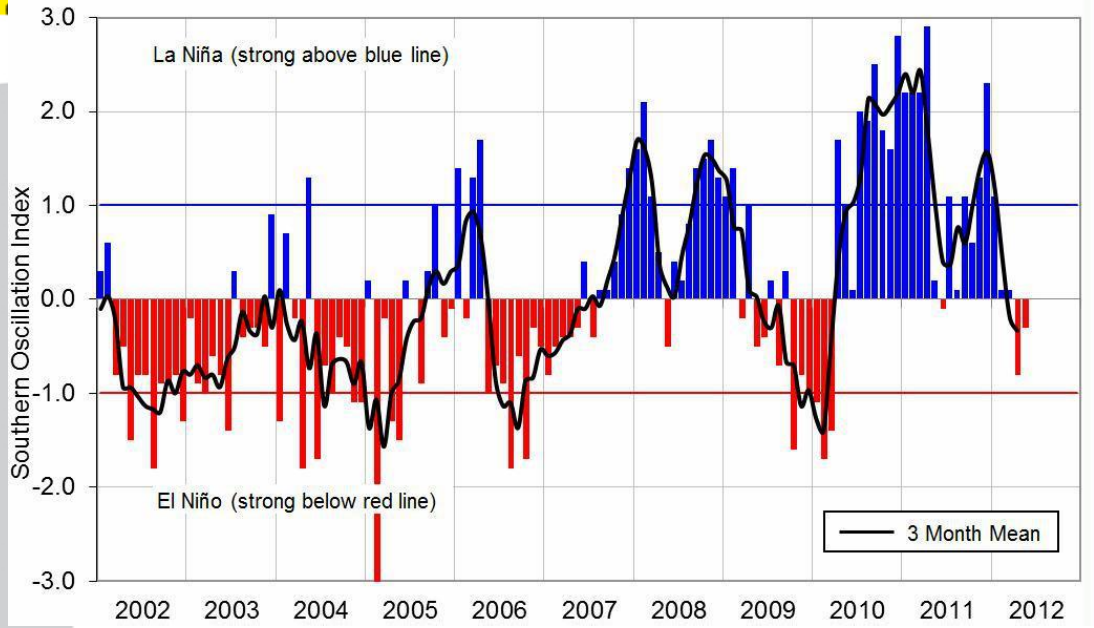
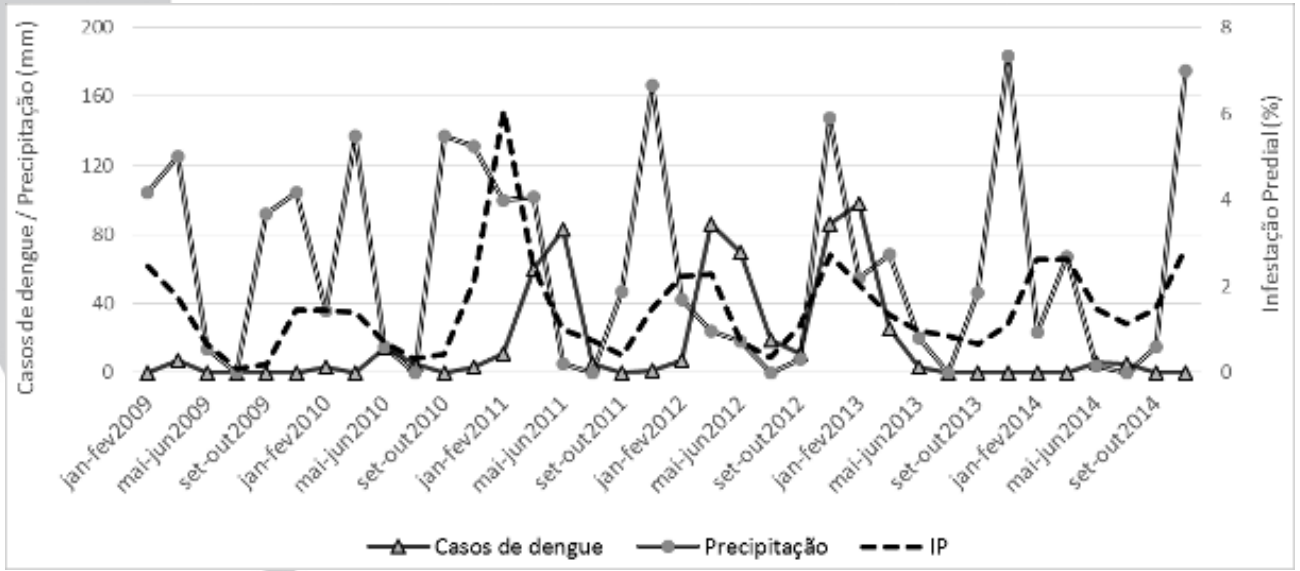
Universidad de los Andes

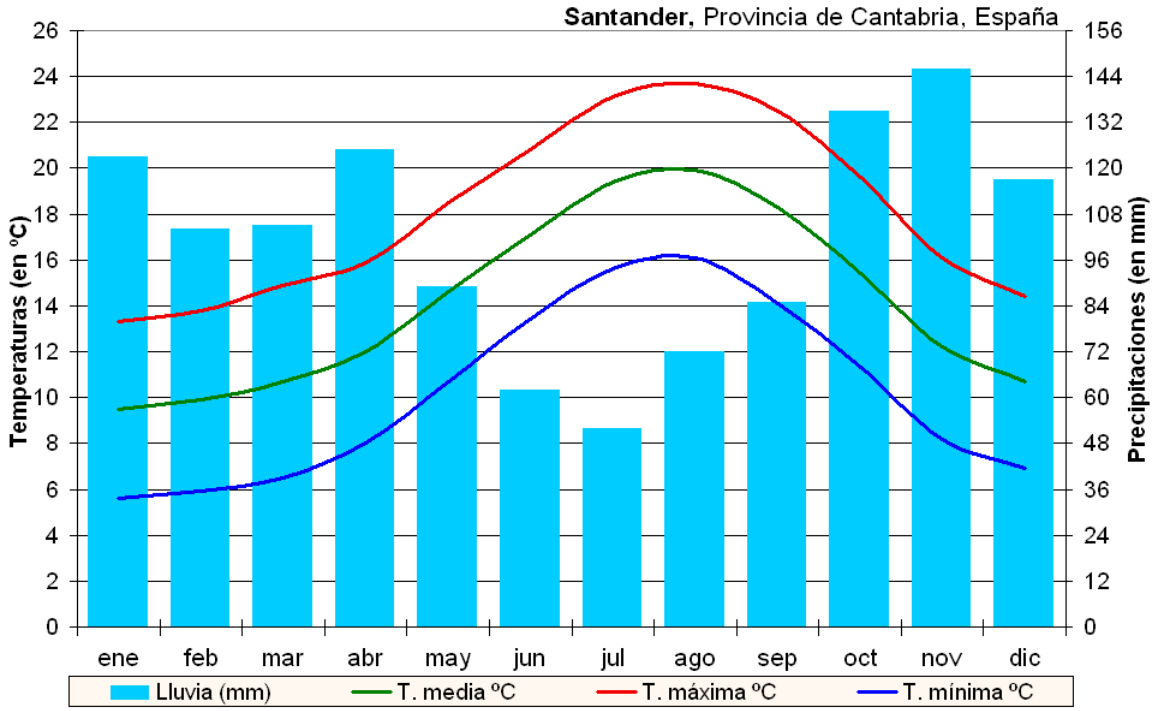
Medidas de correlación de Pearson y Spearman

Universidad de los Andes

Relación entre variables







Universidad de los Andes

```
import numpy as np
import matplotlib.mlab as mlab
import matplotlib.pyplot as plt


N = 100
X = np.linspace(-1,1, N) # genera N valores entre [-1,1]
erro = np.random.uniform(-1,1,N) # Error es incluído en la
relación linear.
sigma= 0.5
Y = 0.8*X + erro*sigma

fig= plt.figure(figsize=(6,4))

plt.scatter(X, Y, marker='o', color = 'black');
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
plt.xlabel("X",fontsize = 15)
plt.ylabel("Y", fontsize = 15)
plt.show(True)
```

Medidas de correlación

- Coeficiente de Pearson:
  - Para dos variables aleatorias X y Y :

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - [E[X]]^2}\sqrt{E[Y^2] - [E[Y]]^2}}$$

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

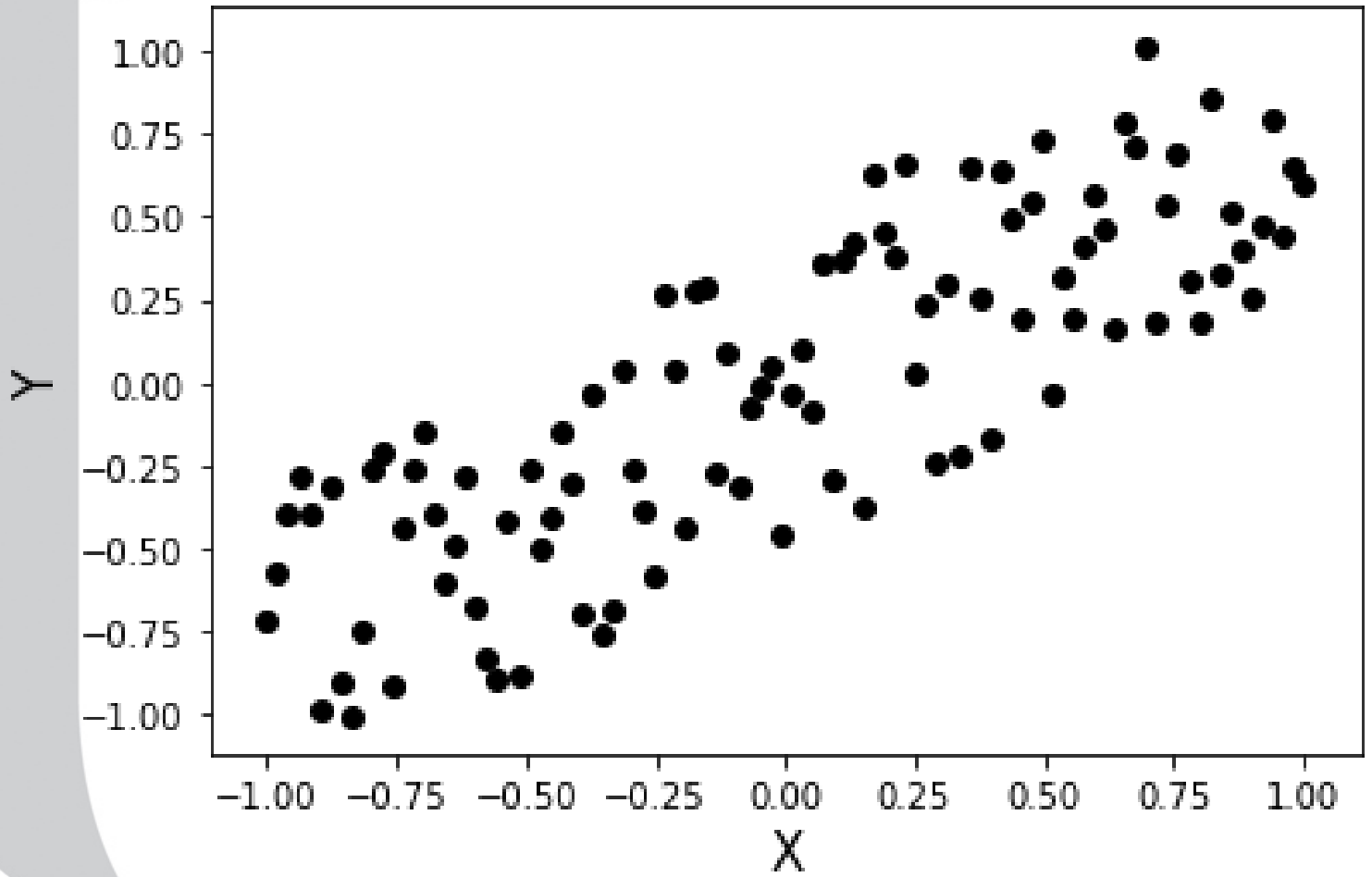
  - Para una muestra:

$$\rho = \frac{\sum_{i=i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=i}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=i}^{n}(y_i - \bar{y})^2}}$$

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}}$$

Universidad de los Andes

Interpretación

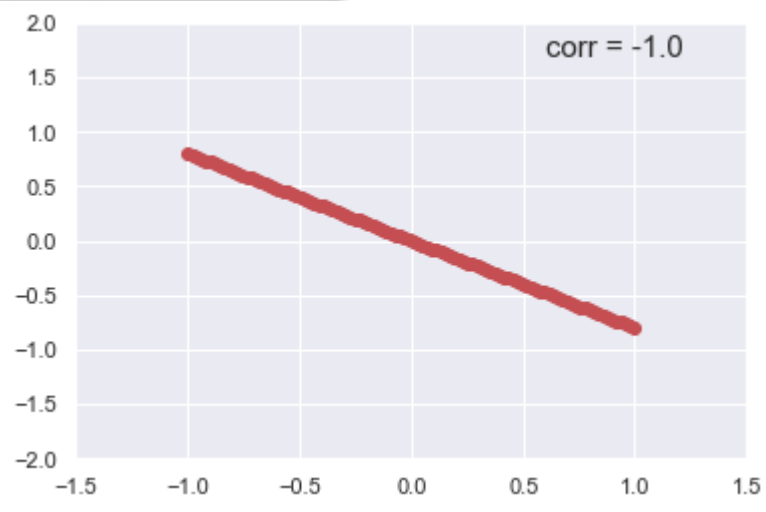


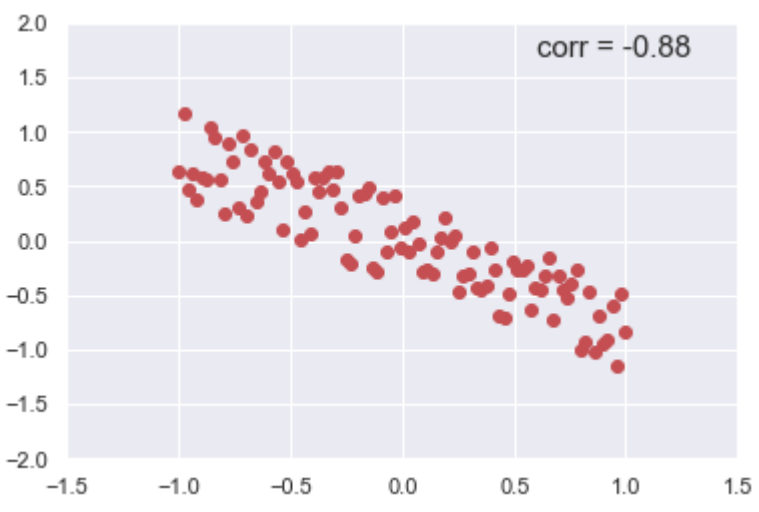
$$\rho = \frac{\sum_{i=i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=i}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=i}^{n}(y_i - \bar{y})^2}}$$

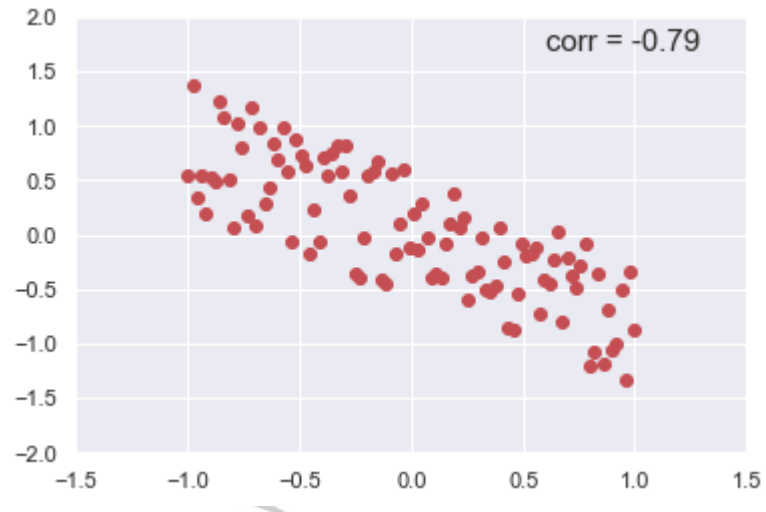
1. Define apenas relaciones lineales.
2. No define independencia entre X y Y.
3. Es afectado por valores atípicos.
4. Debe ser usado solo si X y Y presentan una distribución simétrica.

Universidad de los Andes

Medidas de correlación







Universidad de los Andes

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import pearsonr

N = 100
X = np.linspace(-1,1, N)
erro = np.random.uniform(-1,1,N) # ruído a
ser incluído en la relación lineal.
for sigma in np.arange(0,2,0.2):
    Y = -0.8*X + erro*sigma
    plt.plot(X,Y, 'ro')
    corr, p_value = pearsonr(X, Y) #
calcula la correlación
    corr = int(corr*100)/100
    string = 'corr = '+ str(corr)
    plt.xlim(-1.5,1.5)
    plt.ylim(-2, 2)
    plt.text(0.6,1.7, string, fontsize=15)
    plt.show(True)
```

Universidad de los Andes

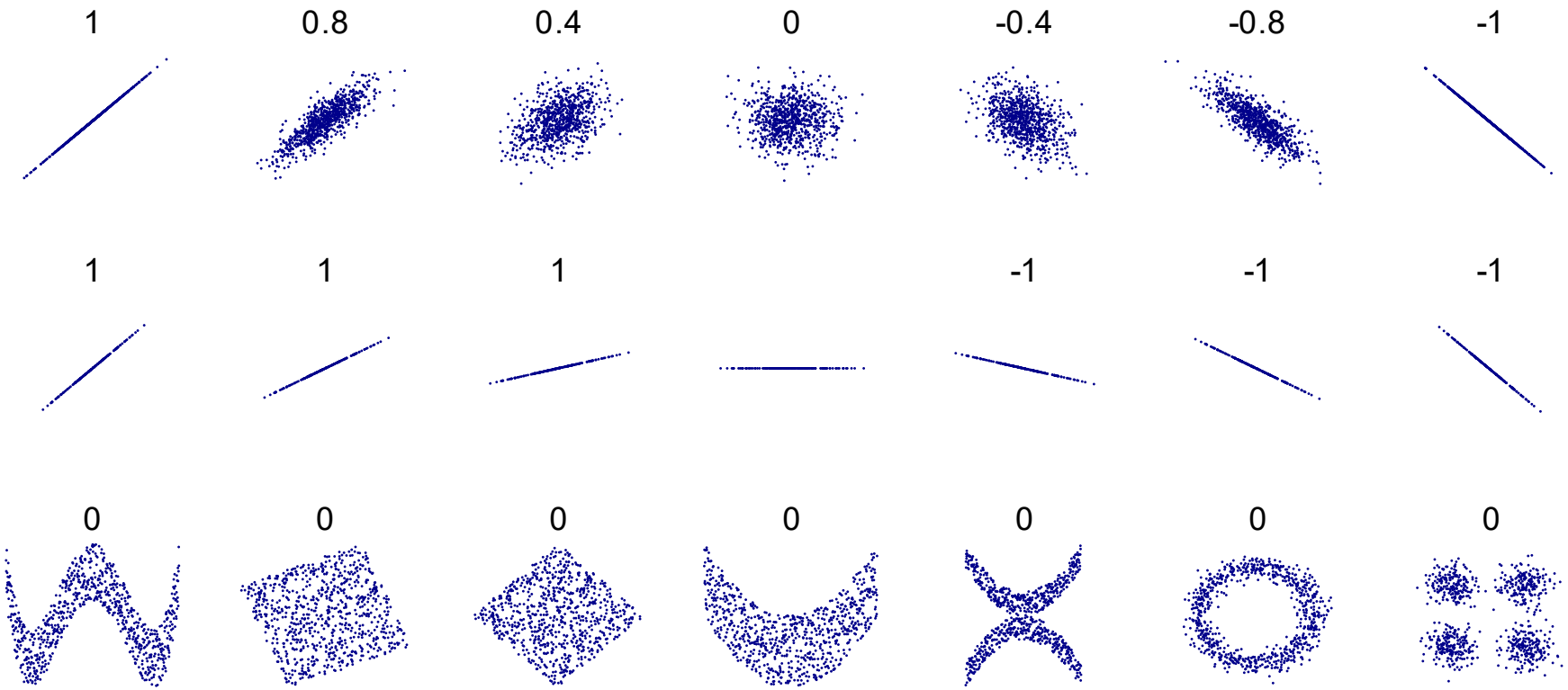
Otro ejemplo:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import pearsonr
import pandas as pd # biblioteca pandas
data = pd.read_csv('iris.csv',
header=(0)) # lee los datos a partir de
un archivo

corr = data.corr()
#Plot Correlation Matrix using
Matplotlib
plt.figure(figsize=(7, 5))
plt.imshow(corr, cmap='Blues',
interpolation='none', aspect='auto')
plt.colorbar()
plt.xticks(range(len(corr)),
corr.columns, rotation='vertical')
plt.yticks(range(len(corr)),
corr.columns);
plt.suptitle('Correlation between
variables', fontsize=15,
fontweight='bold')
plt.grid(False)
plt.show()
```

Universidad de los Andes

Correlación de Pearson



Universidad de los Andes

Correlación de Spearman

- La correlación de Spearman es igual al coeficiente de Pearson aplicado a los valores de orden de dos variables.

$$\rho = \frac{\sum_{i=i}^{n} (rx_i - \overline{rx})(ry_i - \overline{yr})}{\sqrt{\sum_{i=i}^{n} (rx_i - \overline{rx})^2} \sqrt{\sum_{i=i}^{n} (ry_i - \overline{ry})^2}}$$

- Por ejemplo, si X={3,5,1,9,6}, tenemos que rx={2,3,1,5,4}

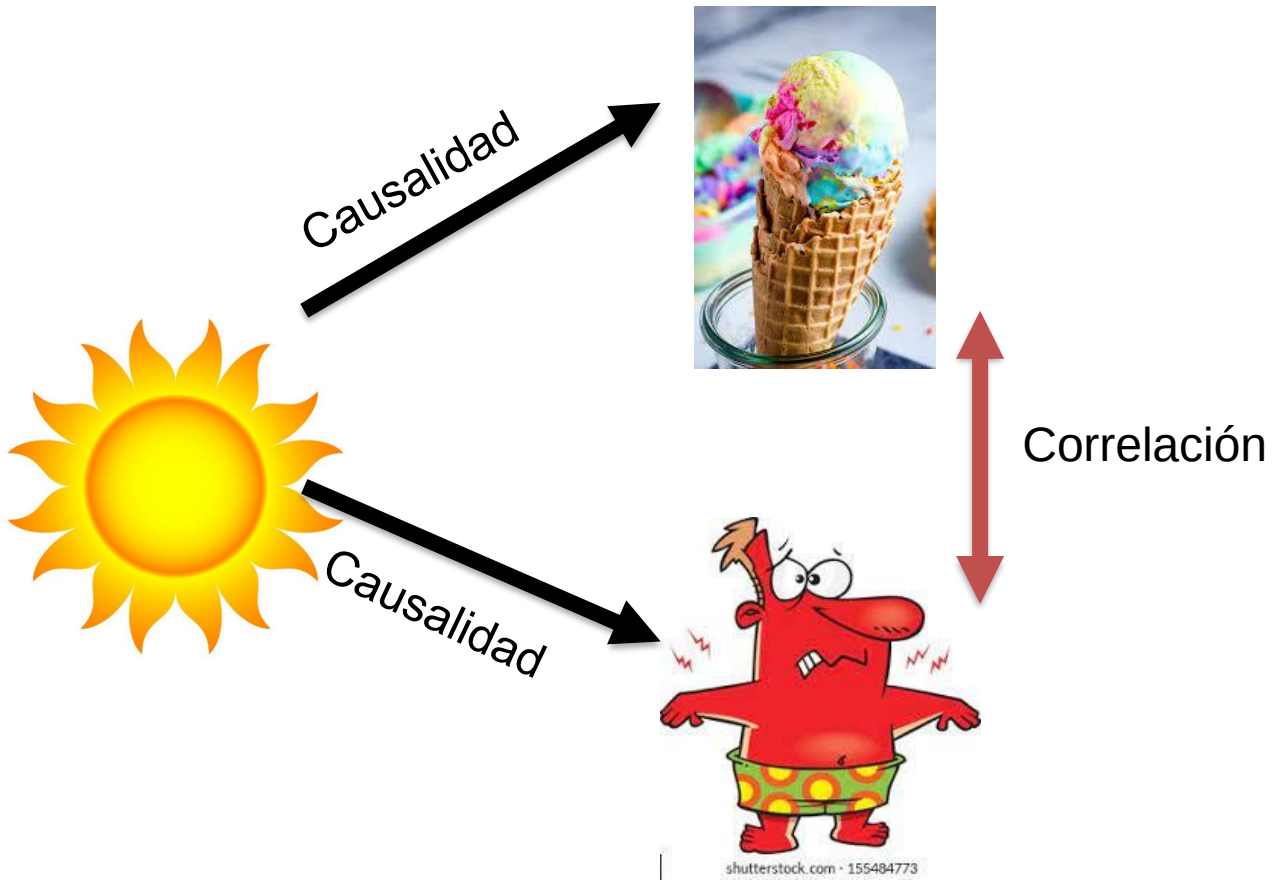
Universidad de los Andes

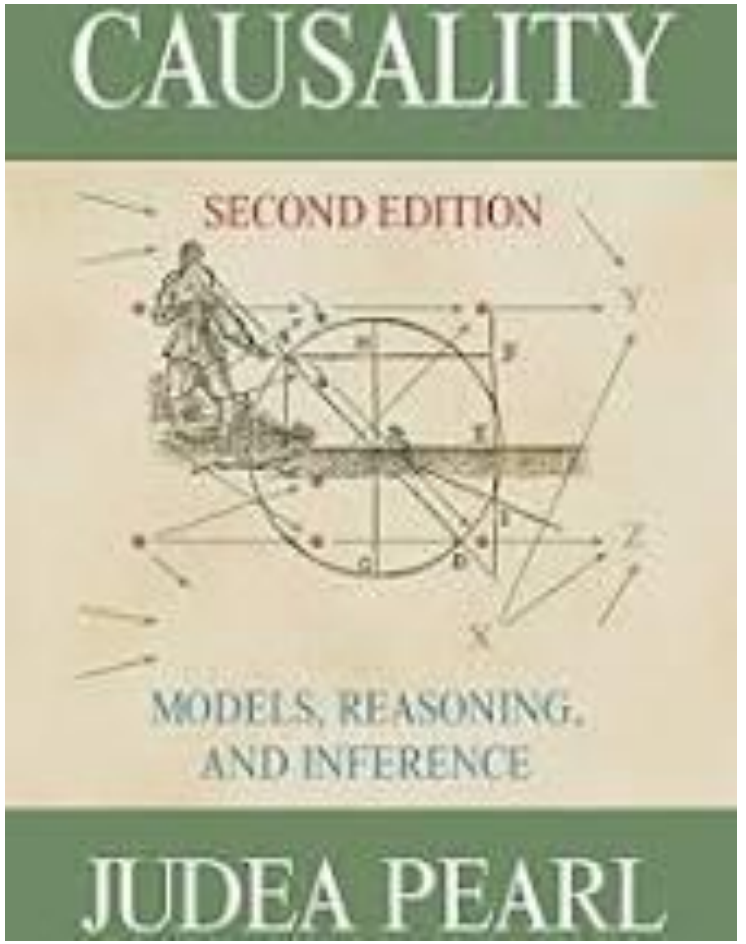
Correlación de Spearman

- Avalia la relación monotónica entre dos variables continuas o ordinales y no es sensible a las asimetrías en la distribución ni a la presencia de valores atípicos.
- En una relación monotónica, las variables tiene a mudar juntas pero no necesariamente a una tasa constante.
- Note que en cuanto la correlación de Pearson mide relaciones lineales, la de Spearman mide apenas relaciones monotónicas (lineales y no lineales).
- **Limitación: Cuando tenemos muchas observaciones con el mismo orden.**

Correlación x Causalidad

Correlación no implica causalidad !!!

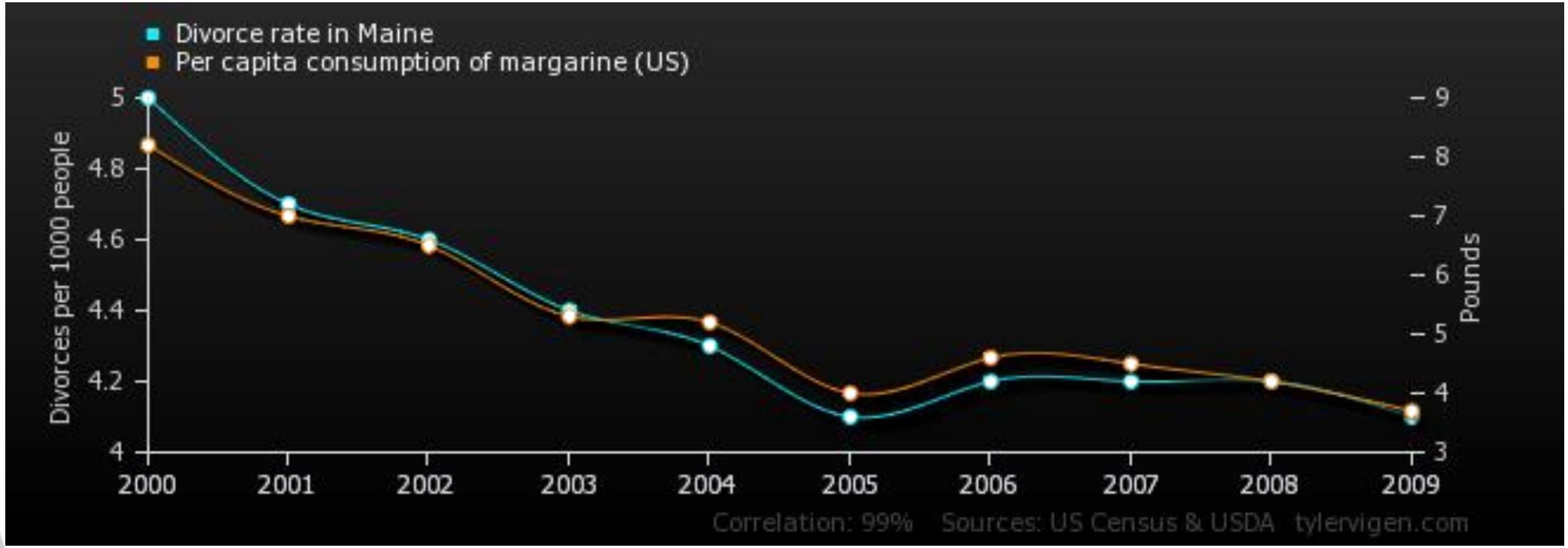




Universidad de los Andes

Correlación x Causalidad

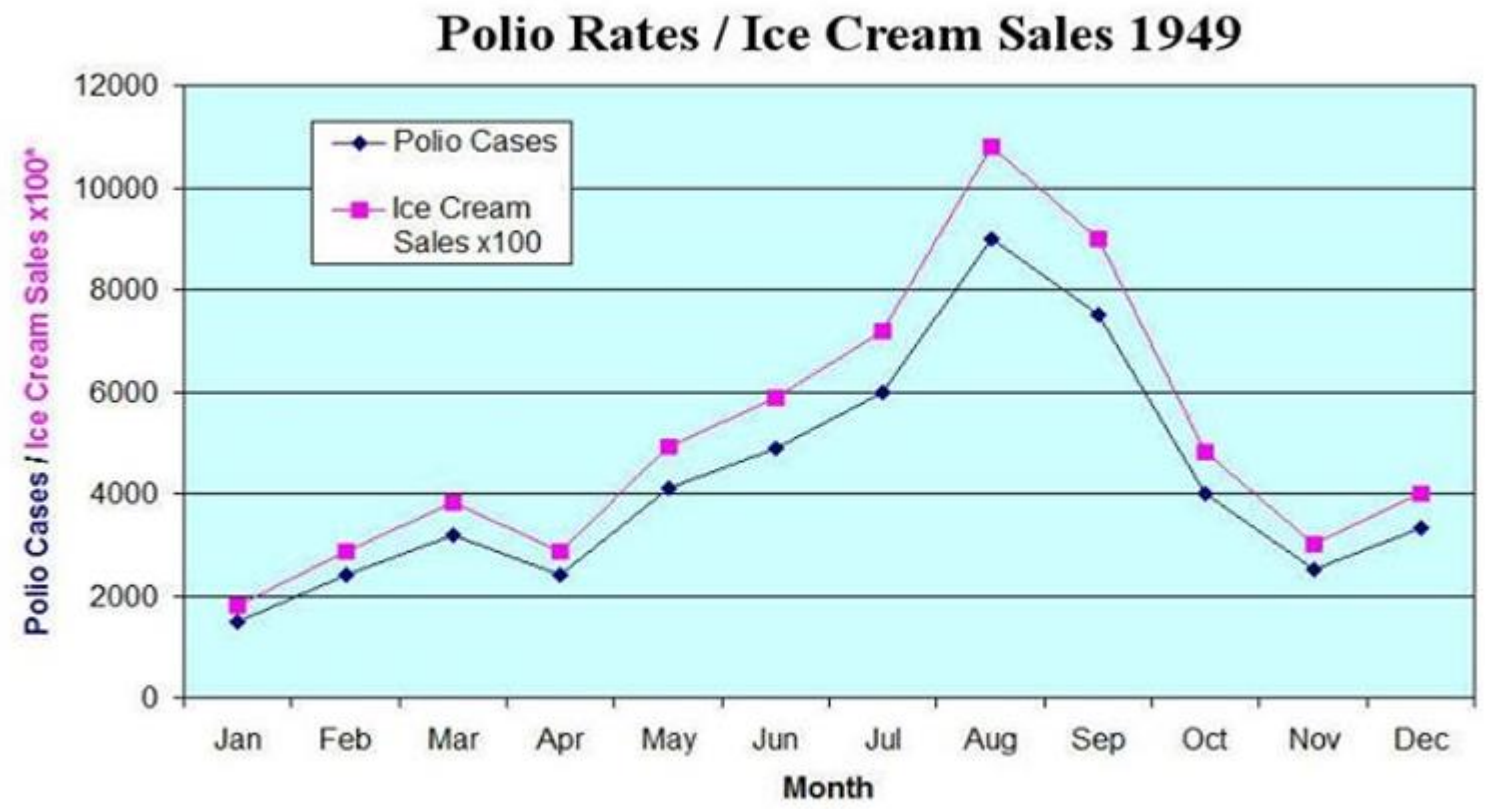
Correlación no implica causalidad !!!



Universidad de los Andes

Correlación x Causalidad

Correlación no implica causalidad !!!



Universidad de los Andes

Otros tópicos importantes

- Coeficiente de correlación de Kendall
- Distancia de Kullback-Leibler
- Información mutua
- Causalidad