# ESTADÍSTICA: MEDIDAS DESCRIPTIVAS II

AUTOR: PABLO VARGAS PEREIRA

FEBRERO: 2021

San Marcos

# Contenido

# Introducción

Los procedimientos de correlación son importantes para el análisis de los datos, la comparación entre diversas fuentes de datos provee a los investigadores herramientas para que se puedan tomar acciones o predicciones sobre diversos comportamientos.
En la presente lectura, se evidenciarán diversas estrategias aplicadas en diversos contextos, que pueden colaborar en muchas situaciones en contextos distintos.

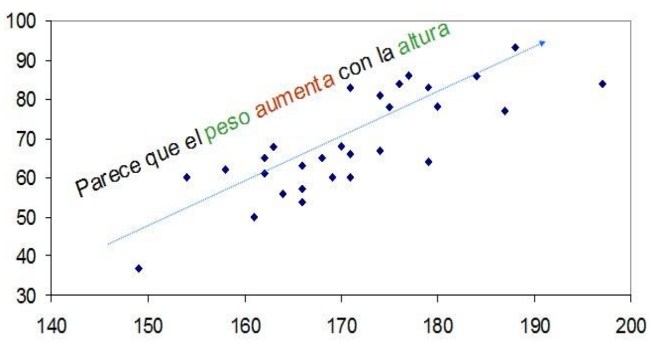
## Dependencia estadística entre tipos de variables

Hasta este punto, en los diversos capítulos se ha abordado, desde los conceptos básicos, tablas de frecuencia, como un método para resumir los datos. Luego los gráficos, como parte de la representación de las tablas, así como la interpretación de estos en los diversos contextos. Además, se abordaron las medidas de tendencia central y de variabilidad, como el inicio de los procesos para el análisis de los datos.

Sin embargo, el análisis de los datos trasciende a solo las características básicas de las medidas de tendencia central o de variabilidad, por lo que “a partir de algunos parámetros, nos proponemos estudiar las relaciones entre variables” (Laguna, 2014, p. 1).

Es importante recordar aspectos relevantes de tema de relaciones y álgebra, donde se definen dos conceptos: dependencia e independencia de variables. En una relación entre dos conjuntos, se puede establecer una dependencia, normalmente definida por la variable $y$ y una independencia, normalmente definida por la variable $x$.

Como lo estipula Laguna (2014) con el gráfico de dispersión, el cual es una herramienta muy útil en el análisis de correlaciones pues coloca la información en forma de puntos en un plano cartesiano, se evidencia la relación entre la altura de varias personas con respecto a su masa. Siendo una variable (masa) dependiente de la otra (altura).




Fuente: Laguna. (2014). Correlación y regresión lineal.

Existen tipos de coeficientes de correlación, como el punto biserial, correlación de Pearson, entre otros. Así como, la regresión lineal como “técnicas estadísticas más utilizadas para investigar la relación entre dos variables continuas $x$ y $y$” (Laguna, 2014, p. 1).

Laguna (2014) enuncia que las técnicas de que demuestran una correlación y las que realizan una regresión son intrísecamente relacionadas, pero que en esencia son distintas.

> Por un lado, el coeficiente de correlación determina el grado de asociación lineal entre X e Y, sin establecer a priori ninguna direccionalidad en la relación

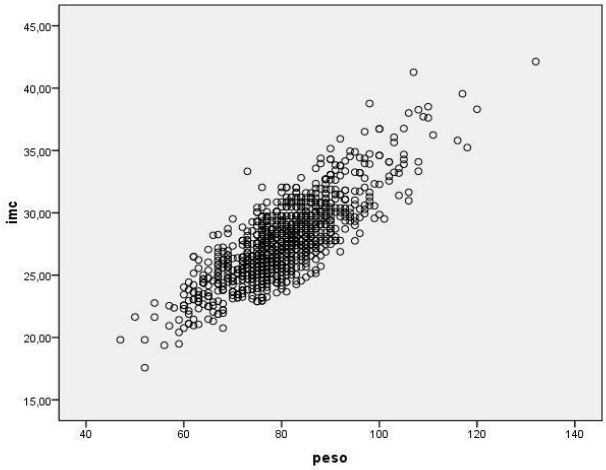
entre ambas variables. Por el contrario, la regresión lineal simple permite cuantificar el cambio en el nivel medio de la variable Y conforme cambia la variable X, asumiendo implícitamente que X es la variable explicativa o independiente e Y es la variable respuesta o dependiente. (Laguna, 2014, p.2)

La regresión se basa en el análisis de la recta que mejor se ajuste a la mayor cantidad de datos, de manera que se puedan predecir comportamientos.

## I. Correlación

De acuerdo con Laguna (2014) “la finalidad de la correlación es examinar la dirección y la fuerza de la asociación entre dos variables cuantitativas.” (p. 2) de manera que a partir de estas se pueda verificar “la relación entre ellas y si, al aumentar el valor de una variable, aumenta o disminuye el valor de la otra variable” (Laguna, 2014, p. 2).

Por ejemplo, Laguna (2014) presenta el siguiente diagrama de dispersión.




Fuente: Laguna. (2014). Correlación y regresión lineal

En este gráfico se puede observar claramente que cuando la masa de una persona aumenta, el índice de masa corporal también lo hará. No hace falta saber el dato exacto correspondiente a cada uno de los puntos, sino que con una observación del gráfico se puede inferir sobre el comportamiento de las variables dadas.

Sin embargo, como todo en matemática, una visualización no es suficiente, se debe realizar algún cálculo que demuestre ineludiblemente que existe tal relación. Por esa razón existen métodos para calcular la correlación y para efectos de la lectura se abordará el coeficiente de correlación punto biserial y el coeficiente de correlación de Pearson.

**Correlación Punto Biserial**

De acuerdo con González (2018) el coeficiente de correlación punto biserial, se utiliza cuando se requiere conocer la relación existente entre dos variables. Se representa con la notación $r_{b_p}$, el cual está definido en el intervalo $[-1, 1]$, suponiendo que los datos presenten un comportamiento normal; si esto no ocurre, se debe recurrir a otro tipo de análisis denominado, no paramétrico.

El uso de las variables es un factor predominante en el uso de este tipo de correlación, este utiliza dos variables, una dicotómica y otra politómica. La dicotómica se define como una variable cualitativa que se codifica en dos valores disjuntos y por otra parte la politómica se define como aquella que puede tomar valores superiores a dos. (Pérez y Merino, 2015)

<u>Ejemplos de variables dicotómicas</u>

La variable sexo: femenino, masculino.

La variable aprobar o no aprobar.

La variable si o no.

La variable acepta o no acepta.

Una vez definidas las variables y el contexto de la relación que se requiere estudiar, se proceden con la aplicación de la fórmula.

**Fórmula con un promedio de proporción**

Dada una variable dicotómica con dos componentes $a$ y $b$. Se define la proporción $p$ como la división de todos los eventos favorables de $a$, con respecto al total.

De forma análoga se define la proporción $q$ como la división de todos los eventos favorables de $b$, con respecto al total.

Además, es importante recalcar que esta fórmula utiliza la desviación estándar de los datos, que, para este caso en particular, por el tipo de variables con las que se trabaja se define de la siguiente manera:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2}$$

De esta forma se define la fórmula como:

$$r_{b_p} = \frac{\bar{x}_p - \bar{x}}{s_x} \cdot \sqrt{\frac{p}{q}}$$

Donde:

$\bar{x}_p$: corresponde al promedio de todas las componentes $a$ de la variable dicotómica.

$\bar{x}$: corresponde al promedio de todos los datos en el estudio.

**Fórmula con ambos promedios de proporción**

Esta fórmula utiliza los promedios de ambas dimensiones de la variable dicotómica, ($a$ y $b$).

Se calcula de la siguiente manera.

$$r_{b_p} = \frac{\bar{x}_p - \bar{x}_q}{s_x} \cdot \sqrt{p \cdot q}$$

Donde:

$\bar{x}_p$: corresponde al promedio de todas las componentes $a$ de la variable dicotómica.

$\bar{x}_q$: corresponde al promedio de todas las componentes $b$ de la variable dicotómica.

Una vez realizados los cálculos requeridos por las fórmulas se consideran resultados dentro del intervalo $[-1, 1]$ y a continuación se detallan las interpretaciones relacionadas al tipo de número encontrado cuando se concluye el procedimiento.

1. **Correlación inversa perfecta**

   Esta se da cuando $r_{b_p} = -1$, es decir, la variable $a$ se relaciona de forma inversa con la variable $b$, por lo que se interpreta que el incremento de $a$ provoca, de forma exacta y precisa, el decrecimiento de $b$ o viceversa, dependiendo del contexto del problema.

2. **Correlación inversa**

   Esta se da cuando $-1 < r_{b_p} < 0$, esta implica que la variable $a$ se relaciona de forma inversa con la variable $b$ en mayor medida si se acerca a $-1$ y en menor medida si se acerca a $0$. Siempre cumpliendo que, si una aumenta, la otra disminuye o viceversa, de acuerdo con el contexto del problema.

3. **No existe evidencia de relación**

   Esta se da cuando $r_{b_p} = 0$, pues no se evidencia una relación entre las variables $a$ y $b$, por lo que el estudio no es concluyente de la forma en la que se plantearon las variables.

4. **Correlación directa**

   Esta se da cuando $0 < r_{b_p} < 1$ e implica que la variable $a$ se relaciona de forma directa con la variable $b$ en mayor medida si se acerca a $1$ y en menor medida si se acerca a $0$, siendo la forma directa, que si una aumenta la otra también lo hace o, al contrario.

5. **Correlación directa perfecta**

   Esta se da cuando $r_{b_p} = 1$, es decir, la variable $a$ se relaciona de forma directa con la variable $b$, por lo que se interpreta que el incremento de $a$ provoca el incremento de $b$ o el decrecimiento de ambos, dependiendo del contexto del problema.

**Ejemplo 1**

Para una Prueba de Actitud Académica en una universidad, el departamento de psicología, junto con la Escuela de Matemática la aplicaron a 11 estudiantes. Una vez aplicada el comité decide analizar la inuencia del ítem 15, con respecto a la calificación final de esta.

El comité defina que utilizarán el coeficiente de correlación entre la variable *nota final* obtenida por estudiante, representada por la variable $y$ la cual es politómica y la variable acierto o no acierto del ítem, representada por la variable $x$, la cual es dicotómica, por lo que deciden codificarla de la siguiente manera: acierto (0) y no acierto (1).

Los datos del estudio se resumen en la siguiente tabla.

| **Acierto del ítem** | **Nota final del estudiante** |
|---|---|
| 0 | **90** |
| 0 | **72** |
| 1 | 80 |
| 1 | 70 |
| 1 | 62 |
| 0 | **85** |
| 1 | 87 |
| 0 | **88** |
| 0 | **80** |
| 0 | **79** |
| 1 | 76 |

Solución 1

Como la variable acierto es dicotómica, $p$ corresponde a la proporción de aciertos (0) en el ítem estudiado.

$$p = \frac{6}{11} \approx 0{,}5454$$

De forma análoga, $q$ corresponde a la proporción de no aciertos (1) en el ítem estudiado.

$$q = \frac{5}{11} \approx 0{,}4545$$

Luego, se calcula el promedio de la proporción $p$, la cual se determina con respecto a las notas obtenidas por todas las personas que acertaron el ítem.

$$\bar{x}_p = \frac{\mathbf{90 + 72 + 85 + 88 + 80 + 79}}{6} \approx 82{,}3333$$

Se continua con el cálculo del promedio de todas las notas, obtenidas por los estudiantes.

$$\bar{x} = \frac{90 + 72 + 80 + 70 + 62 + 85 + 87 + 88 + 80 + 79 + 76}{11} = 79$$

Por último, se calcula la desviación estándar aplicando la fórmula, se tiene que

$$s_x = 8{,}1575$$

De esta forma se tiene que:

$$r_{b_p} = \frac{82{,}33 - 79}{8{,}1575} \cdot \sqrt{\frac{0{,}5454}{0{,}4545}} = 0{,}447$$

Por lo tanto, se concluye que existe una correlación directa donde el acierto del ítem analizado se ve relacionado con una buena nota, en la prueba de actitud.

**Ejemplo 2**

Un director desea realizar un estudio que determine si existe relación entre la nota de presentación de sus 24 estudiantes de undécimo, representados por la variable $y$ con respecto al ingreso a las universidades públicas del país, representada por la variable $x$. Para ello le asigna al departamento de Psicología que realice el estudio.

Para esto, recolectan los datos de ambas variables definiendo, en la variable dicotómica de ingreso como $1$ para los que si ingresaron y $2$ para los que no ingresaron.

Los datos se resumen en la siguiente tabla.

| Ingreso | Nota Presentación | Ingreso | Nota Presentación |
|---|---|---|---|
| 1 | 90 | 1 | 95 |
| 2 | 88 | 2 | 93 |
| 1 | 89 | 1 | 91 |
| 1 | 90 | 1 | 85 |
| 1 | 98 | 1 | 88 |
| 2 | 98 | 2 | 89 |
| 1 | 81 | 2 | 98 |
| 1 | 80 | 2 | 79 |
| 2 | 82 | 1 | 97 |
| 2 | 90 | 2 | 95 |
| 1 | 87 | 1 | 93 |
| 1 | 94 | 1 | 93 |

Solución 2

Como la variable ingreso es dicotómica, $p$ corresponde a la proporción de ingresos (1) en una universidad.

$$p = \frac{15}{24} \approx 0{,}625$$

De forma análoga, $q$ corresponde a la proporción de no ingresos (2) en una universidad.

$$q = \frac{5}{11} \approx 0{,}4545$$

Luego, se calcula el promedio de la proporción $p$, la cual se determina con respecto a las notas de presentación de todas las personas que ingresaron a la universidad.

$$\bar{x}_p = \frac{90 + 89 + 90 + 98 + 81 + 80 + 87 + 94 + 95 + 91 + 85 + 88 + 97 + 93 + 93}{15} \approx 90$$

Se continua con el cálculo del promedio de todas las notas de presentación de los estudiantes de undécimo en ese colegio.

$$\bar{x} = 90{,}125$$

Por último, se calcula la desviación estándar aplicando la fórmula, se tiene que

$$s_x = 5{,}689$$

De esta forma se tiene que:

$$r_{b_p} = \frac{90 - 90{,}125}{5{,}689} \cdot \sqrt{\frac{0{,}625}{0{,}375}} = -0{,}0133$$

Por lo tanto, se concluye que existe una correlación inversa, muy poco evidente por su cercanía a 0, por lo que podría no ser concluyente que una mejor nota influye en el ingreso a una universidad.

**Correlación de Pearson**

El siguiente apartado, está tomado de acuerdo con el análisis de Laguna (2014), quien estipula que el estimador muestral más utilizado para evaluar la asociación lineal entre dos variables $X$ y $Y$ es el coeficiente de correlación de Pearson (r).

Este se define como un "índice que mide si los puntos tienen tendencia a disponerse en una línea recta." (Laguna, 2014, p. 3) Al igual que el coeficiente de correlación punto biserial, puede presentarse con valores comprendidos en el intervalo $[-1, 1]$.

Requiere que los datos tengan un comportamiento normal, pues se trabaja con la varianza, esto se tipifica como un método paramétrico para analizar la correlación.

La definición formal, tal como lo plasma Laguna (2014) "la covarianza muestral entre $X$ y $Y$ dividida por el producto de las desviaciones típicas de cada variable" (p. 3) de acuerdo con la fórmula.

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

Laguna (2014) agrega que:

> La expresión matemática para el coeficiente de correlación de Pearson parece compleja, pero esconde un planteamiento que, en el fondo, es sencillo: "r" estará próximo a 1 (en valor absoluto) cuando las dos variables X e Y estén intensamente relacionadas, es decir, al aumentar una aumenta otra y viceversa. A este concepto de variación al unísono se le llama *covarianza*. (p. 3)

**Covarianza**

Se puede observar que el numerador de la expresión anterior tiene la simbología $S_{xy}$ la cual se define como covarianza muestral entre las variables. Específicamente, "indica si la posible relación entre dos variables es directa o inversa." (Laguna, 2014, p. 3) Se calcula con base en la siguiente fórmula.

$$S_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Laguna (2014) analiza que:

> Así, si valores altos (o bajos) de X tienden a asociarse con valores altos (o bajos) de Y, el producto de las desviaciones tenderá a ser positivo y la covarianza será positiva. Por el contrario, si valores altos de una variable se relacionan con valores bajos de la otra variable, el producto de las desviaciones tenderá a ser negativo y la covarianza será negativa. (p. 3)

En particular $S_{xy}$ puede tener tres posibles escenarios

1. $\boldsymbol{S_{xy} < 0}$

   Cuando una variable crece, la otra tiende a decrecer o viceversa.

2. $\boldsymbol{S_{xy} = 0}$

   No se puede determinar una relación lineal.

3. $\boldsymbol{S_{xy} > 0}$

   Las dos variables crecen o decrecen a la vez.

Una vez realizados los cálculos requeridos por la fórmula se consideran los resultados dentro del intervalo $[-1, 1]$ y a continuación se detallan las interpretaciones relacionadas al tipo de número encontrado cuando se concluye el procedimiento.

1. **Correlación inversa perfecta**

   Esta se da cuando $r = -1$, es decir, la variable $X$ se relaciona de forma lineal muy fuerte pero inversa con la variable $Y$. Se representa con una recta lineal decreciente.

2. **Correlación inversa**

   Esta se da cuando $-1 < r < 0$, esta implica que la variable $X$ se relaciona de forma inversa con la variable $Y$ en mayor medida si se acerca a $-1$ y en menor medida si se acerca a $0$. Siempre siendo una recta decreciente.

3. **No existe evidencia de relación**

   Esta se da cuando $r = 0$, pues no se puede afirmar una relación lineal entre las variables $X$ y $Y$, según Laguna (2014) las variables son incorreladas.

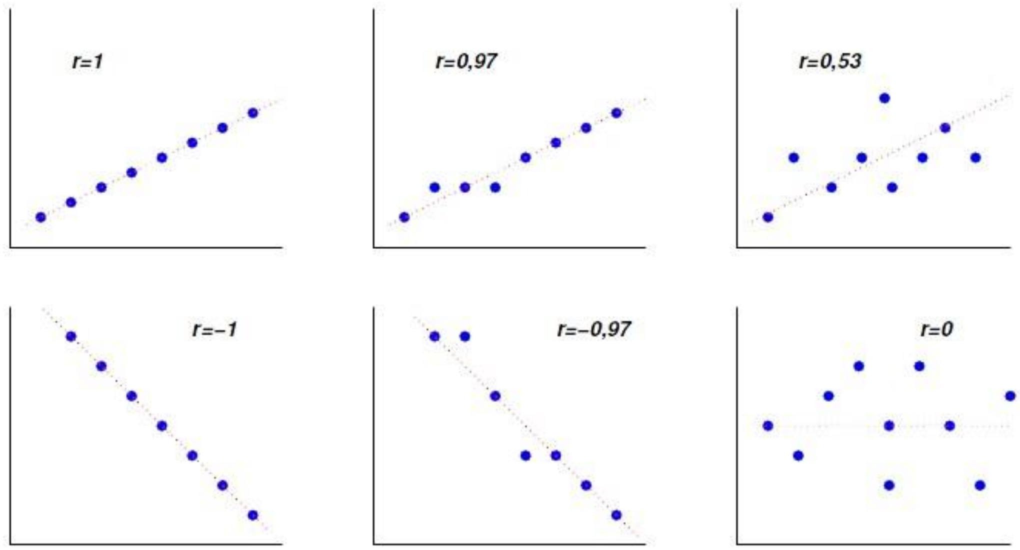
4. **Correlación directa**

   Esta se da cuando $0< r < 1$, esta implica que la variable $X$ se relaciona de forma directa con la variable $Y$ en mayor medida si se acerca a $1$ y en menor medida si se acerca a $0$. Siempre siendo una recta creciente.

5. **Correlación directa perfecta**

   Esta se da cuando $r = 1$, es decir, la variable $X$ se relaciona de forma lineal muy fuerte directamente con la variable $Y$. Se representa con una recta lineal creciente.

La siguiente imagen ejemplifica los puntos anteriores.




Fuente: Laguna. (2014). Correlación y Regresión Lineal

Un punto relevante, dento de lo que enmarca Laguna (2014) “en la correlación no se distingue la variable dependiente de la independiente, la correlación de $X$ con respecto a $Y$ es la misma que la correlación de $Y$ con respecto a $X$” (p. 5).

Se debe recordar, que la interpretación siempre depende del contexto de la situación que se presente. Sin embargo, Laguna (2014) enumera tres puntos de referencia para determinar un parámetro de interpretación.

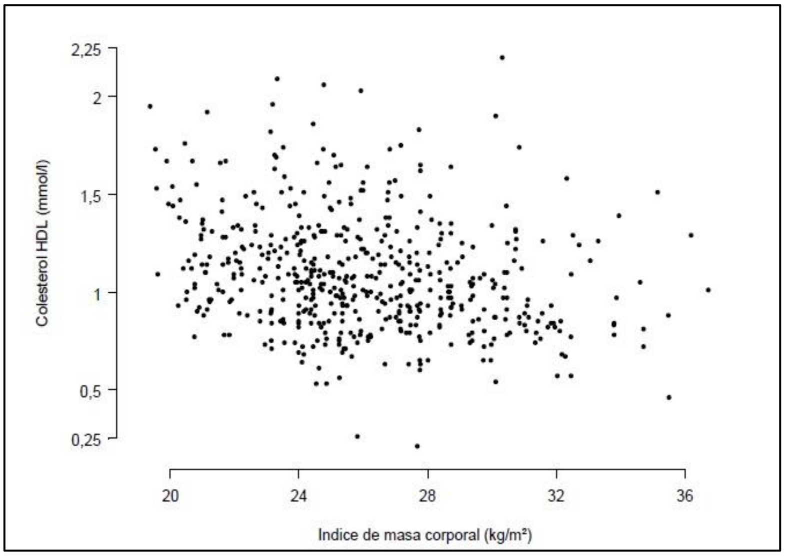
1. Correlación baja: menor al 0,30.
2. Correlación moderada: mayor a 0,30 y menos a 0,70.
3. Correlación alta: mayor a 0,70

**Condiciones para la aplicación del coeficiente de correlación de Pearson**

1. Variables cuantitativas: Ambas variables examinadas han de ser cuantitativas.
2. Normalidad: La normalidad de ambas variables es un requisito en el caso del coeficiente de correlación de Pearson.
3. Independencia: Las observaciones han de ser independientes, es decir, sólo hay una observación de cada variable para cada individuo.

**Ejemplo 3**[1]

En el gráfico, se presenta el diagrama de dispersión entre el índice de masa corporal, medida de obesidad que se obtiene de dividir el peso en kilogramos por la altura en metros al cuadrado, y el colesterol HDL en un estudio realizado a 533 individuos.




Fuente: Laguna. (2014). Correlación y Regresión lineal

[1] Tomado de: Laguna, C. (2014). Correlación y Regresión Lineal. http://www.ics-aragon.com/cursos/salud-publica/2014/pdf/M2T04.pdf

Solución 3

A simple vista, se aprecia un cierto grado de dependencia lineal negativa entre ambas variables; esto es, el colesterol HDL tiende a decrecer conforme aumenta el índice de masa corporal. Esta apreciación visual se confirma mediante el cálculo del coeficiente de correlación muestral de Pearson que indica una asociación lineal negativa moderada entre el índice de masa corporal y el colesterol HDL.

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

$$r = \frac{\frac{1}{532} \cdot \sum_{i=1}^{533}(x_i - \bar{x})(y_i - \bar{y})}{S_x \cdot S_y}$$

$$r = \frac{-0{,}285}{3{,}50 \cdot 0{,}295}$$

$$r = -0{,}276$$

En síntesis, de acuerdo con lo calculado, existe una relación inversa débil entre las variables índice de masa corporal y el colesterol.

## II. Regresión lineal

Un modelo de regresión es un modelo que permite describir cómo influye una variable $X$ llamada independiente, sobre otra variable $Y$ llamada dependiente. El objetivo es obtener estimaciones razonables para las variables a partir de una serie de pares ordenados.

Existen diversos tipos de relaciones entre variables. Algunas son funciones conocidas que no tienen duda de que sean exactamente calculados y que los valores no presentan grado de error.

Ejemplos de aplicación directa.

1. La relación entre la nota obtenida en un examen $X$ con respecto al puntaje obtenido en una prueba de actitud académica $Y$.

$$P(x) = 6x + 200$$

2. La relación en la conversión de grados Celsius a grados Fahrenheit.

$$F(C) = 1{,}8C + 32$$

Los ejemplos anteriores son exactos y precisos, es decir, que todos los puntos muestrales coinciden directamente con la recta dada. Pero no siempre todo se acomoda de tal forma que todos los puntos se encuentren contenidos en una recta.

A esta situación se le denomina no determinnista, puesto que algunos puntos no están sobre la recta y se adiciona una perturbación desconocida.

La idea central de la regresión lineal es buscar una recta lineal, que busca contener la mayor cantidad de puntos, por medio de la búsqueda de los parámetros clásicos de la función lineal.

$$f(x) = mx + b$$

Donde $m$ corresponde a la pendiente de la recta (grado de inclinación) y $b$ representa el punto de intersección con el eje de las ordenadas. (eje $y$)

Una vez definida la recta que contiene la mayor cantidad de puntos, se pueden determinar las siguientes consideraciones:

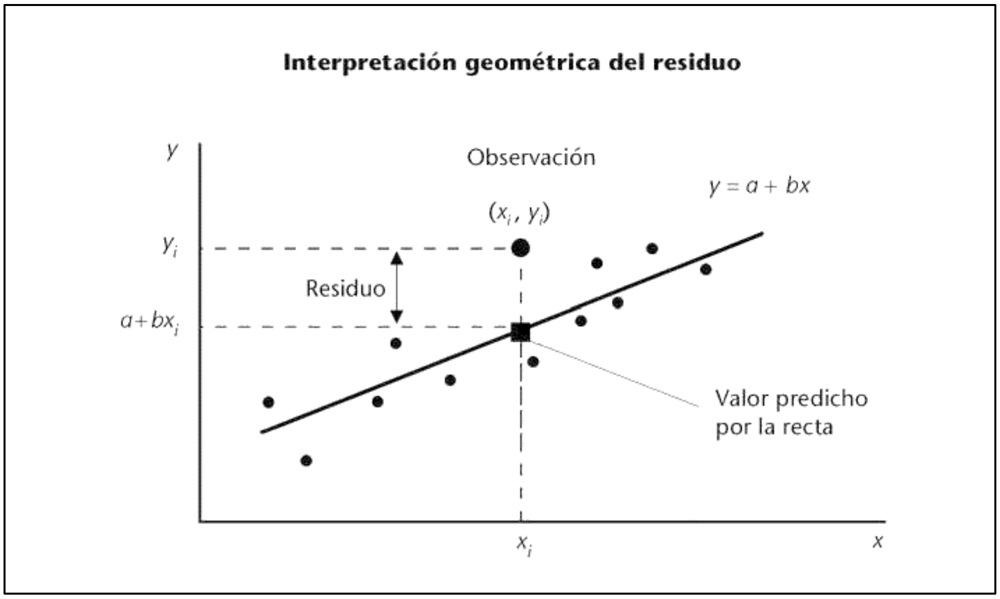
1. Si $m > 0$ se tiene una relación directa, es decir, si $X$ aumenta, $Y$ también lo hace.
2. Si $m < 0$ se tiene una relación inversa, es decir, si $X$ aumenta, $Y$ disminuye.

En muchas ocasiones no se encuentra una función exacta que se ajuste de forma exacta por lo que se puede generar un error.

Para minimizar el error se puede recurrir al método de mínimos cuadrados.

## III. Método de mínimos cuadrados

El error mencionado en el apartado anterior Laguna (2014) lo describe por medio de la siguiente imagen:




Fuente: Laguna. (2014). Correlación y Regresión Lineal

En la imagen se puede apreciar la recta y los puntos que la rodean. Particularmente, en el punto señalado se puede observar que dista una región con respecto a la recta, a eso se le llama residuo o el error en la precisión de la recta.

De acuerdo con Laguna (2014) la definición del método de mínimos cuadrados “consiste en buscar los valores de los parámetros $m$ y $b$ de manera que la suma de los cuadrados de los residuos sea mínima. Esta recta es la recta de regresión por mínimos cuadrados” (p. 9).

Al ser una función lineal, se requiere el cálculo de $m$ y de $b$, tal y como se muestra en la fórmula siguiente.

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y}{n} - m \cdot \frac{\sum x}{n}$$

**Ejemplo 4**

En un colegio se desea realizar una investigación para determinar si existe una relación entre el examen final del curso (x), con el promedio final (y). Los datos que se recolectados se evidencian en la siguiente tabla.

| x | y |
|---|---|
| 95 | 97 |
| 88 | 90 |
| 85 | 88 |
| 89 | 91 |
| 83 | 89 |
| 90 | 95 |
| 80 | 90 |

<u>Solución 4</u>

Con respecto a los datos, se deben realizar unos cálculos necesarios para aplicar las fórmulas.

Primero se realiza el cálculo de la multiplicación de las variables y la suma de estas, es decir, $\sum xy$. También la suma de los $x$ y la suma de los $y$, es decir, $\sum x$, $\sum y$.

| | x | y | xy |
|---|---|---|---|
| | 95 | 97 | 9215 |
| | 88 | 90 | 7920 |
| | 85 | 88 | 7480 |
| | 89 | 91 | 8099 |
| | 83 | 89 | 7387 |
| | 90 | 95 | 8550 |
| | 80 | 90 | 7200 |
| Suma | 610 | 640 | 55851 |

Ahora se realiza el cálculo de la variable independiente al cuadrado y la suma de todas esas, es decir, $\sum x^2$.

| | x | y | xy | $x^2$ |
|---|---|---|---|---|
| | 95 | 97 | 9215 | 9025 |
| | 88 | 90 | 7920 | 7744 |
| | 85 | 88 | 7480 | 7225 |
| | 89 | 91 | 8099 | 7921 |
| | 83 | 89 | 7387 | 6889 |
| | 90 | 95 | 8550 | 8100 |
| | 80 | 90 | 7200 | 6400 |
| Suma | 610 | 640 | 55851 | 53304 |

Con todos estos datos calculados se puede calcular la pendiente de la recta que mejor se adapte a los datos, es decir, $m$.

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$m = \frac{7 \cdot (55851) - (610) \cdot (640)}{7 \cdot (53304) - (610)^2}$$

$$m = 0{,}54182$$

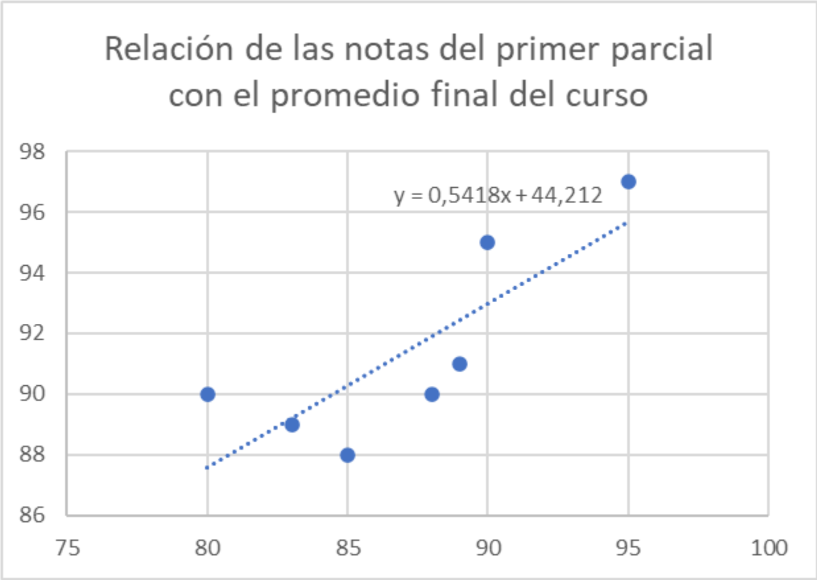
Luego se calcula el punto de intersección con el eje de las ordenadas, es decir, $b$

$$b = \frac{\sum y}{n} - m \cdot \frac{\sum x}{n}$$

$$b = \frac{640}{7} - 0{,}54182 \cdot \frac{610}{7}$$

$$b = 44{,}2120$$

Por último, se realizará el gráfico de dispersión, con la recta calculada.




**Ejemplo 5**

Considerando los siguientes datos, ajuste una recta por medio de mínimo cuadrados.

| | | | | | | |
|---|---|---|---|---|---|---|
| x | -3 | -1 | 1 | 3 | 5 | 7 |
| f(x) | 14 | 4 | 2 | 8 | 22 | 44 |

<u>Solución 5</u>

Con respecto a los datos, se deben realizar unos cálculos necesarios para aplicar las fórmulas.

Primero se realiza el cálculo de la multiplicación de las variables y la suma de estas, es decir, $\sum xy$. También la suma de los $x$ y la suma de los $y$, es decir, $\sum x$, $\sum y$.

| | **x** | **y** | **xy** |
|---|---|---|---|
| | -3 | 14 | -42 |
| | -1 | 4 | -4 |
| | 1 | 2 | 2 |
| | 3 | 8 | 24 |
| | 5 | 22 | 110 |
| | 7 | 44 | 308 |
| Suma | 12 | 94 | 398 |

Ahora se realiza el cálculo de la variable independiente al cuadrado y la suma de todas esas, es decir, $\sum x^2$.

| | x | y | xy | $x^2$ |
|---|---|---|---|---|
| | -3 | 14 | -42 | 9 |
| | -1 | 4 | -4 | 1 |
| | 1 | 2 | 2 | 1 |
| | 3 | 8 | 24 | 9 |
| | 5 | 22 | 110 | 25 |
| | 7 | 44 | 308 | 49 |
| Suma | 12 | 94 | 398 | 94 |

Con todos estos datos calculados se puede calcular la pendiente de la recta que mejor se adapte a los datos, es decir, $m$.

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$m = \frac{6 \cdot (398) - (12) \cdot (94)}{6 \cdot (94) - (12)^2}$$

$$m = 3$$

Luego se calcula el punto de intersección con el eje de las ordenadas, es decir, $b$

$$b = \frac{\sum y}{n} - m \cdot \frac{\sum x}{n}$$

$$b = \frac{94}{6} - 3 \cdot \frac{12}{6}$$

$$b = 9{,}6666$$

## Referencias bibliográficas


Laguna, C. (2014). Correlación y regresión lineal. http://www.ics-aragon.com/cursos/salud-publica/2014/pdf/M2T04.pdf

González, J. (2018). Coeficientes Correlación: Phi, Contingencia, Biserial, Spearman. Universidad Central de Venezuela. http://saber.ucv.ve/bitstream/10872/18490/1/Otros%20Coeficientes%20Correlaci%C3%B3n_FHE_UCV.pdf

Pérez, J. Merino, M. (2015). Definición de variable cualitativa. https://definicion.de/variable-cualitativa/

San Marcos

**www.usanmarcos.ac.cr**
San José, Costa Rica