Introducción al análisis cuantitativo de datos lingüísticos

Bloque 3.2:
Análisis de correlaciones

Ezequiel Koile (MPI-SSH)
Carolina Gattei (IFIBA - CONICET)

Hasta acá:

- Describimos variables numéricas y categóricas
- Comparamos dos grupos

- Siempre hemos hablado de *una sola variable*:
  - Medidas centrales + dispersión
  - Comparar estos valores entre grupos.

¿Y ahora?

- Estudiamos la relación *entre distintas variables*

Correlación

- Dos variables están correlacionadas positivamente *sii* tanto X como Y crecen y decrecen juntas.
- Dos variables están correlacionadas negativamente *sii* crecen en direcciones opuestas
- Dos variables no están correlacionadas *sii* el cambio en una no afecta el cambio en la otra.

Coeficientes de correlación

- Nos interesa definir un número que nos diga
  - Si existe una correlación entre dos variables
  - Cuán fuerte es esta
  - En qué dirección va (positiva o negativa)
- Vamos a definir
  - **Un coeficiente paramétrico para variables intervalo o ratio con distribución normal**
  - **(Un coeficiente no paramétrico para los demás casos)**
- *Elegimos* definir estos coeficientes de manera que:
  - Van de −1 a+1
  - +1 significa correlación positiva perfecta
  - −1 significa correlación negativa perfecta
  - 0 significa variables no correlacionadas
- *¡¡NO RELACIONADOS CON SIGNIFICANCIA ESTADÍSTICA!!*

Coeficiente de correlación de Pearson

- Definido como la covarianza entre ambas variables dividida por ambas desviaciones estándar

Coeficiente de correlación de Pearson

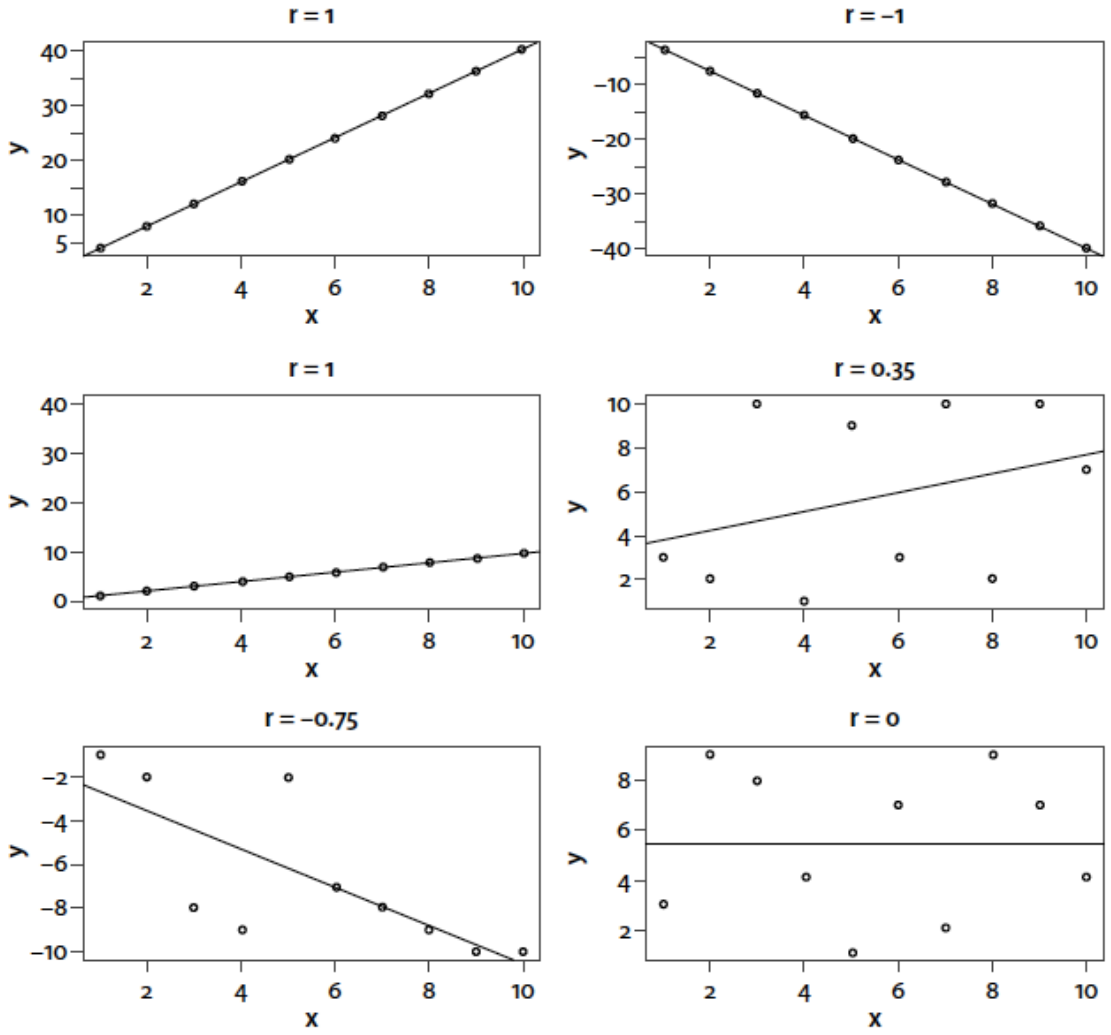


Figure 6.2. Several possible values of Pearson product-moment correlation coefficient *r*

Levshina 2015

Coeficiente de correlación de Pearson

PROS

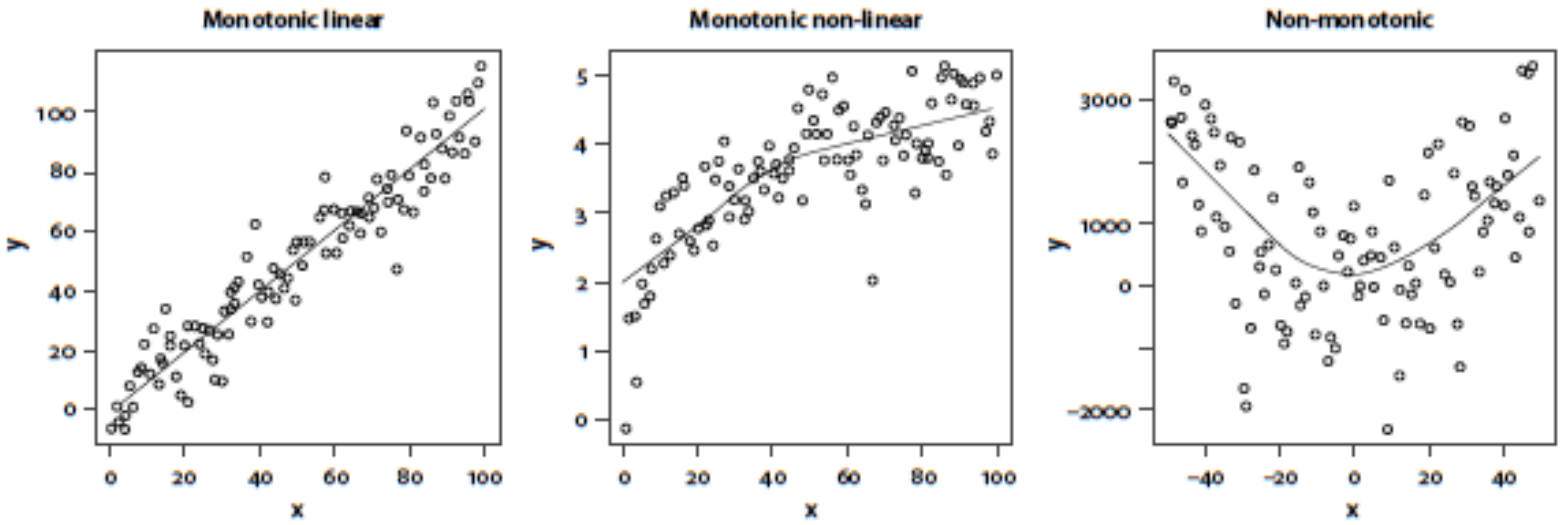
- Si el tamaño de la muestra es moderado o grande y la población es normal bivariada, el CCP es el estimador de mayor verosimilitud (es decir, es imposible construir un coeficiente de correlación mejor que este)

Coeficiente de correlación de Pearson

CONTRAS

- Útil solo si la relación entre las variables es:
  - Monótona
  - Lineal

Coeficiente de correlación de Pearson



Levshina 2015

Coeficiente de correlación de Pearson

CONTRAS

- Útil solo si la relación entre las variables es:
  - Monótona
  - Lineal
- Muy sensible a *outliers* (poco robusto)

Coeficiente de correlación de Pearson

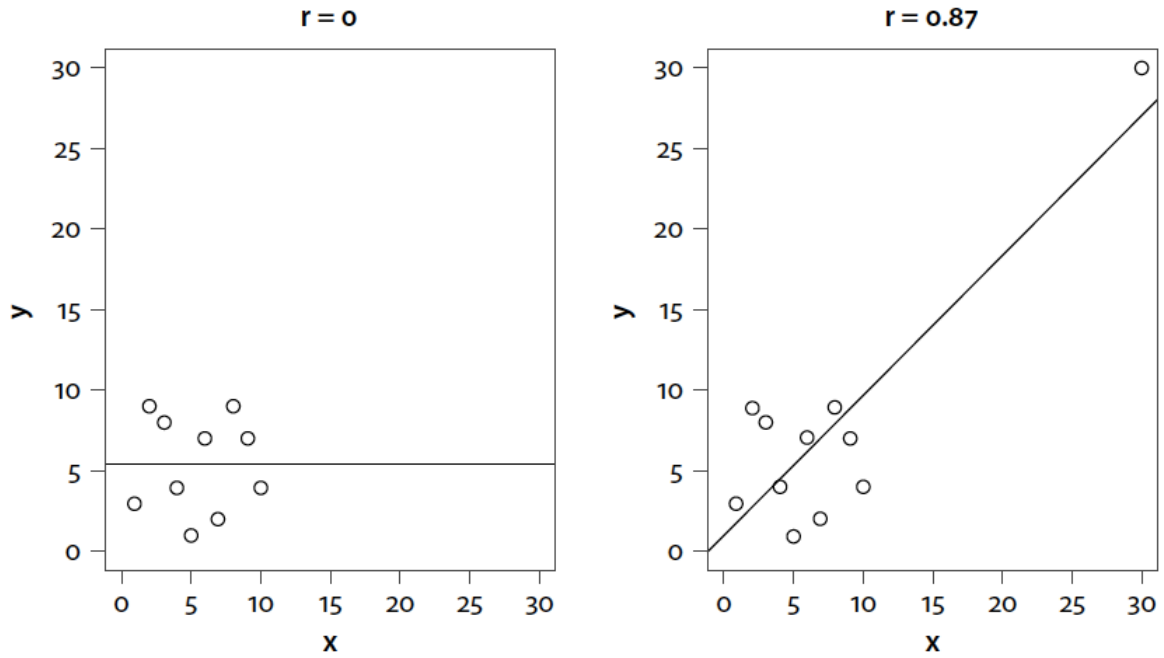


Figure 6.5. Impact of an outlier on the value of the Pearson *r*

Levshina 2015

Coeficiente de correlación de Pearson

¿Cuándo es estadísticamente significativo?

- La muestra se toma aleatoriamente de la población representada
- Ambas variables son al menos tipo intervalo
- Ambas variables forman una distribución normal bivariada y/o el tamaño de la muestra es grande (30 observaciones o más)
- **Homocedasticidad (homoscedasticity) en los residuos**: la relación entre las variables es de igual naturaleza a lo largo del rango de ambas variables.
- **Sin autocorrelación**: el valor de una variable no de**pende de su valor anterior o posterior**

Coeficiente de correlación de Pearson

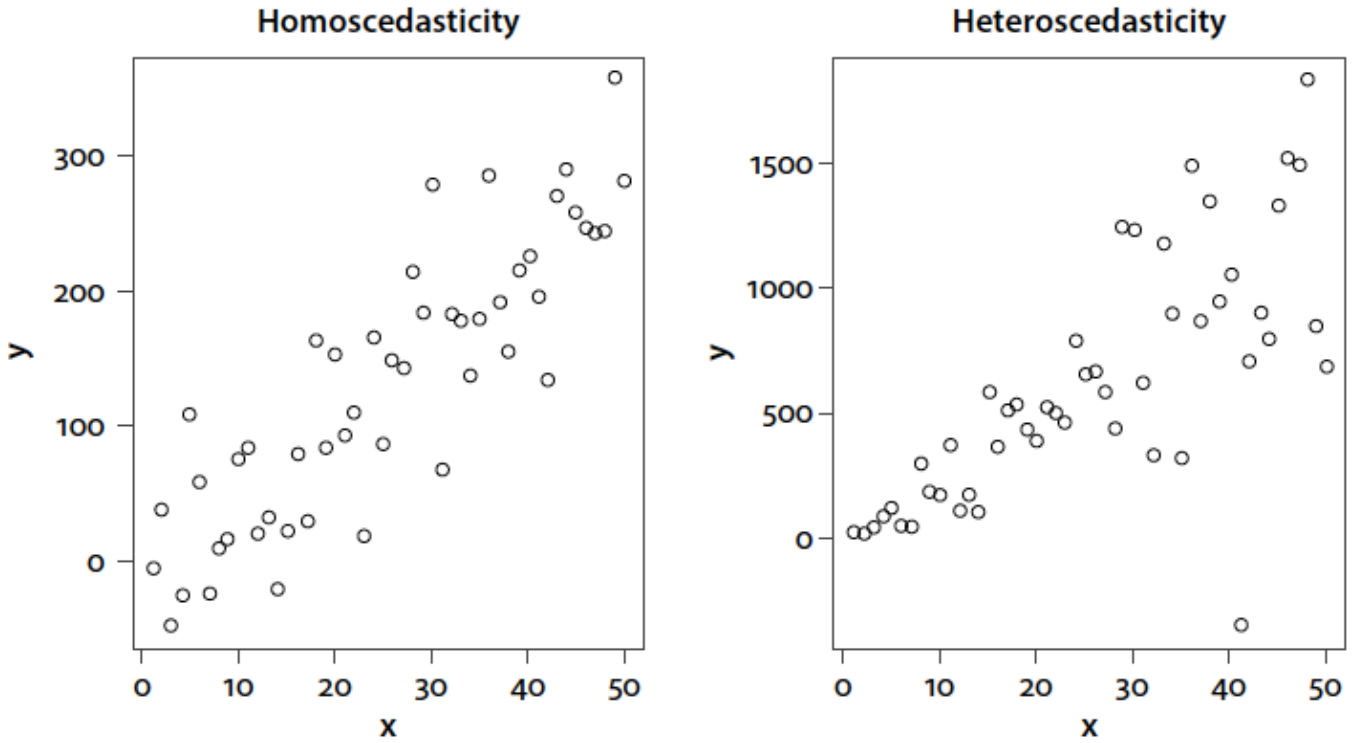
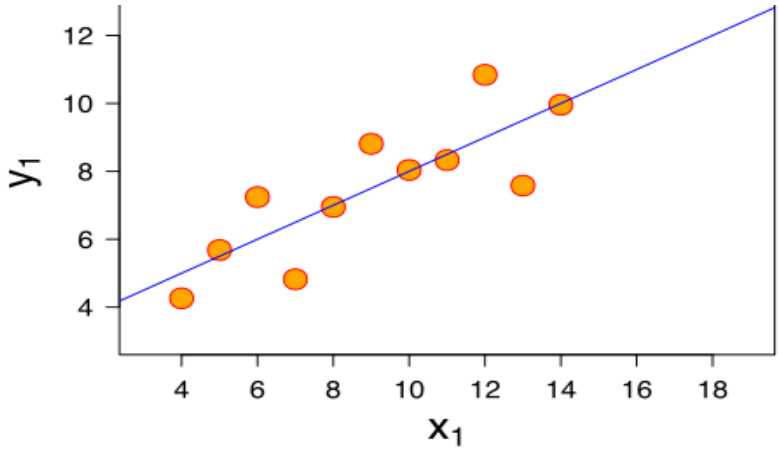
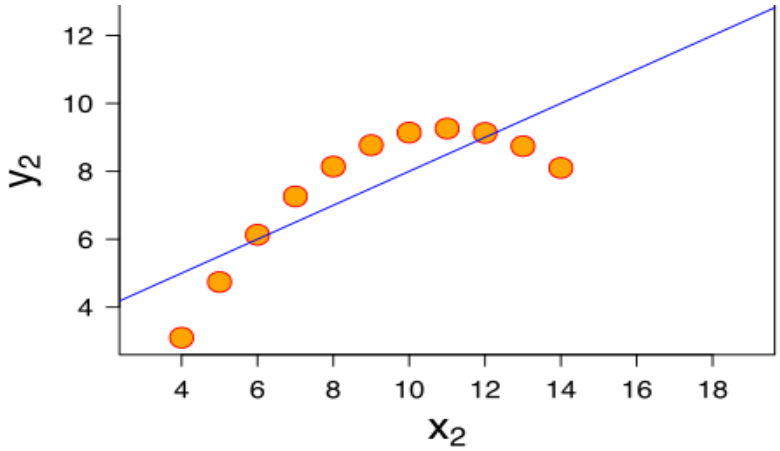


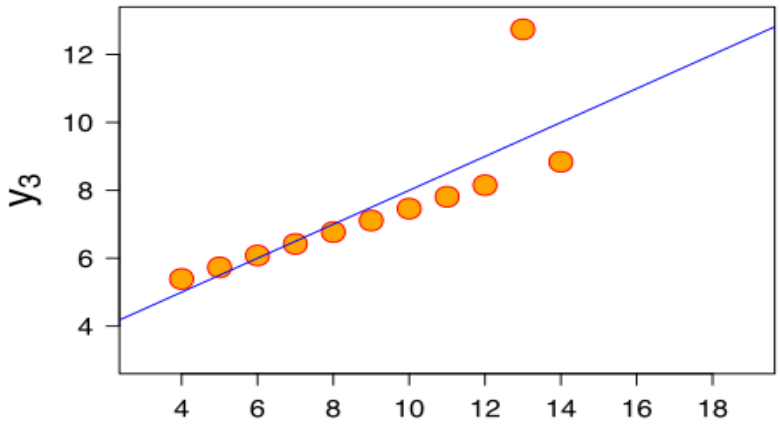
Figure 6.7. Homoscedasticity (left) and heteroscedasticity (right)

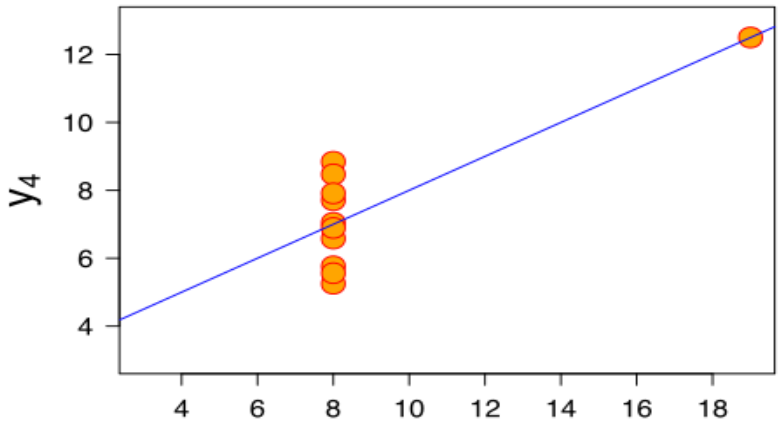
Levshina 2015

Coeficiente de correlación de Pearson





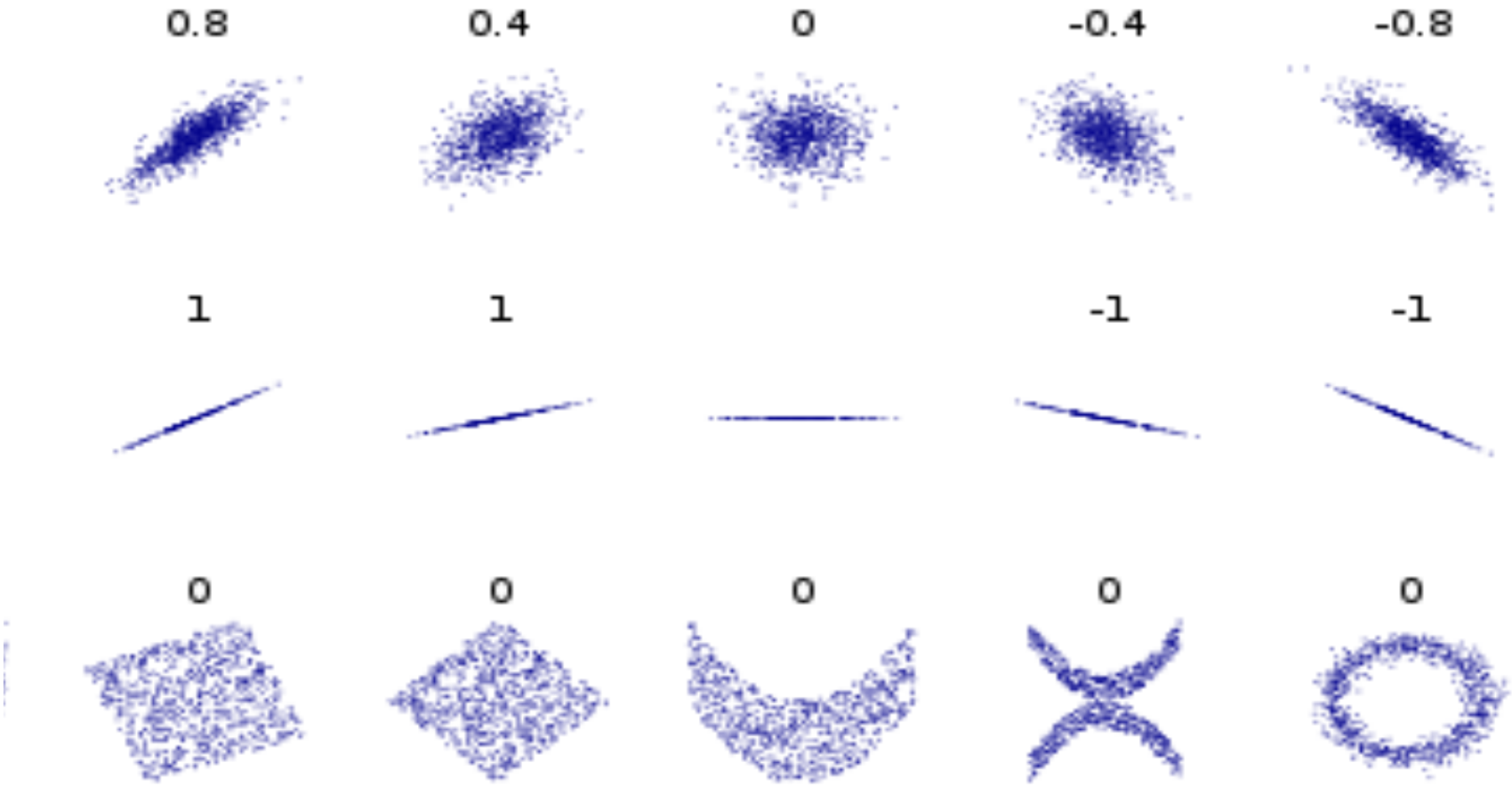




Four sets of data with the same PCC of r = 0.816

Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.

Coeficiente de correlación de Pearson



Source (Wikimedia Commons): DenisBoigelot, original uploader was Imagecreator – Own work, original uploader was Imagecreator

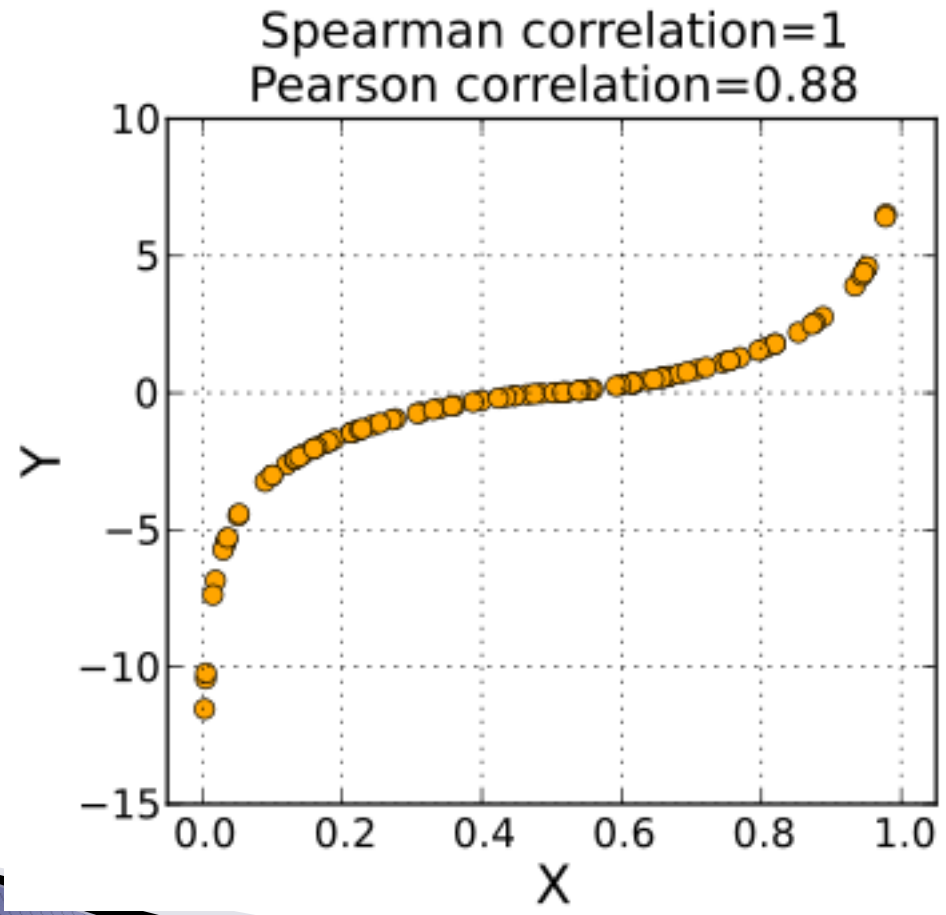
Intensidad de la correlación (CCP)

- $r > 0.7$ (o $r < -0.7$): Fuerte
- $0.3 > r > 0.7$ (o $-0.7 < r < -0.3$): Moderada
- $-0.3 < r < 0.3$: Débil

Coeficientes de correlación no paramétricos

- ρ de Spearman
- τ de Kendall

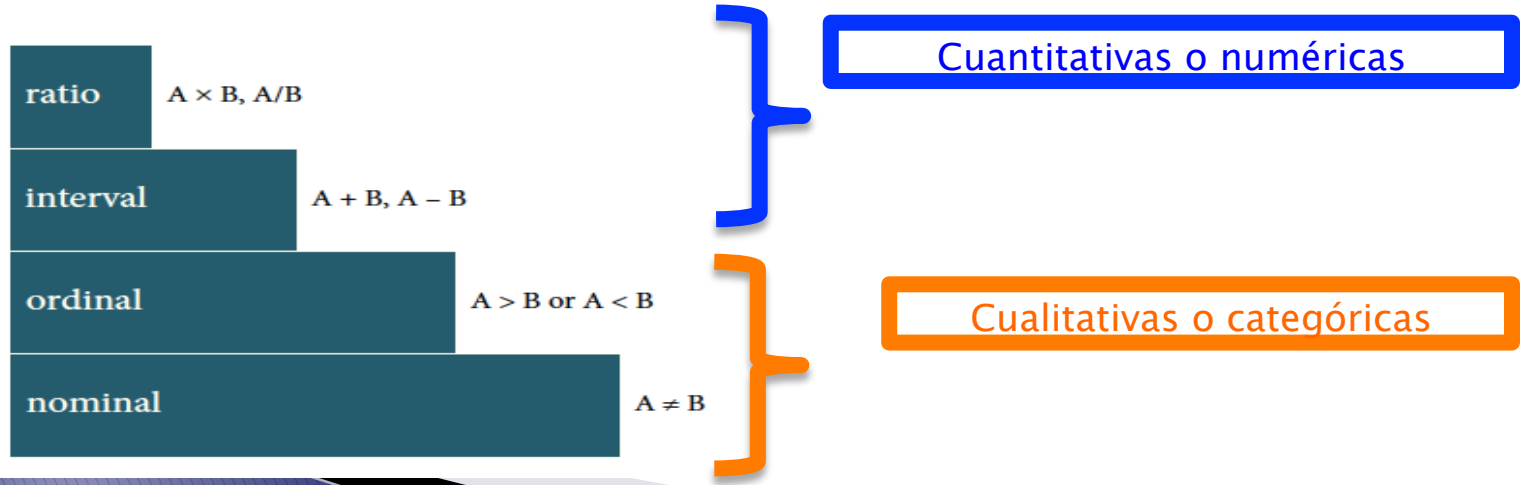
Coeficientes de correlación no paramétricos



Wikimedia commons: Skbkekas – Own work

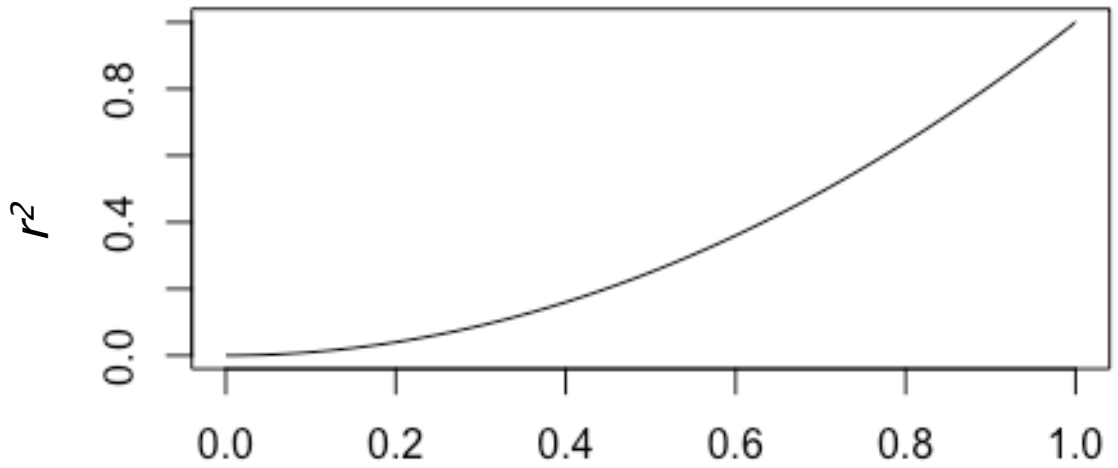
¿Cuándo usar cada CC?

Requisitos	r	ρ or τ
Muestras tomadas aleatoriamente de la población	Sí	Sí
Observaciones independientes	Sí	Sí
Sin autocorrelación	Sí	Sí
Variables al menos _____	intervalo	ordinales
Distribución normal subyacente (o n > 30)	Sí	No
La relación es lineal	Sí	No
Homocedasticidad	Sí	No



R-squared

- En este caso (una variable, relación lineal):
- $R^2 = r^2$



- Explicaciones copadas acá [StatQuest]: *https://youtu.be/2AQKmw14mHM*