Data Analysis (1)

- For a data vector

$$X := (x_1, \ldots, x_n)$$

$$X = (1, 3, 7, 9, 10)$$

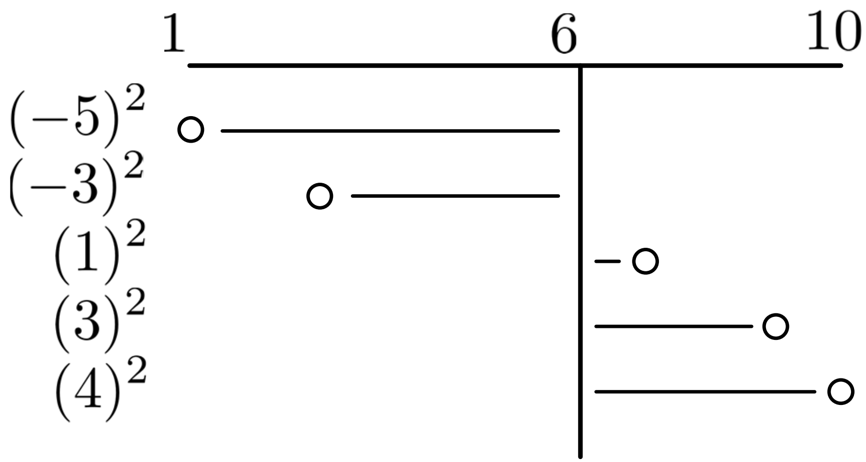
- the mean value is

$$\mu_X = \bar{X} := \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\bar{X} = \frac{1}{5}(1 + 3 + 7 + 9 + 10) = 6$$

- the standard deviation is

$$\sigma_X := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2}$$



$$s = \sqrt{\frac{1}{4}(25 + 9 + 1 + 9 + 16)} \approx 3.87$$

Data Analysis (2)

- For two data vectors

$$x = (x_1, \ldots, x_n), \quad y = (y_1, \ldots, y_n)$$

- the covariance is

$$\sigma_{x,y} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- the Pearson correlation coefficient is $K_{x,y} = \dfrac{\sigma_{x,y}}{s_x \cdot s_y}$

A high positive covariance means a strong correlation, a high negative covariance an inverse correlation

The Pearson correlation coefficient normalizes covariance and makes the values comparable to each other

$$\begin{array}{cccc} 1 & 2 & (-5)(-4.2) & = & 21 \\ 3 & 3 & (-3)(-3.2) & = & 9.6 \\ 7 & 8 & (1)(1.8) & = & 1.8 \\ 9 & 9 & (3)(2.8) & = & 8.4 \\ 10 & 9 & (4)(2.8) & = & 11.2 \end{array}$$

$$\bar{x} = 6,\ \bar{y} = 6.2,\ \sigma_{x,y} = \frac{52}{4} = 13$$

$$K_{x,y} \approx \frac{13}{3.87 \cdot 3.42} \approx 0.98$$

$$\begin{array}{cccc} 1 & 10 & (-5)(4) & = & -20 \\ 3 & 9 & (-3)(3) & = & -9 \\ 7 & 7 & (1)(1) & = & 1 \\ 9 & 3 & (3)(-3) & = & -9 \\ 10 & 1 & (4)(-5) & = & -20 \end{array}$$

$$\bar{x} = 6,\ \bar{y} = 6,\ \sigma_{x,y} = \frac{-57}{4} = -14.25$$

$$K_{x,y} \approx \frac{-14.25}{3.87 \cdot 3.87} \approx -0.95$$

$$\begin{array}{cccc} 1 & 1 & (-5)(-0.2) & = & 1 \\ 3 & 1 & (-3)(-0.2) & = & 0.6 \\ 7 & 1 & (1)(-0.2) & = & -0.2 \\ 9 & 1 & (3)(-0.2) & = & -0.6 \\ 10 & 2 & (4)(0.8) & = & 3.2 \end{array}$$

$$\bar{x} = 6,\ \bar{y} = 1.2,\ \sigma_{x,y} = \frac{4}{4} = 1$$

$$K_{x,y} \approx \frac{1}{3.87 \cdot 0.45} \approx 0.57$$

Data Analysis (3)

- For a data vector

$$x = (1, 3, 7, 9, 10)$$

$$x := (x_1, \ldots, x_n)$$

- the median is

$$\tilde{x} = \frac{1}{2}(x_{\lfloor \frac{n+1}{2} \rfloor} + x_{\lceil \frac{n+1}{2} \rceil})$$

$$\tilde{x} = \frac{1}{2}(7 + 7) = 7$$

- Empirical percentiles are the generalization of a median. The median separates the data into two parts at 50%.
  - The upper quantile is at 75%
  - The lower quantile is at 25%

Box-Whisker-Plot

- A popular visualization for data that has a ordinal scale is the Box-Plot or the Box-Whisker–Plot. It shows

$$data = (1, 1, 2, 3, 4, 5, 10)$$

  - Outliers
  - Maximum
  - Upper quantile
  - Median
  - Lower Quantile
  - Minimum