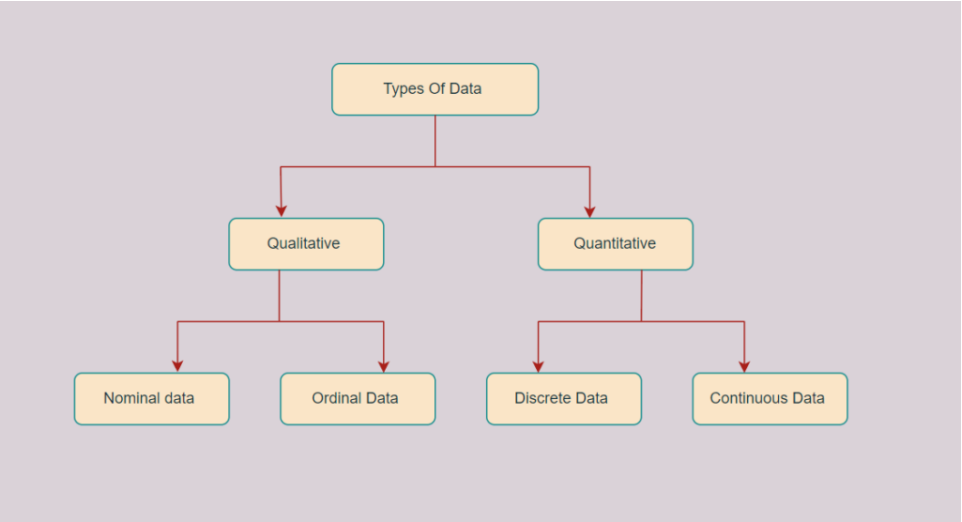
Descriptive Statistics

is a branch of statistics that involves summarizing and organizing data in a meaningful way. It provides a comprehensive$^{(\text{umfassend})}$ overview of the characteristics of a dataset, including its central tendency$^{(\text{zentraler Tendenz})}$, dispersion$^{(\text{Streuung})}$, and shape$^{(\text{Form})}$.

Data types



https://www.mygreatlearning.com/blog/types-of-data/

In statistics, data can be classified into different types based on the nature of the measurements and the level of measurement. The types of data are generally divided into two main categories: **Quantitative** and **Qualitative** data.

Quantitative Data (Numerical Data)

Quantitative data represent quantities and can be measured on a numerical scale. They are further classified into two subtypes:

a. **Discrete Data**: Discrete data can only take specific values (often integers) and cannot be subdivided.
   - **Example**: A dataset recording the number of cars sold by a dealership each day (1, 2, 3, etc.)

b. **Continuous Data**: Continuous data can take any value within a given range and can be divided into finer subdivisions. These are "measurable"

quantities and often involve measurements.

- **Example**: A dataset measuring the amount of rainfall in a city over a year, in millimeters (25.4 mm, 50.8 mm, etc.)

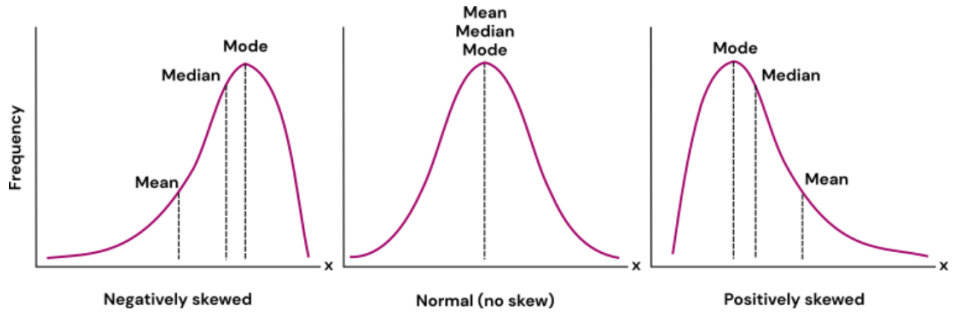
Qualitative Data (Categorical Data)

Qualitative data represent characteristics or attributes and can be categorized based on qualities or characteristics that are not numerical. They are further classified into two subtypes:

a. **Nominal Data**: represent categories without any intrinsic ordering.
   - **Example**: The blood types of individuals in a study (A, B, AB, O)

b. **Ordinal data**: represent categories with a meaningful order or ranking, but the differences between the ranks are not necessarily equal.
   - **Example**: The educational level of respondents in a survey (High School, Bachelor's, Master's, Ph.D.)

Measures of central tendency



https://lledidi.com/academy/measures-of-central-tendency-mean-median-and-mode

are statistical values that describe the **middle** or the **average** of a dataset. The 3 most common measures of central tendency are the `mode`, `median`, and `mean`.

Dataset: 2, 2, 3, 5, 5, 7, 11

Mean

The mean is the most commonly used measure of central tendency. It is calculated by summing all the values `{x_1, ..., x_i}` in the dataset and then dividing by the number of values `N`.

- Formula: $\frac{\sum_{i=1}^{n} x_i}{n}$
- Example: $\overline{x} = \frac{2+2+3+5+5+7+11}{7} = 5$
- **Mean** is useful for quantitative data that is symmetrically distributed without outliers.

Median

The median is the **middle value** of a dataset when it is arranged$^{(\text{aufgeordnet})}$ in **ascending** or **descending** order. If there is an even number of observations, the median is the average of the **two middle numbers**.

- Formula: $\begin{cases} X\left[\frac{n+1}{2}\right] & \text{if } n \text{ is odd} \\ \frac{X\left[\frac{n}{2}\right]+X\left[\frac{n}{2}+1\right]}{2} & \text{if } n \text{ is even} \end{cases}$
- Example: $\tilde{x} = X[\frac{7+1}{2}] = X[4] = 5$
- **Median** is preferred in skewed distributions or when the data contains outliers, as it is not as affected by extreme values.

Mode

The mode is the value that **appears most frequently**$^{(\text{erscheint am häufigsten})}$ in a dataset.

- Formula: $Mode(X) = \{x_i \mid f_i > f_j \ \forall \ j \neq i, j = 1, 2, \ldots, n\}$
- Example: $Mode(X) = \{2, 5\}$
- **Mode** is used for nominal data or to identify the most frequent occurrence in a dataset.

Measures of Dispersion (Variability)

- **Range**: The difference between the highest and lowest values in the dataset. It gives a rough idea of the spread.
    - Formula: Max Value - Min Value

- **Variance**: Measures the dispersion of a set of data points around their mean. It's calculated as the average of the squared differences from the Mean.
  - Formula: Variance = $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$
    - s^2 → sample | σ → population
    - μ → mean value
- **Standard Deviation (SD)**: The square root of the variance, providing a measure of dispersion that is in the same units as the data.
  - Formula: SD = $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$
    - s^2 → sample | σ → population
- **Interquartile Range (IQR)**: The difference between the 75th percentile (Q3) and the 25th percentile (Q1). It measures the spread of the middle 50% of the data, reducing the impact of outliers.
  - $IQR = Q3 - Q1$

Shape of the Distribution

- **Skewness**: A measure of the asymmetry of the probability distribution of a real-valued random variable. Positive skew indicates a tail on the right side, while negative skew indicates a tail on the left.
- **Kurtosis**: A measure of the "tailedness" of the distribution. High kurtosis means more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly sized deviations.

Data Visualization

in statistics is a crucial method used to convey$^{(\text{vermitteln})}$ **information** encoded in data using visual elements such as charts, graphs, maps, and infographics. This approach facilitates$^{(\text{erleichtern})}$ understanding by leveraging$^{(\text{hebeln})}$ the human brain's ability to process visual information more efficiently than textual or numerical data.

Charts

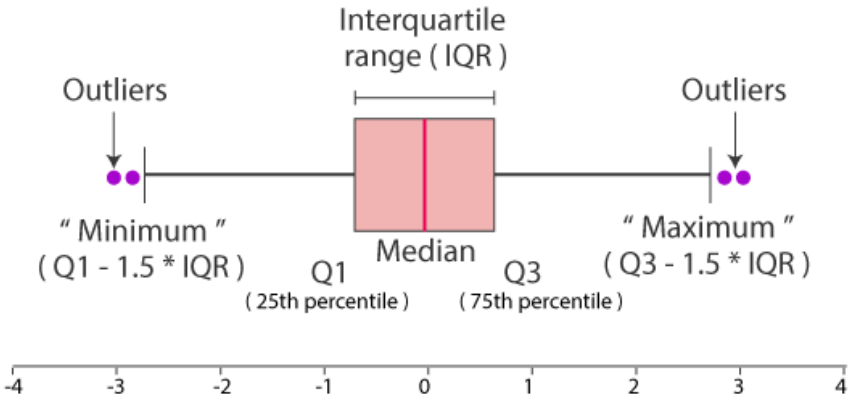
They can be categorized into several types, including bar charts, line charts, pie charts and histograms.

- **Bar Charts**: Used to **compare** different categories or groups. The length or height of the bar represents the measured value or frequency.
- **Line Charts**: Show trends over time. Points are plotted for each time period and connected with lines, highlighting changes and trends.
- **Pie Charts**: Represent parts of a whole as slices of a pie. The **size** of each slice is **proportional to the percentage** it represents.
- **Histograms:** Used to show the distribution of a dataset and identify the **central tendency**, **dispersion**, and shape of the data's distribution.

Box Plots (Box-and-Whisker Plots)

are a standardized way of displaying the distribution of data based on a **five-number summary**: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

- **Interquartile Range (IQR)**: $IQR = Q3 - Q1$, which measures the spread of the middle 50% of the data.
- **Whiskers**: Extend from the box to the highest and lowest values, excluding outliers. Outliers are typically defined as any data point more than 1.5 IQRs below the first quartile or above the third quartile.
- **Box**: The central box spans from Q1 to Q3, with a line at the median (Q2).

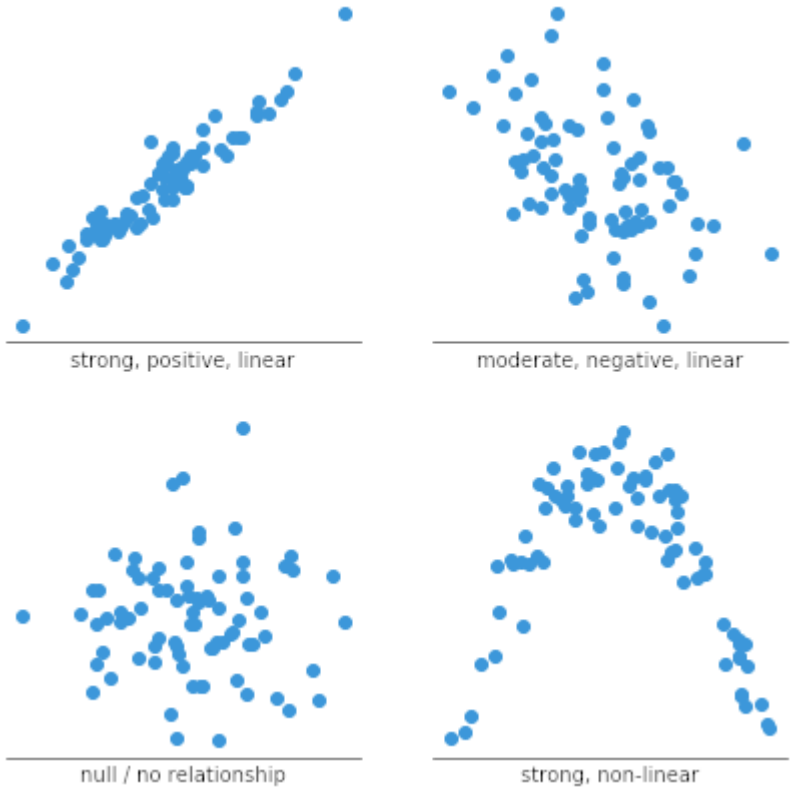


Different parts of boxplot

Scatter Plots

Scatter plots are used to determine the relationship between two quantitative variables. Each point on the plot corresponds to a single data point in the dataset, with one variable determining the position on the x-axis and the other variable the position on the y-axis.

- **Correlation**: Scatter plots can be used to visually assess$^{(\text{beurteilen})}$ whether there is a linear relationship, no relationship, or a non-linear relationship between the two variables.
- **Trend Lines**: Often, a trend line (like a line of best fit) is added to a scatter plot to summarize the relationship between the variables.



https://www.atlassian.com/data/charts/what-is-a-scatter-plot