a: \rho \neq 0$

3. Elegir el valor de alfa crítico

El alfa crítico elegido es 0,01.

4. Estadístico de prueba y “p” calculado

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

Hypothesis Tests for ρ	
$H_0: \rho =$	0
Alternative	
☒	$\neq$
☐	$>$
☐	$<$
$H_1: \rho \neq$	0
p-value =	1.41E-10

Observe que se realiza una prueba de dos colas.

El valor de p calculado es 1,41 E-10 o sea 0,00000000014.

5. Decisión

Dado que el valor de “p” calculado es menor que “P” crítico se rechaza H_0 .

6. Conclusión

El valor de “p” calculado (1,41 E-10) es inferior al valor del “p” crítico (0.01) y por lo tanto se concluye que las variables están linealmente correlacionadas.

8. Coeficiente de correlación de órdenes de Spearman

El coeficiente de correlación de Spearman (Rho) se utiliza para cuantificar la intensidad y dirección de la correlación cuando las variables se miden o se transforman a un nivel de medición ordinal. Este es un estadístico no paramétrico, ya que su distribución muestral exacta se puede obtener sin conocer los parámetros de la distribución de probabilidad conjunta de X y Y (como sí es un requisito con “r” de Pearson). Su valor, al igual que “r” de Pearson se encuentra en el ámbito $[-1 \leq \text{Rho} \leq 1]$. La interpretación del valor de Rho de Spearman sigue la lógica expuesta para la “r” de Pearson.

Dados “n” pares de observaciones (X1,Y1), (X2,Y2),, (Xn,Yn) para las variables aleatorias X y Y; para estimar la correlación entre X y Y primero se ordenan los pares de observaciones de mayor a menor (utilizando X o Y) y luego se les asigna el orden respectivo a cada par X,Y; de esta manera generamos un conjunto de “n” pares de órdenes, los cuales denotamos como (Rx1,Ry1), (Rx2,Ry2),..., (Rxn, Ryn).

Si dos observaciones tienen el mismo orden (empates) se les asigna el orden medio (e.g. si las observaciones correspondientes a los ordenes 8 y 9 son iguales, se les asigna el orden 8,5). La correlación entre X y Y se obtiene calculando el coeficiente de correlación de Pearson para los pares de ordenes. La expresión matemática básica para Rho es:

$$R_b = \frac{n E R_x * R_y - E R_x * E R_y}{n E R_x^2 - (E R_x)^2 * n E R_y^2 - (E R_y)^2} \quad (10)$$

En ausencia de empates, las siguientes formulas facilitan el cálculo de Rb:

$$R_a = \frac{12 \sum R_x * R_y}{n(n^2 - 1)} - 3 \frac{n+1}{n-1} \quad (11)$$

$$R_s = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)} \quad (12)$$

En donde, n es el número de pares y D_i es la diferencia entre R_{xi} y R_{yi} .

Cuando se presentan empates el numerador de la ecuación diez se debe modificar de la siguiente manera:

$$R_s = 1 - \frac{6(\sum D_i^2 + U + V)}{n(n^2 - 1)} \quad (13)$$

$$U = \frac{1}{12} \sum m_j(m_j^2 - 1) \quad (14)$$

$$V = \frac{1}{12} \sum n_j(n_j^2 - 1) \quad (15)$$

En donde m_j es el número de empates en X y n_j es el número de empates en Y.

La expresión trece sólo considera el efecto de empates en $\sum D_i^2$. El denominador de la ecuación once también puede modificarse para compensar por el efecto de los empates, esto se logra sustituyendo el término $n^2 - 1$ por la media geométrica de U y V como se muestra a continuación:

$$R_s = \frac{n(n^2 - 1) - 6(\sum D_i^2 + U + V)}{\{n(n^2 - 1) - 12U\} * \{n(n^2 - 1) - 12V\}^{0.5}} \quad (16)$$

El coeficiente Rb es un estimador apropiado para evaluar la correlación entre X y Y cuando la presencia de empates expresan concordancia (i.e. opiniones de dos personas sobre un determinado número de productos o situaciones). Por otro lado, R_a es apropiado para medir correlación cuando una variable representa el ordenamiento de ítems por un individuo y la otra representa un orden objetivo y conocido. Finalmente, cuando una variable representa tiempo, R_a puede utilizarse como una medida de tendencia de la serie.

Para observaciones a un nivel de medición de intervalo o razón el coeficiente de correlación de Spearman brindará resultados algo diferentes comparados con el coeficiente de correlación de Pearson. Las diferencias se reducen conforme aumenta el tamaño de la muestra.

8.1. Prueba de hipótesis

El estadístico “t” puede utilizarse para someter a prueba la hipótesis $H_0: \rho=0$. Dado que H_0 sea verdadera, dicho estadístico sigue aproximadamente la distribución t de Estudiante con $n-2$ grados de libertad.

$$t = r \sqrt{\frac{n-2}{1-r^2}}, \text{ donde “r” corresponde al valor de } \rho. \quad (17)$$

Si se desea calcular un intervalo de confianza para ρ o someter a prueba una hipótesis para un valor de ρ diferente de cero, se puede utilizar la transformación de Fisher y la distribución Z:

$$F(r) = \frac{1}{2} \ln \frac{1+r}{1-r} = \text{arctanh}(r). \quad (18)$$

Donde “ln” es el logaritmo natural y “arctanh” es el inverso de la función hiperbólica.

$$z = \sqrt{\frac{n-3}{1.06}} F(r) \quad (19)$$

A continuación se ilustra el uso de XLStatistics para calcular el valor de ρ para los datos de diámetro (cm) y altura de Jaúl (m) analizados en la sección anterior.

EJEMPLO

Se desea determinar el coeficiente de correlación de Spearman para el set de datos de diámetro (cm) y altura de Jaúl (m) analizado previamente. El procedimiento de cálculo es el mismo sugerido para el coeficiente de correlación de Pearson.

1. Diagrama de dispersión

A continuación se muestran los valores originales y el respectivo orden para cada uno de los pares de observaciones.

Cuadro 5: Valores originales de diámetro (cm) y altura total (m) y el respectivo orden para cada uno de los pares de observaciones.

Diámetro (cm)	Altura total (m)	Orden diámetro *	Orden altura total *	Orden Diámetro**	Orden Altura total**
40	30	19	15	16.5	19
10	15	46	46	46	46
33	33	24	9	10	25.5
45	33	11	9	10	13.5
33	29	24	19	21	25.5
41	27	18	28	29	18
13	17	45	44	44.5	45

diámetro (cm)	altura total (m)	orden diámetro	orden altura total	Diámetro*	Altura total**
45	29	11	19	21	13.5
39	29	20	19	21	21
33	35	24	6	6.5	25.5
17	20	39	38	39.5	41
17	22	39	36	36	41
28	30	28	15	16.5	29
26	20	31	38	39.5	31
22	27	35	28	29	35
17	18	39	43	43	41
17	20	39	38	39.5	41
17	17	39	44	44.5	41
39	31	20	13	13.5	21
42	28	17	24	25.5	17
56	29	5	19	21	5.5
54	30	8	15	16.5	8
62	33	3	9	10	3
45	28	11	24	25.5	13.5
33	24	24	32	33	25.5
34	23	23	35	35	23
25	24	32	32	33	32
67	42	2	1	1	2
68	36	1	5	5	1
55	32	7	12	12	7
45	28	11	24	25.5	13.5
51	37	10	3	3.5	10
45	37	11	3	3.5	13.5
20	20	36	38	39.5	36.5
23	30	33	15	16.5	33.5
14	24	44	32	33	44
23	21	33	37	37	33.5
20	26	36	31	31	36.5
28	35	28	6	6.5	29
19	19	38	42	42	38
39	29	20	19	21	21
45	27	11	28	29	13.5
28	31	28	13	13.5	29
56	34	5	8	8	5.5
59	39	4	2	2	4
52	28	9	24	25.5	9

Nota: la función jerarquía de Excel permite calcular del orden de cada observación.

* No considera empates. ** Considera empates.

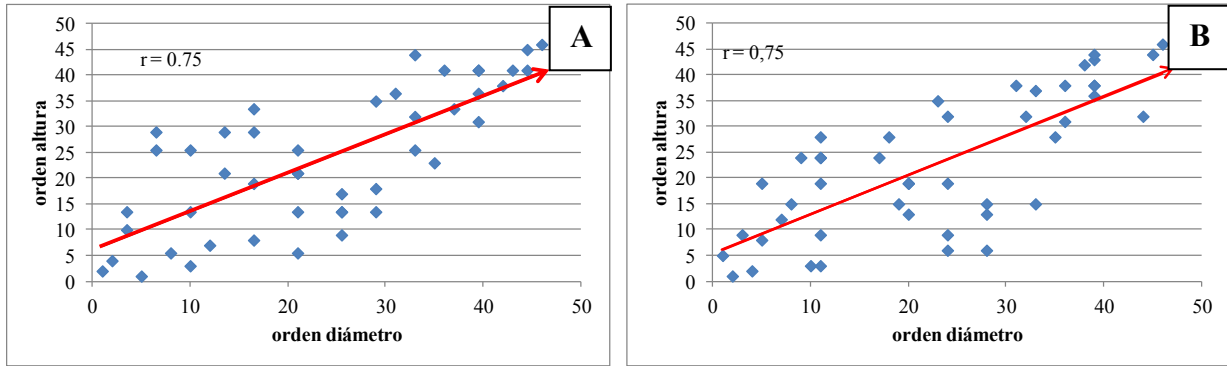


Figura 9: Diagrama de dispersión para órdenes de diámetro (cm) y altura total (m) de Jaúl.

A. Considera empates. B. No considera empates.

2. Calcular Rho

Spearman's Rank Coefficient (rho)

Correlation Coeff	
ρ	0.74987615

El coeficiente de correlación lineal de Spearman para los órdenes de diámetro (cm) y altura total (m) es 0,75. Note su similitud con el valor del coeficiente de correlación de Pearson. El coeficiente de determinación para el coeficiente de correlación de Spearman puede aproximarse como R_b^2 . En este caso R_b^2 es igual a 0,56. Los resultados indican que aproximadamente el 56 de la variación en altura puede atribuirse a la variación en diámetro.

3. Plantear las hipótesis nula y alternativa

$H_0: \rho = 0$

$H_a: \rho \neq 0$

3. Elegir el valor de alfa crítico

El alfa crítico elegido es 0,01.

4. Estadístico de prueba y “p” calculado

Large sample Hypothesis Test for ρ	
$H_0: \rho = 0$	
Alternative	
☒ $\neq$	☐ $>$ ☐ $<$
$H_1: \rho \neq 0$	
T	7.51856
DF	44
p-value	2E-09

El valor de p calculado es 2 E-10 o sea 0,0000000002.

5. Decisión

Dado que el valor de “p” calculado es menor que “P” crítico se rechaza H_0 .

6. Conclusión

El valor de “p” calculado ($1,41 \times 10^{-10}$) es inferior al valor del “p” crítico (0.01) y por lo tanto se concluye que las variables están linealmente correlacionadas.

9. Coeficiente de correlación de órdenes de Kendall's (Tau-a y Tau-b)

El coeficiente de correlación de rangos de Kendall, conocido como el coeficiente tau de Kendall (τ), es otro estadístico no paramétrico utilizado para medir la asociación entre los órdenes de dos variables. Este coeficiente es particularmente útil para medir la concordancia (o ausencia de ella) al evaluarse un set idéntico de condiciones por dos personas o grupos. Por ejemplo, dados cuatro métodos de impartir una clase (M1, M2, M3, M4) cómo serían ordenados por dos alumnos en una escala de uno a cuatro, donde uno es mejor y cuatro peor.

La fórmula de tau de Kendall es:

$$\tau = (\text{número de pares concordantes} - \text{número de pares no concordantes}) / 0,5n * (n-1) \quad (20)$$

A la formula veinte se le conoce como Tau-a y no hace ningún ajuste por empates entre ordenes. El denominador corresponde al número total de pares y por tanto el valor del coeficiente se encuentra en el ámbito $-1 \leq \tau \leq 1$. Cuando existe una total concordancia entre los órdenes de ambas variables el valor de τ es 1 y cuando el desacuerdo el total, el valor es igual a -1. Si las variables son independientes, el valor será cercano a cero.

Bajo la hipótesis nula de que X y Y son independientes, la distribución muestral de τ tendrá un valor esperado de cero. La distribución precisa no puede caracterizarse en términos de distribuciones comunes, pero se puede calcular exactamente para muestras pequeñas; para muestras más grandes (de al menos 10 pares de observaciones), es común utilizar una aproximación a la distribución normal, con media cero y varianza igual a:

$$\sigma_{\tau}^2 = \frac{2(2N+5)}{9N(N-1)} \quad (21)$$

Y por lo tanto, para “N” mayor que diez, $H_0: \rho=0$ puede someterse a prueba transformando ρ en un valor de Z, el cual tiene una distribuye normal con una media igual a cero y desviación estándar igual a uno ($N \sim 0,1$).

$$Z_{\tau} = \frac{\tau}{\sigma_{\tau}} = \frac{\tau}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}} \quad (22)$$

El percentil de Z_{τ} debe compararse con el valor del “p” crítico para decidir sobre H_0 .

Coefficiente de correlación de órdenes de Kendall (tau-b)

Este coeficiente considera los empates en los órdenes de los datos bajo análisis; su fórmula es:

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \quad (23)$$

Donde:

$$n_0 = n(n-1)/2$$

$$n_1 = n(n-1)/2$$

$$\sum_i t_i (t_i - 1) / 2$$

$$\sum_j t_j (t_j - 1) / 2$$

n_c : número de pares concordantes

n_d : número de pares discordantes

t_i = número de empates para el i^{th} grupo de empates para la primera variable.

u_j = número de empates para el j^{th} grupo de empates para la segunda variable.

EJEMPLO

Se desea determinar el coeficiente de correlación de rangos de Kendall (Tau-b) para el set de datos de diámetro (cm) y altura de Jaúl (m) analizado previamente. El procedimiento de cálculo es el mismo sugerido para el coeficiente de correlación de Spearman. El programa XLStatiscis ofrece los siguientes resultados:

Correlation Coeff τ_b 0.58019112	Large sample Hypothesis Test for τ_b $H_0: \tau_b = 0$ Alternative ☒ ? ☐ > ☐ < $H_1: \tau_b \neq 0$ Z 5.663047 p-value = 1.49E-08	La hipótesis nula es: $H_0: \tau = 0$ $H_A: \tau \neq 0$ Nivel de significancia o “p” crítico = 0,001. Observe que el programa también le permite hacer una prueba de una cola.

Dado que el valor de “p” calculado (1,49 E-08) es menor que “p” crítico (0,001), se concluye que el valor del coeficiente de correlación de rangos de Kendall (Tau-b) es diferente de cero y por lo tanto las variables diámetro y altura están correlacionadas.

9. Bibliografía

Abdi Hervé. The Kendall Rank Correlation Coefficient. In: Neil Salkind (Ed.) (2007). Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage. Disponible en <http://www.utdallas.edu/~herve/Abdi-KendallCorrelation2007-pretty.pdf>. Visitado 10 julio 2012.

Chatillon, C. 1985. Reply on "Some remarks on quick estimation of the correlation coefficient". The American Statistician 38(4):230-231.

Cleveland, W.S.; Diaconis, P. and McGil, R. 1982. Variables on scatterplots look more highly correlated when the scales are increased. Science 216 (4550): 1130-1141.

Cui Zhongmin, Li Dongmei, and Tao Wei. 2008. Constructing Bootstrap Confidence Intervals for Pearson's r . 10p. Disponible en http://www.divms.uiowa.edu/~kcowles/s166_2008/taoproject_report.pdf. Visitado 10 julio 2012.

Filliben, J. J. 1975. The probability plot correlation coefficient test for normality. Technometrics 17:111-112.

Goodman, L. A. 1972. Measures of association for cross-classifications, IV: simplification of asymptotic variances. Jour. of the American Statistical Association 67:415-421.

-----, 1963. Measures of association for cross-classifications, III: approximate sampling theory. Jour. of the American Statistical Association 58: 310-364.

-----, 1959. Measures of association for cross-classification, II. further discussion and references. Jour. of the American Statistical Association 54 (285):123-163.

-----, 1954. Measures of association for cross-classification. Jour. of the American Statistical Association 49 (268):732-764.

Helsel, D.R., Mueller, D.K., and Slack, J.R., 2006, Computer program for the Kendall family of trend tests: U.S. Geological Survey Scientific Investigations Report 2005-5275, 4 p. Disponible en <http://pubs.usgs.gov/sir/2005/5275/pdf/sir2005-5275.pdf>. Visitado 15 julio 2012.

Koch, G. G. 1985. A basic demonstration of the -1,1 range for the correlation coefficient. The American Statistician. 39(3):201-202.

Kruskall, W. H. 1958. Ordinal measures of association. Jour. of the American Statistical Association 53:814-861.

Kvalseth, T. O. 1985. Cautionary note about R^2 . The American Statistician. 39(4):Pt.1:279-285.

Pearson, K. 1896. Mathematical contributions to the theory of evolution, III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A*. 187:253-318.

Raveh, A. 1985. On quick estimates of Pearson's R from scatter diagrams. *The American Statistician* 39(3):239-240.

Schilling, M. F. 1984. Some remarks on quick estimation of the correlation coefficient. *The American Statistician* 38(4):330.

Somers, R.H. 1968. On the measurement of association. *American Sociological Review*. 33:291-292.

-----, 1962. A new asymmetric measure of association for ordinal variables. *American Sociological Review* 27:799-811.

Stuart, A. 1963. The estimation and comparison of strengths of association in contingency tables. *Biometrika* 40:105-110.

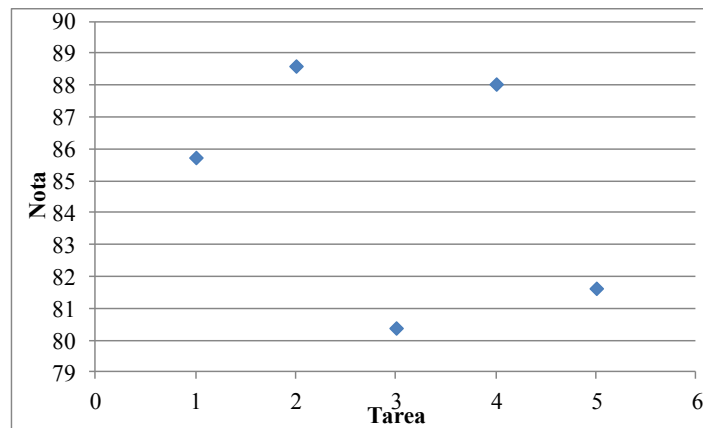
-----, 1963. Calculation of Spearman's rho for ordered two way classifications. *American Statistician* 17(4):23-24.

Wagstaff David A., Elek Elvira, Kulis Stephen and Marsiglia Flavio. 2009. Using a Nonparametric Bootstrap to Obtain a Confidence Interval for Pearson's r with Cluster Randomized Data: A Case Study. *J Prim Prev*.30(5): 497-512. Disponible en <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2792874/pdf/nihms148193.pdf>. Visitado 10 julio 2012.

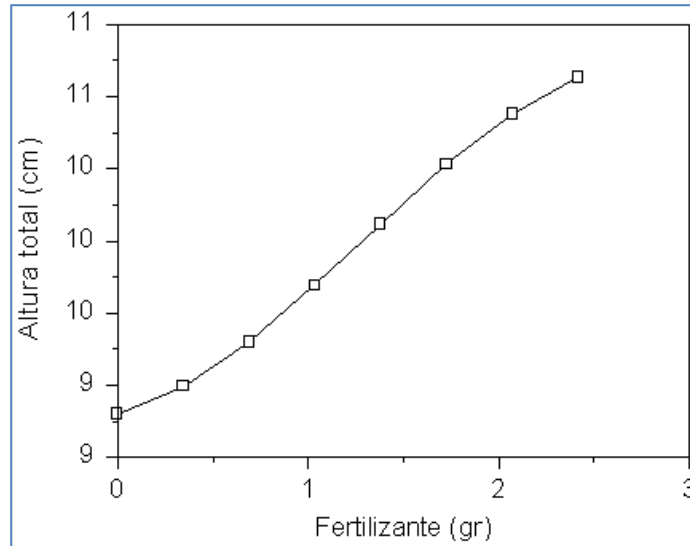
Yule, G. U. 1912. On the methods of measuring association between two attributes. *Jour. of the Royal Statistical Society* 75:579-642.

10. Ejercicios

1. Defina los siguientes conceptos: correlación, distribución bivariable, diagrama de dispersión, correlación positiva, correlación negativa, dirección de la correlación, principio de linealidad, varianza explicada, relación no lineal, Rho, Tau.
2. Brinde una lista de pares de variables que en su opinión muestren una correlación positiva.
3. Brinde una lista de pares de variables que en su opinión muestren una correlación negativa.
4. Basado en la gráfica que se muestra a continuación, responda a la siguiente pregunta: ¿Están las variables “nota” y “tarea” correlacionadas? Argumente su respuesta. ¿Cuál sería su pronóstico para las siguientes tareas?



5. Brinde tres ejemplos en los que usted esperaría que la relación entre X y Y fuera no lineal.
6. Una investigadora selecciona una muestra de 50 árboles al azar de un bosque húmedo Tropical (bh-T) en la península de Nicoya. Para cada árbol mide el diámetro a la altura del pecho (cm) y su altura total (m). Luego obtiene un valor de “r” igual a 0.55 y un valor de “p” calculado igual a 0.04. ¿Es el coeficiente de correlación estadísticamente significativo? Otro investigador que estudia un bosque húmedo Tropical (bh-T) cerca de Peñas Blancas, frontera norte, decide utilizar dicho valor de “r” para caracterizar la relación diámetro-altura de su sitio de estudio. ¿Cuál es su opinión al respecto?
7. Suponga que en un estudio sobre cantidad de fertilizante y crecimiento de plántulas en vivero se obtienen los siguientes resultados:



El valor de r es 0,99 y el valor de “ p ” calculado es 0,0001. Basado en dichos resultados recomendaría usted aplicar fertilizante a las plántulas. ¿Por qué? ¿Indica la gráfica que existe una relación entre la cantidad de fertilizante aplicado y el crecimiento de las plántulas? ¿Se podría argumentar que esta es una relación de causa-efecto?

8. Grafique los datos del archivo ppt_5_estaciones.xlsx. Luego, calcule el coeficiente de correlación lineal de Pearson y Rho de Spearman para Coronado-La Marina; Coronado-Quesada; Coronado-Sta. María. Compare sus resultados. ¿Es la correlación en todos los casos estadísticamente significativa?

9. El archivo biomasa_pochote_laurel.xlsx contiene datos de biomasa seca, diámetro a la altura del pecho (cm) y altura total (m) para pochote y laurel.

- Grafique diámetro a la altura del pecho (cm) versus altura total (m) por especie y diámetro a la altura del pecho (cm) versus biomasa seca por especie. Describa cada gráfica. ¿Se podrían combinar los datos en una sola muestra o se deben analizar por especie? ¿Por qué?
- Calcule e interprete el coeficiente de correlación lineal de Pearson y Rho de Spearman para el diámetro y biomasa seca. Grafique los datos.
- Calcule e interprete el coeficiente de correlación lineal de Pearson y Rho de Spearman para la altura total y biomasa seca. Grafique los datos.
- Si usted tuviese que estimar biomasa en función de diámetro o altura total, ¿Cuál variable utilizaría y por qué?

10. El Ministro de Ambiente realizará una conferencia de prensa para anunciar que durante su gestión se redujo la tala de bosque. Para esto cuenta con los siguientes datos. Usted como asesor ¿Qué le recomendaría? ¿Por qué?

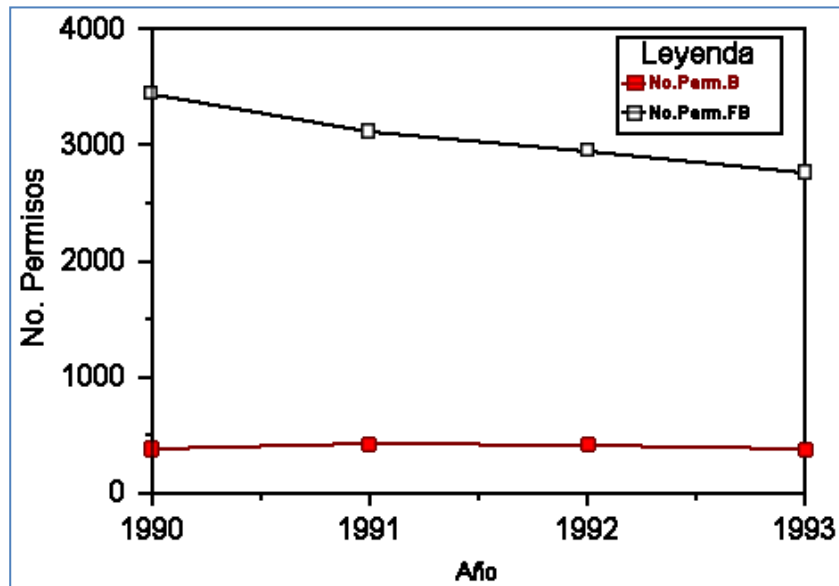
A continuación se muestra la correlación entre el número de permisos de aprovechamiento forestal para áreas ubicadas fuera del bosque (No. Perm. FB) y para áreas boscosas (No. Perm. B) para los últimos 4 años. ¿Cuál es el tamaño de la muestra? ¿Qué observaciones/comentarios puede hacer de los datos? ¿Cuáles pares de variables muestran correlaciones estadísticamente significativas?

 Pearson Correlation Table
 =====

	Año	No.Perm.FB	No.Perm.B
Año		-0.99 (4) 0.01	-0.19 (4) 0.81
No.Perm.FB	-0.99 (4) 0.01		0.04 (4) 0.96
No.Perm.B	-0.19 (4) 0.81	0.04 (4) 0.96	

Note: the values in the table are listed as follows:

correlation value
(sample size)
probability



Fuente: Basados en datos de SINAC,1999.

A continuación se muestra el volumen autorizado para extracción forestal en áreas fuera del bosque (Vol m³ FB) y para bosques (Vol m³ B) para el período 1990-1998. También se muestra la respectiva gráfica. ¿Cuál es el tamaño de la muestra? ¿Existe correlación lineal entre algunas de las variables en estudio? ¿Qué le indica el signo de r? A qué conclusión se llegaría analizando por separado los datos de las preguntas 1 y 2. Observe la línea correspondiente a los datos de volumen aprobado para explotación fuera del bosque; ¿nota algo diferente en dicha línea? ¿Cuál sería su explicación?

Pearson Correlation Table

	Año	Vol (m3) FB	Vol (m3) B
Año		-0.59 (5) 0.30	-0.57 (9) 0.11
Vol (m3) FB	-0.59 (5) 0.30		0.50 (5) 0.40
Vol (m3) B	-0.57 (9) 0.11	0.50 (5) 0.40	

Note: the values in the table are listed as follows:

correlation value
(sample size)
probability

Nota: Basados en datos de SINAC, 1999.

