2025

Statistic
Lecture Notes

BEYZA KÜÇÜK

Table of Contents

Section 1: Basic Statistical Concepts

- **Meaning of Statistics**

  Statistics is the science of collecting, analyzing, interpreting, and presenting data. It is used to draw meaningful conclusions from data.

- **Statistics as a Method**

  It is the entirety of techniques used in the collection, processing, analysis, and interpretation of quantitative data pertaining to events that can be the subject of statistics.

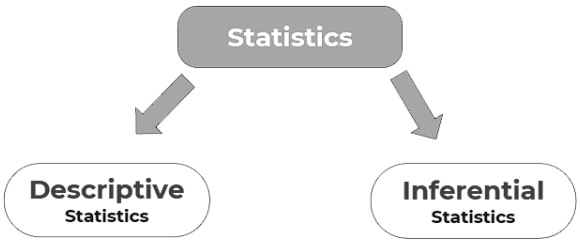
- **Eisenhower Matrix** It is a tool used for prioritization. It classifies tasks according to their urgency and importance.

- **Characterization** It is the process of determining and defining the characteristics of data. This process includes the following steps:

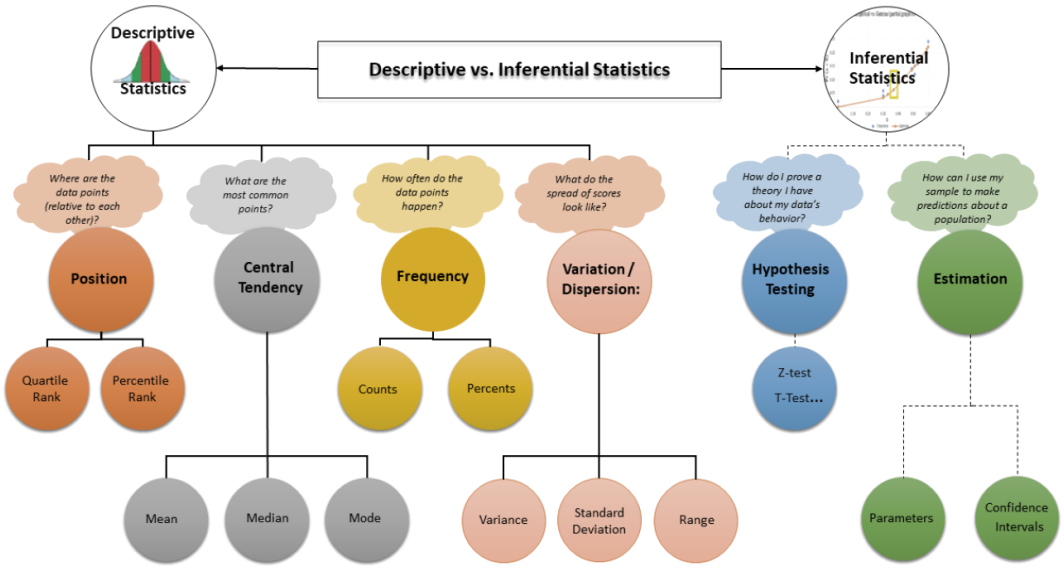
  - **Data Collection:** The systematic collection of data.
  - **Analysis:** The examination of data using statistical methods.
  - **Visualization:** Presenting data in the form of graphs or tables.
  - **Inference:** Drawing meaningful conclusions from the data.
  - **Presentation:** Sharing the findings in an understandable way.
  - **Reasons for Statistics:** To correctly interpret data and guide decision-making processes.
  - **Importance of Statistics:** It is critical for making accurate decisions in scientific research, the business world, and daily life.

How Much Statistical Knowledge is Needed? Data scientists extract meaningful information from data using statistical methods. Basic statistical knowledge is necessary for data science.

Types of Statistics

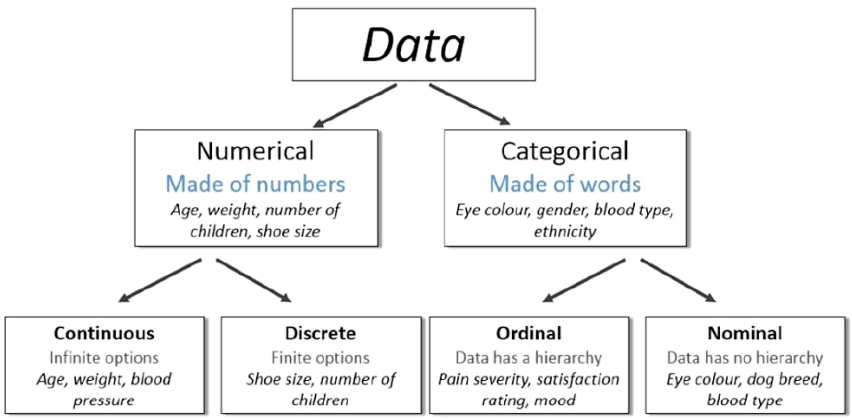


- **Descriptive Statistics**
    - Used to summarize and describe data. Measures such as mean, median, mode, and standard deviation are used.
- **Inferential Statistics**
    - Used to draw general conclusions from data. Hypothesis tests and confidence intervals are within this scope.



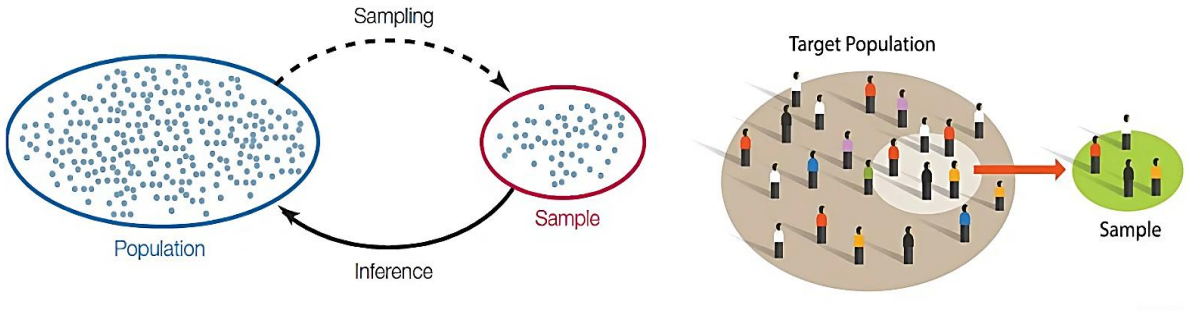
Types of Data:

- Quantitative (numerical) and qualitative (categorical) data.



Parameters and Statistics:

- Parameter: Describes the characteristics of a population.
- Statistic: Describes the characteristics of a sample.



Probability vs. Statistics:

- **Probability:** Examines the likelihood of events occurring.
- **Statistics:** The outcome is known and used to make an inference about the process.

Levels of Measurement:

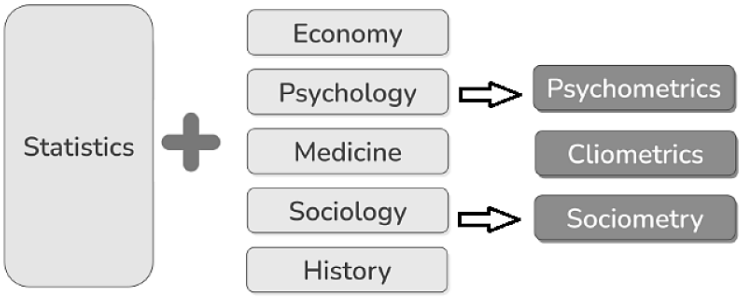
Different levels at which data is measured:

- **Nominal:** Categorical data (e.g., gender).
- **Ordinal:** Ordered categorical data (e.g., education level).
- **Interval:** Numerical data, zero point is arbitrary (e.g., temperature).
- **Ratio:** Numerical data, absolute zero point (e.g., weight).

Levels of Measurement

Incremental Progress	Measure Property	Mathematical Operators	Advanced Operations	Central Tendency
Nominal	Classification, Membership	=, !=	Grouping	Mode
Ordinal	Comparison, Level	>, <	Sorting	Median
Interval	Difference, Affinity	+, -	Yardstick	Mean, Deviation
Ratio	Magnitude, Amount	*, /	Ratio	Geometric Mean, Coeff. of Variation

Real-World Applications of Statistics



- **Medical Studies**

Statistics is used in medical research to test the effectiveness of drugs, measure the prevalence of diseases, and evaluate treatment methods. **Example:** The effectiveness of COVID-19 vaccines was tested using statistical methods.

- **Weather Forecasts**

Meteorologists make weather forecasts using statistical models. **Example:** The probability of rain or temperature predictions are based on statistical data.

- **Quality Control**

Companies use statistical methods to test the quality of their products and offer the best quality. **Example:** An automobile manufacturer tests the reliability of each vehicle.

- **Stock Market**

Investors use statistical analyses to predict stock prices and assess risks. **Example:** Stock market trends and volatility analyses.

- **Consumer Goods**

Companies improve their products by analyzing consumer preferences. **Example:** Testing a new product before it is launched.

- **Government**

Governments use statistics in census taking, economic planning, and policy-making processes.

Example: Evaluating the impact of tax policies.

- **Emergency Preparedness**

Statistics is used to predict the effects of natural disasters and create emergency plans. **Example:** Earthquake risk analyses.

- **Political Campaigns**

Politicians analyze voter preferences to determine campaign strategies. **Example:** Polls and election predictions.

- **Sports**

Athletes and teams use statistical data to improve their performance. **Example:** A basketball player's shooting percentage.

- **Research**

In scientific research, statistics is used for data analysis and interpretation of results. **Example:** Clinical trials of a new drug.

- **Education**

Educational institutions use statistics to evaluate student performance and improve education policies. **Example:** Analysis of exam results.

- **Prediction**

Statistics is used to predict future events. **Example:** Sales forecasts or economic growth projections.

- **Predicting Disease**

Epidemiologists use statistics to predict the spread and effects of diseases. **Example:** Modeling the spread rate of COVID-19.

- **Insurance**

Insurance companies use statistics to assess risks and determine premiums. **Example:** Calculation of car insurance premiums.

- **Financial Markets**

Financial analysts use statistics to evaluate market trends and risks. **Example:** Portfolio management and risk analysis.

- **Business Statistics**

Companies use statistics to analyze data in decision-making processes. **Example:** Analysis of sales data and marketing strategies.

Importance of Statistics in Daily Life

- **Decision Making:** Statistics enables us to analyze data to make accurate and informed decisions.
- **Predictions:** It is used to predict future events and make plans.
- **Quality and Efficiency:** Statistical methods are used to improve the quality of products and services.
- **Risk Management:** Statistical analyses are performed to assess and reduce risks.

Summary Table

Concept	**Description**
Meaning of Statistics	The science of collecting, analyzing, interpreting, and presenting data.
Eisenhower Matrix	A tool used for prioritization.
Characterization	The process of determining and defining the characteristics of data.
Data Science vs Statistics	Data science extracts information from data using statistical methods.
Descriptive Statistics	Summarizing and describing data.
Inferential Statistics	Drawing general conclusions from data.
Types of Data	Quantitative and qualitative data.
Levels of Measurement	Nominal, Ordinal, Interval, Ratio.

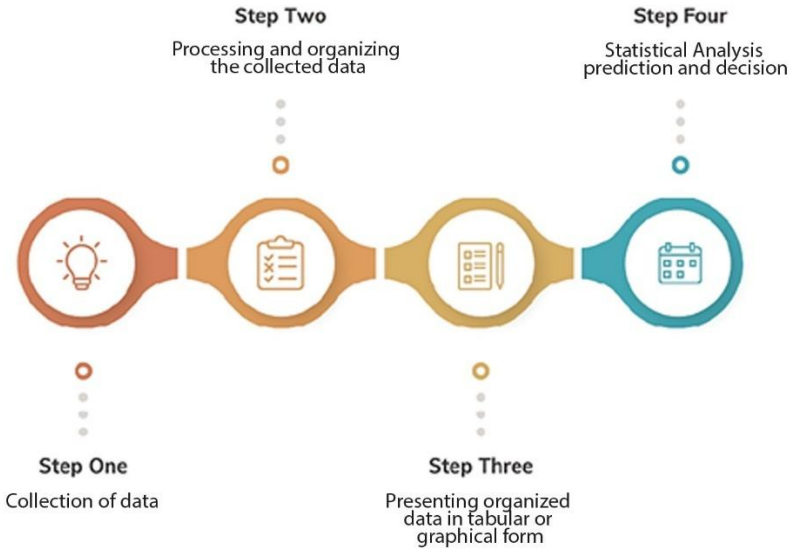
Data Science vs Statistics

Criterion	Data Science	Statistics
Mathematical Understanding	Uses mathematical models and algorithms.	Focuses on mathematical theories and statistical methods.
Problem Examination	Investigates problems in large datasets.	Focuses on the analysis and interpretation of data.
Exploratory Data Analysis (EDA)	Used to explore data and perform preliminary analysis.	Used to understand the distribution and relationships in data.
Trend Analysis	Analyzes trends and patterns in data.	Examines trends in data using statistical methods.
Generating Predictions	Makes predictions using machine learning models.	Makes predictions using statistical models.
Visualization	Uses tools and libraries to visualize data.	Uses statistical methods for the graphical presentation of data.
Reporting to Non-Technical Users	Presents findings understandably to non-technical users.	Reports findings and explains them to non-technical users.

Data Science: Used to analyze large datasets, create machine learning models, and develop automated systems.

Statistics: Used for data analysis, hypothesis testing, and creating statistical models.

Rank Tracked in Statistics



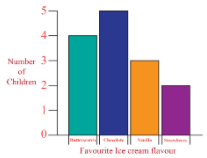
2.Section: Data Visualization and Fundamental Analysis

Graphical Summarization for Data

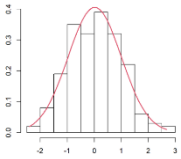
Frequency Tables — Discreate and Continous Data

Colour	Tally marks	Frequency
Black	|	1
Blue	𝍸	5
Pink	||	2
White	||||	4
		Total = 12

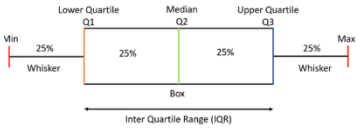
Bar Charts — Discreate Data



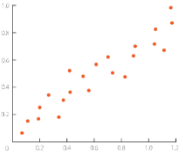
Histograms — Continous Data; Variatin, Skewnes, Outliers



Box Plots — Continous Data; Variatin, Skewnes, Outliers



Scatter Plots — Two Continous Variable

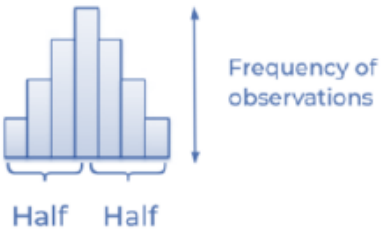


Data Visualization - Graphical Representation

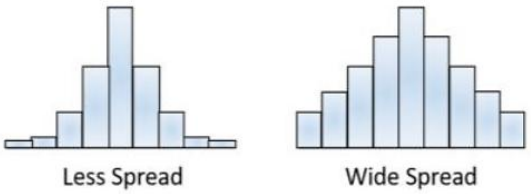
- **Definition:** It is the visual representation of data using graphs or charts.
- **Example:** Bar charts, pie charts, histograms.

Data Patterns

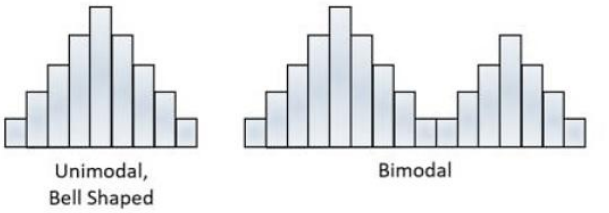
- **Definition:** Patterns or trends observed in the distribution of data.
- **Types:**
  - **Center:** The center of the distribution is graphically located at the median of the distribution.



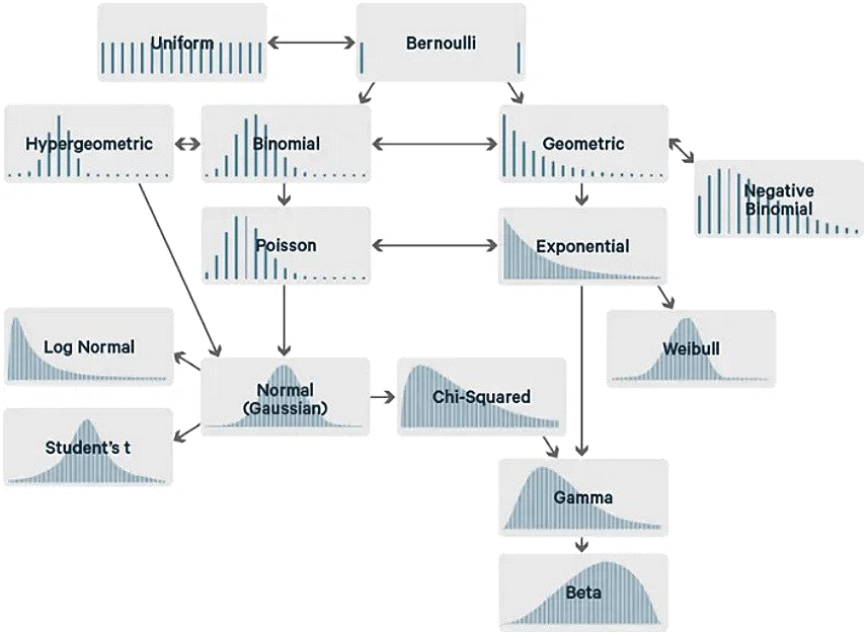
- **Spread:** We examine how widely the data is spread over a range.



- **Shape:** The shape of the data distribution (e.g., symmetric or skewed).
  - **Symmetric:** Whether the data is symmetric about the mean.
  - **Number of Peaks:** How many peak points are in the distribution.
  - **Skewness:** Whether the data is skewed to one side.
  - **Uniform:** The even distribution of the data.

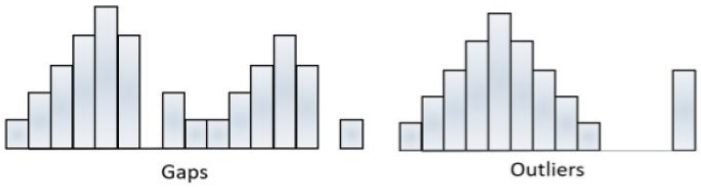


Probility Distributions



Unusual Features

- **Definition:** Unexpected or unusual characteristics in the data.
- **Types:**
    - **Gaps:** Missing or empty areas in the data.
    - **Outliers:** Extreme values in the data.



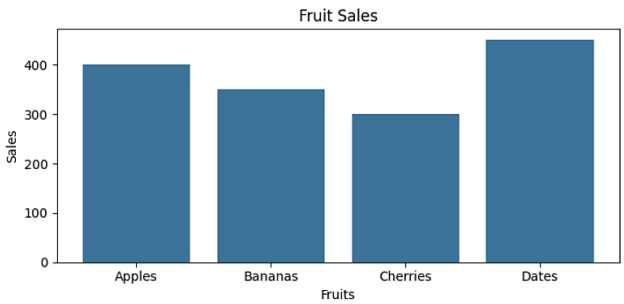
Frequency Table

- **Definition:** It is a table that shows how many times data repeats within specific intervals. It is used to summarize data graphically.
- **Area of Use:** Suitable for discrete and continuous data.
- **Features:**
    - Divides the data into categories and shows the frequency of each category.
    - A frequency table alone does not provide much information but forms the basis for other graphs.
- **Types:**
    - **Relative Frequency:** The ratio of each interval to the total data.
    - **Cumulative Frequency**: The total frequency up to a certain interval.

Colour	Tally marks	Frequency
Black	|	1
Blue	𝍸	5
Pink	||	2
White	||||	4
		Total = 12

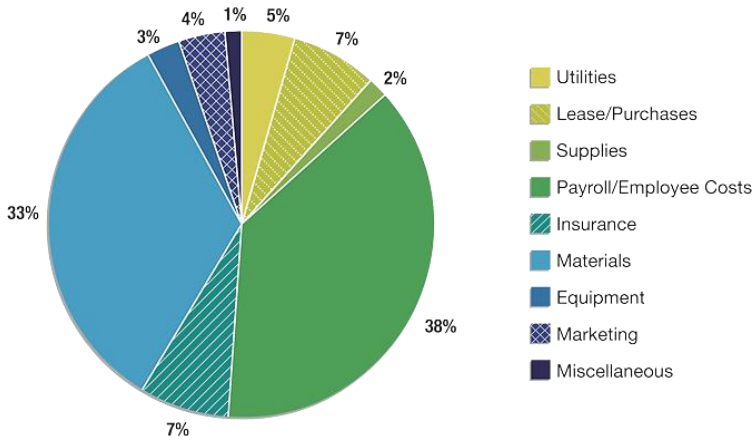
Bar Chart

- **Definition:** It is a graph in which categorical data is represented by bars. The height of each bar shows the frequency of each attribute.
- **Example:** Showing the sales amounts of different products.



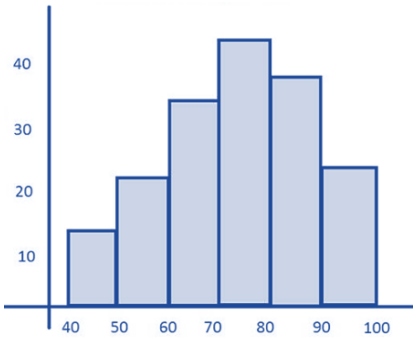
Pie Chart

- **Definition:** It is a graph in which data is represented as slices within a circle. It is generally used for nominal and ordinal (categorical) variables.
- **Example:** The distribution of a company's revenues by different sources.



Histogram

- **Definition:** A graph showing the frequency distribution of continuous data.
- **Area of Use:** Detection of variation, skewness, and outliers.
- **Features:**
  - There are no gaps between the columns.
  - Shows the regions where data is concentrated and the character of the distribution.
- **Interpretation:**
  - Standard deviation and variance can be calculated.
  - Outliers can be detected.
  - Median and mean can be interpreted.
- **Example:** Distribution of students' exam scores.

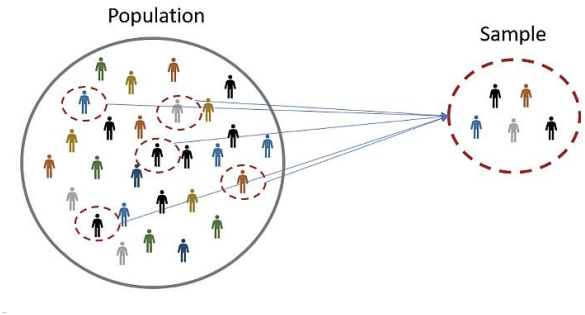


Bar Chart vs Histogram

Criterion	Bar Chart (Çubuk Grafik)	Histogram
Categories	There are categories.	There are no categories, there are intervals.
Variable Type	Discrete variables are used.	Continuous variables are used.
Data Type	Presents categorical data.	Presents numerical data.
Space Between Bars	There is space between bars.	There is no space between bars.
Graphic Representation	Makes a schematic comparison of categorical data.	Shows the frequency distribution of continuous data.
Area of Use	Used for comparing categorical data.	Used to show the distribution of numerical data.

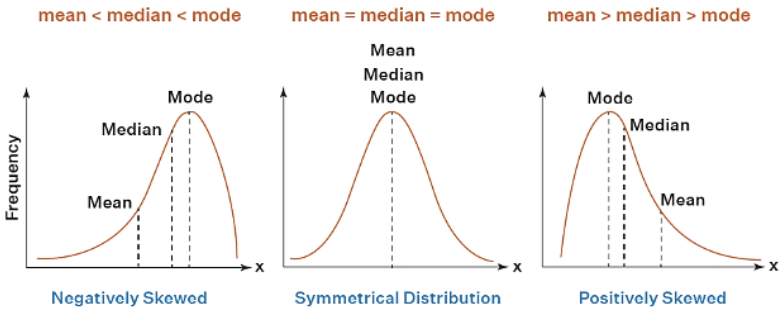
Populations and Samples

- **Definition:** Population represents the entire dataset, while a sample is a subset drawn from the population.
- **Types:**
    - **Parameters:** Describe the characteristics of the population.
    - **Statistics:** Describe the characteristics of the sample.



Measure of Centre

- **Definition:** These are measures that indicate the central point of the data.
- **Types:**
- **Mean:** It is the arithmetic average of the data.
- **Median:** It is the middle value of the data.
- **Mode:** It is the value that repeats most frequently in the data. It is also called the peak value.

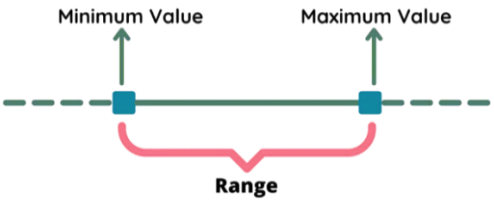


Mean vs Median

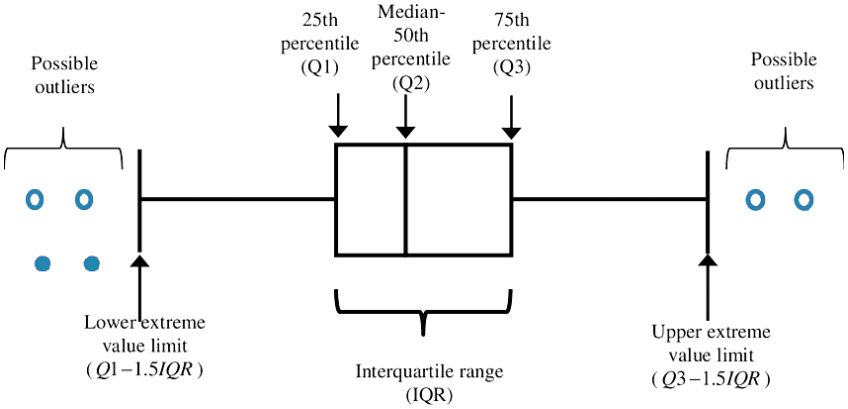
Criterion	Mean	Median
Definition	Found by dividing the sum of the data by the number of data points.	The middle value when the data is ordered.
Sensitivity to Outliers	Highly affected by outliers.	Not affected by outliers.
Large Datasets	Can be misleading if there are outliers.	More reliable if there are outliers.
Büyük Veri Setleri	A better measure if there are no outliers.	As good as the mean if there are no outliers.
Area of Use	Symmetric distributions and situations without outliers.	Skewed distributions and situations with outliers.
Example	The average of students' exam scores.	Median can be used in salary distribution.

Dispersion (Measure of Spread)

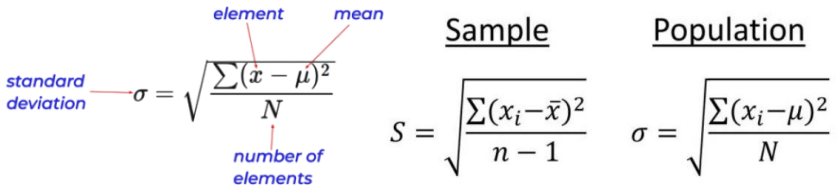
- **Definition:** These are measures that show how widely the data is spread over a range.
- **Types:**
    - **Range:** It is the difference between the largest (maximum) and smallest (minimum) values of the data.

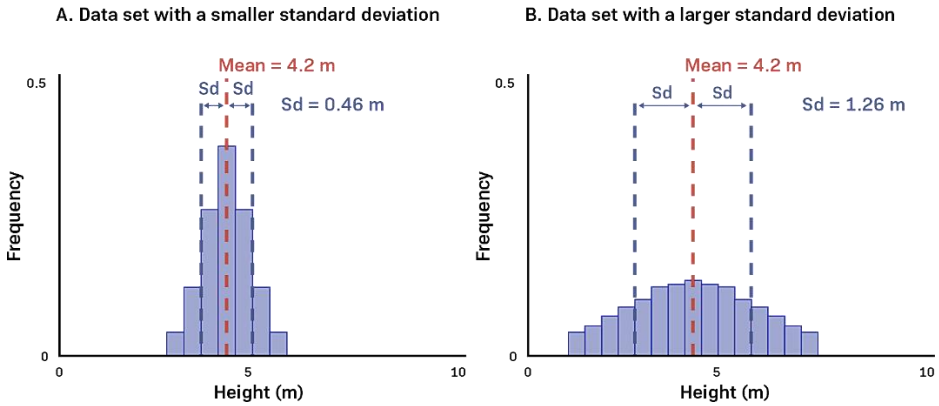


    - **Interquartile Range (IQR):** It is the value that divides a group of numbers into four equal parts. It is the spread of the middle 50% of the data.



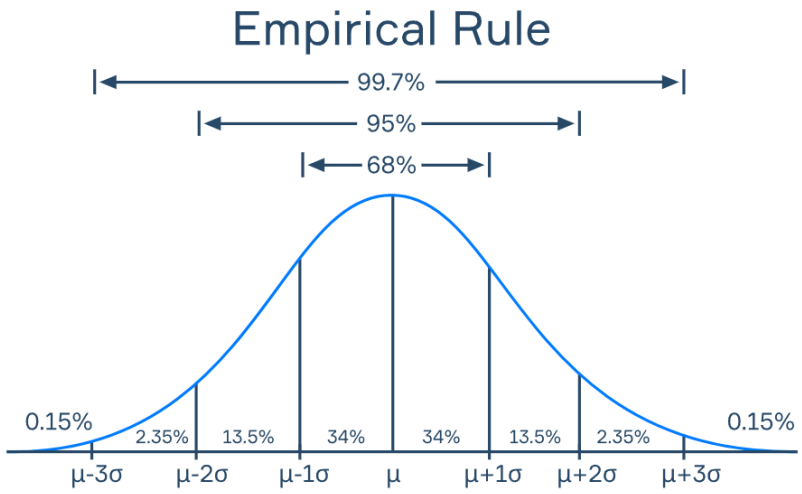
- **Standard Deviation:** It shows how much the data deviates from the mean. The more the data is spread, the larger the standard deviation.





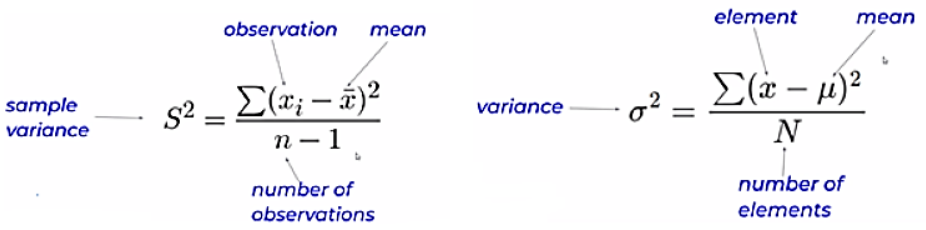
- **Empirical Rule (68-95-99.7 Rule):** Also known as the 3 sigma rule, it states that in a normal distribution, 68%, 95%, and 99.7% of the data falls within certain standard deviation intervals of the mean.

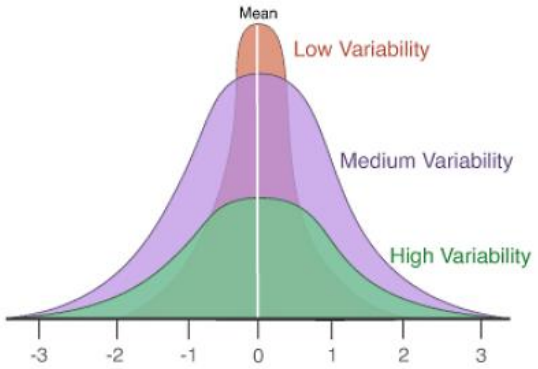
    1. At 68%: (Mean - 1 standard deviation) and (Mean + 1 standard deviation)
    2. At 95%: (Mean - 2 × standard deviation) and (Mean + 2 × standard deviation)
    3. At 99.7%: (Mean - 3 × standard deviation) and (Mean + 3 × standard deviation)



Variation

- **Definition:** Variance is defined as the average of the squared differences from the mean. It is a measure that shows how variable the data is.
- **Example:** Low variation is desired in the quality control of a product.





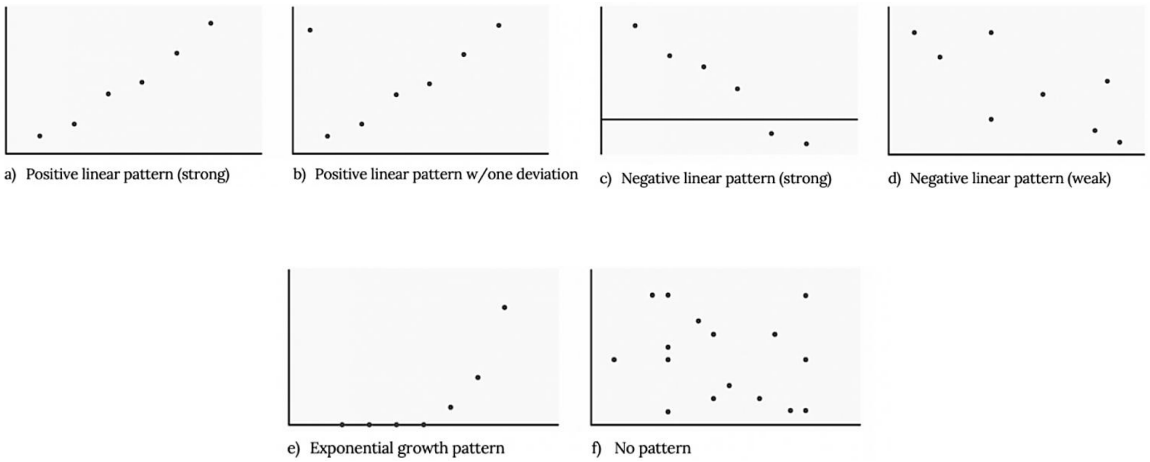
Summary Table

Concept	Description
Data Visualization	Representation of data with graphs.
Patterns	Patterns observed in the distribution of data.
Unusual Features	Unexpected features in the data (gaps, outliers).
Frequency Table	Table showing how many times data repeats within specific intervals.
Bar Chart	Representation of categorical data with bars.
Pie Chart	Representation of data as slices within a circle.
Histogram	Graph showing the frequency distribution of continuous data.
Populations and Samples	Differences between population and sample.
Measure of Centre	Measures indicating the central point of the data (mean, median, mode).
Measure of Spread	Measures showing how widely the data is spread over a range.
Variation	Measure showing how variable the data is.

3.Section : Relationship Analysis

- **Scatter Plot**
    - **Definition:** Used to show the relationship between two continuous variables. It indicates the direction and strength of the relationship.
    - **Area of Use:** Correlation analysis and detection of linear relationships.
    - **Features:**
        - **Positive Correlation:** Points are distributed upwards to the right. There is a positive relationship if both variables increase.
        - **Negative Correlation:** Points are distributed downwards to the right. There is a negative relationship if one variable increases and the other decreases.
        - **Strong/Weak Relationship:** Related to the linear arrangement of the points. In a strong relationship, there is a linear orderly trend. In a weak relationship, there is a more scattered arrangement close to linear.

Example Visual:



- If there is no specific pattern, and the graph has a rounded, circular appearance, it indicates a "no pattern" arrangement. Relationships with no pattern cannot be analyzed. (Example: The relationship between increased ice cream sales in summer and increased house prices.)
- When we look at a scatter plot, we want to see the general pattern and how much deviation there is from that pattern. We analyze the relationship, strength, and direction between two variables from the distributions in the scatter plot.

- If the points in a scatter plot move in the form of a straight line, there is a strong relationship. Being positive or negative does not affect its strength. There can be a strong relationship in the negative direction as well.

Distribution Character and Prediction Processes

- **Distribution Character:**
    - Distribution analysis is used to determine if the dataset is normal.
    - Sector-specific information is used (e.g., Poisson distribution in traffic engineering).

- **Prediction Processes:**
    - When the distribution character is known, how the data will behave in new situations can be predicted.
    - This simplifies prediction processes.

Summary Table

Graph Type	Area of Use	Features
Frequency Table	Discrete and continuous data	Divides data into categories, shows frequencies.
Bar Chart	Discrete data	Compares between categories.
Histogram	Continuous data	Distribution, skewness, outlier detection.
Box Plot	Continuous data	Quartiles, median, outlier detection.
Scatter Plot	Two continuous variables	Shows the direction and strength of the relationship.

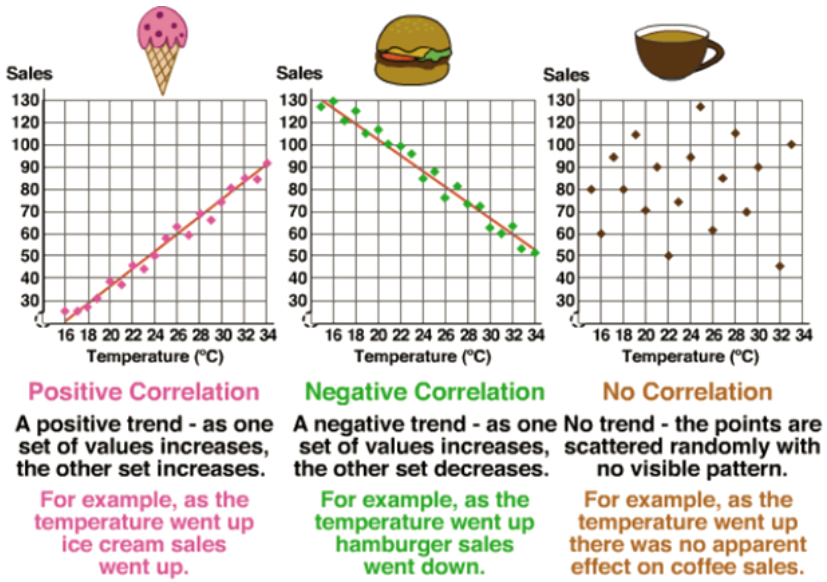
Line of Best Fit

- **Definition:** A line (trend line) drawn on a scatter plot that best represents the general trend of the data.
- **Area of Use:** To model the linear relationship between variables and make future predictions.
- **Formula:**

$$y=ax+b$$

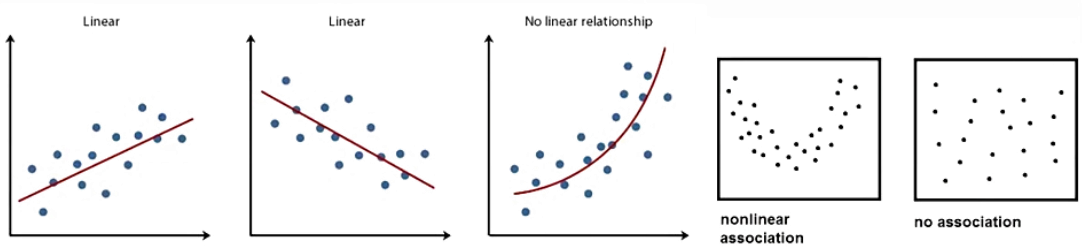
    - **a (Slope):** Indicates how much a 1-unit increase in the X variable affects Y.
    - **b (Intercept):** The value Y takes when X=0.

- **Example Visual:**



Linearity

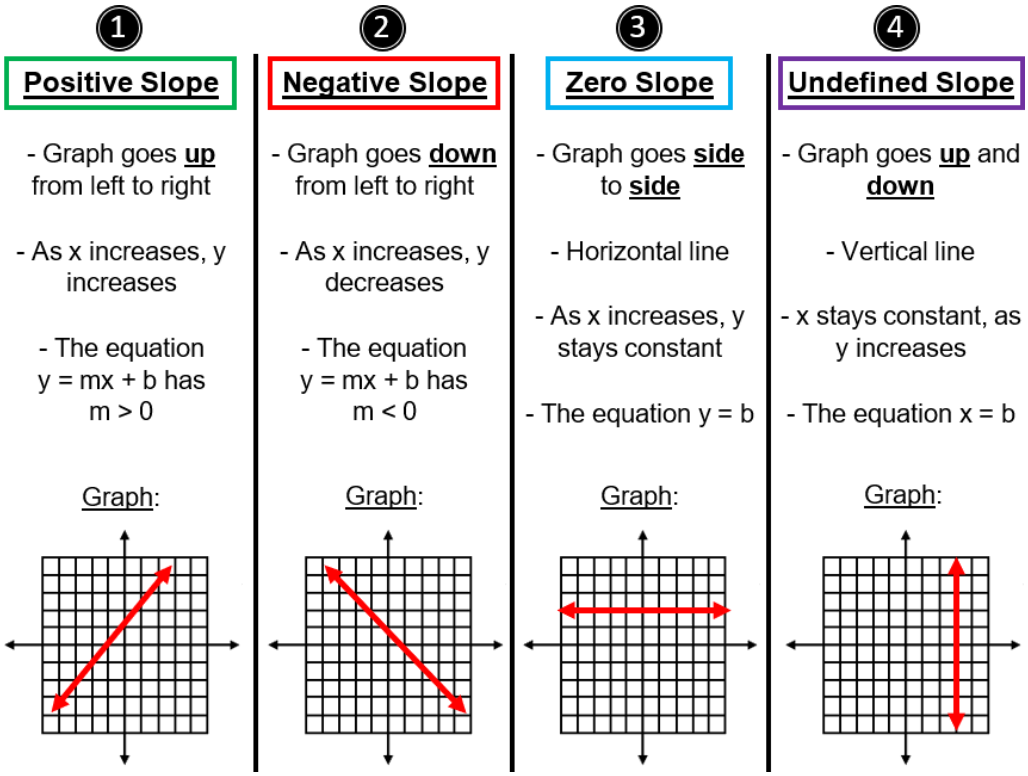
- **Linear:** If the relationship between two variables can be expressed with a straight line, this relationship is linear.
- **Non-Linear:** If the relationship cannot be expressed with a straight line, it is non-linear.
- **Perfect Linearity Does Not Exist:** The correlation coefficient (r) does not exactly equal 1.0 or -1.0, but it can approach these values.
- **Non-Linear Relationships:** Can be examined by dividing them into two through advanced analyses.



Slope

- **Definition: Shows how the change in the X variable affects the Y variable.**
- **Types:**
    - **Positive Slope:** Y increases as X increases.
    - **Negative Slope:** Y decreases as X increases.
    - **Undefined Slope:** Vertical line (Y changes while X remains constant).
    - **Zero Slope:** Horizontal line (Y is constant, X changes).

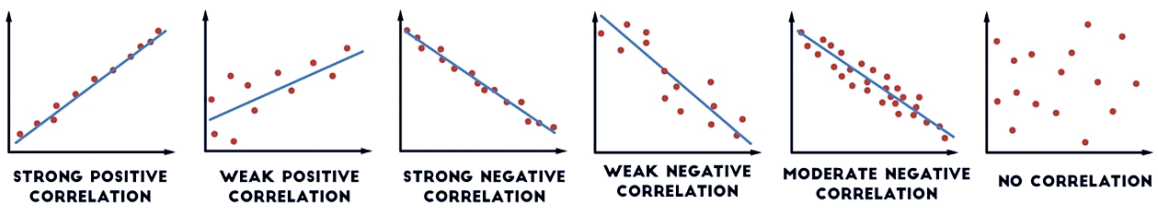
- **Example Visual:**



Strength

- **Definition:** Expresses how strong a relationship the distribution in the graph shows.
- **Interpretation:**
    - **Strong Relationship:** Points are close to the line.
    - **Weak Relationship:** Points are scattered.

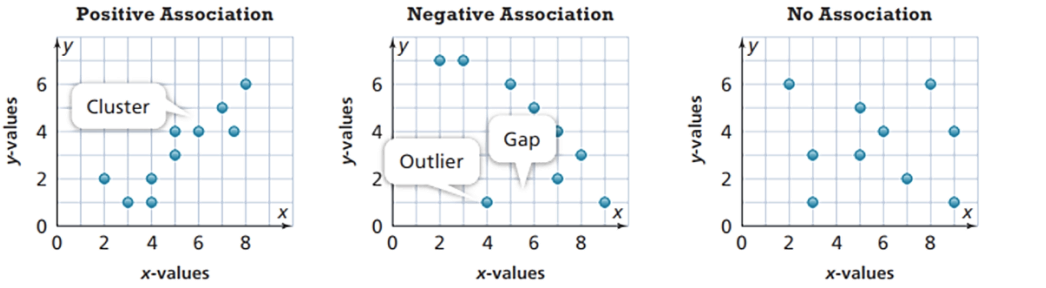
- **Example Visual:**



Unusual Features

- **A. Clusters**
    - **Definition:** The grouping of data showing similar behavior together.
    - **Example:** Similar shopping behaviors in e-commerce data.

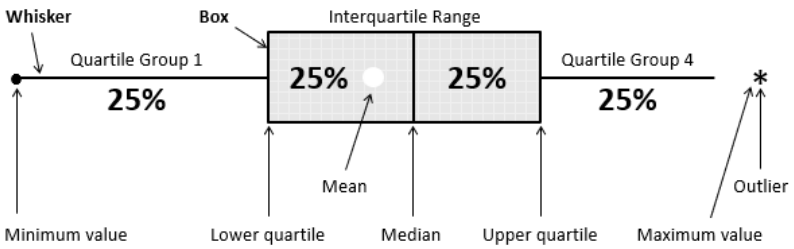
- **B. Gaps**
    - **Definition:** Missing or empty spaces in the data.
    - **Solution:** Missing data should be filled or the analysis should be done separately.
    - **Example:** We should analyze by separating the periods before and after the Covid period when no entries were made.
- **C. Outliers**
    - **Definition:** Values that are significantly different from other data.
    - **Impact:** Weakens the prediction power.
    - **Solution:** Outliers should be detected and corrected or removed from the analysis.



NOTE: There is no such thing as garbage data; the data is thoroughly investigated until all possible insights are extracted.

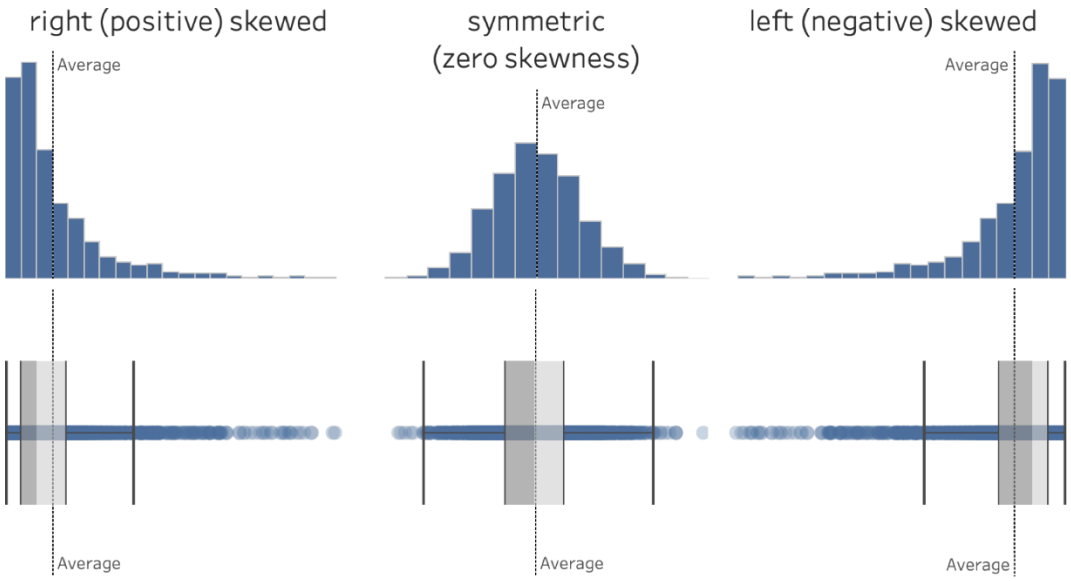
Box Plot

- **Definition:** A graph showing the distribution, quartiles, and outliers of a dataset.
- **Components:**
    - **Q1 (25th percentile):** Lower quartile.
    - **Q3 (75th percentile):** Upper quartile.
    - **Median:** Middle line.
    - **IQR (Interquartile Range):** Q3−Q1.
    - **Outliers:** Values outside Q1−1.5×IQR and Q3+1.5×IQR.
    - The outer part of the whisker is the IQR.
- **Example Visual:**



Skewness

- **Definition:** The state of a data distribution not being symmetrical.
- **Types:**
    - **Positive Skew:** Skewed to the right (long right tail).
    - **Negative Skew:** Skewed to the left (long left tail).
- **Example Visual:**



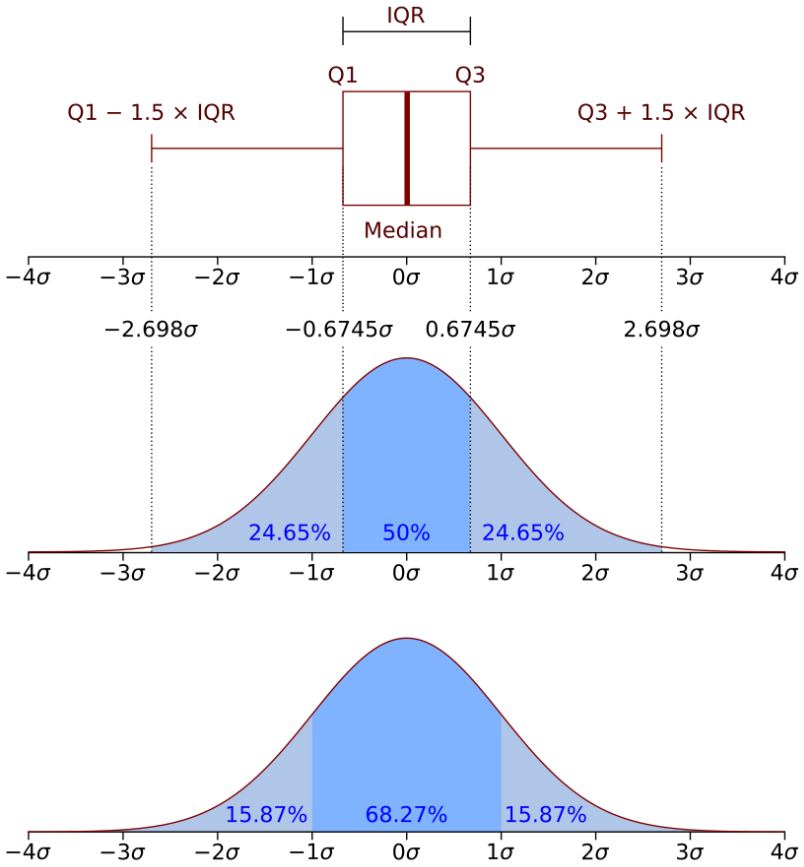
Box Plot - Min & Max Values

- **Definition:** The minimum and maximum values in a dataset.
- **Features:**
    - Data must be ordered.
    - Min and max values are not used for outlier detection; IQR is used.

IQR Rule

- **Definition:** A rule used to detect outliers in a dataset.
- **Proposed By:** Suggested by statistician John Tukey.
- **How It Works:**
    - **IQR (Interquartile Range):** Q3−Q1 (Upper quartile - Lower quartile).
    - **Lower Bound:** Q1−1.5×IQR
    - **Upper Bound:** Q3+1.5×IQR
    - **Values outside these bounds are considered outliers.**

- **Example:**
    - If Q1 = 25 and Q3 = 75, then IQR = 50.
    - Lower Bound: 25−1.5×50=−50
    - Upper Bound: 75+1.5×50=150
    - Values outside this range are outliers.



Summary Table

Topic	Short Description
Line of Best Fit	The line that best represents the trend of the data.
Linearity	Whether the relationship is linear or not.
Slope	Shows how the change in X affects Y.
Strength	Expresses the strength of the relationship.
Clusters	Grouping of data showing similar behavior.
Gaps	Missing parts or empty spaces in the data.
Outliers	Values significantly different from other data.
Box Plot	Graph showing data distribution, quartiles, and outliers.
Skew	The state of data distribution not being symmetrical.
1.5 IQR Rule	Rule used for outlier detection.

Top 60 Statistics Interview Questions 2024

Question 11: What is the benefit of using box plots?

Answer: Boxplot is a visually effective representation of two or more data sets and facilitates quick comparison between a group of histograms.

Question 12: How to detect outliers?

Answer: The best way to detect outliers is through graphical means. Apart from that, outliers can also be detected through the use of statistical methods using tools such as Excel, Python, SAS, among others. The most popular graphical ways to detect outliers include box plot and scatter plot. We can use IQR. Out of [Q1-1.5*IQR, Q3+1.5*IQR] range show us outliers.

Covariance

- **Definition:** Measures how two variables change together. Covariance indicates the negative or positive relationship between two datasets.
- **Formula:**

Population Covariance

$$Cov(X,Y) = \frac{\Sigma\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{N}$$

Sample Covariance

$$Cov(X,Y) = \frac{\Sigma\left(x_i-\bar{x}\right)\left(y_i-y\right)}{N-1}$$

These are the formula for finding Population and Sample Covariance.

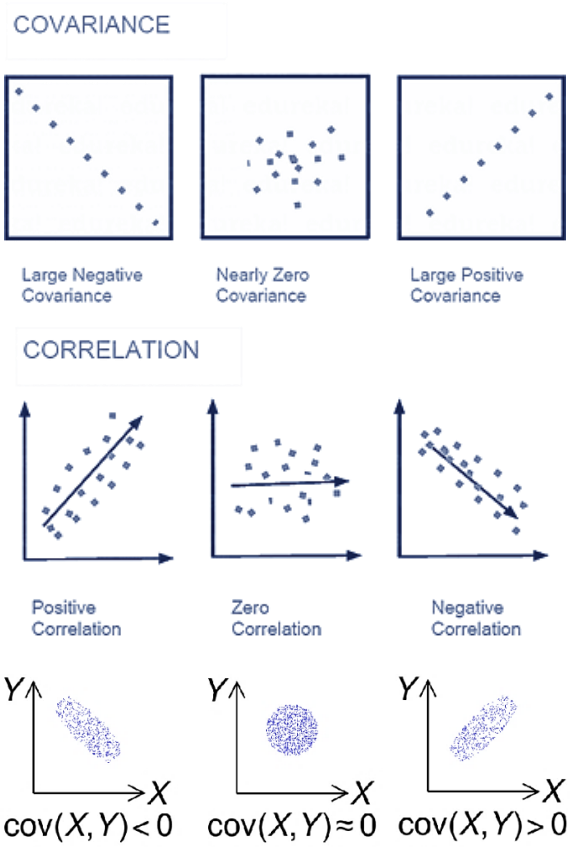
- x_i = data value of x
- y_i = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

Comment:

- **Positive Covariance:** Y increases as X increases.
- **Negative Covariance:** Y decreases as X increases.
- **Zero Covariance:** There is no relationship between the two variables.

Properties:

- Only shows the **direction** of the relationship, not the **strength**.
- The value range is unlimited, making it difficult to interpret.



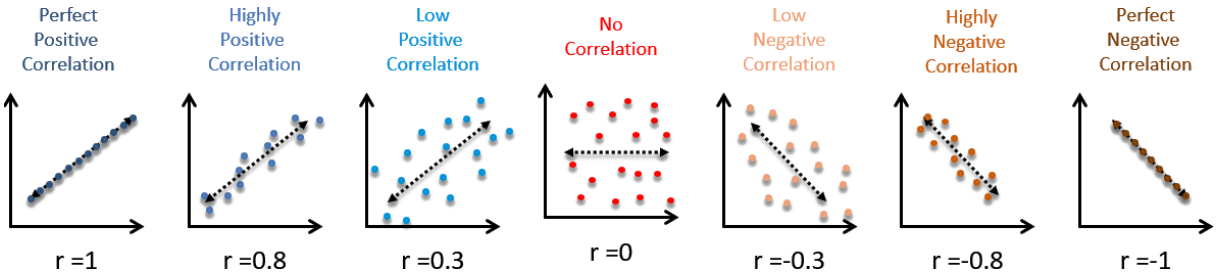
Correlation

- **Definition:** Shows both the direction and the strength of the relationship between two variables.
- **Formula:**

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- **Value Range:** Between -1 and +1. -1 and +1 have the same strength; the difference is the direction. When one increases, the other decreases, or vice versa.
    - **+1:** Perfect positive correlation.
    - **-1:** Perfect negative correlation.
    - **0:** No correlation.
- **Comment:**
    - Correlation is the scaled version of covariance between -1 and +1.
    - **Positive Correlation:** Two variables increase or decrease together.
    - **Negative Correlation:** One variable increases while the other decreases.

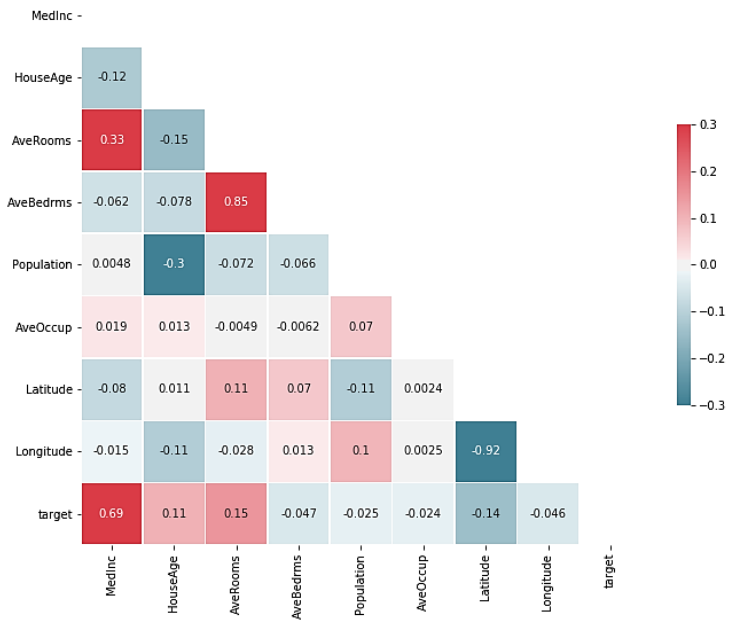
- **Important Note:** Correlation does not imply causation! Correlation between two variables does not mean that one causes the other.



Heatmap and Correlation Matrix

- **Definition:** Used to visualize the correlations between all variables in a dataset.
- **Area of Use:**
    - **Feature Selection:** We use it to determine which variables are related to the target variable.
    - **Multicollinearity:** To identify variables that have the same effect simultaneously.

- **Example Visual:**



Pearson Correlation Coefficient

- **Definition:** The Pearson Correlation Coefficient (usually denoted by r) is a measure that quantifies the strength and direction of a linear relationship between two continuous variables.

- **Value Range:** Between -1 and +1.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

  - xi and yi are the individual data points.
  - $\bar{x}$ and $\bar{y}$ are the means of the x and y variables.
- **Example:**
  - **r = 0.85:** Strong positive correlation.
  - **r = -0.70:** Moderate negative correlation.
  - **r = 0.10:** Weak correlation.
- **Pearson Correlation Coefficient (1.0):** This value indicates a perfect positive linear relationship between age and salary. That is, as age increases, salary increases at a constant rate.
- **P-value (0.0):** This value indicates that the correlation is statistically significant. That is, this relationship is not due to chance.

Correlation Calculation (R Calculation)

- **Steps:**

1. Calculate the covariance.
2. Calculate the standard deviations of both variables.
3. Divide the covariance by the product of the standard deviations.

- **Formula:**

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

- **Example Scenario:** Assume a company wants to examine the relationship between the ages and salaries of its employees. Let's say we have the following data:

Age (x): [25, 30, 35, 40, 45, 50]

Salary (y): [5000, 5500, 6000, 6500, 7000, 7500]

```
import numpy as np
from scipy.stats import pearsonr
age = [25, 30, 35, 40, 45, 50]
salary = [5000, 5500, 6000, 6500, 7000, 7500]
# Calculate Pearson correlation coefficient and p-value
correlation coefficient, p-value = pearsonr(age, salary)
```

Result: Pearson Correlation Coefficient: 1.0 P-value: 0.0

What types of variables are used for Pearson's correlation coefficient?

Answer: Variables (both the dependent and independent variables) used for Pearson's correlation coefficient must be quantitative. It will only test for the linear relationship between two variables.

Question 14: What is the difference between Covariance and Correlation?

Covariance:

- Signifies the direction of the linear relationship between two variables.
- In simple terms, It is a measure of variance between two variables.
- It can take any value from positive infinity to negative infinity.

Correlation:

- It measures the relationship between two variables, as well as the strength between these two variables.
- It can take any value from -1 to 1.

Multicollinearity

- **Definition:** The presence of multiple variables in a model that have the same effect.
- **Impact:** Reduces the efficiency of the model and can cause overfitting.
- **Solution:** Highly correlated variables are identified using a correlation matrix, and unnecessary ones are removed.

Feature Engineering

- **Definition:** Creating new features or improving existing ones by performing operations on the variables.
- **Role of Correlation:**
    - Features with high correlation to the target variable are selected.
    - Features with low correlation or unnecessary features are eliminated.

Summary Table

Topic	Short Description
Covariance	Shows the direction of the change between two variables.
Correlation	Shows the direction and strength of the relationship between two variables.
Heatmap	Visualizes the correlations between variables.
Pearson Coefficient	Measures the strength of the linear relationship between two variables.
Correlation Calculation	Calculated using covariance and standard deviations.
Multicollinearity	Variables with the same effect reduce model efficiency.
Feature Engineering	Improving model performance by performing operations on variables.

4.Section : Regression Analysis

What is Linear Regression?

- **Definition:** A statistical method used to model the linear relationship between two variables and make future predictions based on this relationship.
- **Purpose:** To understand the relationship between the independent variable (X) and the dependent variable (Y) and to predict Y using this relationship.

Basic Concepts

- **A. Independent and Dependent Variables**
  - **Independent Variable (X):** The variable considered as the cause or input.
    - Example: Advertising expenditure (X), Sales quantity (Y).
  - **Dependent Variable (Y):** The variable considered as the result or output.
- **B. Linear Relationship**
  - Formula:

$$y=ax+b$$

  - **a (Slope):** Indicates how much a 1-unit increase in X affects Y.
  - **b (Intercept):** The value Y takes when X=0.

Types of Linear Regression

- **A. Simple Linear Regression**
  - Definition: Modeling the relationship between a single independent variable (X) and a dependent variable (Y).
  - Example: The relationship between advertising expenditure (X) and sales quantity (Y).

- **B. Multiple Linear Regression**
  - **Definition:** Modeling the relationship between multiple independent variables (X1 ,X2,...) and a dependent variable (Y).
  - **Example:** The relationship between advertising expenditure (X1), product price (X2), and sales quantity (Y).

Steps of Linear Regression

1. **Data Collection:** Data for independent and dependent variables is collected.
2. **Model Building:** The line of best fit is drawn.
3. **Making Predictions:** Future predictions are made using the model.

Least Squares Method

- **Definition:** A method to find the line that minimizes the sum of the squares of the differences between the actual values and the predicted values.
- **Formula:**

$$\sum(yi - y\hat{}i)2$$

- yi: Actual values.
- y^i: Predicted values.

Linear Regression Example

- **Example Data**
  - X = [1, 2, 3, 4, 5]
  - Y = [2, 4, 6, 8, 10]
- **Model:**

$$y=2x+0$$

- **Slope (a):** 2
- **Intercept (b):** 0

Error Term

- **Definition:** The error term can be thought of as the deviation between the values predicted by the model and the actual values. Since every model tends to generalize, the values that fall outside the linear line symbolize errors.
- **Important Note:** Error terms should be as small as possible.

Overfitting

- **Definition:** Overfitting is when a model fits the training data too well. This results in the model performing very well on the training data but poorly on new, unseen data.
- **Solution:**
  - Reducing the complexity of the model.
  - Collecting more data.
  - Using cross-validation to better evaluate the model's performance.

Linear Regression and Python

• **Python Libraries:**

Spicy:

```
# Import required libraries

import numpy as np

from scipy import stats

# Define data

Screen_time = np.array([3, 5, 2, 0.5, 5, 3, 1, 4, 3, 4])  # Screen time (hours)

AGNO = np.array([2.7, 2.3, 3.3, 3.4, 2.3, 3.6, 2.4, 3.3, 3.3, 2.6])  #CGPA (Academic Grade Point Average)

# Apply linear regression

reg = stats.linregress(Screen_time, AGNO)

# Print slope and intercept values

print("Intersection (Intercept - b): ", reg.intercept)

print("Slope (Slope - a): ", reg.slope)

# Print the linear regression equation
```

Code Output

```
Intercept - b: 3.293103448275862

Slope - a: -0.1413793103448276

Linear Regression Equation: Y = 3.29 + (-0.141)X
```

Explanations

1. Intercept (b):

b=3.29b=3.29: This is the value YY would take when X=0X=0. In other words, the expected value of GPA when screen time is zero is **3.29**.

2. Slope (a):

a=−0.141: This shows that a 1-hour increase in screen time affects GPA by **-0.141** points. In other words, as screen time increases, GPA tends to decrease.

3. Linear Regression Equation:

Y=3.29+(−0.141)X: This equation expresses the relationship between screen time (X) and GPA (Y). For example, when the screen time is 2 hours, the GPA is estimated as:

Y=3.29+(−0.141)×2

Y=3.29−0.282

Y=3.008

- **Scikit-learn**:

```
from sklearn.linear_model import LinearRegression

model = LinearRegression()

model.fit(X, y)
```

- **Statsmodels**:

```
statsmodels.api as sm

model = sm.OLS(y, X).fit()
```

- If we use the first-order values (not squared or cubed) of the independent variables (features) in our model, these are called Linear Regression models.
-
- If a model has only one feature (independent variable), it is called a SIMPLE LINEAR REGRESSION MODEL.
    - $Y=aX+b$
    - $Y=mX+n$
    - $Y=\beta_0+\beta_1X_1$
    - If a model has more than one feature (independent variable), it is called a MULTIPLE LINEAR REGRESSION MODEL.

- $Y=aX+bZ+c$
- $Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta$<0xE2><0x82><0x96>X<0xE2><0x82><0x96>

These equations show different representations of simple and multiple linear regression models. In summary:

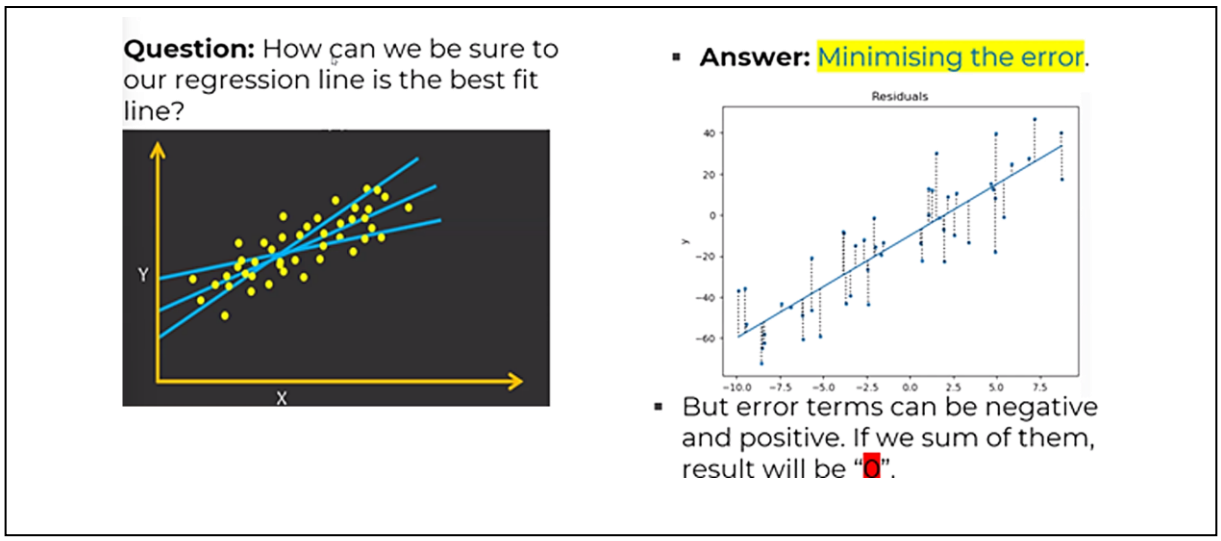
- A Simple Linear Regression model predicts the dependent variable Y using a single independent variable X.
-
- A Multiple Linear Regression model predicts the dependent variable Y using multiple independent variables X1,X2,...,Xk.

Alternative programs for regression calculation:

Regardless of which program is used, the important thing to do for calculations is to interpret the equation.

Summary Table

Topic	Short Description
Linear Regression	Modeling the linear relationship between two variables and making predictions.
Independent Variable (X)	The variable considered as the cause or input.
Dependent Variable (Y)	The variable considered as the result or output.
Least Squares	Method for finding the line that minimizes the sum of squared errors.
Overfitting	The model fitting the training data too well.
Python Implementation	Building models with Scikit-learn or Statsmodels libraries.



Residual Term (Error)

Definition:

The residual term (error term) is a concept that expresses the difference between the values predicted by a regression model and the actual observations.

Purpose:

- To evaluate how accurate the model is and how well it fits the observations.

Example Equation:

- In a regression model, the error term is shown as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where:

- **Yi:** Dependent variable (actual observation).
- **β0:** Y-intercept.

- **β1:** Slope coefficient.
- **Xi:** Independent variable.
- **ϵi:** Error term (residual term).

Pearson's R Calculation (Pearson Correlation Coefficient Calculation)

- **Definition:** It is a statistical method that measures the direction and strength of the linear relationship between two variables.
- **Formula:**

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Where:

- **Cov(X,Y):** Covariance between X and Y.
- ***σX:*** Standard deviation of X.
- ***σY:*** Standard deviation of Y.
- **Value Range: Between -1 and +1.**
  - **+1:** Perfect positive relationship.
  - **-1:** Perfect negative relationship.
  - **0:** No relationship.
- **Comment:**
  - **Positive Correlation: Two variables increase or decrease together.**
  - **Negative Correlation: One variable increases while the other decreases.** [1]

- **Example:**
  - **X = [1, 2, 3, 4, 5]**
  - **Y = [2, 4, 6, 8, 10]**
  - **r = 1 (Perfect positive relationship).**

Residual Term (Error)

- **Definition:** The difference between the actual values and the values predicted by the model.
- **Formula:**

$$(\mathrm{Residual}) = y_i - \hat{y}_i$$

- **yi:** Actual value.

- **y^i:** Predicted value.
- **Comment:**
    - Small residual terms indicate that the model makes predictions close to the actual values.
    - Large residual terms indicate that the model makes inaccurate predictions.

Coefficient of Determination (R^2)-IMPORTANT!

- Definition: It is a measure that shows how much of the total variance in the dependent variable (Y) is explained by the independent variables (X).
- **Formula:**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

- SSRES is the residual sum of squares.
- SSTOT is the total sum of squares.
- yi represents the observed values.
- y^i represents the predicted values.
- $\bar{y}$ is the mean of the observed values.
- The summation sign ∑i indicates summation over all data points.
- **Value Range: Between 0 and 1.**
    - **R2=1:** The model explains all the variance in the dependent variable (perfect fit).
    - **R2=0:** The model explains none of the variance in the dependent variable.
    - **0<R2<1:** The model explains a portion of the variance in the dependent variable.
- **NOTE:** In ML (Machine Learning), when training a model, we use the concept of how well we have fitted (modeled) our model. R2 determines how well we have fitted the model.

Interpretation of R^2

- **Model Fit Quality:**
    - The higher the R2 value, the better the model is said to fit the data.
    - For example, if R2=0.85, it means that the model explains 85% of the variance in the dependent variable.
- **Effect of Independent Variables:**
    - R2 measures the effect of the independent variables (X) on the dependent variable (Y).
    - A high R2 indicates that the independent variables explain Y well.

- **Example**

  Data:

  - X = [1, 2, 3, 4, 5]
  - Y = [2, 4, 6, 8, 10]

- **Model:**

  Y=2X+0

- **R^2 Calculation:**

  - Predictions: Y=[2,4,6,8,10]

$$SS_{\text{res}} = \sum (y_i - \hat{y}_i)^2 = 0$$
$$SS_{\text{tot}} = \sum (y_i - \bar{y})^2 = 40$$
$$R^2 = 1 - \frac{0}{40} = 1$$

Concept	Explanation
Pearson's R	Indicates the direction and strength of the linear relationship between two variables.
Residual Term	The difference between the actual value and the predicted value.
R^2 (Coefficient of Determination)	Indicates how much of the variance in the dependent variable is explained by the independent variables.
$R^2 = 1$	Perfect fit (all variance is explained).
$R^2 = 0$	No variance is explained.
$0 < R^2 < 1$	A portion of the variance is explained.

What is R-squared (R2)?
Answer: R-squared, denoted as R2, is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variable(s) in a regression model.

What does an R-squared value of 0.75 mean?
Answer: An R-squared value of 0.75 means that **75%** of the variance in the dependent variable can be explained by the independent variable(s) in the model, and the remaining 25% is unexplained.

Can R-squared be negative?
Answer: No, R-squared cannot be negative. It will always fall within the range of **0 to 1**. A negative value would not make sense in the context of explaining variance.

What is the range of possible values for R-squared?
Answer: R-squared values range from **0 to 1**. An R2 of 0 indicates that the model does not explain any of the variance in the dependent variable, while an [1] R2 of 1 means that the model explains all of the variance.

What are the limitations of R-squared?
Answer: R-squared has some limitations. It cannot determine causation; it doesn't reveal the significance of individual predictors [1] (features), and a high R-squared does not guarantee a good model fit if the model is overfitted. It's important to consider these limitations when using R-squared in analysis.

When should you use R-squared as an evaluation metric?
Answer: R-squared is commonly used in regression analysis to assess the [1] model's fit. It is useful when you want to understand how well your independent variables explain the variation in the dependent variable.

5.Section : Probability

Probability – The Principle of Likelihood

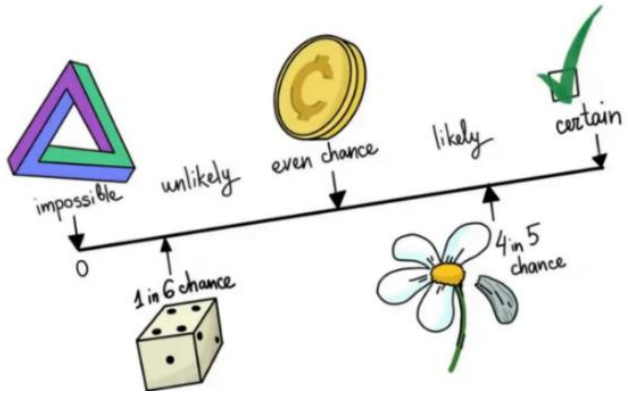
- **Definition**: A numerical value between 0 and 1 that expresses the chance of an event occurring.
    - **0**: The event will definitely not happen.
    - **1**: The event will definitely happen.
- **Example**: The probability of getting heads in a coin toss is 0.5.

Why a Probabilistic Approach?

- **Uncertainty**: Life is full of uncertainties. Probability provides a method for measuring and understanding these uncertainties.
- **Making Predictions**: Probability is used to predict future events. It helps us gain insight into uncertain situations.
- **Importance in Data Science**: Probability is a fundamental tool for making reliable predictions in data analysis and machine learning.

Examples:

- **Weather Forecasting**
  Meteorologists use probabilistic models to forecast the weather. For example, a 70% chance of rain is calculated based on historical data.
- **Health Risk Assessment**
  Doctors assess a patient's risk of developing a particular disease using probabilistic models. For example, calculating the risk of heart disease in a smoker.
- **Finance and Investment**
  Financial analysts use probabilistic models to support investment decisions. For example, predicting the future price of a stock.
- **Machine Learning**
  Machine learning models use probabilistic approaches for classification and prediction tasks. For example, determining whether an email is spam.
- **Optimization and Logistics**
  Companies use probabilistic models for supply chain and logistics planning. For example, estimating the delivery time of products.



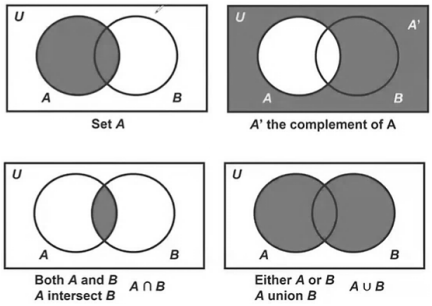
Law of Large Numbers

- **Definition**: The more an experiment is repeated, the closer the results will approach the expected probability.
  - **Example**: The more you flip a coin, the closer the proportion of heads will get to 50%.
  - **Application Areas:** Casinos, stock market predictions, statistical analyses.

What is the Law of Large Numbers?
Answer: It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate.

Sets and Probability (Union, Intersection, Complements)

- **Union:** The probability that at least one of two events occurs.
    - **Example:** P(A∪B)
- **Intersection:** The probability that two events occur at the same time.
    - **Example:** P(A∩B)
- **Complement:** The probability that an event does not occur.
    - **Example:** P(A')=1−P(A)



Permutation

- **Definition:** An arrangement of items in a specific order. Order matters.
- **Formula:**

$$P(n,k) = \frac{n!}{(n-k)!}$$

    - n: Total number of elements
    - k: Number of elements selected
- **Example:** In how many different ways can 3 different books be arranged on a shelf?

$$P(3,3) = \frac{3!}{(3-3)!}$$

Combination

- **Definition:** A selection of items from a group where order does not matter.
- **Formula:**

$$C(n,k) = \frac{n!}{k!(n-k)!}$$

    - n: Total number of elements
    - k: Number of elements selected
- **Example:** In how many different ways can 2 people be selected from a group of 5?

$$C(5,2) = \frac{5!}{2!(5-2)!}$$

Law of Large Numbers

Definition:
The Law of Large Numbers is a statistical principle stating that as an experiment or observation is repeated many times, the observed outcomes converge to the theoretical probability. In other words, the frequency of an event approaches its expected probability as the number of trials increases.

Key Concepts:

1. **Theoretical Probability:** The ideal probability of an event occurring under perfect conditions.
   Example: The theoretical probability of getting heads in a coin toss is 50%.
2. **Empirical (Experimental) Probability:** The probability calculated based on observed outcomes.
   Example: If a coin is tossed 100 times and heads appears 45 times, the empirical probability of heads is 45%.
3. **Relative Frequency:** A measure of how often an event occurs out of a certain number of trials or observations. It is simply the ratio of the number of times the event occurred to the total number of trials.

$$\text{Relative Frequency (RF)} = \frac{\text{Number of Times Event Occurred}}{\text{Total Number of Trials}}$$

Examples:

- **Coin Toss Example:**
  The theoretical probability of getting heads in a coin toss is 50%. In a small number of tosses (e.g., 10), the proportion of heads may vary between 40% and 60%. However, as the number of tosses increases (e.g., 1000), the proportion of heads will approach 50%.
- **Casino Example:**
  In games like roulette, the probability of a specific outcome is fixed. When many games are played, the results converge to these probabilities, allowing the casino to profit in the long run.

Table Example:

Number of Trials	Relative Frequency Range	Percentage (%)
10	0.4 – 0.6	66
100	0.49 – 0.51	92
1,000	0.499 – 0.501	97
10,000	0.4999 – 0.5001	99

Application Areas:

- **Finance:** Used in investment strategies and risk management.
- **Insurance:** Insurance companies use the law of large numbers to calculate risks and determine premiums.
- **Quality Control:** Used to estimate error rates in production processes.

What is the Law of Large Numbers?
Answer: It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate.

Conditional Probability

- **Definition:** The probability of an event occurring **given that** another event has already occurred.
- **Formula:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- P(A|B): The probability of event A occurring given that event B has occurred.
- P(A∩B): The probability that both events A and B occur.
- P(B): The probability of event B occurring.

Question:

At a school, 45% of students fail physics, 35% fail mathematics, and 25% fail both physics and mathematics. For a randomly selected student:

a) What is the probability that the student also failed mathematics, given that they failed physics?
b) What is the probability that the student also failed physics, given that they failed mathematics?

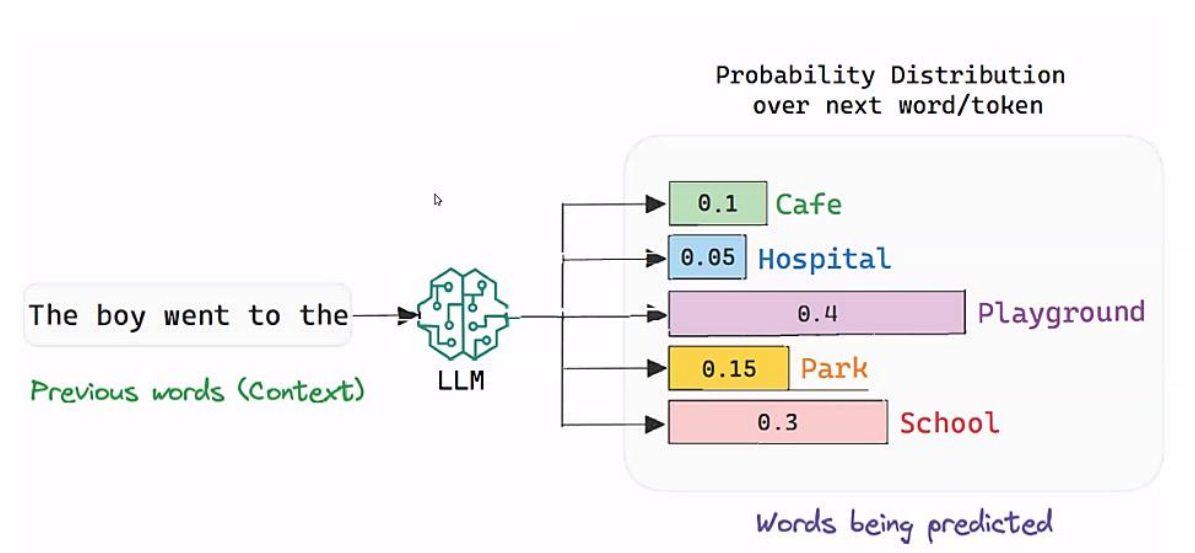
Answer
a) P(M|F)=0.55
The probability that a student failed mathematics **given that** they failed physics is **55%**.

b) P(F|M)=0.71
The probability that a student failed physics **given that** they failed mathematics is **71%**.

- **Example:** The probability of flight cancellations when a storm occurs.

Conditional Probability in LLM Models

- **Definition:** Large Language Models (LLMs) use the probability of word sequences to generate text based on how likely words are to follow each other.
    - **Example:** The probability that the word *"sunny"* follows the phrase *"The weather today is..."*



Bayes' Theorem

- **Definition:** It is used to update prior probabilities when new information becomes available.
- **Formula:**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

    - **P(A|B):** The probability of event A occurring given that event B has occurred.
    - **P(B|A):** The probability of event B occurring given that event A has occurred.
    - **P(A):** The prior probability of A.
    - **P(B):** The prior probability of B.

Example: Medical Test

- **Problem:** A medical test has an accuracy of 99%. The disease occurs in 1% of the population. What is the probability that a person who tested positive actually has the disease?
- **Given:**
    - **P(Disease) = 0.01**
    - **P(Positive | Disease) = 0.99**
    - **P(Positive | No Disease) = 0.01**
- **Bayes' Theorem Calculation:**

$$P(Disease|Positive) = \frac{0.99 \times 0.01}{(0.99 \times 0.01) + (0.01 \times 0.99)} = 0.5$$

- **Result:** There is a 50% chance the person actually has the disease if they tested positive.

Example: Fire Detection with Bayes' Theorem

Given:

1. Probability of a fire (P(Fire)) = 0.01
2. Smoke is commonly seen during picnics (P(Smoke)) = 0.10
3. In 90% of fires, smoke is present (P(Smoke | Fire)) = 0.90

Bayes' Theorem Application:

P(Fire | Smoke) = The probability of a fire given that smoke is observed.

$$P(Fire|Smoke) = \frac{P(Smoke|Fire) \cdot P(Fire)}{P(Smoke)} = \frac{0.90 \times 0.01}{0.10} = 0.09$$

Result: The probability of a fire when smoke is observed is **9%**.

Summary:

- Only 1% of events are fires.
- The probability of observing smoke is 10%.
- In the case of fire, smoke is observed 90% of the time.
- Using Bayes' Theorem, the probability of fire given smoke is 9%.

Applications of Bayes' Theorem

Machine Learning:

- **Classification:**
  - Naive Bayes classifiers are based on Bayes' Theorem and are used in tasks like text classification and spam filtering.
  - In probabilistic models, Bayes' Theorem is used to predict which class a given data point belongs to.
- **Decision-Making Under Uncertainty:**
  - It is used to measure uncertainty in model predictions and to incorporate this uncertainty into decision-making processes.

Statistical Inference:

- **Bayesian Statistics:**
  - Used to update probability distributions of parameters and to test hypotheses.
- **Anomaly Detection:**
  - Helps in building probabilistic models to detect unexpected or abnormal events.
- **Parameter Estimation:**
  - Used to estimate model parameters from data.
- **Time Series Analysis:**
  - Helps predict future values and analyze changes over time.

Other Application Areas:

- **Artificial Intelligence (AI):**
  - Used in probabilistic reasoning and decision-making systems.
- **Natural Language Processing (NLP):**
  - Applied in text classification, language modeling, and information extraction.
- **Medicine:**
  - Used in disease diagnosis, risk assessment, and genetic analysis.
- **Finance:**
  - Utilized in risk management, portfolio optimization, and fraud detection.

Summary:

Bayes' Theorem can be applied to any problem involving probabilistic reasoning. It is especially effective in situations requiring decision-making under uncertainty, probability estimation, and extracting insights from data.

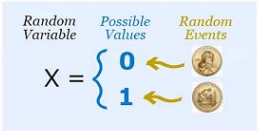
Summary Table

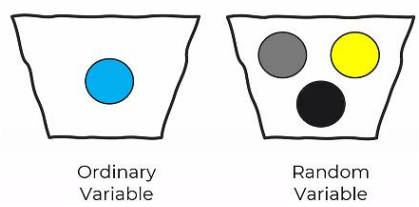
Concept	Description
Probability	Represents the chance of an event occurring.
Random Variable	Values that vary depending on the outcome of an experiment.
Law of Large Numbers	As an event is repeated, the results converge toward the expected probability.
Permutation	Arrangements where order matters.
Combination	Selections where order does not matter.
Conditional Probability	The probability of one event given that another event has occurred.
Bayes' Theorem	Used to update probabilities based on new information.
LLMs and Probability	Language models use the probability of word sequences to generate text.

6.Section : Random Variables and Distributions

Random Variables

- **Definition:** Variables whose values depend on the outcome of a statistical experiment. Random variables are based on probability. They are governed by chance, cannot be known with certainty, but can be predicted.
  Example: The number that appears when a die is rolled or the number of cars sold in a month.
- **Example:** The number that appears when a die is rolled (1, 2, 3, 4, 5, 6) is a random variable.





Random variables are divided into two types: **Discrete** and **Continuous**.

Discrete Random Variables

- **Definition:** Discrete random variables are those that can take on a finite or countable number of values. These variables take values only at specific points within a range.
- **Characteristics:**
  - There are gaps between possible values (e.g., 1, 2, 3).
  - Probability calculations use the **Probability Mass Function (PMF).**
- **Examples:**
  - The number that appears when rolling a die (1 to 6).
  - The number of students in a classroom.
  - The number of customers visiting a store in one day.

Discrete Probability Distributions

Binomial Distribution

- **Definition:** Used for experiments with two possible outcomes (success/failure). It applies to repeated trials and calculates the probability of obtaining a specific number of successes in multiple trials.
- **Formula:**

$$P(X = k) = C(n, k) \cdot p^k \cdot (1 - p)^{n-k}$$

  - P(X=k)Probability of getting exactly k successes in n trials
  - C(n,k): Combination of n items taken k at a time
  - p: Probability of success
  - n: Number of trials
  - k: Number of successes
- **Example:** The probability of getting heads exactly 3 times in 10-coin tosses.

Bernoulli Distribution

- **Definition:** A special case of the binomial distribution, used for a single trial with two possible outcomes.
- **Examples:**
  - Tossing a coin once: probability p for heads, 1−p for tails.
  - A political candidate winning or losing an election (only one election is held, and there are two possible outcomes).

Poisson Distribution

- **Definition:** Used for modeling rare events that occur in a fixed interval of time or space.
- **Formula:**

$$P(X) = \frac{\lambda^X . e^{-\lambda}}{X!}$$

- **Example Scenarios:**
  - If a city experiences an average of 3 major floods per year, what is the probability that there will be 4 floods next year?

    Using the **Poisson formula**:

$$P(X = 4) = \frac{e^{-3} \cdot 3^4}{4!} = \frac{0.0498 \cdot 81}{24} \approx 0.1681$$

    Result: There is a **16.8% chance** that 4 major floods will occur next year.

  - If on average 5 patients visit a hospital per day, what is the probability that **8 patients** arrive on a specific day?
  - λ=5
  - k=8k

$$P(X = 8) = \frac{e^{-5} \cdot 5^8}{8!} \approx 0.065$$

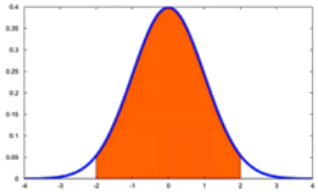
    So, there's about a 6.5% chance of 8 patients arriving in a day.

Continuous Random Variables

• **Definition:**
Continuous random variables are variables that can take **an infinite number of values** within a given range. These variables take values on a **continuous scale**. In these distributions, the probability of taking an exact value is **zero**; instead, the **probability is calculated over an interval**.

• **Properties:**

- There are **no gaps between the values** (e.g., 1.5, 2.3, 3.7, etc.).
- Probability calculations use the **Probability Density Function (PDF)**.
  - Probability Density Function (PDF)
    - Y; X random değişkenin bir fonksiyonudur
    - Y; tüm X değerleri için 0'a eşit veya büyüktür
    - Eğri altındaki kalan alan 1 e eşittir.



Alan = genişlik*Uzunluk = 1 * 1 = 1

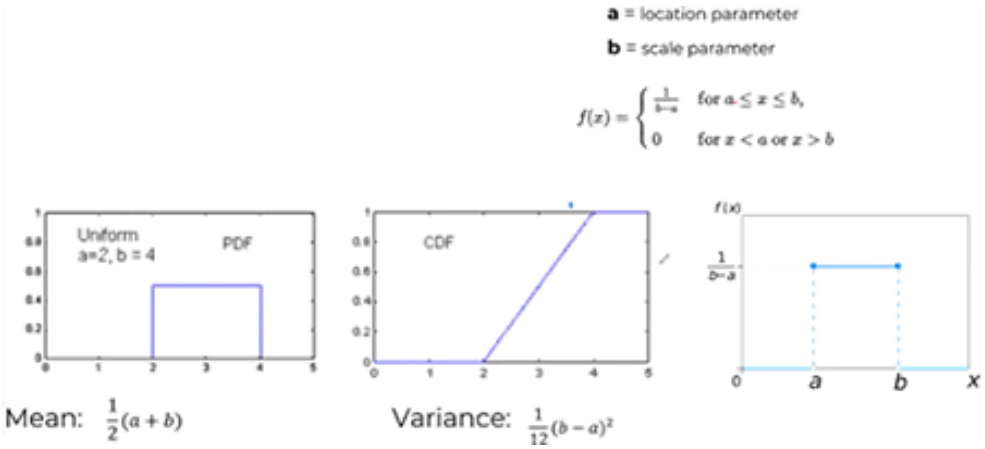
• **Examples:**

- A student's height (e.g., 1.70 m, 1.75 m, 1.80 m).
- A car's speed (e.g., 50 km/h, 60.5 km/h).
- The weight of a product (e.g., 1.2 kg, 1.25 kg)

• **Uniform Distribution**

Definition:
In a uniform distribution, all values are equally likely to occur.

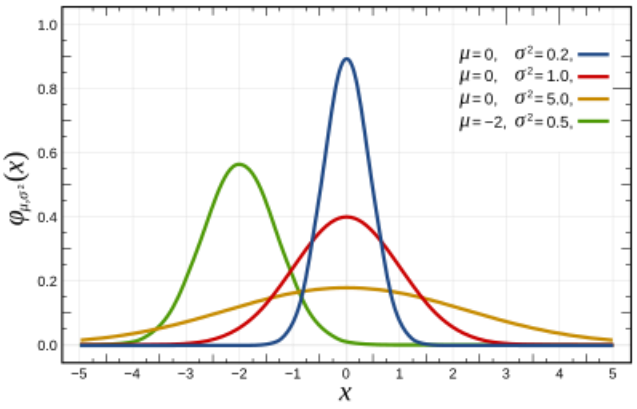
Example:
Suppose a bus arrives at a bus stop every 10 minutes, and the arrival times are completely random. Then, the waiting time for the bus has an equal chance of being any value between 0 and 10 minutes. This is an example of a **continuous uniform distribution**.



• **Normal Distribution (Gaussian Distribution)**

Definition:
Most natural phenomena follow this distribution. It is **symmetric** and shaped like a **bell curve**. The **mean**, **median**, and **mode** are all equal.

Example:
The distribution of people's height.



The **area under the curve equals 1**, representing total probability.

What is the 3-Sigma Rule? (Also known as the 68-95-99.7 Rule)

The **3-Sigma Rule** is a statistical rule that shows how much of the data lies within certain standard deviations from the mean in a normally distributed dataset.

Normal Distribution and Sigma (σ):

- **Normal Distribution**: A symmetrical bell-shaped curve. Many natural events and datasets follow this distribution.
- **Sigma (σ)**: Also called standard deviation, it measures how spread out the data is from the mean.

Meaning of the 3-Sigma Rule:

- **μ ± 1σ**: Approximately **68%** of the data lies within **1 standard deviation** of the mean.
- **μ ± 2σ**: Approximately **95%** of the data lies within **2 standard deviations** of the mean.
- **μ ± 3σ**: Approximately **99.7%** of the data lies within **3 standard deviations** of the mean.

Applications of the 3-Sigma Rule:

• **Quality Control**: Used in manufacturing to check whether products stay within tolerance limits.
• **Finance**: Applied in risk management and portfolio analysis.
• **Engineering**: Used in design and analysis processes.
• **Data Analysis**: Helps in identifying **outliers** and understanding the spread of data.

Example:

Assume the length of screws produced in a factory follows a normal distribution.

- The **mean** length is **10 cm**, and
- The **standard deviation (σ)** is **0.1 cm**.

Then:

- Approximately **68%** of screws will be between **9.9 cm and 10.1 cm**.
- Approximately **95%** will be between **9.8 cm and 10.2 cm**.
- Approximately **99.7%** will be between **9.7 cm and 10.3 cm**.

Summary: The **3-Sigma Rule** helps us understand how data is distributed around the mean in a normal distribution. It is a powerful tool for **decision-making and analysis** in various fields.

What is Normal Distribution?
Normal Distribution is a probability distribution that is symmetric about the mean. It is also known as Gaussian Distribution. The distribution appears as a Bell-shaped curve which means the mean is the most frequent data in the given data set.
In Normal Distribution:

- Mean = Median = Mode
- Total area under the curve is 1.

What is the empirical rule?

The Empirical Rule is often called the 68-95-99.7 rule or Three Sigma Rule. It states that on a Normal Distribution:

- 68% of the data will be within one Standard Deviation of the Mean.
- 95% of the data will be within two Standard Deviations of the Mean.
- 99.7% of the data will be within three Standard Deviations of the Mean.

What is a bell-curve distribution?

A bell-curve distribution is represented by the shape of a bell and indicates normal distribution. It occurs naturally in many situations especially while analyzing financial data. The top of the curve shows the mode, mean and median of the data and is perfectly symmetrical. The key characteristics of a bell-shaped curve are -

- The empirical rule says that approximately 68% of data lies within one standard deviation of the mean in either of the directions.
- Around 95% of data falls within two standard deviations and
- Around 99.7% of data fall within three standard deviations in either direction.

What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

Additional Information:

• Characteristics of the Bell Curve (Normal Distribution):

- **Symmetrical**: It is symmetrical around the mean.
- **Mean, Median, and Mode are Equal**: The center and the most frequent value of the distribution are the same.
- **68-95-99.7 Rule**: Shows how much of the data falls within specific standard deviation ranges.
- **Common in Financial Data**: Frequently observed in financial datasets such as asset returns and risk measurements.

• Why is it Important?

- It helps to **understand the distribution** of data and make **predictions**.
- Widely used in **statistical analysis** and **machine learning**.
- Useful in **risk management** and **decision-making processes**.

What do you understand by the term Normal Distribution?

Normal distribution, also known as the Gaussian distribution, is a bell-shaped frequency distribution curve. Most of the data values in a normal distribution tend to cluster around the mean.

Z Distribution (Standard Normal Distribution)

Definition:

A normal distribution with a **mean of 0** and a **standard deviation of 1**.

Use Cases:

- When the **population standard deviation (σ)** is known.
- When the **sample size (n)** is large (typically n > 30).

Properties:

- Symmetrical and bell-shaped.
- About **68%** of the data lies within **±1** standard deviation from the mean,
  95% within **±2σ**, and
  99.7% within **±3σ**.

Formula:

$$z = \frac{x - \mu}{\sigma}$$

- **x**: Data point
- **μ**: Mean
- **σ**: Standard deviation

Example 1: Z Distribution

Data: The average height of students in a class is 170 cm, with a standard deviation of 10 cm.
Question: What proportion of students are taller than 180 cm?

Solution:

$$z = \frac{180 - 170}{10} = 1$$

From the Z-table, the probability corresponding to **z = 1** is **0.8413**.
So, the proportion taller than 180 cm is: 1−0.8413=0.1587 or 15.87%

T Distribution (Student's t-distribution)

Definition:

A probability distribution used when the **sample size is small** and the **population standard deviation is unknown**.

Use Cases:

- When **σ (population standard deviation)** is unknown.
- When **n < 30** (small sample size).

Properties:

- Similar to the normal distribution but has **heavier tails**.
- As the sample size increases, it approaches the normal distribution.
- Defined by **degrees of freedom (df)**.

Formula:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- **x¯\bar{x}x¯**: Sample mean
- **μ**: Population mean
- **s**: Sample standard deviation
- **n**: Sample size

Example 2: T Distribution

Data: The average height of 10 students is 172 cm, with a sample standard deviation of 8 cm.
Question: Assuming the population mean is 170 cm, what is the t-value?

Solution:

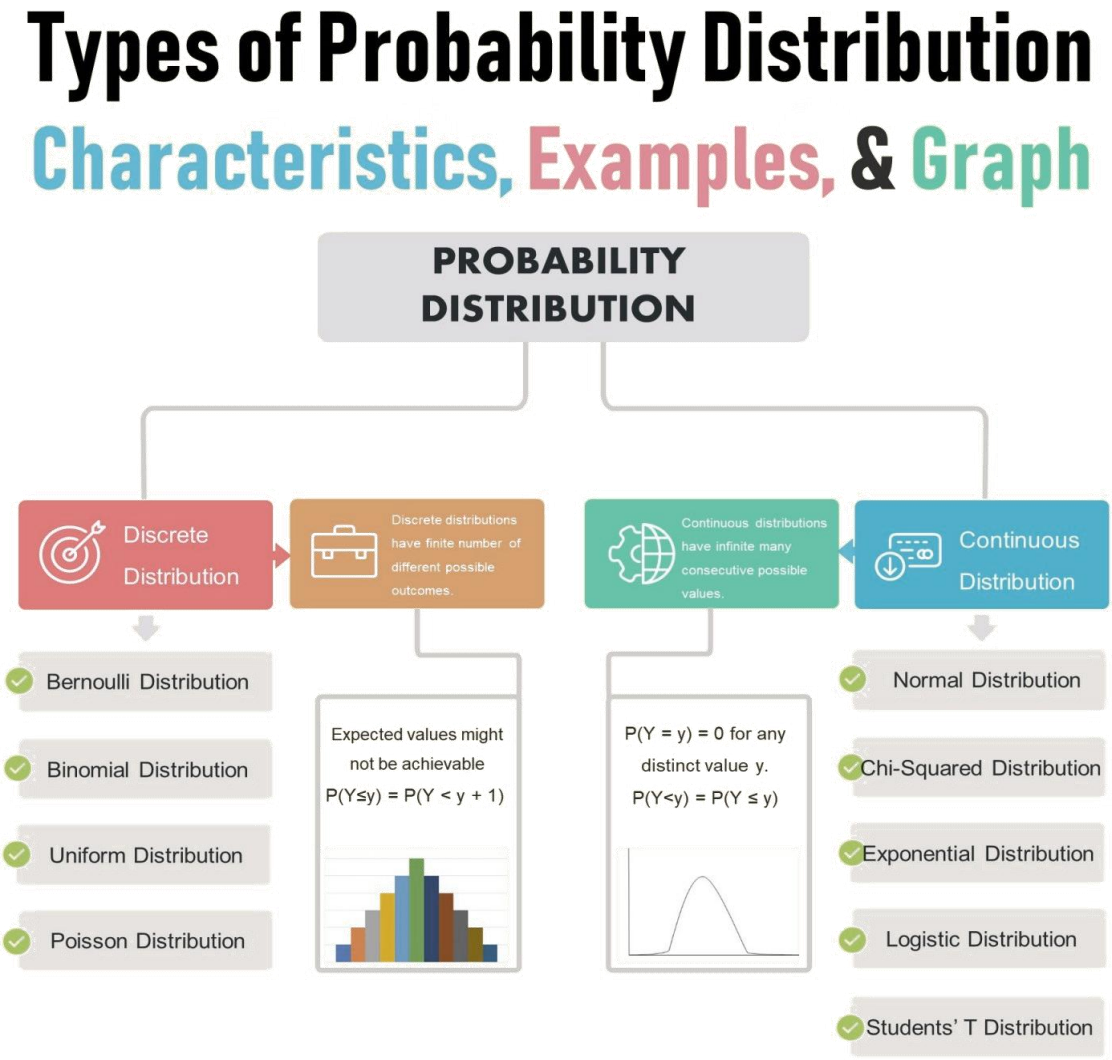
$$t = \frac{172 - 170}{8/\sqrt{10}} = \frac{2}{2.53} \approx 0.79$$

Degrees of freedom (df) = 10 − 1 = **9**
Find the corresponding probability from the **t-table** for **t = 0.79** and **df = 9**.

Why Are Probability Distributions Important?

1. **Probability Predictions**: Help estimate the likelihood of events.
2. **Statistical Analysis**: Crucial for understanding and analyzing data.

3. **Machine Learning**: Many machine learning models are based on assumptions involving probability distributions.



Common Concepts Used as Functions in Python Statistical Libraries

1. PMF (Probability Mass Function):

- **What Is It For?**
  - PMF is only concerned with **discrete random variables**, meaning variables that can take on specific, separate values (e.g., the outcome of a dice roll, or a coin flip with heads or tails).
  - PMF calculates the **probability of a random variable being exactly equal to a specific value**.

- **Example:**
  - What is the probability of rolling a 3 on a die?
    PMF answers this question.

$$PMF(X = 3) = \frac{1}{6}$$ (if the die is fair)

- **Summary:**
  - PMF answers the question: "What is the probability of **exactly this value**?"

2. CDF (Cumulative Distribution Function):

- **What Is It For?**
  - CDF works with **both discrete and continuous random variables**.
  - It calculates the **probability that a random variable takes on a value less than or equal to a specific value**.
- **Example:**
  - What is the probability of rolling a number **less than or equal to 3** on a die? CDF answers this question.

$$CDF(X \leq 3) = PMF(1) + PMF(2) + PMF(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

- **Summary:**
  - CDF answers the question: "What is the probability of **this value or less**?"

Key Differences:

Feature	PMF	CDF
What it measures	Probability of one specific value	Probability of a value or anything less
Applies to	Only discrete variables	Both discrete and continuous variables

Sample solution in Python:

Dice Roll

```
from scipy.stats import binom

# Dice roll example (6 sides, each face has a probability of 1/6)

n = 6  # Number of trials (number of die faces)

p = 1/6  # Probability of success (probability of any one face showing up)

# PMF: Probability of getting exactly 3 successes

pmf_3 = binom.pmf(k=3, n=n, p=p)

print(f"PMF (X = 3): {pmf_3}") # Çıktı: 0.0536

# CDF: 3 veya daha küçük bir değer gelme olasılığı

cdf_3 = binom.cdf(k=3, n=n, p=p)  print(f"CDF (X ≤ 3): {cdf_3}")
```

Summary Table

Concept	Description
Random Variables	Values that change depending on the outcome of an experiment. Example: Result of a dice roll.
Discrete Probability Distributions	Probability distributions of discrete random variables. Example: Rolling a die.
Binomial Distribution	Used for experiments with two outcomes (success/failure). Example: Coin toss.
Bernoulli Distribution	A special case of the binomial distribution, used for a single trial. Example: A single coin toss.
Poisson Distribution	Used for rare events within a fixed time interval. Example: Number of customers arriving in a day.
Continuous Probability Distributions	Probability distributions of continuous random variables. Example: Human height.
Uniform Distribution	All values occur with equal probability. Example: Probability of any face of a fair die.
Normal Distribution	Most natural phenomena follow this distribution. Symmetric and bell-shaped. Example: Human height.
Standard Distribution	A normal distribution with mean 0 and standard deviation 1. Example: Standard normal distribution table.
T Distribution	Used with small sample sizes; similar to normal distribution but with heavier tails. Example: Mean estimates in small samples.
Z Distribution	Used for large samples when population standard deviation is known. Example: Hypothesis testing with large samples.

7. Section : Sampling Distributions and Confidence Intervals

Sampling Distributions

- **Definition**: A sampling distribution refers to the distribution of a statistic (e.g., mean, standard deviation) calculated from multiple samples drawn from the same population.
- **Features**:
    - The sampling distribution depends on both the sample size and the population distribution.
    - As the sample size increases, the sampling distribution becomes more similar to the population distribution.
- **Example**: If 100 different samples are drawn from a population, the mean of each sample will differ. The distribution of these means forms the sampling distribution.

Weibull Distribution and Samples

- **Weibull Distribution**: A continuous probability distribution commonly used in reliability analysis and life data modeling. It is defined by the shape parameter (k) and the scale parameter (λ).
- **Samples**: Different samples drawn from a population may have different means and standard deviations. This shows how well the samples represent the population.

Sample Error

- **Definition**: Sample error measures how consistent the samples are with each other and how well they estimate population parameters.
- **Features**:
    - If sample error is small, the samples are close to each other and estimates are more accurate.
    - If sample error is large, the samples differ significantly, making the estimates less reliable.
- **Calculation**: Sample error depends on the standard deviations of the samples and their sizes.

Sampling Distribution and Weibull Distribution

- **Weibull Distribution Chart**: A Weibull distribution chart may include 6 different tabs created using different parameters (e.g., different k and λ values). Each tab represents a different sample distribution.
- **Difference in Means**: The mean values of the samples in each tab may vary. This shows how well each sample represents the population.

Sample Error and Standard Error

- **Standard Error**: Indicates how close the sample mean is to the population mean. It is a measure of sample error.
- **Formula**:

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}}$$

    - σ: Population standard deviation
    - n: Sample size
- **Sample Error**: If the standard deviations among samples vary greatly, the sample error is high, indicating less reliable estimates.

Example: Weibull Distribution and Sample Error

- **Population**: The lifespan of a product is distributed according to a Weibull distribution.
- **Samples**: 6 different samples are drawn from the population. Each has a different mean and standard deviation.
- **Weibull Distribution Chart**: Contains 6 tabs, each representing the distribution of a different sample.
- **Difference in Means**: The sample means differ across tabs, showing how well they represent the population.
- **Standard Error Calculation**:
    - **Sample 1**: Mean = 100, Standard Deviation = 10, Sample Size = 30

$$SE_1 = \frac{10}{\sqrt{30}} \approx 1.83$$

    - **Sample 2**: Mean = 105, Standard Deviation = 15, Sample Size = 30

$$SE_2 = \frac{15}{\sqrt{30}} \approx 2.74$$

- **Sample Error**: Since Sample 1 has a smaller standard error, its estimates are more reliable. Sample 2 has a larger standard error, making its estimates less reliable.

Conclusion

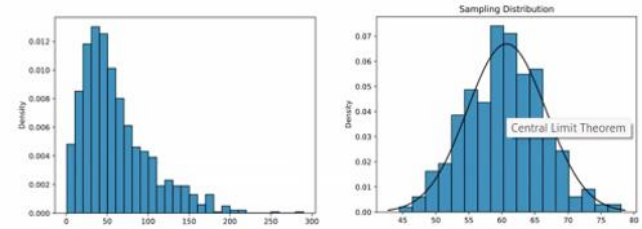
- **Weibull Distribution**: Demonstrates how samples drawn with different parameters can form different distributions.
- **Sample Error**: Measures how consistent the samples are and how accurate the estimates are.
- **Standard Error**: Shows how close a sample mean is to the population mean. A smaller standard error indicates more accurate estimates.

2. Central Limit Theorem (CLT)

- **Definition**: For large samples, the distribution of sample means approaches a normal distribution.
    - **Features**: As sample size (n) increases, the distribution of the sample means becomes approximately normal.
    - Even if the population distribution is not normal, the CLT holds when the sample size is large enough (typically n > 30).
- **Example**: If 50 samples are drawn from a population, the distribution of their means will approximate a normal distribution.

What is the Central Limit Theorem?

The Central Limit Theorem (CLT) states that, given a sufficiently large sample size from a population with a finite level of variance, the sampling distribution of the mean will be normally distributed regardless of if the population is normally distributed.



Sampling Distribution: (Graph description)

• In the image, the original population distribution (Salmon Weight) is shown on the left. This distribution does not resemble a normal distribution.

• On the right, the distribution of sample means (Sampling Distribution) is shown. As can be seen, the distribution of sample means converges to a normal distribution, regardless of the original population distribution.

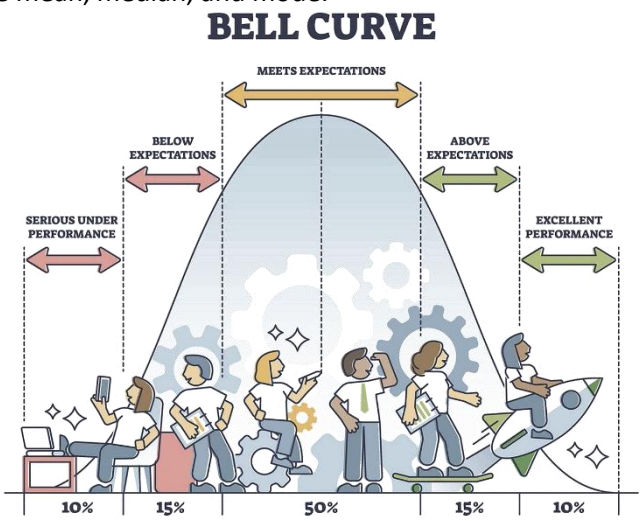
What is the Central Limit Theorem?
The Central Limit Theorem (CLT) states that, given a sufficiently large sample size from a population with a finite level of variance, the sampling distribution of the mean will be normally distributed regardless of if the population is normally distributed.

Advantages of the Normal Distribution

- **Ease of Analysis and Interpretation**: The normal distribution provides a simplified framework for statistical analysis and interpretation.
- **Use of Parametric Tests**: With normally distributed data, powerful and widely used parametric tests can be applied.
- **Central Limit Theorem (CLT)**: With large sample sizes, many distributions converge toward a normal distribution, facilitating statistical inference.
- **Distribution of Error Terms**: In regression analysis, normally distributed error terms are important for the validity of the model.
- **Detection of Anomalies and Outliers**: Deviations from a normal distribution help in identifying anomalies or outliers.
- **Prediction and Forecasting**: The normal distribution supports the development of probability-based prediction and forecasting models.

Bell Curve

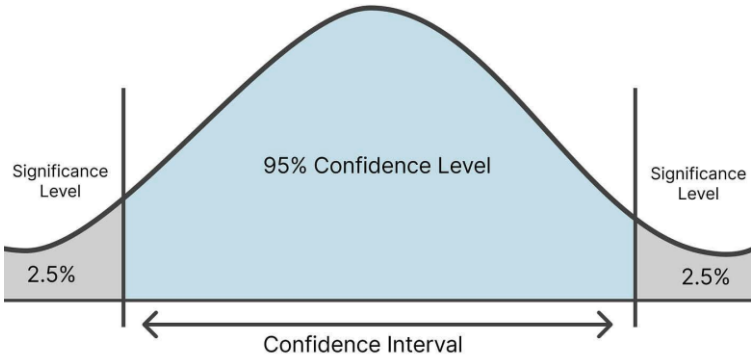
- **Bell Curve**: The graphical representation of the normal distribution.
- **Center**: Most data points cluster around the mean, median, and mode.
- **Symmetry**: The curve is symmetrical around the mean.



- **Distribution**:
  - **10%**: Below Expectations
  - **15%**: Seriously Below Expectations
  - **50%**: Meets Expectations
  - **15%**: Above Expectations
  - **10%**: Excellent Performance

3. Confidence Interval (CI)

- **Definition**: A confidence interval indicates the range within which a population parameter is likely to fall, given a specific confidence level (e.g., 95%).
- **Features**:
  - The confidence interval shows how reliable an estimate is.
  - A confidence level (e.g., 95%) represents the probability that the population parameter lies within the interval.
- **Example**: If a survey yields a result of 50% with a margin of error of ±3%, the actual value is likely between 47% and 53%.



What is the difference between Point Estimates and Confidence Interval?
Point Estimation:
Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

Confidence Interval: A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by 1 - alpha, where alpha is the level of significance.

4. Margin of Error (MOE)

- **Definition**: Indicates how accurate a statistical estimate obtained from a sample is.
- **Features**:
  - The margin of error determines the width of the confidence interval.
  - A smaller margin of error means a more accurate estimate.
- **Example**: If a survey result shows 50% with a margin of error of ±3%, the true value is likely between 47% and 53%.

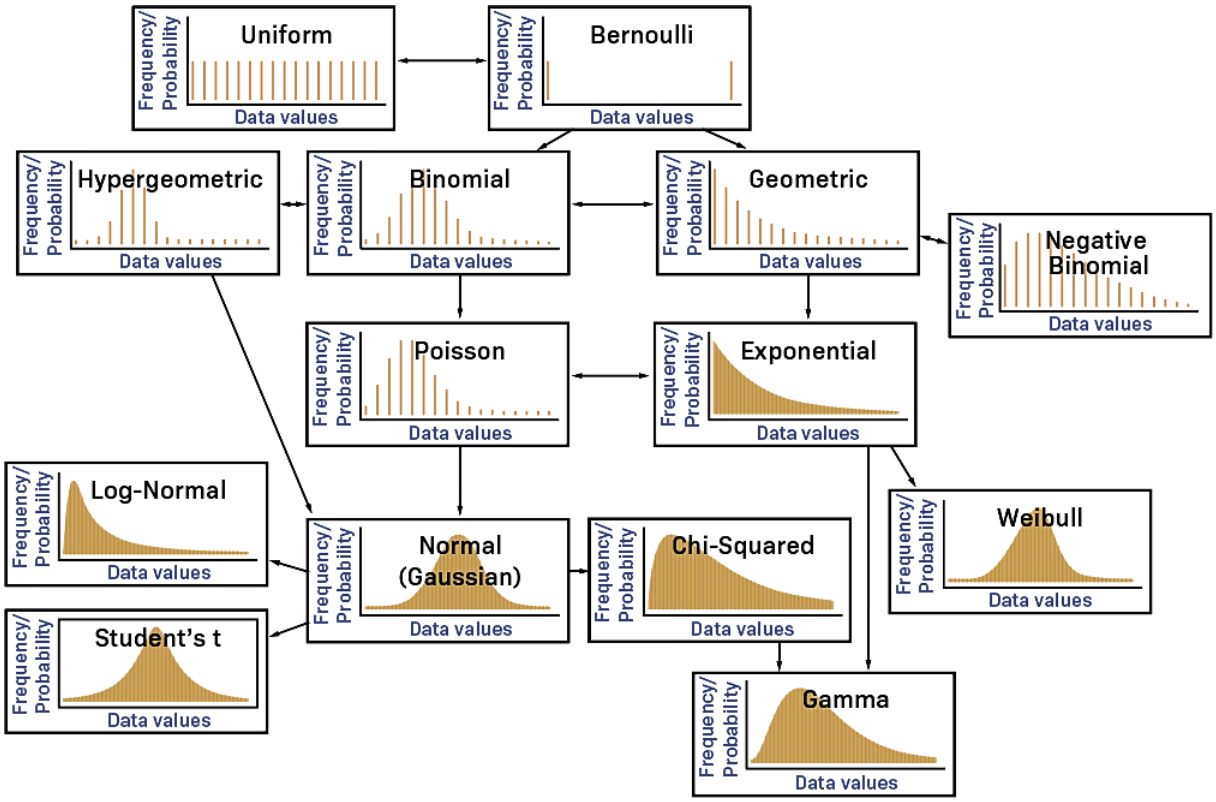
Point Estimates

- **Definition**: The estimation of a population parameter (e.g., mean, variance) using a single value.
- **Features**:
  - Considered the best guess for the population parameter.
  - It gives a single value and therefore does not account for uncertainty.
- **Example**: The sample mean ($\bar{x}$) is the point estimate of the population mean (μ).

Summary Table

Concept	Description
Sample Distributions	The distribution of statistics calculated from samples. Example: Distribution of sample means.
Central Limit Theorem	With large sample sizes, the distribution of sample means approaches a normal distribution.
Confidence Interval (CI)	An interval that likely contains the population parameter at a given confidence level (e.g., 95%).
Margin of Error (MOE)	Indicates the accuracy of a statistical estimate from a sample. Example: ±3% margin of error.
Point Estimates	A single-value estimate of a population parameter. Example: Sample mean.

Chart types according to scatter type:



What is Cluster Sampling? Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

What is sampling? How many sampling methods do you know?
"Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined."

8.Section : Hypothesis Testing

Significance Tests

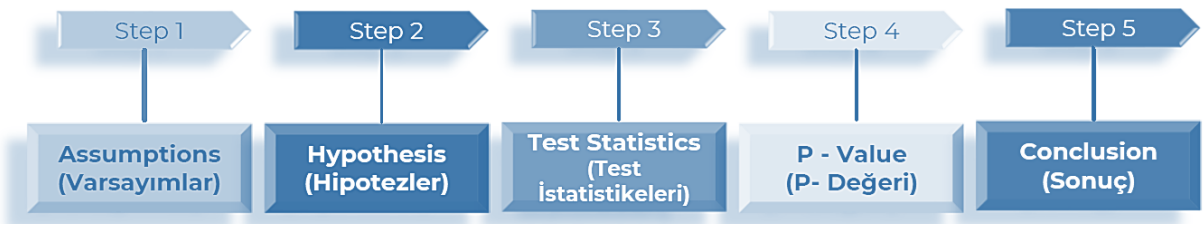
Significance tests are statistical methods used to determine whether the results obtained from a sample can be generalized to the broader population. These tests help us assess whether observed differences represent a real effect or are simply due to random variation.

Main Purposes:

- **Generalization**: To test whether results from a sample are valid for the population.
- **Evaluating Significance of Differences**: To determine whether an observed difference is real or negligible.

Examples:

1. **"I lost weight, but is this weight loss statistically significant?"**
    - This example questions whether a personal experience (weight loss) is statistically meaningful.
    - A significance test helps determine if the weight loss is due to random fluctuation or represents a true effect.
2. **"I followed different diets and lost weight with each. Are these differences statistically significant?"**
    - This examines whether the outcomes of multiple experiences (weight loss from different diets) are statistically significant.
    - A significance test helps determine whether each diet has a real effect on weight loss and whether the differences between diets are meaningful.
3. **"Three people saw Dietitian A and lost weight; three saw Dietitian B and also lost weight. Is the difference between these two dietitians statistically significant?"**
    - This compares two groups (those who saw Dietitian A vs. Dietitian B) in terms of weight loss outcomes.
    - A significance test helps determine whether there's a statistically meaningful difference between the two groups.



Hypothesis Test

- **Purpose**:
  To test whether the results obtained from a sample can be generalized to the population.
- **Main Steps:**

  1. Assumptions

  - In this step, assumptions required for the validity of the test are checked.
  - For example, whether the data are normally distributed or whether samples are independent.
  - Violating these assumptions may affect the reliability of the test results.

  2. Formulate Hypotheses:

  - **Null Hypothesis (H_o)**: Assumes "no change" or "status quo" is true.
    Example: "The lead concentration in the sea is 10 ppm."
  - **Alternative Hypothesis (H_1)**: Represents the researcher's claim.
    Example: "The lead concentration in the sea is greater than 10 ppm."

  3. Choose the Test

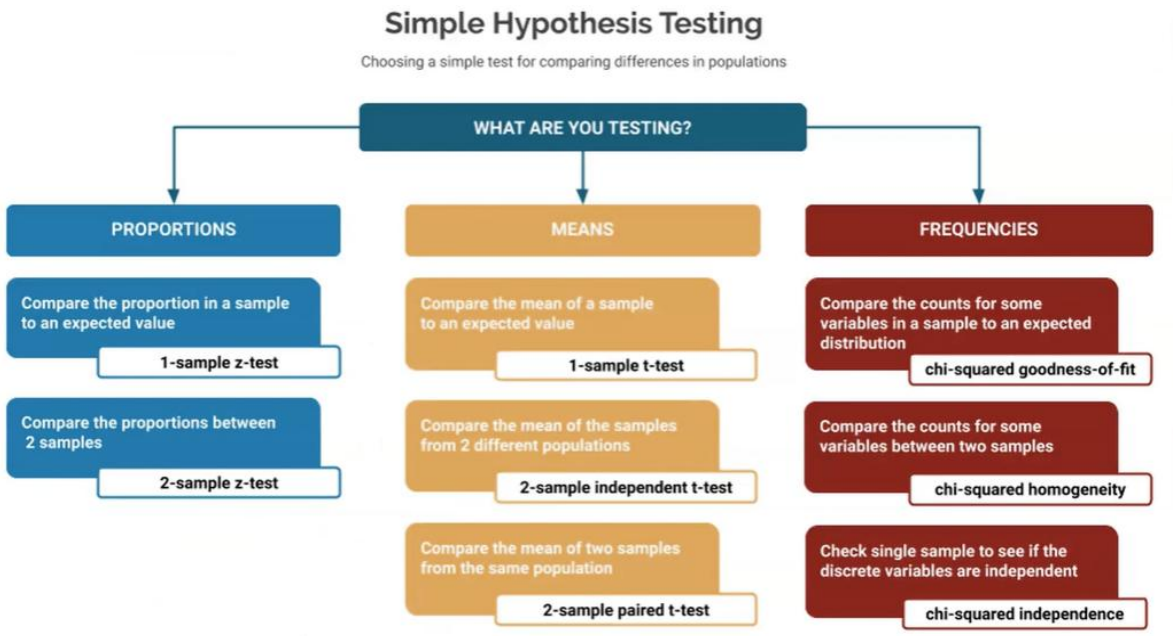
  - Depending on the data type (numerical/categorical) and assumptions, the appropriate test is selected (e.g., z-test, t-test, chi-square test).
  - A statistical value is calculated using the sample data.

  4. Determine the P-Value

  - The p-value represents the probability of obtaining the observed result, or more extreme, assuming the null hypothesis is true.
  - A low p-value indicates a high probability that the null hypothesis is incorrect.

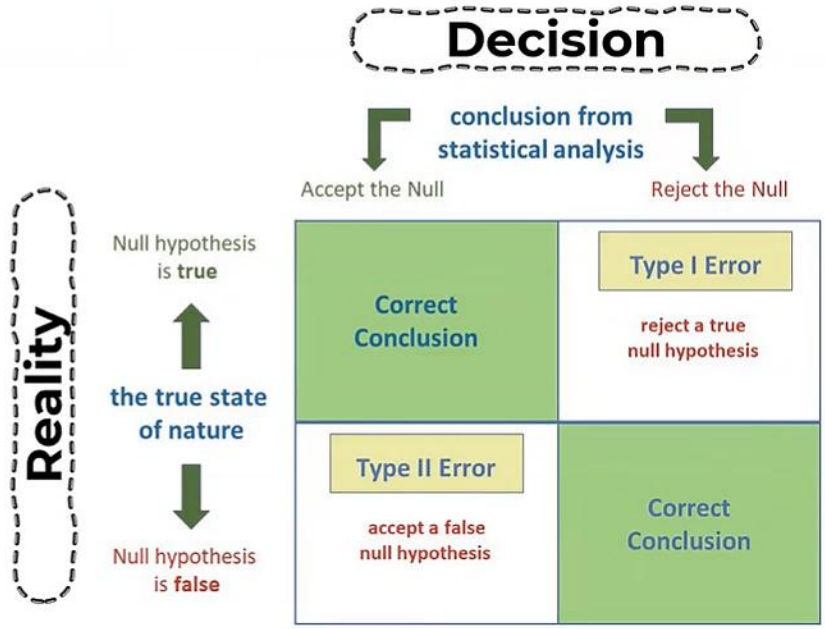
  5. Conclusion

  - Compare the p-value to the significance level (α):
    - **If $p < \alpha \rightarrow$ Reject H_o**: This supports the alternative hypothesis.
    - **If $p \geq \alpha \rightarrow$ Fail to Reject H_o**: There is not enough evidence to reject the null hypothesis.



Type I – Type II Errors

- **Type I Error (α)**: Incorrectly rejecting the null hypothesis (H_0) when it is actually true.
  Example: "Concluding that a drug is effective when it is actually not."
- **Type II Error (β)**: Failing to reject the null hypothesis when it is actually false.
  Example: "Concluding that a drug is ineffective when it actually works."



One-Tailed vs. Two-Tailed Tests

- **One-Tailed Test**: The alternative hypothesis specifies a direction (e.g., "greater than" or "less than").
  Example: "The new drug is more effective than the current drug."
- **Two-Tailed Test**: The alternative hypothesis states there is a difference without specifying a direction.

Example: "There is a difference between the two groups."
P-value Calculation: In two-tailed tests, the p-value accounts for both directions of deviation.

Example Scenario: Lead Levels in the Sea

Problem:

- The expected lead concentration in a sea is 10 ppm (μ_o).
- The population is normally distributed with a standard deviation $\sigma = 1.5$.
- A sample of 40 measurements was taken, and the sample mean lead concentration was 10.5 ppm.
- Does this difference indicate that the lead level is significantly higher than the population mean at $\alpha = 0.05$ (95% confidence)?

Solution:

Step 1: Assumptions

- **Normal Distribution**: Lead levels in the population follow a normal distribution, allowing the use of a z-test.
- **Independent Samples**: The 40 samples are independently selected, which increases the reliability of the sample mean.
- **Known Standard Deviation**: The population standard deviation ($\sigma = 1.5$) is known, permitting a z-test.
- **Large Sample Size**: n = 40, which qualifies as a large sample.

Step 2: Hypotheses

- **Null Hypothesis (H_o)**: The mean lead level in the sea is 10 ppm. ($\mu = 10$)
- **Alternative Hypothesis (H_1)**: The mean lead level in the sea is significantly greater than 10 ppm. ($\mu > 10$)

Step 3: Test Statistic

- Since σ is known, we use the **z-test**:
- $z = (\bar{x} - \mu_o) / (\sigma / \sqrt{n})$
- $z = (10,5 - 10) / (1,5 / \sqrt{40})$
- $z \approx 2,108$

Step 4: P-value

- From the z-table or a statistical calculator, for z = 2.108, the p-value is approximately **0.0175**.

Step 5: Conclusion

- Compare p-value with $\alpha = 0.05$:
  $0.0175<0.050 \rightarrow$ Reject H_o

Step 6: Final Result
With 95% confidence, we conclude that the lead level in the sea is significantly greater than 10 ppm.

Additional Information:

- This is a **one-tailed (right-tailed)** hypothesis test because the alternative hypothesis claims the mean is greater than a specific value.
- The **p-value** shows the probability of observing results as extreme as, or more extreme than, the sample result under the assumption that H_o is true.
- The **significance level (α)** indicates the probability of incorrectly rejecting a true null hypothesis (Type I Error).

Significance Level (α) vs. P-Value

- **α (Significance Level)**: Commonly set at 0.05 (5% margin of error).
  Why 0.05? It's a conventional standard, but it may vary depending on the field (e.g., 0.01 in medicine).
- **P-Value**:
  - Acts like a referee: indicates the likelihood that the observed results are due to chance.
  - *Example*: If **p = 0.03**, and **α = 0.05**, then reject H_o.

Another Example Scenario

Problem: A pharmaceutical company claims its new drug is more effective in reducing fever than the existing drug.

- **H_o**: There is no difference between the new and existing drugs.
- **H_1**: The new drug is more effective. (One-tailed test)
- **Test Used**: **t-test** (population variance unknown)
- **Result**: If p = 0.02 and α = 0.05 → Reject H_o

T-Test vs. Z-Test

- **T-Test**:
  - Used when the population standard deviation is **unknown** and for **small samples (n < 30)**.
  - *Example*: Comparing the means of two groups.
- **Z-Test**:
  - Used when the population standard deviation is **known** and for **large samples (n ≥ 30)**.
  - *Example*: Comparing a sample mean to a population mean.

Summary Table

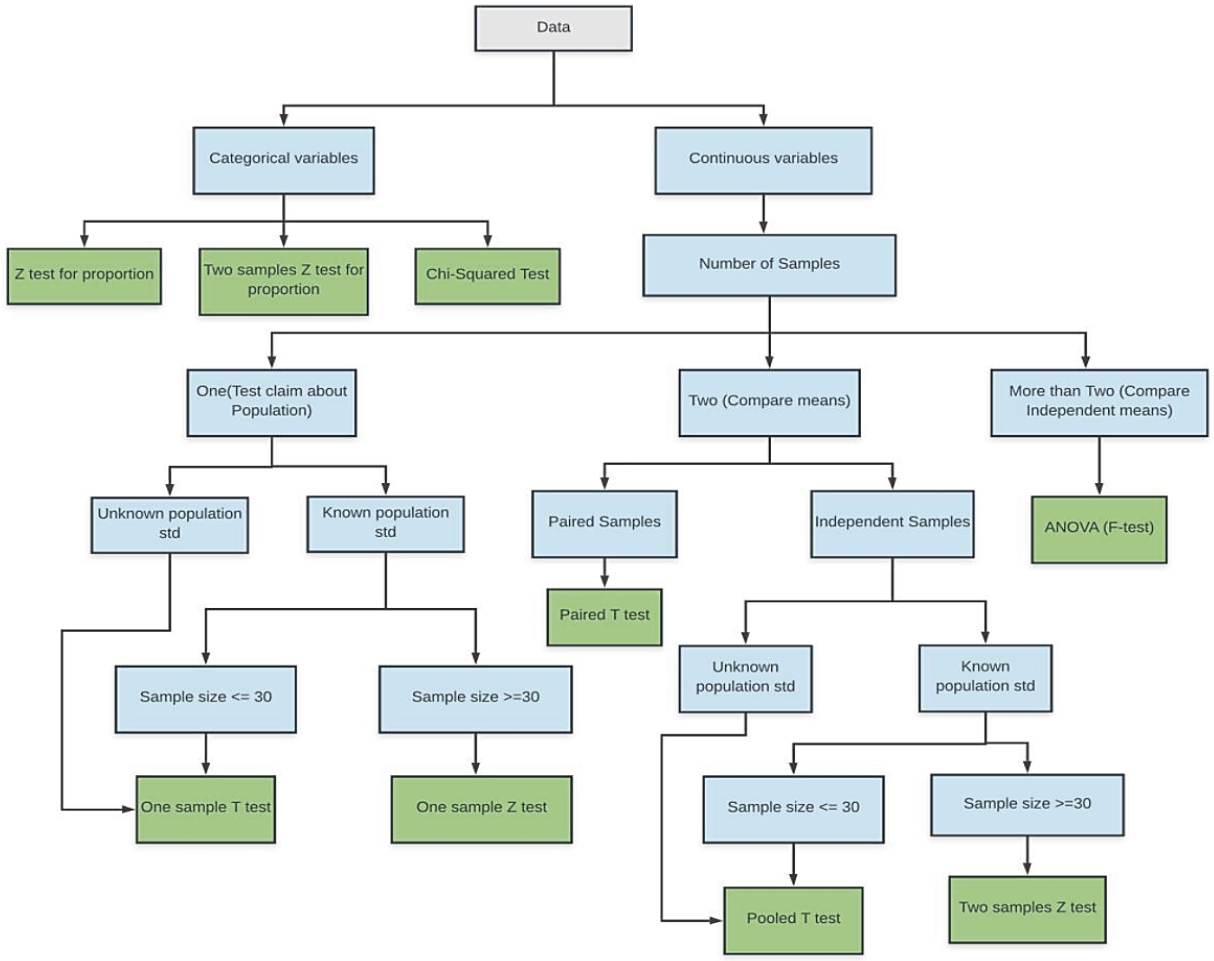
Concept	Description
H_o (Null Hypothesis)	The hypothesis that assumes no change or effect.
H_1 (Alternative Hypothesis)	The hypothesis representing the researcher's claim.
P-value	The probability of observing the given result assuming the null hypothesis is true.
α (Significance Level)	The acceptable margin of error (commonly 0.05).

Concept	Description
One/Two-Tailed Test	Determined based on the direction of the alternative hypothesis.
T-Test	Used when the sample size is small and the population standard deviation is unknown.
Z-Test	Used when the sample size is large and the population standard deviation is known.

Conclusion

Significance tests are critical tools for supporting scientific claims with statistical evidence. Proper formulation of hypotheses, appropriate test selection, and accurate interpretation of p-values determine the reliability of the results. These methods are used to derive meaningful insights from data and guide decision-making processes.

The Path We Will Follow in Hypothesis Testing:



9.Section : Advanced Hypothesis Testing

Independent Samples T-Test

- **Purpose:** To test whether the mean difference between two independent groups is statistically significant.
- **Example:** Is there a significant difference in math scores between male and female students?

Steps:

1. **Assumptions:**
    - The data should follow a normal distribution.
    - Group variances should be approximately equal (homogeneity of variance).
2. **Hypotheses:**
    - $\mathbf{H_o}$**:** There is no difference in means between the two groups.
    - $\mathbf{H_1}$**:** There is a difference in means between the two groups.
3. **Test Statistic:** Calculate the t-statistic.
4. **P-Value:** Determine the p-value corresponding to the t-statistic.
5. **Decision:** If $P < \alpha$, reject H_o.

Dependent Samples T-Test (Paired T-Test)

- **Purpose:** To test whether the mean difference between two related groups is statistically significant.
- **Example:** Comparing patient measurements before and after taking a medication.

Steps:

1. **Assumptions:**
    - The differences should follow a normal distribution.
2. **Hypotheses:**
    - $\mathbf{H_o}$**:** The mean difference is zero.
    - $\mathbf{H_1}$**:** The mean difference is not zero.
3. **Test Statistic:** Calculate the t-statistic.
4. **P-Value:** Find the p-value corresponding to the calculated t-statistic.
5. **Decision:** If $P < \alpha$, reject H_o.

One-Way ANOVA

- **Purpose:** To test whether the means of three or more independent groups differ significantly.
- **Example:** Comparing the effects of three different types of medication on patients.

Steps:

1. **Assumptions:**
    - Groups should be normally distributed.
    - Homogeneity of variances should be satisfied.
2. **Hypotheses:**
    - $\mathbf{H_o}$**:** All group means are equal.
    - $\mathbf{H_1}$**:** At least one group mean differs from the others.

3. **Test Statistic:** Calculate the F-statistic.
4. **P-Value:** Find the p-value corresponding to the F-statistic.
5. **Decision:** If $P < \alpha$, reject H_o.

ANOVA Table and Test Statistics

- **SSR (Sum of Squares for Regression):** Variance explained by the model.
- **SSE (Sum of Squares for Error):** Variance not explained by the model.
- **SST (Total Sum of Squares):** Total variance (SST = SSR + SSE).
- **MSR (Mean Square Regression):** SSR / degrees of freedom.
- **MSE (Mean Square Error):** SSE / degrees of freedom.
- **F-Statistic:** MSR / MSE.

Categorical Data Analysis

- **Purpose:** To examine the relationship between categorical variables.
- **Example:** Is there a relationship between smoking habits and lung cancer?

Chi-Square Test

Steps:

1. **Assumptions:**
    - Observed frequencies are compared with expected frequencies.
2. **Hypotheses:**
    - **H_o:** No relationship exists between the two variables.
    - **H_1:** A relationship exists between the two variables.
3. **Test Statistic:** Calculate the chi-square statistic.
4. **P-Value:** Find the p-value corresponding to the calculated chi-square value.
5. **Decision:** If $P < \alpha$, reject H_o.

Example Scenarios and Solutions

Example 1: Independent T-Test

- **Problem:** Is there a significant difference in math scores between male and female students?

Steps:

1. **Assumptions:** Normal distribution and equal variances.
2. **Hypotheses:**
    - **H_o:** Mean scores are equal.
    - **H_1:** Mean scores are not equal.
3. **Test Statistic:** T = 2.45
4. **P-Value:** P = 0.015
5. **Result:** $P < 0.05 \rightarrow$ Reject H_o. There is a significant difference between male and female students' scores.

Example 2: ANOVA

- **Problem:** Comparing the effects of three different drugs on patients.

- **Steps:**

1. **Assumptions:** Normal distribution and homogeneity of variance.
2. **Hypotheses:**
    - **H_o:** The effects of all three drugs are equal.
    - **H_1:** At least one drug has a different effect.
3. **Test Statistic:** F = 4.67
4. **P-Value:** P = 0.012
5. **Result:** P < 0.05 → Reject H_o. At least one drug has a significantly different effect.

Example 3: Chi-Square Test

- **Problem:** Is there a relationship between smoking and lung cancer?

Steps:

1. **Assumptions:** Observed frequencies compared with expected frequencies.
2. **Hypotheses:**
    - **H_o:** No relationship exists.
    - **H_1:** A relationship exists.
3. **Test Statistic:** $\chi^2 = 9.82$
4. **P-Value:** P = 0.002
5. **Result:** P < 0.05 → Reject H_o. There is a significant relationship between smoking and lung cancer.

Which Test Should I Use?

Data Type	Number of Groups	Test
Continuous (Numeric)	2	Independent T-Test
Continuous (Numeric)	2 (Matched)	Paired T-Test
Continuous (Numeric)	≥3	ANOVA
Categorical	Any	Chi-Square Test

Conclusion

These tests are used to analyze relationships in data and draw meaningful conclusions. Correct test selection and proper hypothesis formulation are essential for ensuring the reliability of results.

What do you understand by Hypothesis Testing?

In Statistics, Hypothesis Testing is mainly used to see if a certain experiment generates meaningful results. It helps assess the statistical significance of insight by finding the odds of the results occurring by chance. In Hypothesis Testing, the first thing is to know the null hypothesis and then specify it. After that, the p-value is calculated, and if the null hypothesis is true, the other values are also determined. The alpha value specifies the significance, and you can adjust it accordingly. If the p-value is less than the alpha value, the null hypothesis is rejected, but the null hypothesis is accepted if the p-value is greater than the alpha value. If the null hypothesis is rejected, it indicates that the results obtained are statistically significant.

What is the relationship between the significance level and the confidence level in Statistics?

In Statistics, the significance level is the probability of getting a completely different result from the condition where the null hypothesis is true. On the other hand, the confidence level is used as a range of similar values in a population.
We can specify the similarity between the significance level and the confidence level by the following formula:
Significance level = 1 - Confidence level

What is p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.
Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis. p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,
High P values: your data are likely with a true null.
Low P values: your data are unlikely with a true null.

How do you assess the statistical significance of an insight?

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

What is the significance of p-value?

- **p-value typically ≤ 0.05**

  This indicates that there is strong evidence against the null hypothesis, so you reject the null hypothesis.

- **p-value typically > 0.05**

  This indicates that there is weak evidence against the null hypothesis, so you accept the null hypothesis.

What is the Null and Alternate Hypothesis?

A null and alternate hypothesis is used in statistical hypothesis testing.

Null Hypothesis (Sıfır Hipotezi):

- It states that the population parameter is equal to the assumed value.
- It is an initial claim based on previous analysis or experience.

Alternate Hypothesis (Alternatif Hipotez):

- It states that population parameters are equal or different to the assumed value.
- It is what you might believe to be true or want to prove true.

What is Hypothesis Testing?

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.
There are 3 steps in Hypothesis Testing:

1. State Null and Alternate Hypothesis
2. Perform Statistical Test
3. Accept or reject the Null Hypothesis

What do you understand by Hypothesis Testing?

In Statistics, Hypothesis Testing is mainly used to see if a certain experiment generates meaningful results. It helps assess the statistical significance of insight by finding the odds of the results occurring by chance. In Hypothesis Testing, the first thing is to know the null hypothesis and then specify it. After that, the p-value is calculated, and if the null hypothesis is true, the other values are also determined. The alpha value specifies the significance, and you can adjust it accordingly. If the p-value is less than the alpha value, the null hypothesis is rejected, but the null hypothesis is accepted if the p-value is greater than the alpha value. If the null hypothesis is rejected, it indicates that the results obtained are statistically significant.

What are a p-value and its role in Hypothesis Testing?

P-value is the probability that a random chance generated the data or something else that is equal or rare.
P-values are used in hypothesis testing to decide whether to reject the null hypothesis or not.

- **p-value < alpha – value**
  Means results are not in favor of the null hypothesis, reject the null hypothesis.

📌 **These results are based on 300+ statistics interview questions from 50+ companies.**