# COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

# COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

## 1.1- Introducción

Antes de introducirnos en el modelo de regresión lineal, que hace referencia a la naturaleza de la relación entre distintas variables, pasaremos a exponer el estadístico utilizado para medir la magnitud de la relación (supuestamente lineal) entre dichas variables. Tiene sentido darle un tratamiento aparte por su importancia y las continuas referencias que ofreceremos a lo largo de este texto. Comenzaremos su desarrollo, por razones de simplicidad, para el caso particular de dos variables.

## 1.2.- Coeficiente de correlación lineal de Pearson

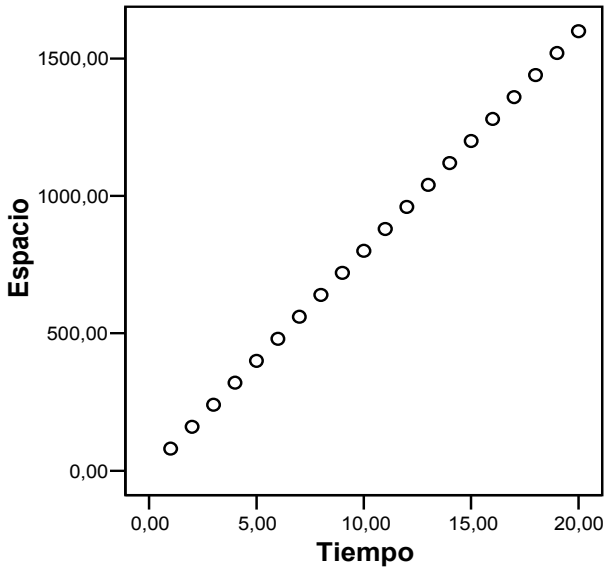
El coeficiente de correlación de Pearson, pensado para variables cuantitativas (escala mínima de intervalo), es un índice que mide el grado de covariación entre distintas variables relacionadas linealmente. Adviértase que decimos "variables relacionadas linealmente". Esto significa que puede haber variables fuertemente relacionadas, pero no de forma lineal, en cuyo caso no proceder a aplicarse la correlación de Pearson. Por ejemplo, la relación entre la ansiedad y el rendimiento tiene forma de U invertida; igualmente, si relacionamos población y tiempo la relación será de forma exponencial. En estos casos (y en otros muchos) no es conveniente utilizar la correlación de Pearson. Insistimos en este punto, que parece olvidarse con cierta frecuencia.

El coeficiente de correlación de Pearson es un índice de fácil ejecución e, igualmente, de fácil interpretación. Digamos, en primera instancia, que sus valores absolutos oscilan entre 0 y 1. Esto es, si tenemos dos variables X e Y, y definimos el coeficiente de correlación de Pearson entre estas dos variables como $r_{xy}$ entonces:

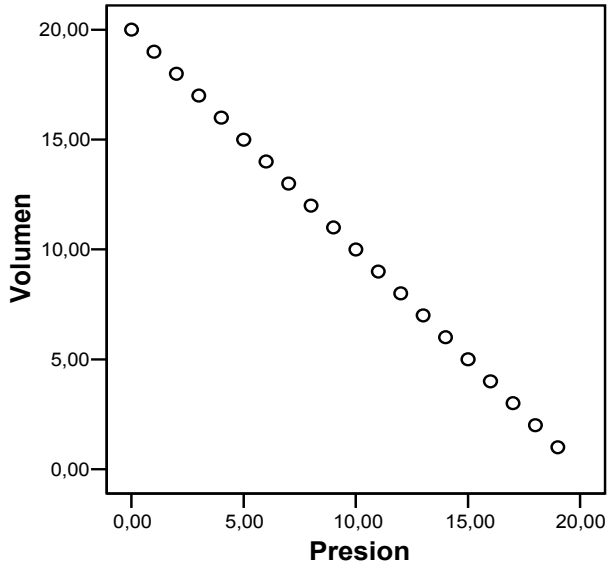
$$0 \leq r_{xy} \leq 1$$

Hemos especificado los términos "valores absolutos" ya que en realidad si se contempla el signo el coeficiente de correlación de Pearson oscila entre –1 y +1. No obstante ha de indicarse que la magnitud de la relación vienen especificada por el valor numérico del coeficiente, reflejando el signo la dirección de tal valor. En este sentido, tan fuerte es una relación de +1 como de -1. En el primer caso la relación es *perfecta positiva* y en el segundo *perfecta negativa*. Pasamos a continuación a desarrollar algo más estos conceptos.

Decimos que la correlación entre dos variables X e Y es perfecta positiva cuando exactamente en la medida que aumenta una de ellas aumenta la otra. Esto sucede cuando la relación entre ambas variables es funcionalmente exacta. Difícilmente ocurrirá en psicología, pero es frecuente en los ciencias físicas donde los fenómenos se ajustan a leyes conocidas, Por ejemplo, la relación entre espacio y tiempo para un móvil que se desplaza a velocidad constante. Gráficamente la relación ser del tipo:

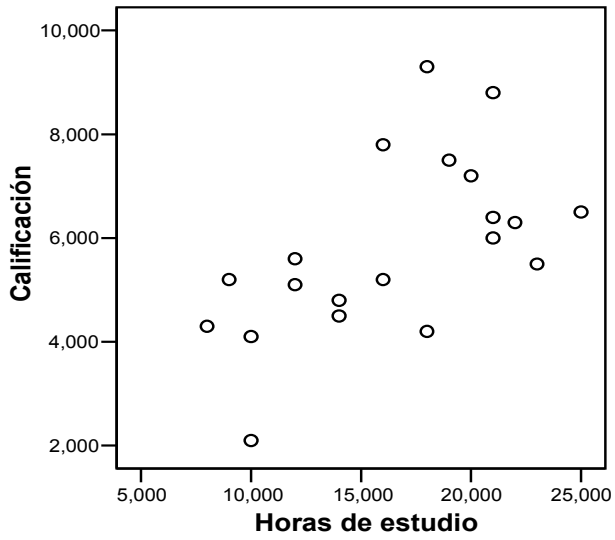



Se dice que la relación es perfecta negativa cuando exactamente en la medida que aumenta una variable disminuye la otra. Igual que en el caso anterior esto sucede para relaciones funcionales exactas, propio de las ciencias físicas. Por ejemplo, la relación entre presión y volumen se ajusta a este caso. El gráfico que muestra la relación sería del tipo:



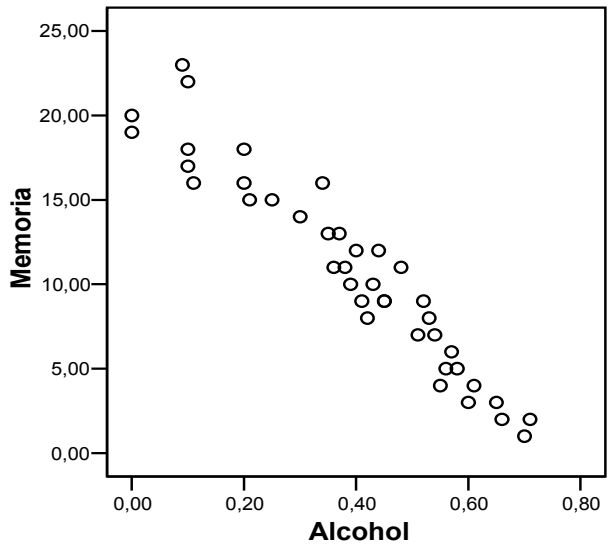

En los fenómenos humanos, fuertemente cargados de componentes aleatorios, no suelen ser posible establecer relaciones funcionales exactas. Dado un cierto valor en la variable X no encontraremos uno y solo un único valor en la variable Y. Por ejemplo, si relacionamos horas de estudio con el rendimiento académico obtendremos mayor rendimiento a mayor inteligencia, pero será prácticamente imposible saber con exactitud la puntuación que obtendrá un sujeto para unas horas determinadas. Dado un cierto número de personas con un mismo número de horas, por ejemplo 10, no todos

obtendrán exactamente la misma puntuación en rendimiento. Unos obtendrán más o menos en función de otras variables, tales como motivación o personalidad. Si relacionásemos ambas variables dada una muestra de sujetos tendríamos un gráfico de las siguientes características:



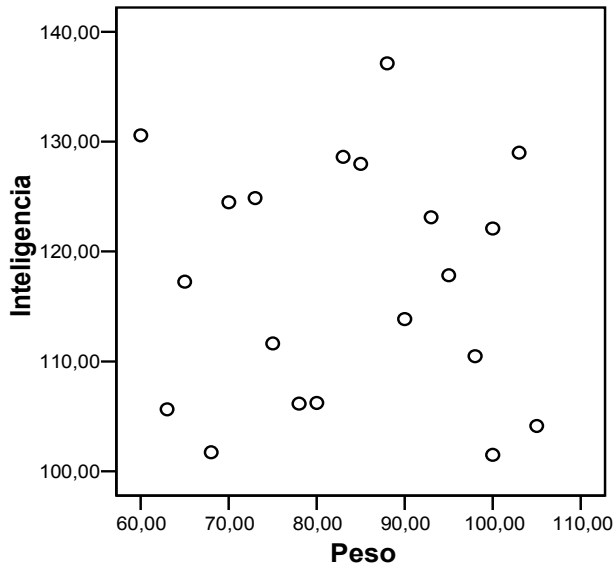

Se observa que para un mismo valor en inteligencia existen diferentes posibles valores en rendimiento. Se trata de una correlación positiva pero no perfecta. Este conjunto de puntos, denominado diagrama de dispersión o nube de puntos tiene interés como primera toma de contacto para conocer la naturaleza de la relación entre dos variables. Si tal nube es alargada -apunta a una recta- y ascendente como es el caso que nos ocupa, es susceptible de aplicarse el coeficiente lineal de Pearson. El grosor de la nube da una cierta idea de la magnitud de la correlación; cuanto más estrecha menor será el margen de variación en Y para los valores de X, y por tanto, más acertado los pronósticos, lo que implica una mayor correlación.

Si la nube de puntos es alargada y descendente nos encontramos con una correlación negativa. Supongamos, en este sentido, que relacionásemos la cantidad de alcohol ingerida y el grado de memorización ante determinados estímulos. Obtendríamos un gráfico como el siguiente:



Se observa que a mayor cantidad de alcohol ingerida menor material recordado. Igual que anteriormente no puede establecerse con exactitud el grado de memorización en función del alcohol ingerido, aunque queda claro la tendencia existente.

Por último, si la nube de puntos adopta una configuración más o menos redondeada de tal forma que no pueda especificarse ningún tipo de relación, nos encontramos con una correlación nula. Supongamos que relacionásemos *peso* con *inteligencia.* Obtendríamos el siguiente gráfico:




Se observa que las personas con poco peso obtienen en inteligencia tanto puntuaciones bajas como medias o altas. Lo mismo sucede con personas de peso alto. No puede establecerse, pues, ningún tipo de relación. Ambas variables son independientes entre sí; la variación de una de ellas no influye para nada en la variación de la otra.

### 1.3.1.- Formula utilizada

El coeficiente de correlación de Pearson viene definido por la siguiente expresión:

$$r_{xy} = \frac{\sum Z_x Z_y}{N}$$

Esto es, el coeficiente de correlación de Pearson hace referencia a la media de los productos cruzados de las puntuaciones estandarizadas de X y de Y. Esta formula reúne algunas propiedades que la hacen preferible a otras. A operar con puntuaciones estandarizadas es un índice libre de escala de medida. Por otro lado, su valor oscila, como ya se ha indicado, en términos absolutos, entre 0 y 1.

Téngase en cuenta que las puntuaciones estandarizadas muestran, precisamente, la posición en desviaciones tipo de un individuo respecto a su media. Reflejan la medida en que dicho individuo se separa de la media. En este sentido, supongamos que para cada individuo tomamos dos medidas en X e Y. La correlación entre estas dos variables será

perfecta positiva cuando cada individuo manifieste la misma superioridad o inferioridad en cada una de ellas. Esto se cumple cuando su posición relativa sea la misma, es decir, cuando sus puntuaciones tipo sean iguales ($Z_x = Z_y$). En este caso la formula de la correlación se transforma en:

$$r_{xy} = \frac{\sum Z_x Z_y}{N} = \frac{\sum Z_x Z_x}{N} = \frac{\sum Z_x^2}{N} = 1$$

ya que tal expresión equivale a la varianza de $Z_x$ , que como se sabe vale la unidad.

Cuando la correlación es perfecta negativa los valores de $Z_x$ y $Z_y$ son exactamente iguales pero de signo contrario, resultando los productos cruzados de $Z_x$ y $Z_y$ negativos. En este caso, el valor de la correlación es el mismo que anteriormente pero de signo negativo:

$$r_{xy} = \frac{\sum Z_x Z_y}{N} = \frac{\sum Z_x Z_x}{N} = \frac{\sum Z_x^2}{N} = 1$$

Cuando la correlación es nula, para un valor obtenido de X se podrá obtener cualquier valor de Y; es decir, para un valor determinado de $Z_x$ la misma cantidad de valores positivos y negativos de $Z_y$. De todo ello resulta que la suma de productos cruzados valdrá cero ya que habrá tantos productos positivos como negativos. Así pues:

$$r_{xy} = \frac{\sum Z_x Z_y}{N} = 0$$

La formula (1.5) puede expresarse de forma más sencilla de la siguiente manera:

$$r_{xy} = \frac{\frac{\sum XY}{N} - \overline{X}\,\overline{Y}}{S_x S_y}$$

Efectivamente:

$$r_{xy} = \frac{\sum Z_x Z_y}{N} = \frac{\sum \left( \frac{X - \overline{X}}{S_x} * \frac{Y - \overline{Y}}{S_y} \right)}{N} = \frac{\sum \left( X - \overline{X} \right)\left( Y - \overline{Y} \right)}{N S_x S_y} = \frac{\sum \left( XY - X\overline{Y} - \overline{X}Y + \overline{X}\,\overline{Y} \right)}{N S_x S_y} =$$

$$= \frac{1}{S_x S_y} \left( \frac{\sum XY}{N} - \frac{\overline{Y} \sum X}{N} - \frac{\overline{X} \sum Y}{N} + \frac{N \overline{X}\,\overline{Y}}{N} \right) = \frac{1}{S_x S_y} \left( \frac{\sum XY}{N} - \overline{X}\,\overline{Y} - \overline{X}\,\overline{Y} + \overline{X}\,\overline{Y} \right) = \frac{\frac{\sum XY}{N} - \overline{X}\,\overline{Y}}{S_x S_y}$$

Esta formula es especialmente útil cuando se conocen las medias de X e Y así como sus desviaciones tipo, lo cual es relativamente frecuente. Si por cualquier circunstancia no dispusiéramos de la información de estos estadísticos podríamos calcular $r_{xy}$ recurriendo a la expresión en *puntuaciones directas*:

$$r_{xy} = \frac{\frac{\sum XY}{N} - \overline{XY}}{S_x S_y} = \frac{\frac{\sum XY}{N} - \frac{\sum X}{N}\frac{\sum Y}{N}}{\sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}\sqrt{\frac{\sum Y^2}{N} - \left(\frac{\sum Y}{N}\right)^2}} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

Podemos expresar, igualmente, el coeficiente de correlación de Pearson en puntuaciones diferenciales o centradas mediante la siguiente formula:

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

donde $x = X - \overline{X}$ e $y = Y - \overline{Y}$. Para su demostración partamos de (1.5):

$$r_{xy} = \frac{\sum Z_x Z_y}{N} = \frac{\sum\left(\frac{X-\overline{X}}{S_x} * \frac{Y-\overline{Y}}{S_y}\right)}{N} = \frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{NS_xS_y} =$$

$$= \frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{N\sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{N}}\sqrt{\frac{\sum\left(Y-\overline{Y}\right)^2}{N}}} = \frac{\sum xy}{N\sqrt{\frac{\sum x^2}{N}}\sqrt{\frac{\sum y^2}{N}}} = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

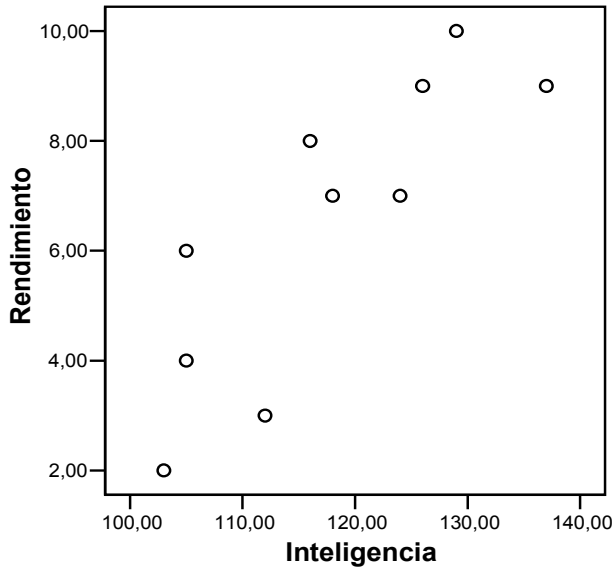
**Ejemplo 1.1.-** Tengamos las siguientes puntuaciones en las variables X (inteligencia) e Y (rendimiento académico):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| X: | 105 | 116 | 103 | 124 | 137 | 126 | 112 | 129 | 118 | 105 |
| Y: | 4 | 8 | 2 | 7 | 9 | 9 | 3 | 10 | 7 | 6 |

Calcular el coeficiente de correlación de Pearson: a) en puntuaciones directas, b) puntuaciones diferenciales y c) puntuaciones estandarizadas.

SOL:

Antes de calcular el coeficiente de correlación de Pearson hemos de comprobar si existe una tendencia lineal en la relación. Aunque más adelante ofreceremos procedimientos analíticos que permitan verificar con exactitud la Hipótesis de linealidad, por el momento, recurriremos a procedimientos gráficos, que en una primera instancia, pueden resultar suficientes:




Se observa la existencia de una cierta tendencia lineal en la relación. Podemos, en consecuencia, proceder a calcular el coeficiente de correlación de Pearson.

*a) Puntuaciones directas.*

Configuremos la siguiente tabla:

| X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|
| 105 | 4 | 11025 | 16 | 420 |
| 116 | 8 | 13456 | 64 | 928 |
| 103 | 2 | 10609 | 4 | 206 |
| 124 | 7 | 15376 | 49 | 868 |
| 137 | 9 | 18769 | 81 | 1233 |
| 126 | 9 | 15876 | 81 | 1134 |
| 112 | 3 | 12544 | 9 | 336 |
| 129 | 10 | 16641 | 100 | 1290 |
| 118 | 7 | 13924 | 49 | 826 |
| 105 | 6 | 11025 | 36 | 630 |
| 1175 | 65 | 139245 | 489 | 7871 |

De donde:

$$\overline{X} = \frac{\sum X}{N} = \frac{1175}{10} = 117.5$$
$$\overline{Y} = \frac{\sum Y}{N} = \frac{65}{10} = 6.5$$

$$S_x = \sqrt{\frac{\sum X^2}{N} - \overline{X}^2} = \sqrt{\frac{139245}{10} - 117.5^2} = 10.874$$
$$S_x = \sqrt{\frac{\sum Y^2}{N} - \overline{Y}^2} = \sqrt{\frac{489}{10} - 6.5^2} = 2.579$$

Aplicando (1.9):

$$r_{xy} = \frac{\frac{\sum XY}{N} - \overline{X}\overline{Y}}{S_x S_y} = \frac{\frac{7871}{10} - 117.5 * 6.5}{10.874 * 2.579} = 0.8327$$

*b) Puntuaciones diferenciales o centradas*

Hagamos las siguientes transformaciones:

x = X - X
y = Y – Y

| X | Y | x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|---|---|
| 105 | 4 | -12.50 | -2.50 | 156.25 | 6.25 | 31.25 |
| 116 | 8 | -1.50 | 1.50 | 2.25 | 2.25 | -2.25 |
| 103 | 2 | -14.50 | -4.50 | 210.25 | 20.25 | 65.25 |
| 124 | 7 | 6.50 | .50 | 42.25 | .25 | 3.25 |
| 137 | 9 | 19.50 | 2.50 | 380.25 | 6.25 | 48.75 |
| 126 | 9 | 8.50 | 2.50 | 72.25 | 6.25 | 21.25 |
| 112 | 3 | -5.50 | -3.50 | 30.25 | 12.25 | 19.25 |
| 129 | 10 | 11.50 | 3.50 | 132.25 | 12.25 | 40.25 |
| 118 | 7 | .50 | .50 | .25 | .25 | .25 |
| 105 | 6 | -12.50 | -.50 | 156.25 | .25 | 6.25 |
| 1175 | 65 | 0 | 0 | 1182.5 | 66.5 | 233.5 |

Apliquemos (1.10):

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sqrt{\sum y^2}}} = \frac{233.5}{\sqrt{1182.5}\sqrt{66.5}} = 0.8327$$

*c) Puntuaciones estandarizadas*

Hagamos las oportunas transformaciones:

$$Z_x = \frac{X - \overline{X}}{S_x}$$
$$Z_y = \frac{Y - \overline{Y}}{S_y}$$

Y configuremos la siguiente tabla:

| X | Y | $Z_x$ | $Z_y$ | $Z_xZ_y$ |
|---|---|---|---|---|
| 105.0 | 4.0 | -1.15 | -.97 | 1.11 |
| 116.0 | 8.0 | -.14 | .58 | -.08 |
| 103.0 | 2.0 | -1.33 | -1.74 | 2.33 |
| 124.0 | 7.0 | .60 | .19 | .12 |
| 137.0 | 9.0 | 1.79 | .97 | 1.74 |
| 126.0 | 9.0 | .78 | .97 | .76 |
| 112.0 | 3.0 | -.51 | -1.36 | .69 |
| 129.0 | 10.0 | 1.06 | 1.36 | 1.44 |
| 118.0 | 7.0 | .05 | .19 | .01 |
| 105.0 | 6.0 | -1.15 | -.19 | .22 |
| 1175 | 65 | 0 | 0 | 8.327 |

Apliquemos la formula (1.5):

$$r_{xy} = \frac{\sum Z_x Z_y}{N} = \frac{8.327}{10} = 0.8327$$

## 1.4.- Significación del coeficiente de correlación

Una vez calculado el valor del coeficiente de correlación interesa determinar si tal valor obtenido muestra que las variables X e Y están relacionadas en realidad o tan solo presentan dicha relación como consecuencia del azar. En otras palabras, nos preguntamos por la significación de dicho coeficiente de correlación.

Un coeficiente de correlación se dice que es significativo si se puede afirmar, con una cierta probabilidad, que es diferente de cero. Más estrictamente, en términos estadísticos, preguntarse por la significación de un cierto coeficiente de correlación no es otra cosa que preguntarse por la probabilidad de que tal coeficiente proceda de una población cuyo valor sea de cero. A este respecto, como siempre, tendremos dos hipótesis posibles:

$H_0$: $r_{xy} = 0 \Rightarrow$ El coeficiente de correlación obtenido procede de una población cuya correlación es cero ( $\rho = 0$ ).

$H_1$ : $r_{xy} = 0 \Rightarrow$ El coeficiente de correlación obtenido procede de una población cuyo coeficiente de correlación es distinto de cero ( $\rho \neq 0$ ).

Desde el supuesto de la Hipótesis nula se demuestra que la distribución muestral de correlaciones procedentes de una población caracterizada por una correlación igual a cero ( $\rho = 0$ ) sigue una ley de *Student* con N-2 grados de libertad, de media el valor poblacional y desviación tipo:

$$S_r = \sqrt{\frac{1 - r_{xy}^2}{N - 2}}$$

En consecuencia, dado un cierto coeficiente de correlación $r_{xy}$ obtenido en una determinada muestra se trata de comprobar si dicho coeficiente es posible que se encuentre dentro de la distribución muestral especificada por la Hipótesis nula. A efectos prácticos, se calcula el número de desviaciones tipo que se encuentra el coeficiente obtenido del centro de la distribución, según la formula conocida:

$$t = \frac{r_{xy} - 0}{\sqrt{\frac{1 - r_{xy}^2}{N - 2}}}$$

y se compara el valor obtenido con el existente en las tablas para un cierto nivel de significación $\alpha$ y N-2 grados de libertad - $t_{(\alpha, N-2)}$ -, que como se sabe, marca el límite (baja probabilidad de ocurrencia, según la Hipótesis nula) de pertenencia de un cierto coeficiente $r_{xy}$ a la distribución muestra de correlaciones procedentes de una población con $\rho = 0$. De esta forma si:

$t > t_{(\alpha,N-2)} \Rightarrow$ Se rechaza la Hipótesis nula. La correlación obtenida no procede de una población cuyo valor $\rho_{xy} = 0$. Por tanto las variables están relacionadas.

$t \leq t_{(\alpha,N-2)} \Rightarrow$ Se acepta la Hipótesis nula. La correlación obtenida procede de una población cuyo valor $\rho_{xy} = 0$. Por tanto ambas variables no están relacionadas.

**Ejemplo 1.2.-** Determinar la significación del coeficiente de correlación del ejemplo 1.1.

SOL:

Apliquemos (1.12):

$$t = \frac{r_{xy} - 0}{\sqrt{\frac{1 - r_{xy}^2}{N-2}}} = \frac{0.8327}{\sqrt{\frac{1 - 0{,}8327^2}{10-2}}} = 4.21$$

Buscamos en la tabla de *t de Student* para $\alpha = 0.05$ y 10-2 = 8 grados de libertad, tal como se observa a continuación donde se muestra un fragmento de dicha tabla:

**Critical Values of the t Distribution**

| df | 2-tailed testing | | | 1-tailed testing | | |
|---|---|---|---|---|---|---|
| | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| 5 | 2.015 | 2.571 | 4.032 | 1.476 | 2.015 | 3.365 |
| 6 | 1.943 | 2.447 | 3.707 | 1.440 | 1.943 | 3.143 |
| 7 | 1.895 | 2.365 | 3.499 | 1.415 | 1.895 | 2.998 |
| 8 | 1.860 | 2.306 | 3.355 | 1.397 | 1.860 | 2.896 |
| 9 | 1.833 | 2.262 | 3.250 | 1.383 | 1.833 | 2.821 |
| 10 | 1.812 | 2.228 | 3.169 | 1.372 | 1.812 | 2.764 |
| 11 | 1.796 | 2.201 | 3.106 | 1.363 | 1.796 | 2.718 |
| 12 | 1.782 | 2.179 | 3.055 | 1.356 | 1.782 | 2.681 |
| 13 | 1.771 | 2.160 | 3.012 | 1.350 | 1.771 | 2.650 |
| 14 | 1.761 | 2.145 | 2.977 | 1.345 | 1.761 | 2.624 |

El valor marcado con una elipse:

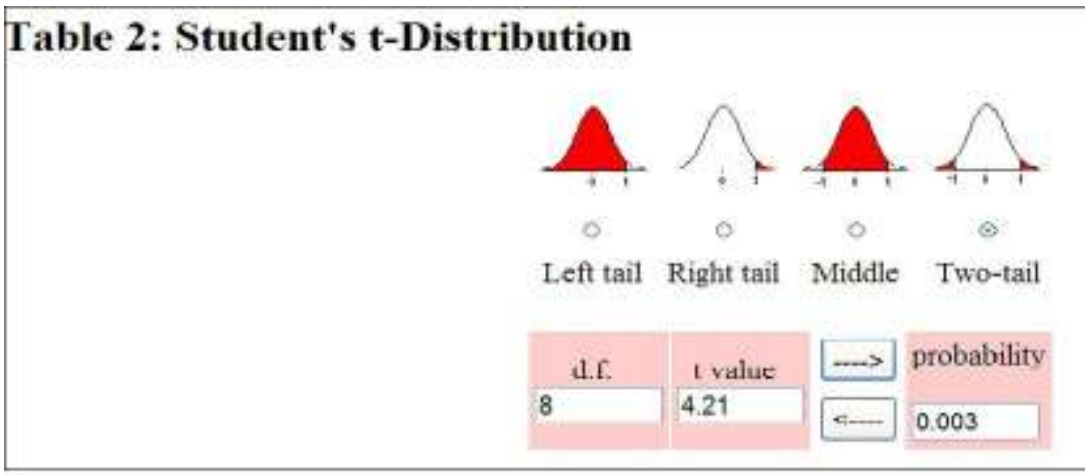
$$t_{(0.05,8)} = 2.306$$

Comparamos el valor *t* obtenido con el de las tablas:

$$4.21 > 2.306$$

Rechazamos la Hipótesis nula con un riesgo (máximo) de equivocarnos de 0.05. La correlación obtenida no procede de una población caracterizada por una correlación de cero. Concluimos, pues, que ambas variables están relacionadas.

Una manera más exacta de conocer el riesgo asociado (y no el genérico 0.05 que se toma como referencia máxima) es recurrir a las tablas interactivas:




Hemos seleccionado el gráfico superior derecho, ya que nos interesa la probabilidad correspondiente a la zona de rechazo de la Hipótesis nula.. Los grados de libertad son 8 (degre of freedom) y el valor de t (t value) de 4.21. La probabilidad asociada es 0.003, lo que nos indica la probabilidad de obtener en una muestra de tamaño 10 una correlación de 0.8237 procedente de una población cuya correlación es 0. Al afirmar que esta correlación no procede de tal población no estaremos equivocando un 0.003 de las veces, que es precisamente nuestro riesgo de equivocarnos. Obsérvese que aquí afinamos más, y no concluimos el genérico de "0.05 como máximo" sino que concretamos a su probabilidad verdadera de 0.003.

## 1.5.- Interpretación del coeficiente de correlación.

Como se ha indicado el coeficiente de correlación de Pearson es un índice cuyos valores absolutos oscilan entre 0 y 1. Cuanto más cerca de 1 mayor ser la correlación, y menor cuanto más cerca de cero. Pero como interpretar un coeficiente determinado? ¿Qué significa un coeficiente de 0.6?. ¿Es alto o bajo?. No puede darse una respuesta precisa. Depende en gran parte de la naturaleza de la investigación. Por ejemplo, una correlación de 0.6 sería baja si se trata de la fiabilidad de un cierto test, pero sin embargo, sería alta si estamos hablando de su validez.

No obstante, intentaremos abordar el tema desde dos perspectivas distintas. Por un lado, ya ha sido tratado desde la perspectiva de la significación estadística mencionada en el apartado anterior. Desde este enfoque una correlación es *efectiva* si puede afirmarse que es distinta de cero. Pero ha de decirse que una correlación significativa no necesariamente ha de ser una correlación fuerte; simplemente es una correlación diferente de cero. O en otros términos, es una correlación que es poco probable que proceda de una población cuya correlación es cero. Tan solo se está diciendo que se ha obtenido "algo" y que ese "algo" es (probablemente) más que "nada". La significación de $r_{xy}$ depende en gran medida del tamaño de la muestra, tal como puede observarse en (1.12); una correlación de 0.01 puede ser significativa en una muestra suficientemente grande y otra de 0.9 no serlo en una muestra pequeña. Aquí se cumple la ley de los grandes números; tendencias débiles son muy improbables, desde la Hipótesis nula, en grandes masas de datos, mientras que tendencias fuertes pueden ser relativamente probables en un tamaño pequeño de muestra.

Más interés tiene la interpretación del coeficiente de correlación en términos de *proporción de variabilidad compartida o explicada*, donde se ofrece una idea más cabal de la magnitud de la relación. Nos referimos al coeficiente de determinación. Dicho coeficiente se define como el cuadrado del coeficiente de correlación; esto es, dada dos variable X e Y, hace referencia a $r_{xy}^2$, y se entiende como una proporción de variabilidades (lo demostraremos más adelante). Por ejemplo, si la correlación entre inteligencia y rendimiento académico es de 0.8, significa que $0.8^2 = 0.64$ es la proporción de varianza compartida entre ambas variables. Puede interpretarse como que un 64% del rendimiento académico es debido a la inteligencia -variabilidad explicada-, o bien, y esto es más exacto si hemos de ser estrictos, que inteligencia y rendimiento académico comparten un 64% de elementos, o lo que es lo mismo, tanto la inteligencia como el rendimiento ponen en juego un 64% de habilidades comunes.

En estas circunstancias, si tomamos como variable dependiente o a *explicar* el rendimiento académico y elegimos la inteligencia como variable predictora o explicativa, tendremos que tal variable da cuenta de un 64% de la variabilidad en rendimiento. Queda, por ello, 1-0.64=0.36, un 36% del rendimiento que queda sin explicar. A este valor (0.36) se le denomina coeficiente de no determinación o coeficiente de alienación, y se define como $1 - r_{xy}^2$. Un término más adecuado y que proporciona mayor compresión es el de *proporción de variabilidad no explicada*. Si incrementásemos el número variables explicativas con otras variables como la motivación o la personalidad probablemente logremos aumentar la proporción de variabilidad explicada en rendimiento, obteniendo, si es eso lo que nos interesa, un

mayor control en la variable a predecir. De esto nos ocuparemos cuando tratemos la correlación múltiple.

El planteamiento de la correlación en términos de proporción variabilidad es, en nuestra opinión, la forma más comprensiva de afrontar la correlación lineal. Si acordamos que la variable dependiente Y corresponde a un cierto aspecto de la conducta que deseamos conocer, y definimos su variabilidad total, se trata de encontrar un conjunto de variables $X_1$ , $X_2$ , ... $X_k$ que absorban de Y un gran porcentaje de su variabilidad. De esta forma, interviniendo sobre el conjunto de variables independientes podremos dar cuenta de lo que sucede en Y, y modificarlo, si fuera el caso.

## 1.6.- Correlación y causalidad

En sentido estricto, correlación entre dos variables tan solo significa que ambas variables comparten información, que comparten variabilidad. Determinar el origen de la información, la fuente de la variabilidad -la causa- es una cuestión que no puede resolverse mediante recursos exclusivamente matemáticos.

Existen diferentes procedimientos para determinar, dada un serie de variables, la posible causa de ellas. Depende del tipo de contexto en el que nos encontremos. En *los contextos experimentales,* donde las variables pueden ser manipuladas a voluntad del investigador (tiempo de presentación de un determinado estímulo, cantidad de droga suministrada, ..etc) no existe especial dificultad en localizar las causas. Basta con mantener constantes todas las variables implicadas excepto la que nos interesa para determinar la posible fuente de variación. Se impone en estos casos, lo que se denomina control experimental -manipulación de variables-.

En los denominados estudio de campo donde el investigador ha de conformarse con los valores de las variables tal como vienen asignados (edad, sexo, nivel social, ingresos, hábitat ...etc) la determinación de las causas exige un proceso algo más complicado. Son en estos casos, el conocimiento que tengamos de la materia en cuestión, la lógica ciertas dosis de sentido común las claves a considerar. Existen casos sencillos como el que se ilustra a continuación que muestran lo dicho. Bernard Shaw afirmaba, en tono irónico, que llevar paraguas hacía a la gente más ilustrada, vista en aquella época, la relación entre ambas variables. Es evidente que llevar paraguas no hace a la gente más inteligente, sin embargo, ambas variables frecuentemente se presentaban unidas (en el siglo XIX), ya que era propio de la gente adinerada llevar paraguas. Nos encontramos que es el nivel social lo que hace que se utilice tal instrumento, y que es también, el nivel social el factor relevante en cuanto el nivel educativo alcanzado. La misma consideración puede formularse ante el hecho comprobado de una correlación negativa entre el número de mulas y el de licenciados universitarios en las distintas regiones españolas. Está claro que no lograremos aumentar el número de licenciados simplemente suprimiendo mulas. En todos estos casos sucede lo que gráficamente podría quedar ilustrado de la siguiente manera:

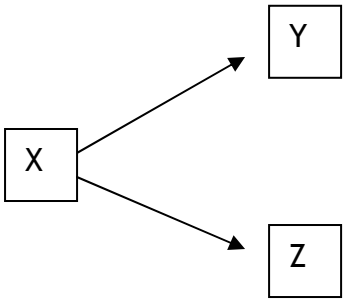
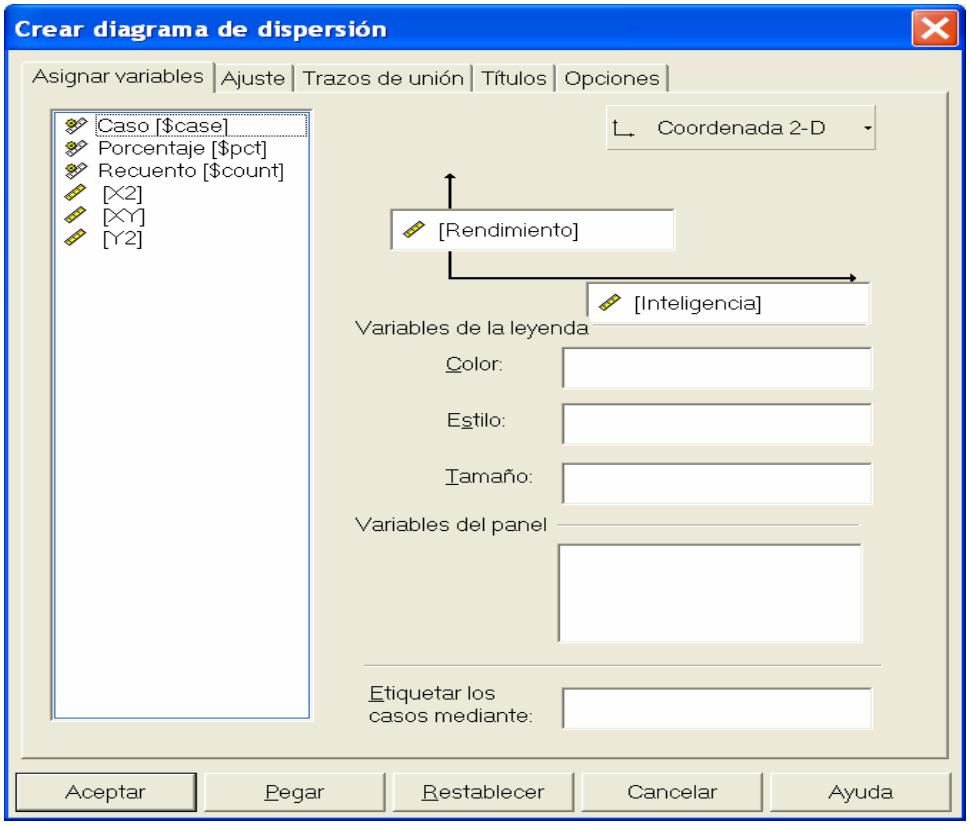



*Figura 1.9* ^

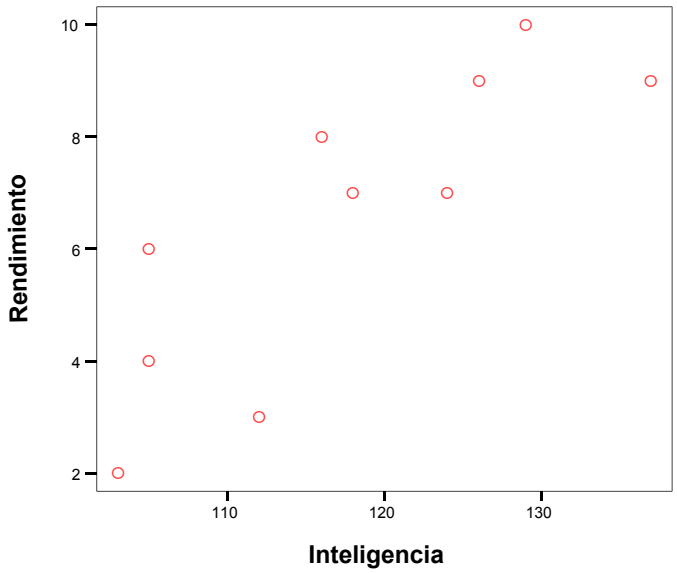
La variable Y (número de mulas, presencia de paraguas) correlaciona con la variable Z (número de licenciados, inteligencia) debido a la variable X, causa común de ambas, (desarrollo económico de la región, nivel social de la persona). Se dice, en estos casos, que la relación entre Y y Z es una relación espúrea. Se observa, de esta forma, cómo dos variables están relacionadas sin que haya una relación directa de una sobre la otra, sino debido al influjo de una tercera variable. Se concluye, pues, que correlación entre dos variables no implica necesariamente causalidad entre ambas. Dejaremos para próximos capítulos, en el tema de las correlaciones parciales y semiparciales, y especialmente, en el dedicado al *Path Análisis*, una discusión extensa sobre este tema.

## 1.6.- Aplicación Informática

Procedemos en las próximas páginas a desarrollar los ejemplos realizados en este capítulo mediante los recursos que nos ofrece el paquete estadístico SPSS. A este respecto, elaboremos primeramente el diagrama de dispersión, que nos dará cuenta de la adecuación del coeficiente lineal de Pearson. Para ello vayamos primeramente a **Gráficos/ Interactivos/Diagrama de dispersión:**




Obtendremos:




Para el cálculo del coeficiente de correlación de Pearson, vayamos a **Analizar/ Correlaciones/Bivariadas:**

**Correlaciones**

| | | Inteligencia | Rendimiento |
|---|---|---|---|
| Inteligencia | Correlación de Pearson | 1 | ,833** |
| | Sig. (bilateral) | | ,003 |
| | N | 10 | 10 |
| Rendimiento | Correlación de Pearson | ,833** | 1 |
| | Sig. (bilateral) | ,003 | |
| | N | 10 | 10 |

**. La correlación es significativa al nivel 0,01 (bilateral).

Donde se nos ofrece el valor de la correlación con sus probabilidades asociadas (Sig. Bilateral)

## Bibliografía


Achen, C. H. (1982). *Interpreting and using regression*. London: Sage.

Amon, J. (1990). *Estadística para psicólogos (1)*. Estadística Descriptiva. Madrid: Pirámide. (*)

Amon, J. (1990). *Estadística para psicólogos (2)*. Probabilidad. Estadística Inferencial. Madrid: Pirámide.

Berry, W. D., & Feldman, S. (1985). *Multiple Regression in Practice.* Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-050). Newbury Park, CA: Sage.

Botella y Sanmartin, R. (1992). *Análisis de datos en Psicología I*. Madrid: Pirámide.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cook, R. D. and Weisberg S. (1982). *Residual and influence in regression. New York*: Chapman & Hall.

Chatterjee, S. (1977). *Regression analysis by example*. New York: Wiley

Domenech, J. M. (1985). *Métodos estadísticos: modelo lineal de regresión*. Barcelona:

Etxeberria, J. (1999). *Regresión Múltiple*. Cuadernos de Estadística. Ed. La Muralla S.A. Espérides, Salamanca

Pedhazur, E. J., (1997). *Multiple Regression in Behavioral Research* (3rd ed.). Orlando, FL:Harcourt Brace.

Wonnacott, T. H. and Wonnacott, R. J. (1981). *Regression: a second course in statistics.* New York: Wiley.

## Internet

Correlación en Wikipedia (español): http://es.wikipedia.org/wiki/Correlaci%C3%B3n

Relación entre variables cuantitativas:
http://www.fisterra.com/mbe/investiga/var_cuantitativas/var_cuantitativas2.pdf

Correlation en Wikipedia (inglés): http://en.wikipedia.org/wiki/Correlation

Electronic Statistics Textbook: http://www.statsoft.com/textbook/stathome.html

Stat notes: An Online Textbook, by G. David Garson of North Carolina State University:
http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

Página de Karl Wünsch sobre correlación:
http://core.ecu.edu/psyc/wuenschk/docs30/corr6430.doc