Correlación

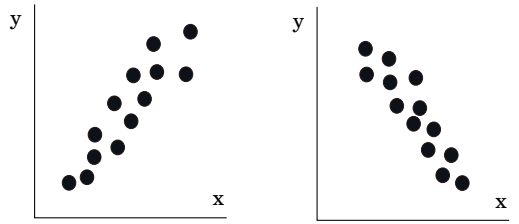
Josemari Sarasola

sigmalitika.hirusta.io

Concepto de correlación

La correlación es la covariación o variación conjunta de dos variables cuantitativas, por ejemplo el fenómeno que observamos cuando rentas familiares altas generalmente (estamos en estadística, y por tanto hablamos de pautas generales y no de leyes deterministas) conllevan mejores notas o rendimientos escolares de los hijos.

Tipos de correlación

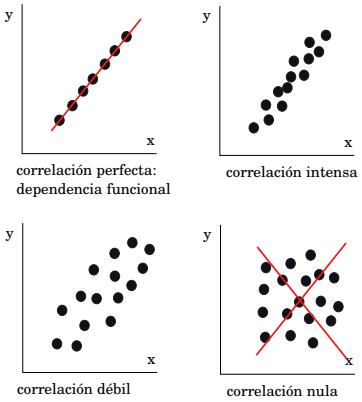


correlación directa o positiva: x sube, y sube

correlación inversa o negativa: x sube, y baja

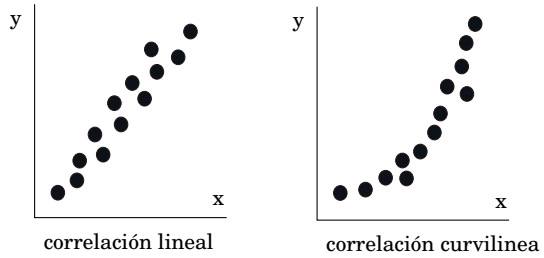
Correlación según dirección

Tipos de correlación



Correlación según intensidad

Tipos de correlación



Correlación según forma

Medidas de correlación

Covarianza

Fórmulas para la covarianza:

$$s_{xy} = \underbrace{\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n}}_{\text{formula original}} = \underbrace{\frac{\sum x_i y_i}{n} - \overline{xy}}_{\text{formula rápida}}$$

- Mide la dirección de correlación, pero no su intensidad.
- $s_{xy} > 0 \rightarrow$ correlación directa
- $s_{xy} < 0 \rightarrow$ correlación inversa
- Puede tomar cualquier valor real, positivo o negativo.
- Sus unidades son el producto de unidades de las variables correspondientes.

Medidas de correlación

Por qué la covarianza no mide la intensidad de correlación

Tenemos datos pesos en kg y alturas en metros de un grupo de personas. Calculamos la covarianza: 28 $kg \times m$. Pasamos los datos de altura a cms. Calculamos la covarianza: 2800 $kg \times cm$. Se ha multiplicado la covarianza por 100, ¿pero ha aumentado la intensidad de la correlación? Evidentemente, no.

Medidas de correlación

El coeficiente de Pearson

Para poder medir la intensidad de correlación, debemos estandarizar las variables, dejándolas de esta forma sin unidades:

$$z_x = \frac{x_i - \overline{x}}{s_x} \; ; \; z_y = \frac{y_i - \overline{y}}{s_y}$$

Y calculamos la covarianza entre las variables estándar: $s_{z_x z_y}$
Si desarrollamos esa covarianza, nos da el coeficiente de Pearson:

$$s_{z_x z_y} = r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Medidas de correlación

El coeficiente de Pearson

Lleva el nombre de (haz click) Karl Pearson, prominente estadístico inglés a caballo entre los siglos XIX y XX.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- $-1 \leq r_{xy} \leq 1$
- Cuidado: el coeficiente de Pearson indica sólo correlación lineal.
- Mide todo: la dirección y la intensidad de correlación.

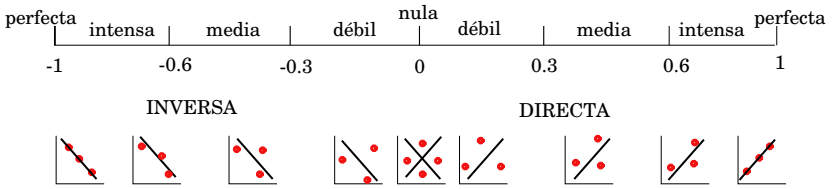
Medidas de correlación

El coeficiente de Pearson

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- Lo interpretamos así respecto a la dirección:
  - $r_{xy} > 0 \rightarrow$ correlación directa
  - $r_{xy} < 0 \rightarrow$ correlación inversa
- Y así respecto a la intensidad:
  - $r_{xy} = 0 \rightarrow$ correlación nula
  - $|r_{xy}| < 0{,}3 \rightarrow$ correlación débil
  - $0{,}3 < |r_{xy}| < 0{,}6 \rightarrow$ correlación media
  - $|r_{xy}| > 0{,}6 \rightarrow$ correlación intensa
  - $|r_{xy}| = 1 \rightarrow$ correlación perfecta

Medidas de correlación



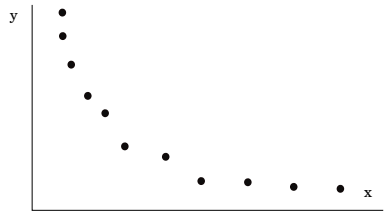
Interpretación del coeficiente de Pearson

Pero una matización: la interpretación hay que hacerla también teniendo en cuenta los resultados obtenidos en estudios similares al que se está realizando (de modo que un coeficiente de 0.85 puede ser considerado bajo si en otros estudios se ha obtenido un coeficiente siempre por encima de 0.9). Además hay que considerar el error muestral (haz click).

Medidas de correlación

Covarianza y Pearson son medidas de corr. lineal

La covarianza y el coeficiente de Pearson son medidas de correlación lineal. De esta forma, una débil correlación no implica falta de correlación, sino estrictamente que la correlación lineal es de nivel bajo, sin descartar que pueda existir una correlación curvilinea intensa.



$r_{xy} = 0{,}13$, indicando correlación lineal baja, pero la correlación curvilinea es intensa.

Medidas de correlación

Correlación no lineal y monotónica: el coeficiente de Spearman

Cuando la correlación es curvilinea y monotónica, esto es, siempre mostrando un perfil creciente o decreciente, para detectar y medir la correlación puede utilizarse el coeficiente de Spearman, que no es más que el coeficiente de Pearson pero aplicado a los rangos o números de orden de los valores dentro de cada variable (p.ej., si una variable tomas losvalores 4-9-6, los rangos correspondientes son 1-3-2). Se interpreta el valor del coeficiente de Spearman como el coeficiente de Pearson pero para correlaciones curvilineales y monotónicas.

Variables dicotómicas y correlación

Variables dicotómicas

Las variables dicotómicas son aquellas variables cualitativas que toman unicamente dos valores (p.ej., hombre/mujer, ha comprado/no ha comprado). Las variables cuantitativas pueden ser dicotomizadas estableciendo un valor umbral (aprobado, ≥ 5; suspendido, < 5).

Variables dicotómicas y correlación

Correlación con variable dicotómicas

Para determinar la correlación cuando intervienen variables dicotómicas, basta asignar los valores 0 y 1 de forma arbitraria a las dos categorías.

Variables dicotómicas y correlación

Correlación item-test

Decimos que un item o pregunta de test es adecuado (coherente, consistente) cuando sus resultados (correcto/incorrecto) están correlacionados notablemente con el puntaje (haz click) total (en definitiva, que el que contesta bien a la pregunta tiene tendencia a obtener puntaje alto en el test). La correlación item-test no es más que el coeficiente de Pearson entre la variable dicotómica 'ha contestado bien a la pregunta' y el puntaje total. Se considera que la pregunta es adecuada si tiene una correlación item-test positiva superior a 0.4. Una correlación cercana a 1 implica redundancia (la pregunta no añade nada al test y por tanto puede ser obviada) e inferior a 0.4 o negativa, falta de consistencia.

Correlación parcial

Estamos en una empresa y queremos analizar la relación entre edad y producción individual.

- Hipótesis: a más edad, más producción ($r_{ep} > 0$)
- Calculamos la correlación de Pearson entre ambas variables.
- Resultado: a más edad, menos producción ($r_{ep} < 0$)
- Explicación: el absentismo ejerce una influencia que distorsiona los resultados: a más edad, más absentismo y a más absnetismo, menos producción. Esta influencia supera a la relación original.
- Solución: calculamos la correlación parcial entre edad y producción eliminando el efecto absentismo:

$$r_{ep.a} = \frac{r_{ep} - r_{ea}r_{pa}}{\sqrt{1 - r_{ea}^2}\sqrt{1 - r_{pa}^2}}$$

Correlación parcial

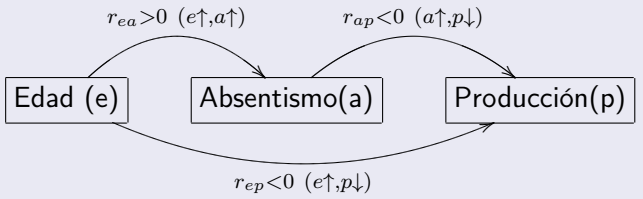


Figura: Efecto indirecto del absentismo (a) en la correlación entre edad (e) y producción (p): a mayor edad, mayor absentismo y a mayor absentismo menor producción (flechas superiores), dando como resultado un efecto global de a mayor edad, menor producción (flecha inferior), frente a lo que cabría esperar en un principio.

Correlación parcial

Fórmula general para la correlación parcial entre x e y, eliminando z:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

Cuando lo que parece, en realidad no es

Supongamos que detectamos una fuerte correlación entre la calificación en lengua inglesa de un grupo de alumnos y la ingesta semanal de cerveza. Es evidente que aún así no puede establecerse una relación entre las dos variables por falta de fundamento teórico. En esos casos, se dice que la correlación es espuria (espurio es sinónimo de falso), expresando de esta forma que ha surgido de forma casual, y que sería equivocado tomarla como real.