



Estadística: Regresión y correlación

Matemáticas I

1.º Bachillerato

Contenidos

Estadística

Regresión y correlación

1. Introducción

Últimamente, y de manera cada vez más frecuente, habrás escuchado en los informativos noticias que nos hablan de **previsiones** sobre la economía de una cierta región, el comercio o la evolución de la población mundial a un medio o largo plazo. Seguro que te has planteado: ¿cómo se puede pronosticar un acontecimiento que aún está por ocurrir?



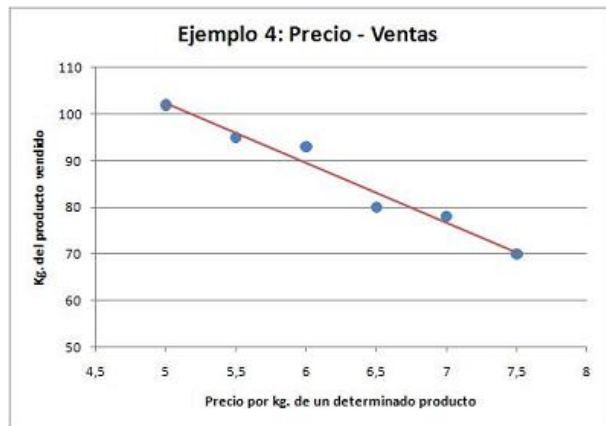
Imagen de Tumisu en [Pixabay](#). [Pixabay License](#)

Por ejemplo, algunas de las noticias que has podido oír referentes a nuestra historia sobre el cambio climático, han sido, entre otras, las siguientes:

- La población mundial en el año 2050 aumentará en más de mil millones de personas, y con ello las emisiones de CO₂, según datos de la ONU.
- La Agencia Estatal de Meteorología prevé que la temperatura en España aumentará hasta en 6°C en 2100.
- El nivel del mar podría aumentar hasta un metro para 2100 según científicos australianos.

En este tema veremos que, si entre dos variables de una determinada situación existe cierta relación, podremos hacer previsiones sobre el comportamiento futuro de éstas.

2. Correlación. Covarianza y Coeficiente de Correlación de Pearson

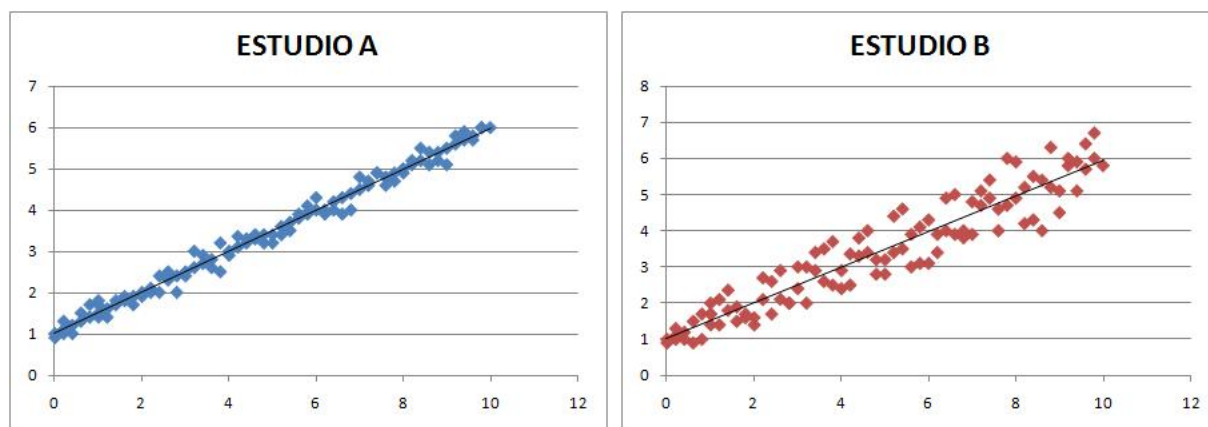


¿Recuerdas las nubes de puntos del tema anterior? Con ellas podíamos determinar si había algún tipo de relación entre dos variables, y en ese caso decidir si la relación era positiva o negativa (al aumentar la primera, aumentaba o disminuía respectivamente la segunda).

Esta relación la definiremos como **correlación**. Y ésta puede ser:

- **Correlación funcional:** si existe una relación funcional entre las variables X e Y. Es decir, podemos calcular los valores de Y a partir de los de X, con una función.
- **Correlación positiva o directa:** existe cierta relación entre ambas variables, y al aumentar los valores de X también aumentan los de Y (primer gráfico).
- **Correlación negativa o inversa:** existe cierta relación entre las variables, pero al aumentar los valores de X disminuyen los de Y (segundo gráfico).
- **Correlación nula:** no existe ningún tipo de relación entre ambas.

Observa ahora las dos siguientes gráficas obtenidas al hacer el mismo estudio sobre dos poblaciones diferentes.



Como puedes ver, en ambas la correlación es positiva. Pero, ¿crees que en los dos casos existe la misma dependencia entre las variables? Por lo que se puede apreciar en las gráficas, en el Estudio A la dependencia parece ser más fuerte que en el Estudio B. Por tanto, debe existir alguna forma para **medir** la correlación.

A continuación, definiremos dos parámetros, la **covarianza** y el **coeficiente de correlación lineal**, que nos servirán para establecer esta medida.



Para saber más

Recuerda, que el símbolo \sum se utiliza para indicar de forma abreviada la suma de varios números.

Por ejemplo, para acortar la siguiente suma $x_1 + x_2 + x_3 + x_4$, escribimos: $\sum_{i=1}^4 x_i$.

Si queremos abreviar la suma $x_1 \cdot f_1 + x_2 \cdot f_2 + x_3 \cdot f_3 + \dots + x_n \cdot f_n$, escribimos: $\sum_{i=1}^n x_i \cdot f_i$.

Antes de seguir, tienes que repasar algunas de las medidas que viste en el tema 2: **la media** y la **desviación típica**. Recuerda que si una variable X toma los valores x_1, x_2, \dots, x_n con frecuencias f_1, f_2, \dots, f_n , entonces:

- La media se calcula:

$$\bar{x} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n x_i f_i}{N}$$

- En tanto que la desviación típica se halla:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n f_i \cdot x_i^2}{N} - \bar{x}^2}$$



Importante

La **Covarianza** de una Variable Estadística Bidimensional (X,Y) la denotaremos como σ_{xy} y se calcula

$$\sigma_{xy} = \frac{\sum_{i=1}^n x_i y_i f_i}{N} - \bar{x} \bar{y}$$

donde f_i es la frecuencia de cada par (x_i, y_i) , N es el total de pares de valores y \bar{x} e \bar{y} son las medias marginales de cada variable.

Interpretación: el signo de la covarianza nos permitirá saber el tipo de correlación.

- Si la covarianza es positiva, la correlación será directa.
 - Si la covarianza es negativa, la correlación será inversa.
-

En la siguiente presentación puedes ver cómo calcularemos la covarianza a partir de una tabla simple.

http://www.slideshare.net/slideshow/embed_code/4879552

[Covarianza](#) from [saulvalper](#)



Importante

b) El **Coefficiente de Correlación de Pearson** es el parámetro que nos va a decir si la correlación es débil o fuerte. Para calcularlo, necesitamos conocer el valor de las desviaciones típicas marginales de cada variable σ_x y σ_y . Su valor siempre estará entre -1 y 1. Su expresión es la siguiente:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Interpretación: Según los valores de r, tenemos cuatro casos:

- $r = 0$: No existe correlación.
- $r = 1$ ó $r = -1$: La correlación es perfecta.
- r próximo a 1 o -1: La correlación es fuerte.
- r próximo a 0: La correlación es débil.

Veamos cómo calcularla con el ejemplo anterior.

http://www.slideshare.net/slideshow/embed_code/4881949

[Cálculo del Coeficiente de Correlación de Pearson](#) from [saulvalper](#)



Comprueba lo aprendido

Vamos a recuperar los Estudios A y B del comienzo de este apartado. Vas a calcular el coeficiente de correlación de Pearson para comprobar si el resultado se corresponde con lo que habíamos supuesto por la gráfica.

Para realizar los cálculos te daremos la suma de los totales de cada columna, como en los ejemplos anteriores. El valor de $N=104$.

ESTUDIO A	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
TOTALES	502,2	354,56	2165,43	3336,04	1437,21

ESTUDIO B	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
TOTALES	502,2	353,83	2157,92	3336,04	1451,59

Observación: por cuestión de redondeo, puede que los resultados que te ofrecemos no coincidan exactamente con los tuyos. Elige los que más se aproximen.

a) ¿Cuáles de los siguientes valores corresponden a las medias marginales de X e Y para ambos estudios?

- ☐ Estudio A: Media X = 4,83
Estudio B: Media Y = 4,53
- ☐ Estudio A: Media Y = 3,41
Estudio B: Media X = 4,83

Solución

- 1. Incorrecto
- 2. Correcto

b) ¿Qué valores corresponden a las desviaciones típicas marginales de X e Y para esos dos estudios?

- ☐ Estudio A: Desviación típica de X = 2,96
Estudio B: Desviación típica de Y = 2,15
- ☐ Estudio A: Desviación típica de Y = 1,48
Estudio B: Desviación típica de X = 2,96

Solución

- 1. Incorrecto
- 2. Correcto

c) ¿Cuáles de las siguientes covarianzas corresponden a la de esos estudios?

- ☐ Estudio A: Covarianza = 4,36
Estudio B: Covarianza = 4,32
- ☐ Estudio A: Covarianza = 4,58
Estudio B: Covarianza = 4,52

Solución

- 1. Correcto
- 2. Incorrecto

d) Calcula los correspondientes Coeficientes de Correlación de Pearson.

☐

Estudio A: $r = 0,99$

Estudio B: $r = 0,95$

- ☐ Estudio A: $r = 0,97$
Estudio B: $r = 0,91$

Solución

1. Correcto
2. Incorrecto

e) Por último, a la vista de los resultados, podemos afirmar que:

- ☐ Ambos estudios tienen correlación positiva, y la del Estudio A es más fuerte que la del Estudio B
- ☐ Ambos estudios tienen correlación positiva, y la del Estudio A es más débil que la del Estudio B

Solución

1. Correcto
 2. Incorrecto
-

3. Regresión. Rectas de regresión y estimación



Imagen de Cinzia A. Rizzo en [Flickr](#). Licencia [CC 2.0 by-nc-nd](#)



Imagen de ATBravo en [Flickr](#). Licencia [CC 2.0 by](#)

Si miramos las hojas o las pisadas de estas fotos como si fueran puntos de nuestras gráficas, ¿podrías decir dónde caería la próxima hoja? ¿o dónde estaría la pisada que sigue el camino? En el primer caso sería mucha casualidad que acertases, pues la correlación es nula. Sin embargo, en el segundo caso la correlación es muy fuerte y seguro que podrías dar una respuesta bastante ajustada.

Eso vamos a hacer en este apartado, predecir resultados a partir de los datos que tenemos. Ten en cuenta que esas predicciones serán más fiables cuanto mayor sea el coeficiente de correlación.

A lo largo de este tema y del anterior, has visto que en algunas gráficas aparecía una recta que se ajustaba a la nube de puntos. Esa línea se llama **recta de regresión** y es la **recta que mejor se ajusta a la nube de puntos**.

En la siguiente escena tienes un ejemplo de una nube de puntos y su **recta de regresión** (de color azul), así como otra recta (de color rojo) que hemos llamado aleatoria. Esta última recta la puedes cambiar moviendo los puntos rojos. En la escena también se expresa la aproximación que se consigue con cada recta de la nube de puntos.

Vamos a hacer un par de pruebas:

- Mueve los puntos rojos e intenta hallar una mejor aproximación a la nube de puntos que la expresada por la recta de regresión. ¿Es posible encontrarla?
- Si mueves los puntos azules, estarás modificando la nube de puntos y por tanto la recta de regresión. Modifica los puntos azules que quieras y repite el intento de mejorar la aproximación que viene dada por la recta de regresión.

Hasta ahora sólo hemos hablado de la recta de regresión desde un punto de vista visual, a continuación veremos cuál es la ecuación de la recta que mejor "ajusta" una nube de puntos.



Importante

Para una variable estadística bidimensional, la **recta de regresión de Y sobre X** viene dada por la ecuación:

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

Una vez que tenemos la ecuación de la recta, podemos "predecir" valores de la variable que no conozcamos sustituyendo el valor de x. Veámoslo en el siguiente ejemplo.



Caso práctico

En el apartado anterior calculaste los parámetros para una Variable en la que se relacionaba el peso y la tensión arterial de los pacientes de un Centro de Salud.

Los valores que obtuviste fueron los siguientes:
 $\bar{x} = 96,4$ $\bar{y} = 147,2$ $\sigma_x = 19,52$ $\sigma_{xy} = 440,42$

a) Calcula la recta de regresión de Y sobre X.

Basta con que sustituyas cada uno de los valores en la expresión de la recta de regresión.

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}) = 147,2 + \frac{440,42}{19,52^2}(x - 96,4)$$

Operando, obtendrás la ecuación de la recta en forma estándar $y = m \cdot x + n$. En nuestro caso, quedará:

$$y = 1,16x + 35,77$$

b) Usando la recta de regresión anterior, haz una estimación sobre la tensión que tendrá una persona que pese $x=100$ kg.

Si sustituimos $x=100$ en la ecuación de la recta anterior, nos quedaría:

$$y = 1,16 \cdot 100 + 35,77 = 151,77$$

Este valor, 151,77 será la estimación de presión arterial para una persona que pese 100 kg.

Como hemos visto, la recta de regresión de Y sobre X es la que mejor aproxima los valores de la variable Y a una recta, pero también podemos aproximar los valores de X. Para ello usamos otra recta, la **recta de regresión de X sobre Y**, que tiene la siguiente ecuación:

$$x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

En la siguiente escena tienes la representación gráfica de las dos rectas. Cuanto más próximo sea el valor del Coeficiente de correlación de Pearson a 1 o -1, mayor coincidencia habrá entre ambas rectas. Puedes modificar los puntos para ver cómo varían ambas rectas.

El punto verde es el **Centro de Gravedad** de la distribución, y sus coordenadas son (\bar{x}, \bar{y}) .

<https://tube.geogebra.org/material/iframe/id/2022873/width/711/height/524/border/888888/rc/false/ai/false/sdz/false/s>



Comprueba lo aprendido

Utilizando los datos del Estudio B del apartado anterior, calcula las ecuaciones de las rectas de regresión de Y sobre X y de X sobre Y.

$$\bar{x} = 4,83$$

$$\bar{y} = 3,40$$

$$\sigma_x = 2,96$$

$$\sigma_y = 1,54$$

$$\sigma_{xy} = 4,32$$

- ☐ Recta de regresión de Y sobre X: $y = 0,49x + 1,02$
- ☐ Recta de regresión de Y sobre X: $y = 0,49x + 2,01$
- ☐ Recta de regresión de X sobre Y: $x = 1,82y - 3,16$
- ☐ Recta de regresión de X sobre Y: $x = 1,82y - 1,36$

Solución

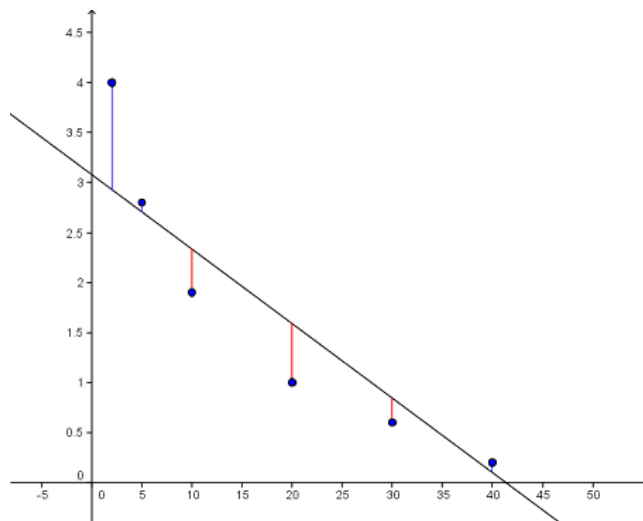
1. Correcto
2. Incorrecto
3. Incorrecto
4. Correcto



Para saber más

¿Cómo se consigue la recta que **mejor se ajusta**? Fíjate en el gráfico de la derecha.

Si tomamos una recta cualquiera, medimos la distancia en vertical desde cada punto hasta la recta. Si el punto está por encima de la recta, el resultado es positivo (segmento azul), y si queda por debajo, el resultado es negativo (segmento rojo).



Lo que queremos es que la suma de esas distancias sea lo menor posible para que el ajuste sea mejor, pero si sumamos números positivos con negativos unos anularán a los otros.

Un método para conseguir que todas las distancias sean positivas es elevarlas al cuadrado. Ahora sí podemos sumar todos los resultados y buscar la recta que hace que esa suma sea lo menor posible.

Este método, que se llama **Método de los mínimos cuadrados**, es el que nos da la recta de regresión que vamos a estudiar en este apartado.

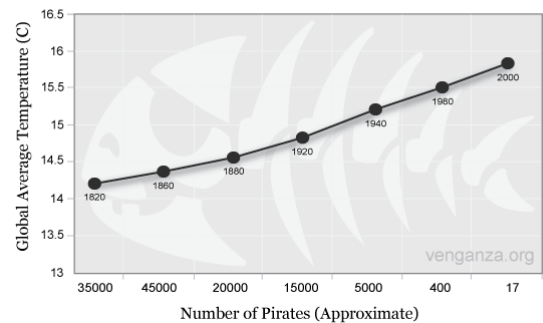


Curiosidad

Aunque hemos visto que la correlación indica que una variable esté relacionada con otra, esto no quiere decir que exista una relación de causa y efecto entre una y otra. Mira los siguientes ejemplos:

- Según la DGT, el 20% de los motoristas fallecidos en accidentes de tráfico no llevaban casco.
- La mayoría de los accidentes de tráfico se producen entre vehículos que ruedan a una velocidad moderada.
- Los días de luna llena se produce un aumento en el número de nacimientos.
- El cambio climático provoca el aumento de tornados en el hemisferio norte.

Global Average Temperature Vs. Number of Pirates



En el primer y segundo caso, no se puede establecer que sea mejor no llevar casco o circular a una velocidad excesiva, pues lo que no se dice es que la mayoría de los motoristas llevan casco, y la mayoría de los coches circulan a una velocidad moderada, de ahí que sea también mayor el número de accidentes.

El tercer ejemplo es el típico caso en el que, a pesar de haber sido refutadas con estudios estadísticos, siguen formando parte de las leyendas urbanas.

El cuarto, es un ejemplo de mal uso de las estadísticas para crear buenos titulares. Los estudios indican que, aunque haya aumentado el número de tornados en el hemisferio norte, no ha sido así a nivel mundial y no se puede establecer que el aumento de tornados sea causa directa del cambio climático.

El gráfico que tienes arriba, de la web www.venganza.org, representa el aumento de las temperaturas globales y la disminución en el número de piratas en los últimos siglos. Es un ejemplo muy curioso de cómo se pueden relacionar dos variables que no tienen nada que ver, y obtener una correlación muy fuerte ¿Han desaparecido los piratas debido al cambio climático?

Resumiendo, como nos explican en [Microsiervos](#), hay que aprender a interpretar los estudios estadísticos y a ser crítico con las noticias que nos llegan a diario.

Resumen



Importante

La relación entre dos variables la definiremos como **correlación**. Y ésta puede ser:

- **Correlación funcional:** si existe una relación funcional entre las variables X e Y. Es decir, podemos calcular los valores de Y a partir de los de X, con una función.
 - **Correlación positiva o directa:** existe cierta relación entre ambas variables, y al aumentar los valores de X también aumentan los de Y (primer gráfico).
 - **Correlación negativa o inversa:** existe cierta relación entre las variables, pero al aumentar los valores de X disminuyen los de Y (segundo gráfico).
 - **Correlación nula:** no existe ningún tipo de relación entre ambas.
-



Importante

La **Covarianza** de una Variable Estadística Bidimensional (X,Y) la denotaremos como σ_{xy} y se calcula

$$\sigma_{xy} = \frac{\sum_{i=1}^n x_i y_i f_i}{N} - \bar{x} \bar{y}$$

donde f_i es la frecuencia de cada par (x_i, y_i) , N es el total de pares de valores y \bar{x} e \bar{y} son las medias marginales de cada variable.

Interpretación: el signo de la covarianza nos permitirá saber el tipo de correlación.

- Si la covarianza es positiva, la correlación será directa.
 - Si la covarianza es negativa, la correlación será inversa.
-



Importante

El **Coeficiente de Correlación de Pearson** es el parámetro que nos va a decir si la correlación es débil o fuerte. Para calcularlo, necesitamos conocer el valor de las desviaciones típicas marginales de cada variable σ_x y σ_y . Su valor siempre estará entre -1 y 1. Su expresión es la siguiente:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Interpretación: Según los valores de r , tenemos cuatro casos:

- $r = 0$: No existe correlación.
- $r = 1$ ó $r = -1$: La correlación es perfecta.
- r próximo a 1 o -1: La correlación es fuerte.
- r próximo a 0: La correlación es débil.



Importante

Para una variable estadística bidimensional, la **recta de regresión de Y sobre X** viene dada por la ecuación:

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

Una vez que tenemos la ecuación de la recta, podemos "predecir" valores de la variable que no conozcamos sustituyendo el valor de x . Veámoslo en el siguiente ejemplo.

La **recta de regresión de X sobre Y**, que tiene la siguiente ecuación:

$$x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

El punto de coordenadas (\bar{x}, \bar{y}) se denomina **Centro de Gravedad** de la distribución.

Aviso legal

Las páginas externas no se muestran en la versión imprimible

<http://www.juntadeandalucia.es/educacion/permanente/materiales/index.php?aviso#space>