



Aprendiendo idiomas VIII

María Jesús Vázquez Gallo



1. **Introducción.**
2. **Correlación.**
3. **Coeficiente de correlación lineal.**
4. **Regresión.**
5. **Continuará...**

Ejemplo. Se realiza un ensayo de marketing para estudiar la relación entre el tiempo que requiere un comprador para decidir su compra y el número de presentaciones distintas del producto exhibidas.

El tiempo empleado en decidir se registra para todos los participantes en el estudio.

¿Aportan los datos evidencia suficiente de que el tiempo empleado en decidir está linealmente relacionado con el número de presentaciones?

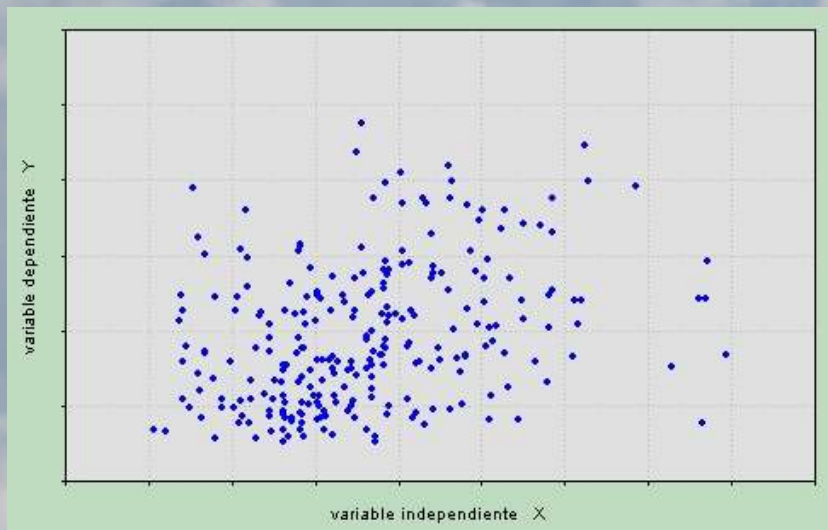
¿Podríamos predecir qué tiempo corresponde a cierto número de presentaciones?



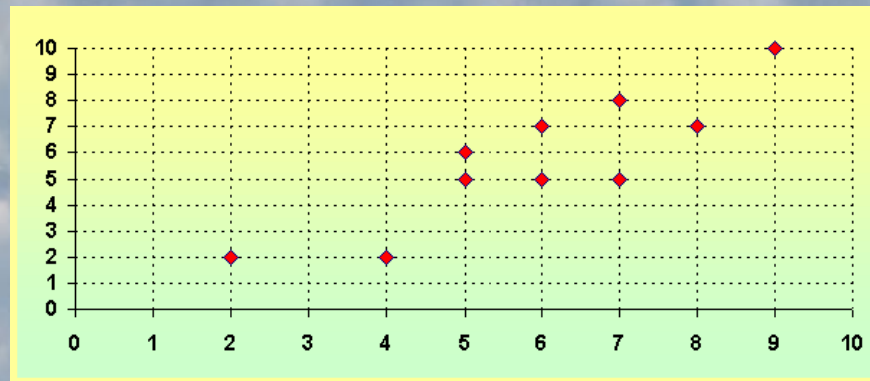
Éste es un ejemplo en el que tratamos de determinar si existe alguna relación entre dos variables, es decir, si el valor de una de ellas influye de algún modo en el de la otra.

Cuando esto ocurre, decimos que existe **correlación** entre las variables y , habitualmente, nos gustaría predecir el valor de una conocido el de la otra.

Los **métodos de regresión** permiten construir **modelos explicativos** de este tipo de relaciones: son modelos que representan la dependencia de una (la **variable respuesta o dependiente, y**) respecto de la otra (la **variable explicativa o independiente, x**).



Una primera forma de averiguar si existe correlación entre dos variables es dibujar en un plano los puntos que corresponden a las parejas (x,y) . El gráfico obtenido recibe el nombre de **nube de puntos** o **diagrama de dispersión**.



Observando este diagrama, ¿te parece que el valor de la variable vertical depende del valor de la variable horizontal?

Parece que la vertical aumenta a medida que la horizontal lo hace. Pero no siempre ocurre: por ejemplo, al pasar en horizontal de 2 a 4, la vertical no varía.

2. Correlación

La apreciación visual de la nube de puntos no parece suficiente para decidir la existencia de correlación entre dos variables.

¿Cómo hacerlo?

¿Recuerdas? En **Aprendiendo idiomas I**, cuando analizábamos una única variable aleatoria considerábamos, entre otras medidas importantes, la **media**, como medida de centralización y la **varianza**, como medida de dispersión.

La **varianza** es igual la media de los cuadrados de las diferencias entre cada valor de la variable y la media. Su raíz cuadrada, la **desviación típica**, se interpreta como la desviación en promedio de los n datos con respecto a la media.

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



Estamos analizando una **muestra de dos variables aleatorias**.
Lo que queremos es medir cómo se desvían los datos de una variable con respecto a otra.

La generalización de la **varianza** al caso de dos variables se llama **covarianza**.

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



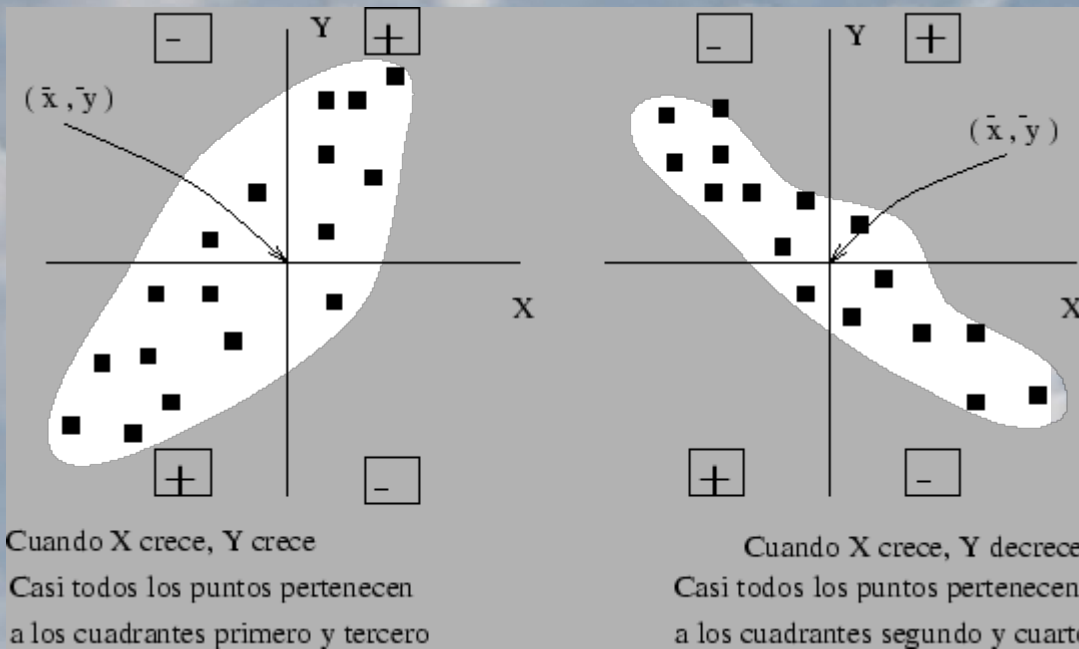
Observa que:

- cuando $x=y$, la **covarianza** se reduce a la **varianza**.
- la **covarianza** se puede expresar como

$$C_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} .$$

Geoméricamente:

En la nube de puntos, el punto correspondiente a las medias muestrales, indica el centro de gravedad de la nube.



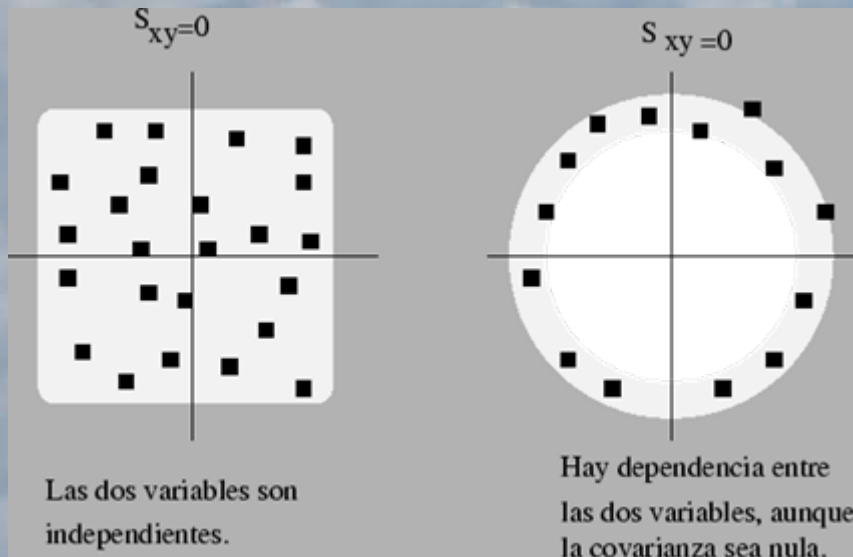
Trasladando el origen de coordenadas al centro de gravedad de la nube, la dividimos en 4 cuadrantes.

Comprueba, que los puntos del 1er y 3er cuadrante contribuyen positivamente al valor de la covarianza, y los del 2º y 4º lo hacen negativamente.



Entonces:

- Si hay mayoría de puntos en el 1er y 3er cuadrante, la **covarianza será no negativa**, y esto se interpreta como que la variable **Y** tiende a aumentar cuando lo hace **X**;
- Si la mayoría de puntos están repartidos entre el 2º y el 4º cuadrante **¿qué ocurre?**;
- Si los puntos se reparten con igual intensidad alrededor del centro de gravedad, entonces la **covarianza será nula**.



Observa la 2ª figura: la covarianza es 0 pero hay dependencia cuadrática entre las variables (los puntos corresp. tienen posiciones cercanas a una circunferencia). La covarianza no detecta dependencias no lineales.

3. Coeficiente de correlación lineal.

Para medir **dependencia lineal** entre dos variables, se utiliza, en lugar de la covarianza, el llamado **coeficiente de correlación lineal de Pearson**, denotado por r .



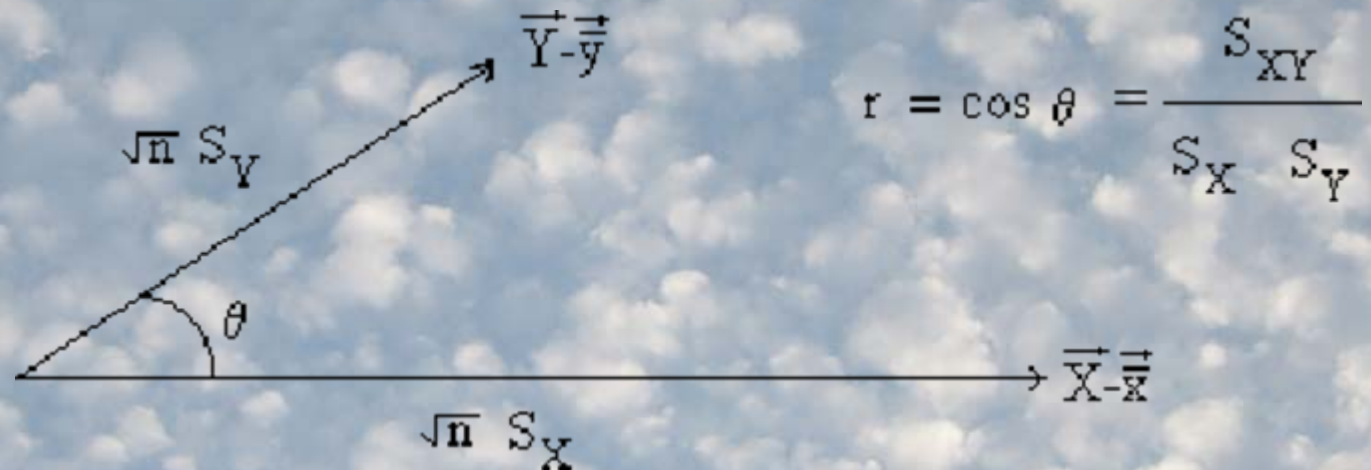
$$r_{xy} = \frac{C_{xy}}{S_x S_y} .$$

Se calcula dividiendo la covarianza por el producto de las desviaciones típicas de cada variable, y así se obtiene un coeficiente adimensional.

Para cada colección de valores de las dos variables, el coeficiente r nos va a permitir valorar si la correlación lineal es fuerte o débil, positiva (cuando crece la horizontal crece la vertical) o negativa (cuando crece la horizontal decrece la vertical).

Geométricamente:

El valor de r se puede interpretar como el **coseno del ángulo** que forman los vectores de las desviaciones de X y de Y con respecto a sus respectivas medias (esta interpretación está relacionada con el concepto de producto escalar).

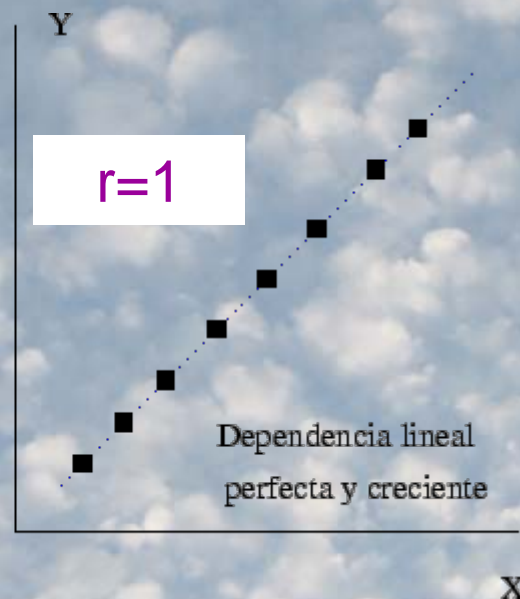
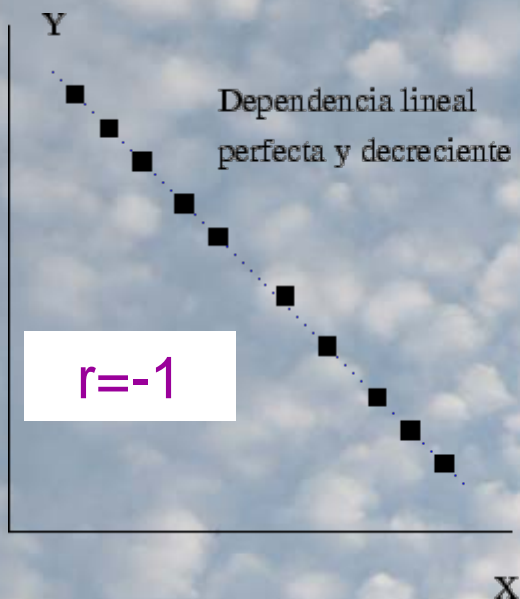


Observa que cuando estos vectores determinan la misma dirección, el ángulo que forman es de 0° o de 180° , es decir, el valor de r es 1 ó -1.

En ese caso, los vectores en cuestión serán proporcionales:

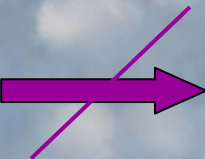
$$(\vec{Y} - \vec{\bar{y}}) = m \cdot (\vec{X} - \vec{\bar{x}})$$

para algún número m .



Pero entonces las desviaciones de los puntos dados por las variables X e Y , con respecto a sus medias son proporcionales, es decir, sucede que los puntos de la nube están alineados.

Cuando el coeficiente de correlación r está cerca de 1 o de -1, se dice que las variables están fuertemente correladas.

Pero correlación  causalidad

Que dos fenómenos estén fuertemente correlados no implica, necesariamente, que uno sea causa del otro.

Es muy frecuente que una correlación fuerte esté indicando que los dos caracteres dependen de un tercero que no ha sido medido. Este tercer carácter se llama **factor de confusión**.

Ejemplo: Que exista una fuerte correlación entre la recaudación de impuestos en Inglaterra y la criminalidad en el Japón, indica que ambos están ligados al aumento global de la población.

3. Coeficiente de correlación lineal.

Puede ser que una fuerte correlación exprese una verdadera causalidad, como entre el número de cigarrillos que se fuma al día y la aparición de un cáncer de pulmón.

Pero no es la correlación la que demuestra la causalidad, en todo caso, puede ser un indicio.

Ejemplo: La influencia del consumo del tabaco en el cáncer de pulmón ha sido científicamente demostrada en la medida en que se han podido analizar los mecanismos que hacen que el alquitrán y la nicotina induzcan errores en la reproducción del código genético de las células.



3. Coeficiente de correlación lineal.

Como el coeficiente de correlación lineal r es el coseno de un ángulo, cualesquiera que sean los valores de las variables, se cumple

$$-1 \leq r \leq 1$$

Si al calcular r en un caso práctico obtienes un valor no comprendido en ese rango, ¿qué harías?

Cuando no haya una relación lineal exacta entre las variables, el ángulo que forman los vectores de las desviaciones de cada variable respecto a la media tendrá un coseno menor que 1 en tamaño y, por tanto, r será un número entre -1 y 1, que no alcanza los valores extremos.

Cuanto más aplastada esté la nube de puntos, es decir, cuanto más se parezca la relación entre las variables a una relación lineal, más cerca de 1 ó de -1 estará el valor de r .

3. Coeficiente de correlación lineal.

En el caso extremo, los vectores en cuestión determinan direcciones perpendiculares entre sí, con lo cual $r=0$ y se dice que las variables son **incorreladas**.

Un coeficiente de correlación nulo o cercano a 0 **significa que no hay relación *lineal* entre los caracteres**, pero **no conlleva ninguna noción de independencia más general**.

Consideremos, por ejemplo, las dos muestras:

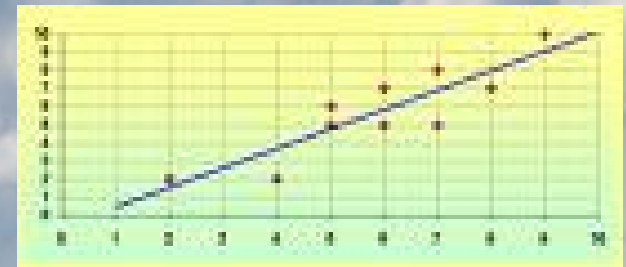
$$\begin{aligned}x &= (-3 \quad -2 \quad -1 \quad 0 \quad 1 \quad 2 \quad 3) \\y &= (3 \quad 2 \quad 1 \quad 0 \quad 1 \quad 2 \quad 3).\end{aligned}$$

Comprueba que su coeficiente de correlación es nulo y, sin embargo, hay una clara dependencia de **X** con respecto a **Y**:
La variable **Y** es el valor absoluto de la variable **X**.

4. Regresión.

Cuando resulte **razonable** suponer una **correlación lineal** entre dos variables, los **modelos de regresión lineal simple** nos permitirán:

- Determinar qué recta se ajusta mejor a la nube de puntos que representa a los datos, la **recta de regresión**.
- Utilizando la recta de regresión, **predecir el valor de la variable respuesta**, conocido el valor de la variable explicativa.



En general, para dependencias lineales entre un número mayor de variables, se utilizarán **modelos de regresión lineal múltiple**. Los **modelos de regresión general simples o múltiples** estudiarán relaciones de dependencia no necesariamente lineales entre dos o más variables.

