Correlation

Dr. Jyotismita Chaki

What is correlation and correlation analysis

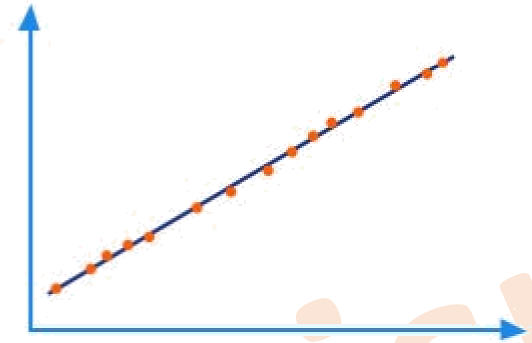
- Correlation is used to test relationships between quantitative variables (numerical variables: counts, percents, or numbers) or categorical variables (descriptions of groups or things, Type of pet owned (e.g. dog, cat, rodent, fish)).
- In other words, it’s a measure of how things are related.
- The study of how variables are correlated is called **correlation analysis.**
- Correlation analysis in research is a statistical method used to measure the strength of the linear relationship between two variables and compute their association.
- Correlation analysis calculates the level of change in one variable due to the change in the other. A high correlation points to a strong relationship between the two variables, while a low correlation means that the variables are weakly related.
- When it comes to market research, researchers use correlation analysis to analyze quantitative data collected through research methods like surveys and live polls.
- They try to identify the relationship, patterns, significant connections, and trends between two variables or datasets.
- There is a positive correlation between two variabls when an increase in one variable leads to the increase in the other.
- On the other hand, a negative correlation means that when one variable increases, the other decreases and vice-versa.

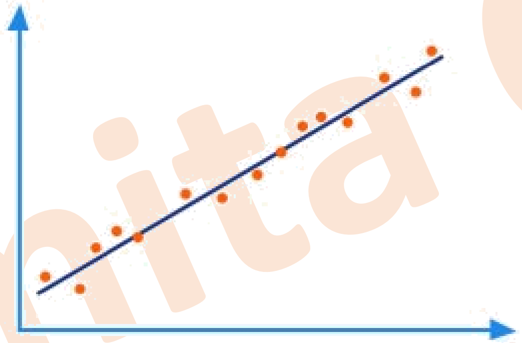
Types

- Correlation between two variables can be either a positive correlation, a negative correlation, or no correlation. Let's look at examples of each of these three types:
- Positive correlation:
  - A positive correlation between two variables means both the variables move in the same direction. An increase in one variable leads to an increase in the other variable and vice versa. For example, spending more time on a treadmill burns more calories.
- Negative correlation:
  - A negative correlation between two variables means that the variables move in opposite directions. An increase in one variable leads to a decrease in the other variable and vice versa. For example, increasing the speed of a vehicle decreases the time you take to reach your destination.
- Weak/Zero correlation:
  - No correlation exists when one variable does not affect the other. For example, there is no correlation between the number of years of school a person has attended and the letters in his/her name.

Types

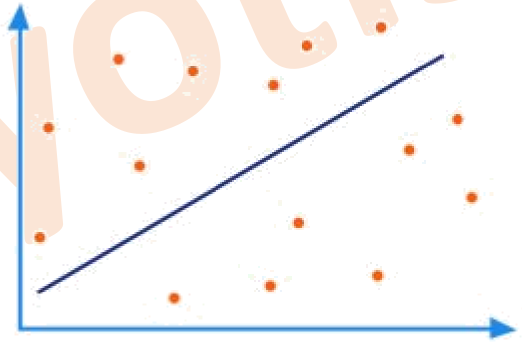
1.
Large positive
correlation

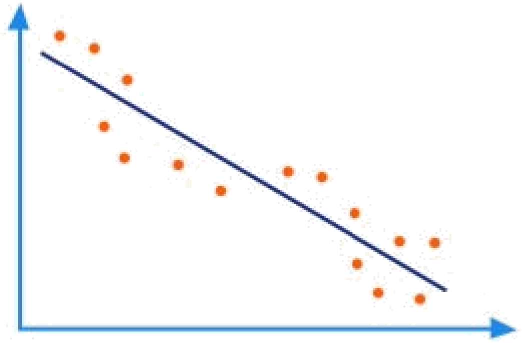




2.
Medium positive
correlation

4.
Weak / no
correlation





3.
Small negative
correlation

Examples

- Some examples of data that have a **high correlation:**
  - Your caloric intake and your weight.
  - Your eye color and your relatives’ eye colors.
  - The amount of time your study and your GPA.
- Some examples of data that have a **low correlation** (or none at all):
  - Your sexual preference and the type of cereal you eat.
  - A dog’s name and the type of dog biscuit they prefer.
  - The cost of a car wash and how long it takes to buy a soda inside the station.

Advantages

- Observe relationships: A correlation helps to identify the absence or presence of a relationship between two variables. It tends to be more relevant to everyday life.
- A good starting point for research: It proves to be a good starting point when a researcher starts investigating relationships for the first time.
- Uses for further studies: Researchers can identify the direction and strength of the relationship between two variables and later narrow the findings down in later studies.
- Simple metrics: Research findings are simple to classify. The findings can range from -1.00 to 1.00. There can be only three potential broad outcomes of the analysis.

Methods: Pearson Correlation

- **Pearson correlation (r)**, which measures a linear dependence between two variables (x and y). It’s also known as a **parametric correlation** test because it depends to the distribution of the data.
- It can be used only when x and y are from normal distribution.
- The plot of y = f(x) is named the **linear regression** curve.

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

- **x** and **y** are two vectors of length **n**
- **m_x** and **m_y** corresponds to the means of x and y, respectively.

Methods: Spearman Correlation

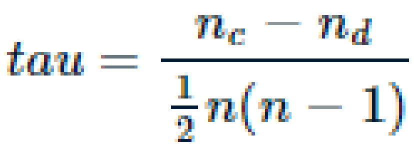
- The **Spearman correlation** method computes the correlation between the rank of x and the rank of y variables.

$$rho = \frac{\sum(x' - m_{x'})(y'_i - m_{y'})}{\sqrt{\sum(x' - m_{x'})^2 \sum(y' - m_{y'})^2}}$$

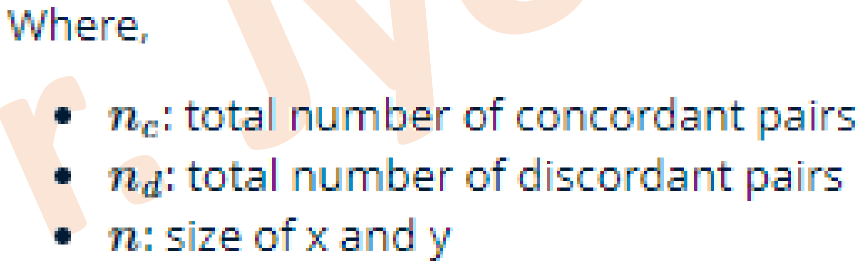
Where x’=rank(x) and y’=rank(y)

Methods: Kendall correlation

- The **Kendall correlation** method measures the correspondence between the ranking of x and y variables. The total number of possible pairings of x with y observations is n(n-1)/2 where n is the size of x and y.



$$tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$



Where,

- n_c: total number of concordant pairs
- n_d: total number of discordant pairs
- n: size of x and y