
scidata

Correlación

Directa o inversa

Instructor: Juan Luis Palacios Soto

palacios.s.j.l@gmail.com

Contenido

El siguiente problema al que nos enfrentaremos es: cómo sabemos si, dada una muestra con dos o más características (variables), hay alguna de ellas que dependa, de alguna manera, de alguna otra.

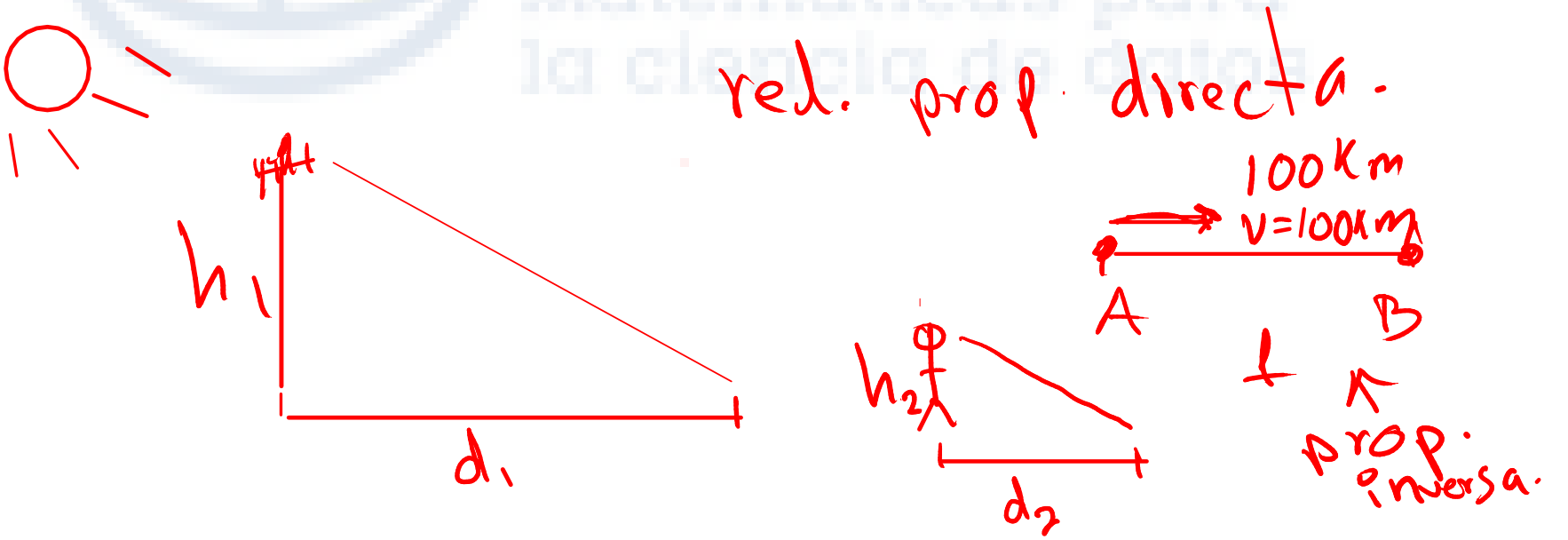
Por ejemplo, si en un hospital se hace un registro de pacientes y se registra su peso y nivel de colesterol en la sangre. ¿Habrá alguna relación matemática entre algunas de estas DOS mediciones?

Sean X_1 y X_2 dos características que medimos de una muestra. Lo que nos hemos preguntado en la diapositiva anterior es: ¿Existe una función f tal que $f(X_1) = X_2$? Más aún: en caso de existir, ¿Cómo es esa función?

La correlación es una medida de qué tan relacionadas están dos características dentro de una misma muestra.

Existe un tipo de correlación, la llamada lineal, que trata de establecer una medida bajo la cual, bajo ciertos criterios subjetivos, podamos decir que dicha relación entre las dos características existe y además es lineal. Es decir, $X_2 = aX_1 + b$ para algunos números a y b.

En otras palabras, la correlación lineal es una medida de qué tan relacionadas están, linealmente, dos características dentro de una misma muestra. Además, una propiedad importante que tiene es que carece de unidades y no se altera al sumar el mismo número a las dos características o multiplicar ambas por el mismo número.

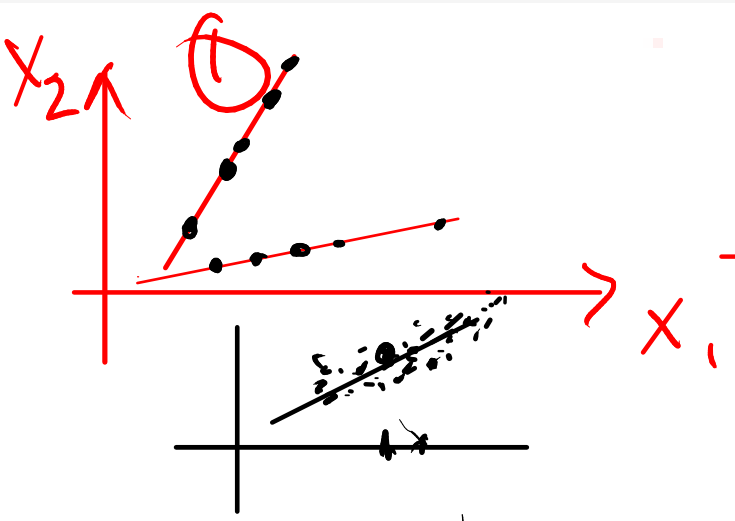


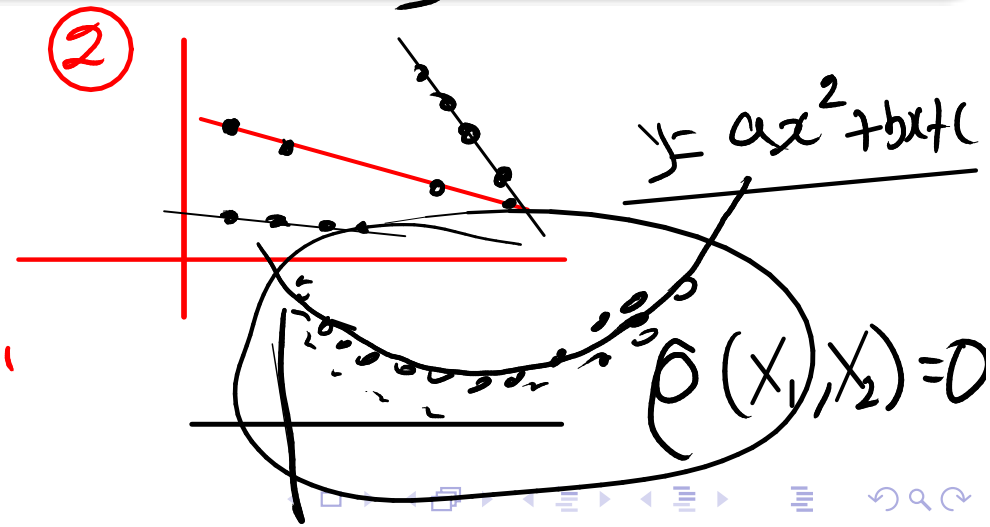
Tipos de correlación

Existe 6 tipos de correlación. Estos se pueden identificar haciendo un gráfico de puntos tomando como abscisas y ordenadas las dos características que se quieren analizar.

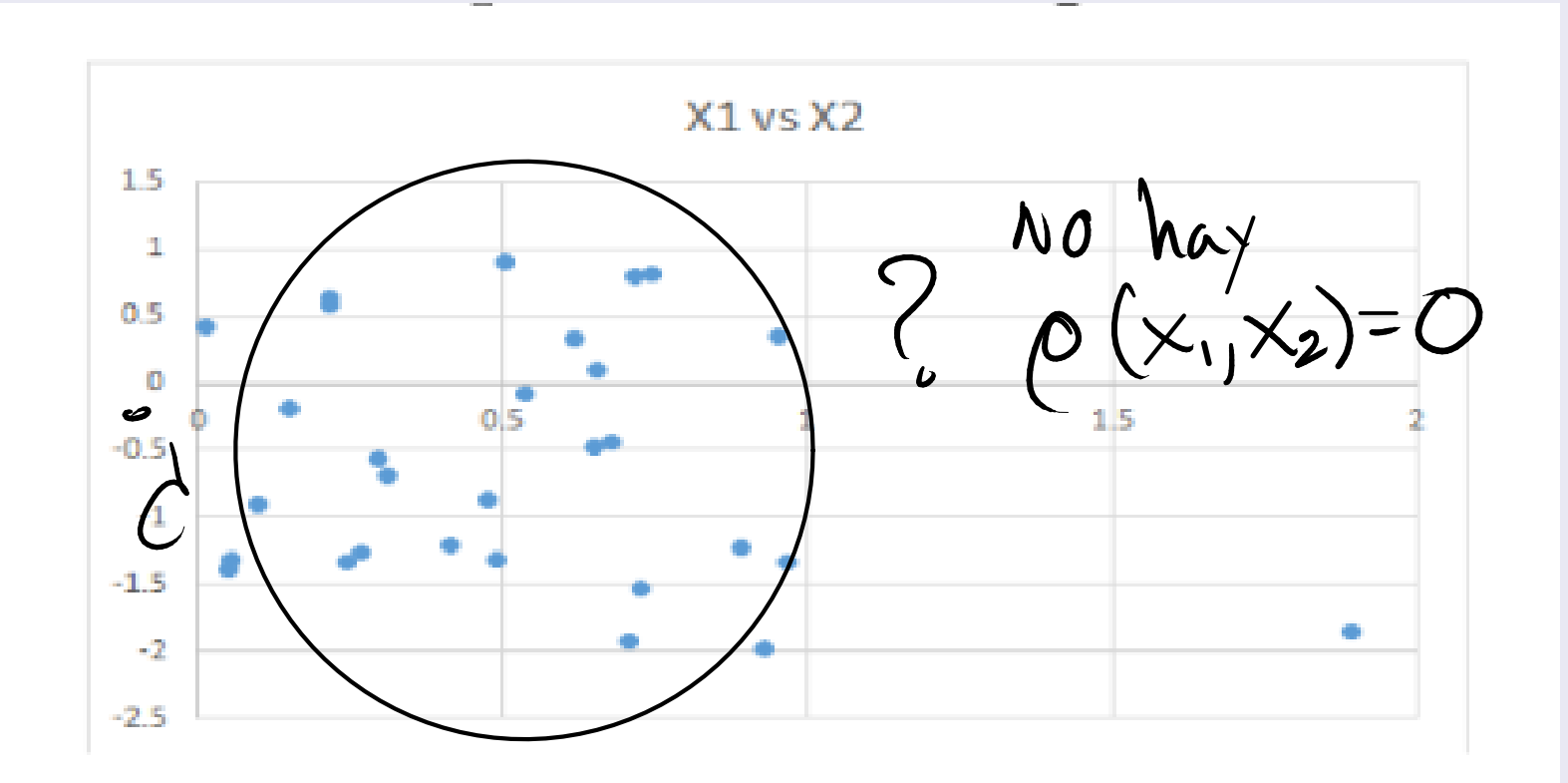
1. **Perfectamente positiva**. Los puntos forman una recta ascendente (si una característica aumenta de valor, la otra también). $\rho(X_1, X_2) = 1$
2. **Perfectamente negativa**. Los puntos forman una recta descendente (si una característica aumenta de valor, la otra disminuye). $\rho(X_1, X_2) = -1$
3. **Parcialmente positiva**. Los puntos no forman una recta, pero si una característica aumenta de valor, la otra también. $0 < \rho(X_1, X_2) < 1$
4. **Parcialmente negativa**. Los puntos no forman una recta, pero si una característica aumenta de valor, la otra disminuye. $-1 < \rho(X_1, X_2) < 0$
5. **Sin correlación**. No se ve una relación funcional de ningún tipo.
6. **Otra tipo de correlación**. Hay una relación funcional no lineal.

$\rho(X_1, X_1) = 0$





Recordemos que una nube de puntos es una gráfica de la forma



En Excel, se pueden crear directamente desde el panel de Gráficos del menú Insertar.

X_1	X_2
2	-1
3	5

Definición (Correlación)

Este es un número que mide el nivel de correlación lineal entre dos características de una misma muestra. Se trata de un número entre -1 y 1. Si su valor absoluto es 1, entonces hay correlación perfecta. Si el valor es 1, la correlación es además positiva; si es -1, es correlación negativa. Finalmente, si es 0, no hay correlación lineal, pudiendo haber otro tipo de correlación.

Sean X_1 y X_2 dos características de la misma muestra. Digamos $X_1 = (x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)})$ y $X_2 = (x_1^{(2)}, x_2^{(2)}, ..., x_n^{(2)})$. Si $\bar{X}_1$ y $\bar{X}_2$ son las medias aritméticas de las características, respectivamente, entonces definimos el coeficiente de correlación de Pearson, denotado por $\rho(X_1, X_2)$, como

$$-1 \leq \rho(X_1, X_2) = \frac{\sum_{i=1}^{n}(x_i^{(1)} - \bar{X}_1)(x_i^{(2)} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{n}(x_i^{(1)} - \bar{X}_1)^2}\sqrt{\sum_{i=1}^{n}(x_i^{(2)} - \bar{X}_2)^2}} \leq 1$$

En Excel es simplemente COEF.DE.CORREL(Característica 1, Característica 2)

X_1	X_2
$x_1^{(1)}$	$x_1^{(2)}$
$x_2^{(1)}$	$x_2^{(2)}$
⋮	⋮
$x_n^{(1)}$	$x_n^{(2)}$

CORREL

$\bar{X}_1$, $\bar{X}_2$

$\text{var}(x)\ \frac{n-1}{n}$

$\frac{1}{n-1}\ \Sigma$

Desventajas de la correlación

- Los coeficientes de correlación más utilizados sólo miden una relación lineal. Por lo tanto, es perfectamente posible que, si bien existe una fuerte relación no lineal entre las variables, ρ está cerca de 0 o igual a 0. En tal caso, una nube de puntos puede indicar aproximadamente la existencia o no de una relación no lineal.
- Hay que tener cuidado al interpretar el valor de ρ. Por ejemplo, se podría calcular ρ entre el número de calzado y la inteligencia de las personas, la altura y los ingresos. Cualquiera sea el valor de ρ, no tiene sentido y por lo tanto es llamado correlación de oportunidad o sin sentido.
- ρ no debe ser utilizado para decir algo sobre la relación entre causa y efecto. Dicho de otra manera, al examinar el valor de ρ podríamos concluir que las variables X_1 y X_2 están relacionadas. Sin embargo, el mismo valor de ρ no nos dice si X_1 causa a X_2 o al revés. La correlación estadística no debe ser la herramienta principal para estudiar la causalidad.