Descriptive Statistics

1. **Introduction to statistics**

- What is statistics?
- Types of statistics - Descriptive and Inferential
- Types of data
- Population and Sample
- Sampling Techniques
- Statistical Data Analysis Steps

2. **Descriptive Statistics**

- Measures of central tendency - Mean, Median and Mode
- Measures of dispersion - Range, Variance, Standard Deviation, Percentiles and Quartiles
- Frequency
- Graphical Representations - Boxplots, Histograms, Scatterplots
- Outliers and understanding their impact
- Correlation and Covariance

What is statistics?

Statistics is a branch of mathematics that involves collecting, analysing, interpreting and drawing conclusions from information/data. It provides methods for making inferences about the characteristics of a population based on a limited set of observations or data points.

Note: Data(plural) are measurements or observations, A datum (singular) is a singular measurement.

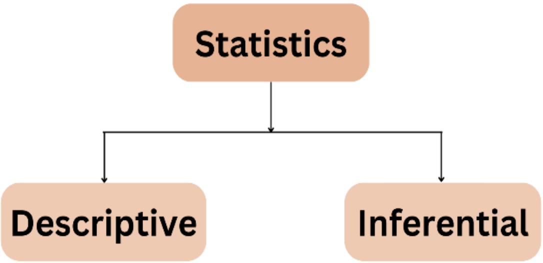
To be more specific, here are some claims that we have heard on several occasions.

- 4 out of 5 dentists recommend Dentine.
- Almost 85% of lung cancers in men and 45% in women are tobacco related.
- There is an 40% chance that in a room full of 30 people that at least two people will share the same birthday
- The average score of the students in the math test was 75 out of 100.

- The average monthly sales of the company increased by 10% compared to the previous year.
- The minimum number of participants required for the workshop is 60.

Types of statistics

Statistics can be broadly categorized into two main types:



- **Descriptive Statistics**

  Descriptive Statistics is a branch of statistics that deals with the collection, presentation, and interpretation of data. The primary goal of this statistics is to summarize and describe the main features of a dataset. This involves organizing and simplifying large amounts of data in a meaningful way to make it more understandable.

  These statistics are the foundation for more advanced statistical analysis and are essential for making informed decisions based on data. It consists of methods for organizing and summarizing information.

  Key features to describe about data:

  - What is the centre of the data? (location)
  - How much does the data vary? (scale)
  - What is the shape of the data? (shape)

  These can be described by summary statistics.

It includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, percentiles.

- **Inferential Statistics**

  Inferential Statistics involves drawing conclusions or inferences about a population based on information obtained from a sample of population. The key idea is to use the information obtained from a representative sample to make generalizations and predictions about the entire population.

  Inferential statistics often involves the use of probability theory and statistical methods to make probabilistic statements about population parameters. It helps researchers make decisions, formulate policies and draw conclusions in situations where it may be impractical or impossible to study an entire population.

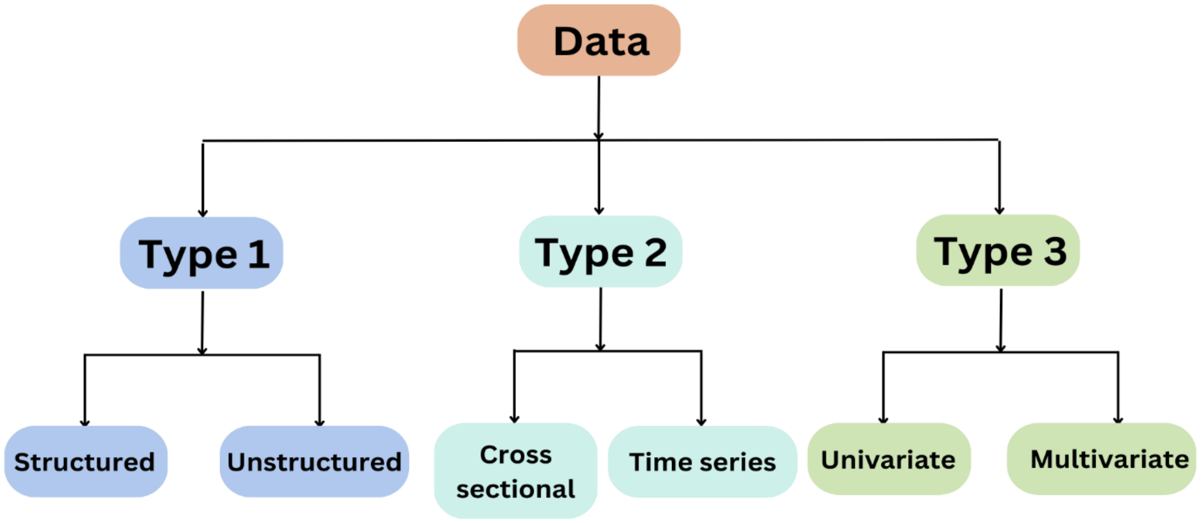
  Key concepts related to inferential statistics in the context of data analysis:

  - Testing whether the average exam scores of two groups are significantly different.
  - Estimating the average height of a population with a 95% confidence interval.
  - Analysing the relationship between hours of study and exam scores.

Descriptive and Inferential statistics are interrelated. It is almost always necessary to use methods of descriptive statistics to organize and summarize the information obtained from a sample before methods of inferential statistics can be used to make more thorough analysis of subject under investigation.

Types of data

Data can be broadly divided into different types based on its nature and characteristics.



- **Type 1:**
  - **Structured:** Structured data refers to data that is organized in rows and columns and well-defined manner.

    Examples: Tables in a relational database, spreadsheet

  - **Unstructured:** It refers to data that lacks a predefined data model or simply not structured.

    Examples: Text documents (emails, articles, social media posts), Multimedia content (images. Videos., audio recordings), web pages and other free form text

- **Type 2:**
  - **Cross-sectional:** This data is collected at a single point in time, or over a very short period, and it involves observations of multiple subjects or entities.

    Examples: Survey data collected from individuals in a city at a specific date, Marks obtained in a test

  - **Time series:** Time series data involves observations taken over a sequence of time intervals.

    Examples: Monthly sales data for a product over several years, daily stock prices for a particular company over a month

- **Type 3:**
    - **Univariate:** It involves data consisting of a single variable.
    - **Multivariate:** It involves data consisting of two or more variables.

After analysing the type of data, identifying the type of variables is necessary and it also comes under the types of data.

- **Nominal**: Nominal data represents categories or labels with no inherent order or ranking.

  Examples: Gender, colours

- **Ordinal**: Ordinal data represents categories with a clear order or ranking, but the intervals between the categories are not uniform or meaningful.

  Examples: Education levels, customer satisfaction ratings, no. of cars owned by a household

- **Categorical**: Categorical data represents categories and can be either nominal or ordinal.

  Examples: Types of cars, product categories

- **Numerical**: Numerical data includes both discrete and continuous data and represents measurable quantities.

  Examples: Temperature, income, age

- **Interval**: Interval data has meaningful interval between values, but there is no true zero point.

  Examples: Temperature measured in Celsius or Fahrenheit, IQ scores,

- **Ratio**: Ratio data has meaningful interval between values and it has a true zero point, indicating the absence of the attribute being measured.

  Examples: Height, income, age, weight

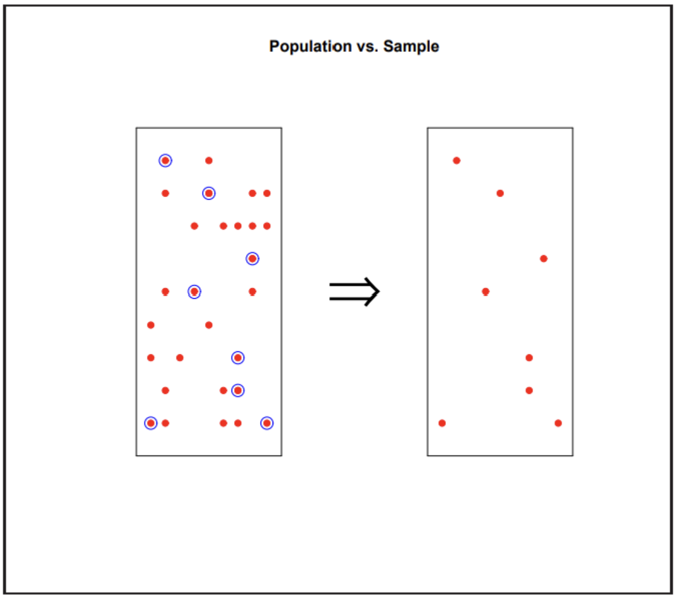
Population and Sample

- **Population:** The population is the entire group of individuals, objects or observations. It includes all possible members that meet the criteria for inclusion in the study.

Example**:** All people in India, All customers of Netflix,

- **Sample:** A sample is a subset of the population, selected for study or analysis. It is a representative group that is used to draw inferences about the larger population.
  Example: 10k people from India, 300 customers of Netflix

Population Sample



In statistical analysis, the goal is often to make inferences about a population based on observations from a sample. Various sampling techniques and statistical methods are employed to ensure that the sample is a fair and accurate representation of the population of interest.

For good statistical analysis, the sample needs to be as similar as possible to the population. If they are similar enough, we say that the sample is representative of population. The sample is used to make conclusions about the whole population. If the sample is not similar enough to the whole population, the conclusions count be useless.

The characteristics of population is known as population parameters and characteristics that describes a sample is called sample statistic.

Why samples are used?

- To reduce cost of data collection
- When a full census cannot be taken

Sampling Techniques

Sampling techniques are methods used to select a subset of elements (a sample) from a larger population for the purpose of making inferences about that population. Here are some common sampling techniques:

- **Simple Random Sampling**: Every individual or element in the population has an equal chance of being included in the sample.

- **Systematic Sampling**: A fixed interval is used to select every kth element from a list after a random starting point is chosen

  Example: Selecting every 10th person from a list of names, first 20 and last 20

- **Stratified Sampling:** Population is divided into subgroups or strata based on certain characteristics (e.g., gender, age) and then random samples are taken from each stratum.

  Example: Dividing a population of students into strata based on grade level and then randomly selecting students from each grade.

- **Clustered Sampling:** Population is divided into clusters and a random sample of clusters is selected. Then, all members within the chosen clusters are included in the sample. Unlike stratified, that selects individuals from each subgroup, it selects entire subgroups

  Example: Dividing a city into neighbourhoods, randomly selecting several neighbourhoods, and surveying all households in the chosen neighbourhoods.

The choice of sampling technique depends on various factors, including the research objectives, the nature of the population, available resources and the desired level of precision. Each sampling method has its advantages and limitations, and researchers must carefully consider the appropriateness of the technique for their specific study.

Statistical Data Analysis

Statistical data analysis involves a systematic process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, reaching conclusions, and support decision – making.

Here are the key steps in statistical data analysis:

1. Define the problem or research question:

   Clearly define the problem or research question you want to address. This step is crucial for guiding the entire analysis process.

2. Data Collection:

   Gather relevant data based on the research question. Data can be collected through surveys, experiments, observations or from existing datasets.

3. Data Cleaning:

   Check for errors, missing values, outliers, and inconsistencies in the data. Clean and preprocess the data to ensure its quality and reliability.

4. Exploratory Data Analysis (EDA):

   Perform initial exploratory analysis to understand the characteristics of the data. This may involve summary statistics, visualizations (histograms, scatter plots), and identifying patterns or trends.

5. Data Transformation:

   If needed, transform the data to meet the assumptions of statistical methods. Common transformations include normalization, standardization, and handling categorical variables.

6. Hypothesis Formulation:

   Formulate a hypothesis based on your research question. Clearly define the null hypothesis (HO) and alternate hypothesis (H1) that you want to test.

7. Statistical Testing:

   Choose appropriate statistical tests based on the nature of your data and research question. Common tests include t-tests, chi-square tests, ANOVA, regression analysis, etc.

8. Interpretation of Results:

Analyse the results of your statistical tests. Determine whether the evidence supports or contradicts your hypothesis. Consider the significance level and confidence intervals.

9. Draw Conclusions:

   Based on the results, draw conclusions regarding the research question. Consider the practical significance of your findings in addition to statistical significance.

10. Document the Analysis Process/ Report Making:

    Document all the steps, methods, and decisions made during the analysis. This documentation is important for transparency, reproducibility, and future reference. Prepare a report based on your conclusions, provide recommendations for future action. Discuss the implications of your findings in the context of the original research question.

Measures of Central Tendency

Measures of central tendency are statistical measures that describe the centre of the data. They provide a single representative value around which the entire data set tends to cluster. The three main measures of central tendency are the mean, median and mode.

1- Mean:

The mean, also known as the average, is calculated by adding up all the values in a data set and then dividing by the number of values.

$$\mu = \frac{sum\ of\ all\ values}{Number\ of\ values} = \frac{x1 + x2 + x3 \ldots .\, xn}{\mathrm{n}}$$

Example: For the dataset {2,4,6,8,10}, the mean is $\frac{2+4+6+8+10}{5}$ = 6

Properties:

a. Meaningful for continuous variables
b. Affected by outliers or extreme values, which can heavily skew the result.

2- Median:

The median is the middle value of a dataset when it is ordered from least to greatest. If there is an even number of values, the median is the average of the two middle values.

Example: For the dataset {3,1,5,7,9}, when ordered, becomes {1,3,5,7,9} and the median is 5.

Properties:

a. Less influenced by extreme values, making it a better measure of central tendency for skewed distributions.
b. Meaningful for ordinal, ratio, and interval data.

3- Mode:

The mode is the value that occurs most frequently in a data set. A data set may have no model (if no value is repeated), one mode (If one value is repeated more than others), or multiple models (if more than one value is repeated with the same frequency)

Example: In the set {4,2,8,6,2,9,2}, the mode is 2 because it appears more frequently than any other value.

Properties:

a. Meaningful for categorical values. For numerical values, if unique values is small then it can also be used.

Summary:

1- Choose the measure based on the distribution of the data.
2- For normally distributed data, the mean is often appropriate.
3- For skewed, or data with outliers, consider the mean.
4- In case of categorical variables, mode is used.

Measures of Dispersion

Dispersion is the degree of variation in the data. Measures of dispersion, also known as variability or spread quantify the extent to which individual data points in a dataset differ from the central tendency (mean, median or mode). They provide important insights into the spread, scatter or distribution of the data. Two datasets of the same variable may exhibit similar positions of center but may be remarkably different with respect to variability. The main measures of dispersion include the range, IQR, variance, quartiles, percentiles and standard deviation.

1. Range:

   The range is the simplest measure of dispersion and is calculated as the difference between the maximum and minimum values in a dataset.

   Range = Max – Min

   Properties:

   a. Sensitive to extreme values
   b. Doesn't consider the distribution of values
   c. Used when a quick assessment of the spread is needed and suitable for small datasets

2. Quartiles

   Quartiles divide a dataset into four equal parts, with three quartiles, Q1, Q2 (median) and Q3. Q1 is the value below which 25% of the data falls, Q2 is the median and the 50% of the data falls below it, and Q3 is the value below which 75% of the data falls.

   Q1 is at position $\frac{n+1}{4}$

   Q2 is at position $\frac{n+1}{2}$

   Q3 is at position $\frac{3(n+1)}{2}$

   Properties:

   a. Useful for identifying the central tendency and spread of specific sections of the data.

3. Percentiles:

   Percentiles divide a dataset into 100 equal parts, with specific percentiles representing the percentage of data below a given value. The 25th, 50th, and 75th percentiles are equivalent to the Q1, Q2 and Q3 quartiles, respectively.

$$P^{th} Percentile: Value = \frac{P}{100} X(Number\ of\ observations + 1)$$

Properties:

a. Useful for comparing the position of a particular data point relative to the entire dataset.

4. IQR (Interquartile Range)

IQR is the range of the middle 50% of the data, representing the spread of the central portion of the distribution.

IQR = Q3-Q1

Properties:

a. Less sensitive to extreme values than the range.
b. Useful for identifying the spread of the central part of the data.

5. Variance

Variance measures the average squared deviation of each data point from the mean.

Variance = $\frac{\sum_{i=1}^{N}(Xi-\mu)^2}{N}$

Where Xi is the individual data points, N is the number of data points and μ is the mean of the data.

Properties:

a. When a detailed understanding of the variability is needed.
b. Sensitive to extreme values.

6. Standard deviation

Standard deviation is the square root of the variance and is expressed in the same units as the original data.

Standard deviation = $\sqrt{Variance}$

It’s good to use std as it is in the same units as x and variance is in square units.

Note:

- Average paints a partial picture of the data

- Average statistics is incomplete without std/var.

Example: Smartphone price analysis

Retailer A: $800

Retailer B: $850

Retailer C: $820

Retailer D: $855

Retailer E: $870

Retailer F: $855

Retailer G: $825

Retailer H: $865

Retailer I: $840

Retailer J: $810

List = [800,850,820,855,870,855,825,865,840,810]

Ordered list = [800, 810, 820, 825, 840, 850, 855, 855, 865, 870]

Mean = 839

Median = 845

Mode = 855

Range = 70

Q1 = 25th percentile = 821.25

Q3 = 75th percentile = 855

Standard Deviation = 24.01

Variance = 576.67

Frequency

Frequency is the number of times a value of the data occurs. It is commonly used for categorical data where you want to know how often each category occurs. Table listing all classes and their frequencies is called a frequency distribution table.

Example: 1,3,3,2,4,1,2,2,1,2,3,5,4,1,2,1,3,1,4,1

Data Value	Frequency
1	7
2	5
3	4
4	3
5	1

A relative frequency is the ratio of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. In other words, percentage or proportion of the data value present in the dataset.

Relative frequency = $\frac{Frequency\ in\ the\ class}{Total\ number\ of\ observation} * 100$

Cumulative frequency is a way to show the running total of frequencies as you move through categories. It provides information about the number of data points that are less than or equal to a certain value or category.

Example: Customer Complaints at a Company

Imagine you work in the customer service department of an e-commerce company, and you have collected data on the types of complaints received over a month. You have categorized the complaints into four main types: Shipping Delays, Product Quality, Billing Issues, and Returns. Here is a summary of the data:

Shipping Delays: 15 complaints

Product Quality: 10 complaints

Billing Issues: 8 complaints

Returns: 12 complaints

Data Value	Frequency	Relative Frequency	Cumulative Frequency
Shipping Delays	15	0.3125	15
Product Quality	10	0.2083	25
Billing Issues	8	0.1667	33
Returns	12	0.3125	45

Graphical Representations

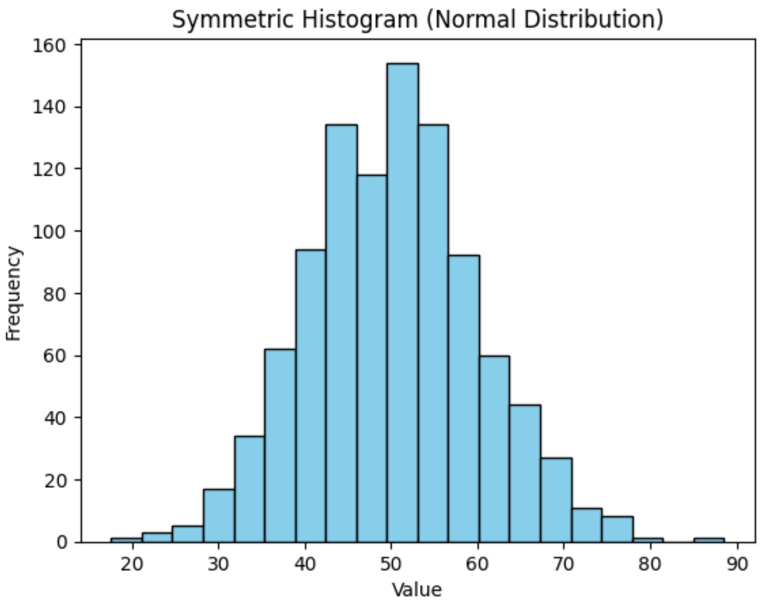
Graphical representations play a crucial role in descriptive statistics, providing visual insights into the distribution, frequency, central tendency and dispersion of data.

1. Histograms

- Display the distribution of continuous data.
- Divides data into intervals (bins) and represents the frequency or density of observations in each bin.
- The number of bins can impact the appearance of the histogram.
- Vertical axis represents the frequency in each bin.
- Helps visualize the shape, central tendency and spread of the data
- Useful for identifying patterns, potential outliers, and skewness.

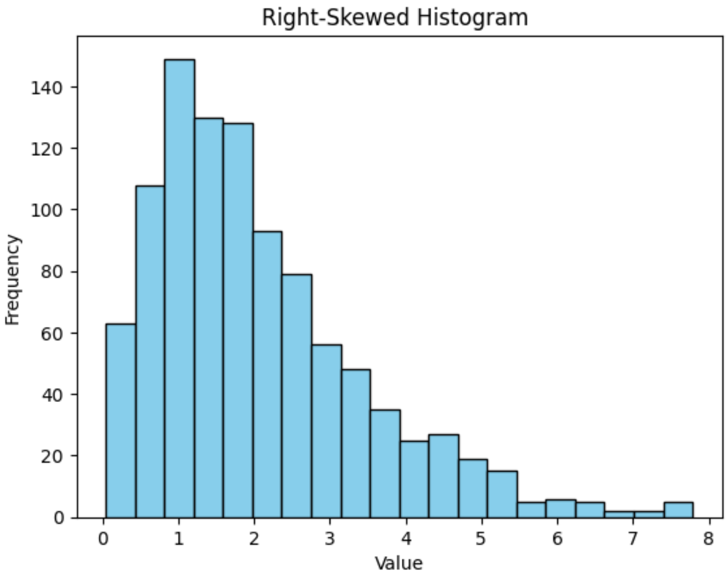
Types of skewed histograms:

- ○ Symmetric (Normal Distribution):
  - ▪ Bell Shaped curve
  - ▪ Mean, median and mode are at the centre.
  - ▪ The data is evenly distributed on both sides of the mean.

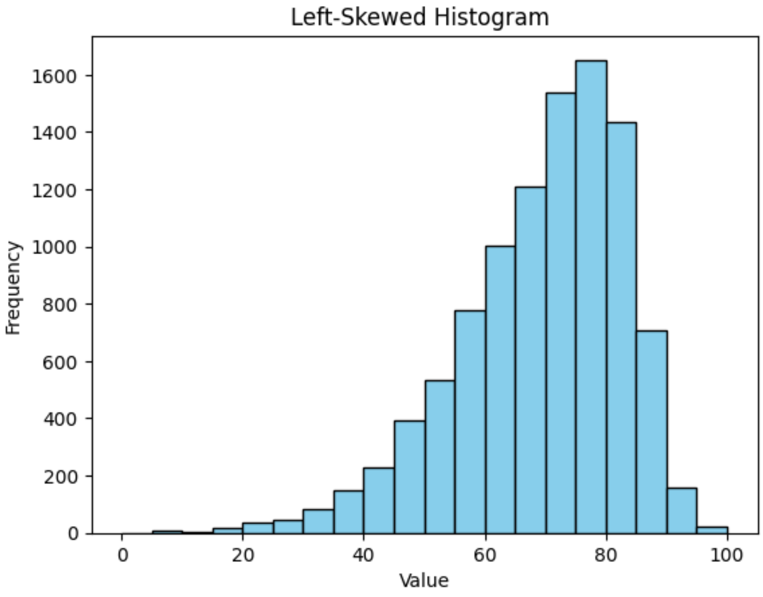


- ○ Positively skewed (Right skewed) Distribution:
  - ▪ Tail on the right side

    - Mean is greater than the median
    - Data concentrates on the left side, with a tail extending to the right.

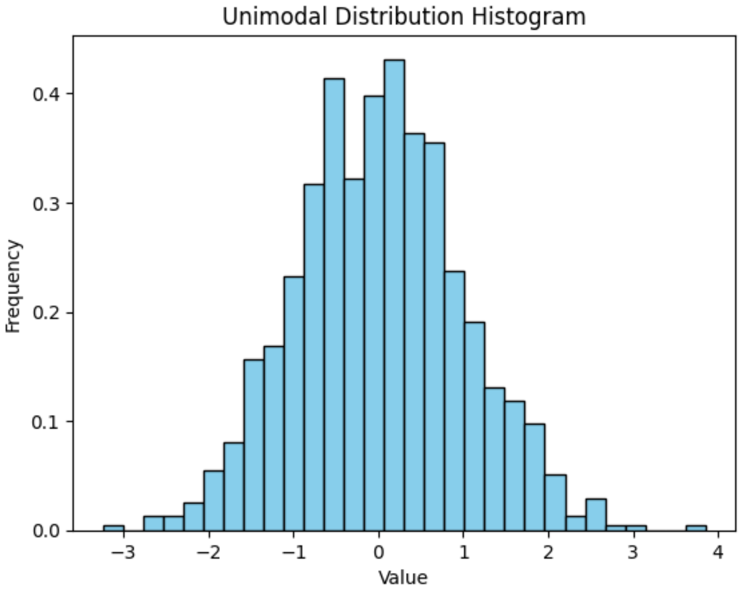


  - Negatively skewed (Left skewed) Distribution:
    - Tail on the left side.
    - Mean is less than the median.
    - Data concentrates on the right side, with a tail extending to the left.

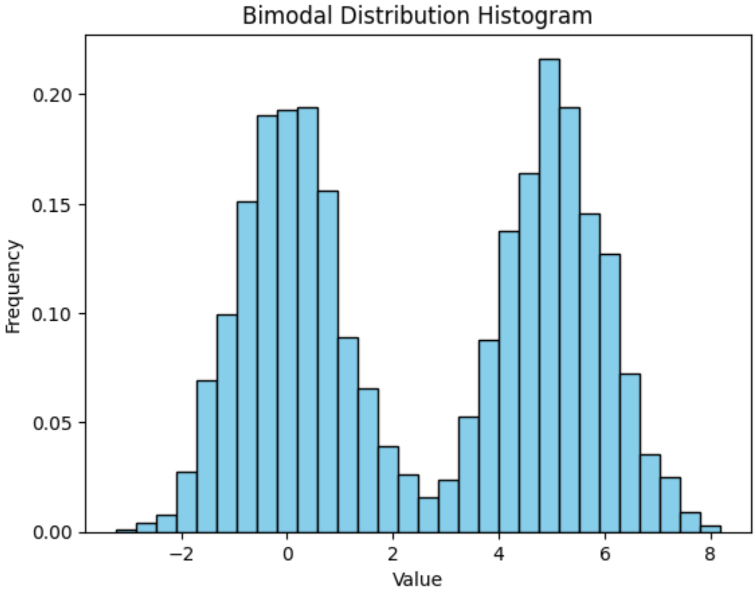


Types based on number of modes:

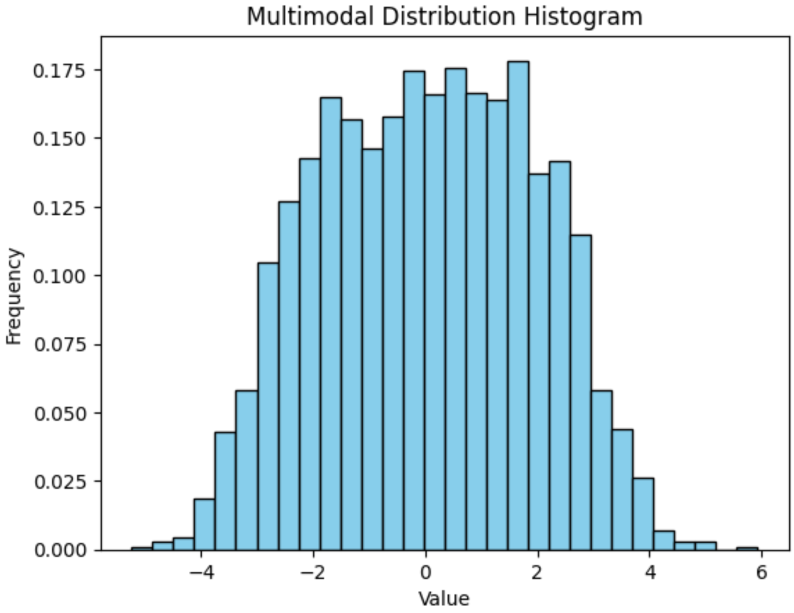
- Unimodal Distribution
    - One clear peak



- Bimodal Distribution
    - Two distinct peaks

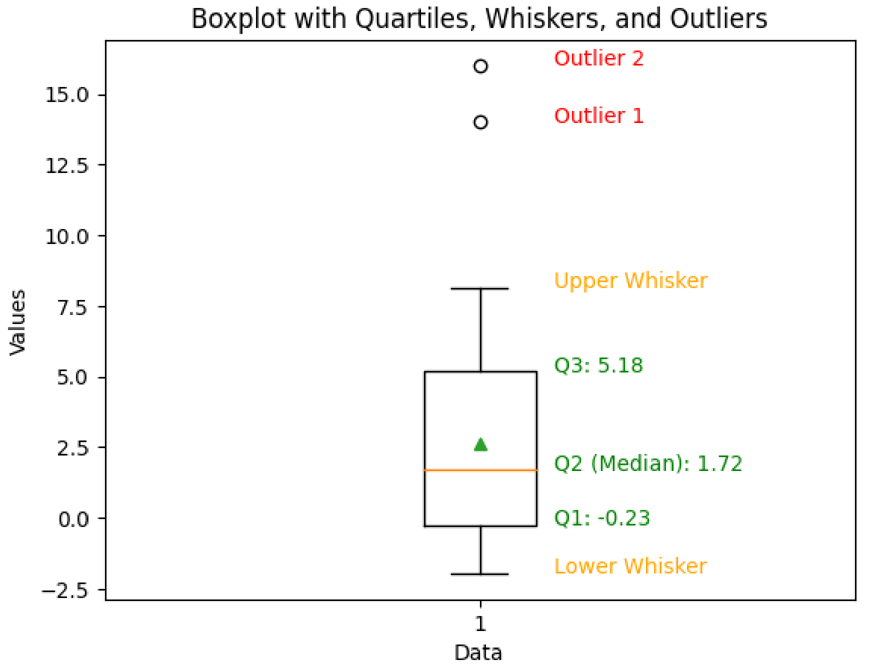


- Multimodal Distribution
    - More than two peaks



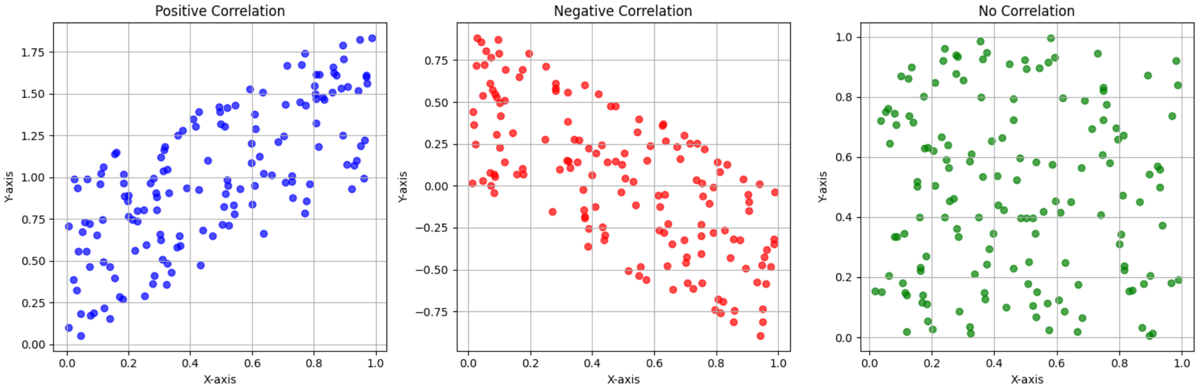
2. Box Plots (Whisker Plots)

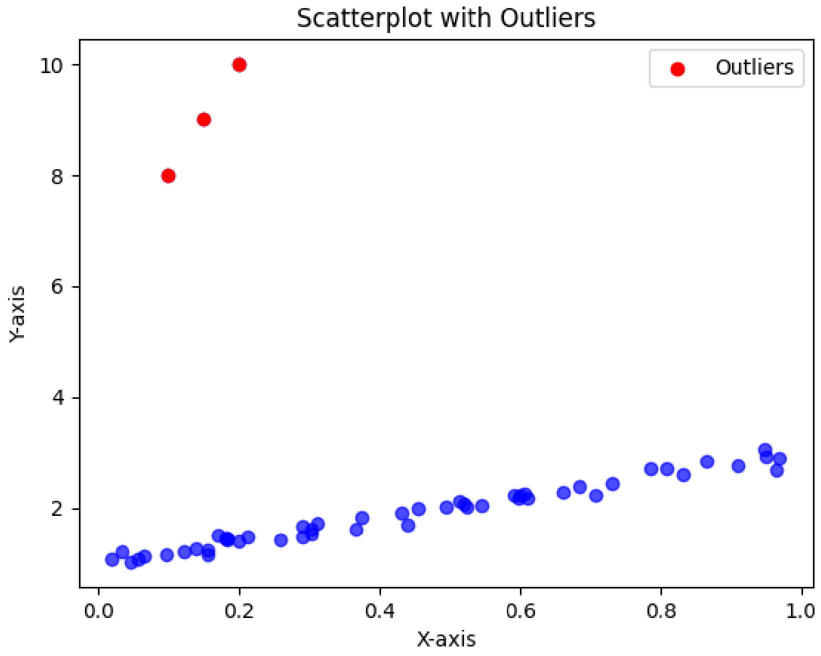
- Show the spread of the data and identify outliers
- Box represents the IQR with the median line inside
- Whiskers extend to the minimum and maximum values within a certain range
- Outliers are plotted individually, aiding in the identification of extreme values.
- Useful for comparing distributions and identifying skewness
- The range is typically set to 1.5 times the IQR. Values beyond this range are considered potential outliers.



3. Scatter Plots

- Visualize the relationship between the two continuous variables.
- Each point represents a pair of values, showcasing the joint distributions.
- Helps identify patterns, trends and correlation between variables
- Useful for detecting outliers and understanding the strength and direction of relationships
- Relationships can be Positive, Negative or No correlation.
- Can be enhanced with regression lines to highlight trends.





Outliers and understanding their impact

Outliers are data points that deviate significantly from the rest of the dataset. These values are notably different from the majority of the observations, can be unusually high or low and often potential to skew the overall interpretation of the data. Identifying and understanding outliers is a crucial aspect of data analysis, as they can have a substantial impact on statistical analysis.

1. Identification of Outliers:

Outliers can be identified through various statistical methods and visualizations. Common techniques include:

- Boxplots: Outliers are often visible as points beyond the "whiskers" of a boxplot.
- Z-Scores: Calculating the z-score for each data point helps identify values that deviate significantly from the mean.
- IQR (Interquartile Range): Outliers can be detected using the IQR by considering values outside a certain range.

2. Impact on Descriptive Statistics:

Outliers can heavily influence summary statistics:

- Mean and Standard Deviation: Outliers, especially those in the tails of a distribution, can distort the mean and inflate the standard deviation.

- Median and Quartiles: Robust statistics like the median and quartiles are less sensitive to outliers and provide a more reliable measure of central tendency and dispersion.

3. Effects on Inferential Statistics:

Outliers can affect the validity of statistical inferences:

- Parametric Tests: Outliers may violate assumptions of normality in parametric tests, leading to inaccurate results.
- Regression Analysis: Outliers can disproportionately impact regression coefficients, affecting the model's predictive performance.

4. Data Distribution and Modelling:

Outliers can impact the distribution of data:

- Skewness and Kurtosis: Outliers can introduce skewness and kurtosis, altering the shape of the distribution.
- Model Assumptions: Outliers may violate assumptions of linear models, leading to biased predictions and inaccurate model evaluations.

Correlation and Covariance

Covariance is a statistical measure that describes how much two variables change together. It indicates whether an increase in one variable is associated with an increase or decrease in another variable.

If the variables tend to increase or decrease together, the covariance is positive. If one variable tends to increase as the other decreases, the covariance is negative.

$$Cov(x,y) = \frac{\sum_{i=1}^{N}(Xi - \mu x)(Yi - \mu y)}{N}$$

It indicates the direction of the relationship; it does not quantify the strength of the relationship.

Correlation is a statistical measure that describes the extent to which two variables change together. In other words, it quantifies the degree to which a change in one variable is associated with a change in another variable. Correlation does not imply causation, but it helps us understand the relationship between variables in a dataset. It is a statistical technique used to measure the strength and direction of a linear relationship between two variables. The result is a correlation coefficient, a value between -1 and 1.

- Positive Correlation (1): When one variable increases, the other variable tends to increase. Correlation coefficient closer to 1 indicates a strong positive correlation.
- Negative Correlation (-1): When one variable increases, the other variable tends to decrease. Correlation coefficient closer to -1 indicates a strong negative correlation.
- No Correlation (0): There is no pattern or relation between the variables.

How to calculate correlation?

The most common method to calculate correlation is Pearson's correlation coefficient. It is calculated by dividing the covariance of the two variables by the product of their standard deviations.

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{var(x)}\,\sqrt{var(y)}} = \frac{Cov(x,y)}{std(x)std(y)}$$

Considerations and Cautions:

- Correlation does not imply causation: Just because two variables are correlated does not mean that one causes the other.
- Outliers: Extreme data points can influence correlation, so its essential to check for outliers.
- Non-linear Relationships: Correlation specifically measures linear relationships. Non-linear relationships may not be accurately represented.

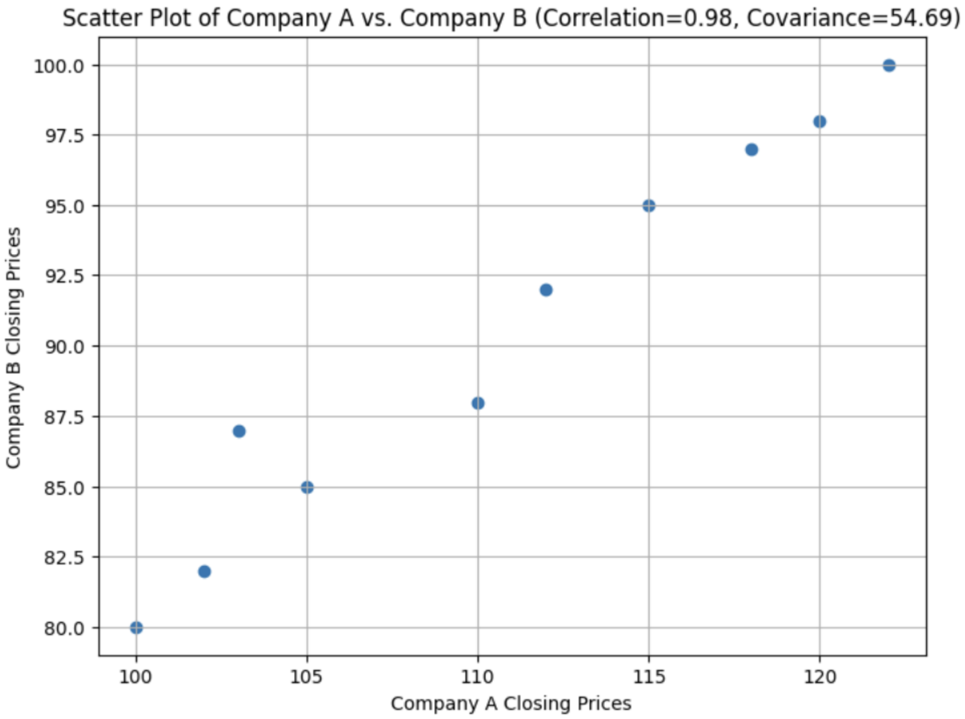
Note: Causation refers to the relationship between cause and effect. In a causal relationship, a change in the independent variable is directly responsible for a change in the dependent variable. Unlike correlation, which simply describes a relationship between two variables, causation implies a direct connection in which one variable influences the other.

Example: Correlation and Covariance in Stock Market Analysis

Suppose you are a financial analyst studying the relationship between the stock prices of two technology companies, Company A and Company B, over the past year. You have collected daily closing prices for both companies and want to analyze whether there is a correlation or covariance between their stock performance.

Company A prices = [100, 102, 105, 103, 110, 112, 115, 118, 120, 122]

Company B Prices = [80, 82, 85, 87, 88, 92, 95, 97, 98, 100]


Scatter Plot of Company A vs. Company B (Correlation=0.98, Covariance=54.69)
Company B Closing Prices
Company A Closing Prices