Correlation

What is Correlation

- Correlation is a standardized measure that quantifies the strength and direction of a linear relationship between two variables

- We can visualize the correlation using a scatter plot
  Examining the scatter plot makes it possible to detect if there is a particular structure of association between the two variables.

- the correlation coefficient help us quantify the strength of this association

- The regression line detect if the association between the two variables actually reflects a statistical dependence between Y and X

Linear Correlation Coefficient

Covariance

- Covariance measures the degree to wich two variables change together.
- It can be positive (both variables increase or decrease together), negative (one variable increases as the other decreases), or close to zero (little linear relationship)

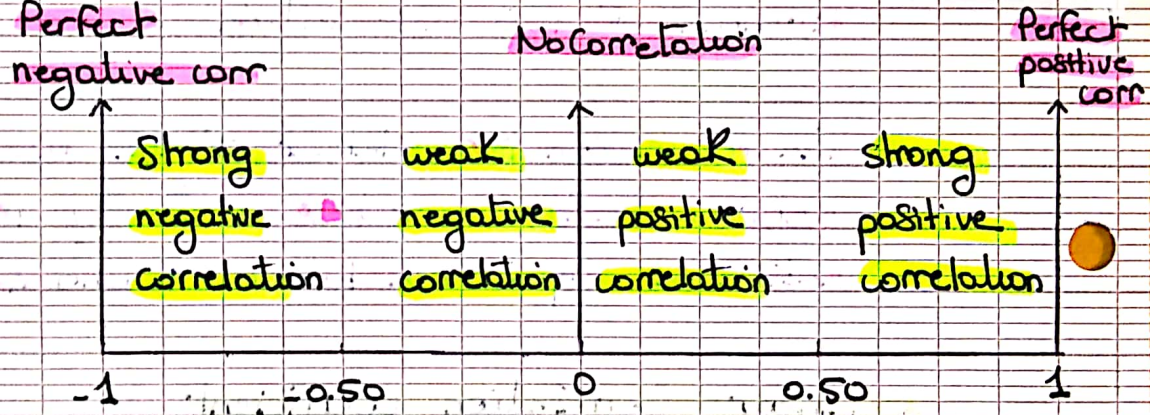
S_{xy}

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

↳ **Correlation :**

- Correlation measures the intensity of relationship between 2 variables it is kind of normalized covariance by standarized of vanables
- The correlation coefficient (r) range from -1 to 1 A value of (1) indicate a perfect positive linear relationship; (-1) indicates a perfect negative linear relationship, and (0) indicates no linear relationship

- Formula: $\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$ (For population)



- The sample correlation coefficient (r) can be estimated using sample data

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

→ Correlation coefficient and regression

- In simple linear regression, the correlation coefficient (r) is the square root of the coefficient of determination (R^2)

$$r = \sqrt{R^2}$$

- We can estimate the slope of the regression line using the correlation coefficient (r)

$$\hat{\beta}_1 = r \cdot \frac{S_y}{S_x}$$

→ Test on the correlation coefficient

- we test the significance of r to determine wether the observed correlation between two variables is statistically significant

Step 1 : Hypothesis

$H_0 : \rho(x,y) = 0$

$H_1 : \rho(x,y) \neq 0$

Step 2 : Test statistic (T-test)

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \rightsquigarrow t(n-2)$$

Step 3 : Make the decision

if $|t| > t_{critical}$ ⇒ reject H_0

⇒ There exist a linear relationship between x and y.