DATA ANALYTICS

What is DATA?

- Data is a collection of facts or numbers that are organized to be analyzed, categorized, or used to help make decisions.
- It can be numerical or non-numerical, and can include discrete or continuous values.
- Quantitative Data: Numerical data that can be measured and quantified. It is further divided into:

❖Discrete Data: Countable data, like the number of students in a class.

❖Continuous Data: Data that can take any value within a range, like height or temperature.

- Qualitative Data: Non-numerical data that describes qualities or characteristics. It includes:

❖Nominal Data: Categorical data without a specific order, like colors or names.

❖Ordinal Data: Categorical data with a specific order, like rankings or grades.

- **Data Analytics:** The process of examining raw data to uncover patterns, draw conclusions, and make informed decisions. It involves techniques such as descriptive, diagnostic, predictive, and prescriptive analytics. The goal of data analytics is to transform data into actionable insights that can guide decision-making in various fields such as business, healthcare, and finance.
- **Data Mining:** In this it focused on discovering hidden patterns and relationships in large datasets using methods from statistics, machine learning, and database systems. It involves tasks such as clustering, classification, regression, and association rule learning. The primary aim of data mining is to extract previously unknown and potentially useful information from data.
- **Data Science:** It is an interdisciplinary field that combines computer science, statistics, and domain expertise to extract knowledge and insights from structured and unstructured data. It encompasses a broader scope than data analytics and data mining, including data collection, cleaning, analysis, and visualization. Data science also involves the development of algorithms and models to solve complex problems and predict future trends.

- **Datasets:** A dataset is a collection of data that is organized in a structured format, often presented in a tabular form consisting of rows and columns. Each row represents a record, and each column represents a variable or attribute of the data. Datasets are fundamental to data analysis, machine learning, and data science as they provide the raw information needed for analysis and model training. E.g. Iris dataset, MINST dataset.

- **Features:** They are individual measurable properties or characteristics of the phenomena being observed. Features are also referred to as attributes, variables, or predictors. They are the input variables used to build models that can predict outcomes or uncover patterns in the data. It can be in numerical and categorical form.

- **Data Scales:** Data scales refer to the levels of measurement that describe the nature of information within the values assigned to variables. They are essential in determining the appropriate statistical analysis and visualizations that can be applied to the data. There are four primary scales of measurement: nominal, ordinal, interval, and ratio.

Numerical and Categorical data:

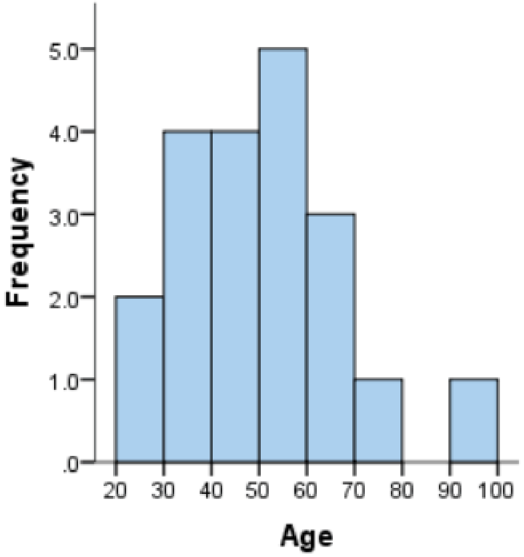
- **Numerical data** refers to data that represents measurable quantities and can be expressed as numbers. It is often used for quantitative analysis and can be further classified into two types: discrete and continuous. Discrete numerical data consists of countable values, such as the number of students in a class or the number of cars in a parking lot. Continuous numerical data, on the other hand, can take any value within a given range and includes measurements such as height, weight, and temperature.
- **Categorical data** represents characteristics or attributes that can be grouped into categories. Unlike numerical data, categorical data does not imply a quantity but rather a quality or classification. This type of data can be divided into nominal and ordinal categories. Nominal data consists of categories without any intrinsic order, such as gender, nationality, or hair color. Ordinal data, while still categorical, includes a meaningful order or ranking, such as levels of education (high school, bachelor's, master's) or customer satisfaction ratings (satisfied, neutral, dissatisfied).

Cross-sectional and Time series data:

- **Cross-sectional data** refers to data collected at a single point in time, capturing a snapshot of multiple subjects or entities. This type of data is used to analyze variations across different subjects, such as individuals, companies, countries, or any other units of analysis, at a particular moment. For example, a survey collecting data on the income levels, educational background, and employment status of different individuals at one point in time is cross-sectional.
- **Time series data**, in contrast, consists of data points collected or recorded at successive points in time, often at regular intervals. This type of data tracks the evolution of specific variables over time, enabling analysis of trends, cycles, and patterns. Examples include daily stock prices, monthly unemployment rates, or yearly GDP growth rates. Time series data is crucial for forecasting and understanding temporal dynamics.

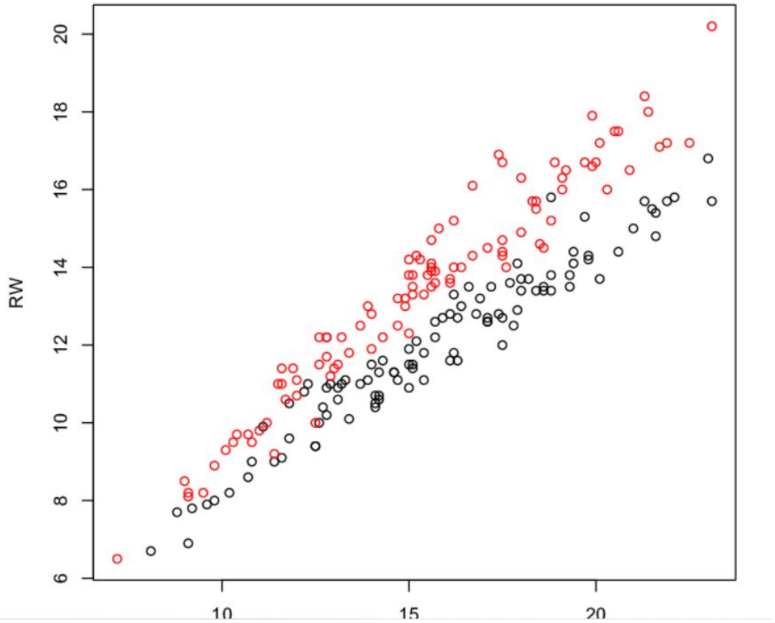
Univariate:

- **Univariate analysis** involves the examination and analysis of a single variable. The primary objective is to understand the distribution and characteristics of this variable within a dataset. Univariate analysis includes calculating measures of central tendency (such as mean, median, and mode) and measures of dispersion (such as range, variance, and standard deviation). It also involves creating visualizations like histograms, bar charts, and box plots. This type of analysis is fundamental in statistics as it provides a foundational understanding of each variable independently before considering their interactions with other variables.



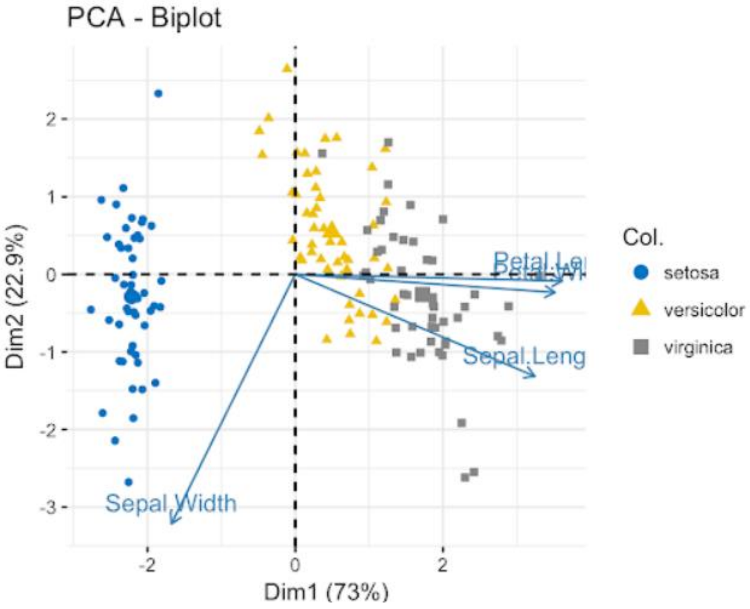
Bivariate:

- **Bivariate analysis** examines the relationship between two variables. This type of analysis aims to determine the association or correlation between the variables, exploring whether and how one variable influences or is related to another. Techniques used in bivariate analysis include scatter plots for visualizing relationships, correlation coefficients for measuring the strength and direction of linear relationships, and cross-tabulations for categorical data. Regression analysis, particularly simple linear regression, is also a common tool in bivariate analysis to model the relationship between an independent variable and a dependent variable. Bivariate analysis is crucial for identifying and understanding direct interactions and dependencies between two variables.



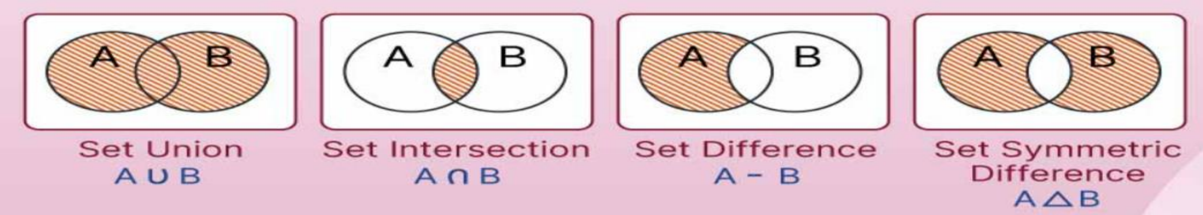
Multivariate:

- **Multivariate analysis** extends beyond two variables to simultaneously analyze three or more variables. This type of analysis aims to understand complex relationships and interactions among multiple variables, often in a more holistic manner. Techniques for multivariate analysis include multiple regression analysis, principal component analysis (PCA), factor analysis, and cluster analysis. These methods help in identifying patterns, reducing dimensionality, and uncovering underlying structures in the data. Multivariate analysis is particularly powerful in fields like data science, market research, and social sciences, where understanding the interplay among multiple factors is essential for comprehensive insights and decision-making.

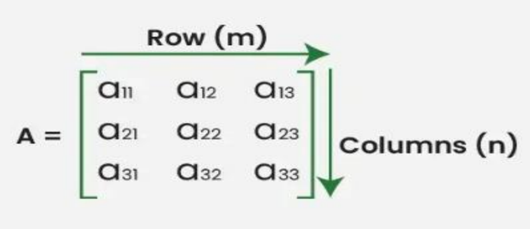


Set and matrix representations ,relations:

- A set is a fundamental concept that refers to a collection of distinct objects, considered as an object in its own right. These objects, called elements or members, can be anything: numbers, people, letters, other sets, etc. Sets are typically denoted using curly braces, such as {a,b,c} where a, b, and c are the elements of the set.



- A matrix representation is a way to encode algebraic structures, such as groups, rings, and vector spaces, into matrices, thereby facilitating their manipulation and analysis. By transforming abstract elements into concrete arrays of numbers, matrix representations allow for the application of linear algebra techniques, making complex problems more tractable.

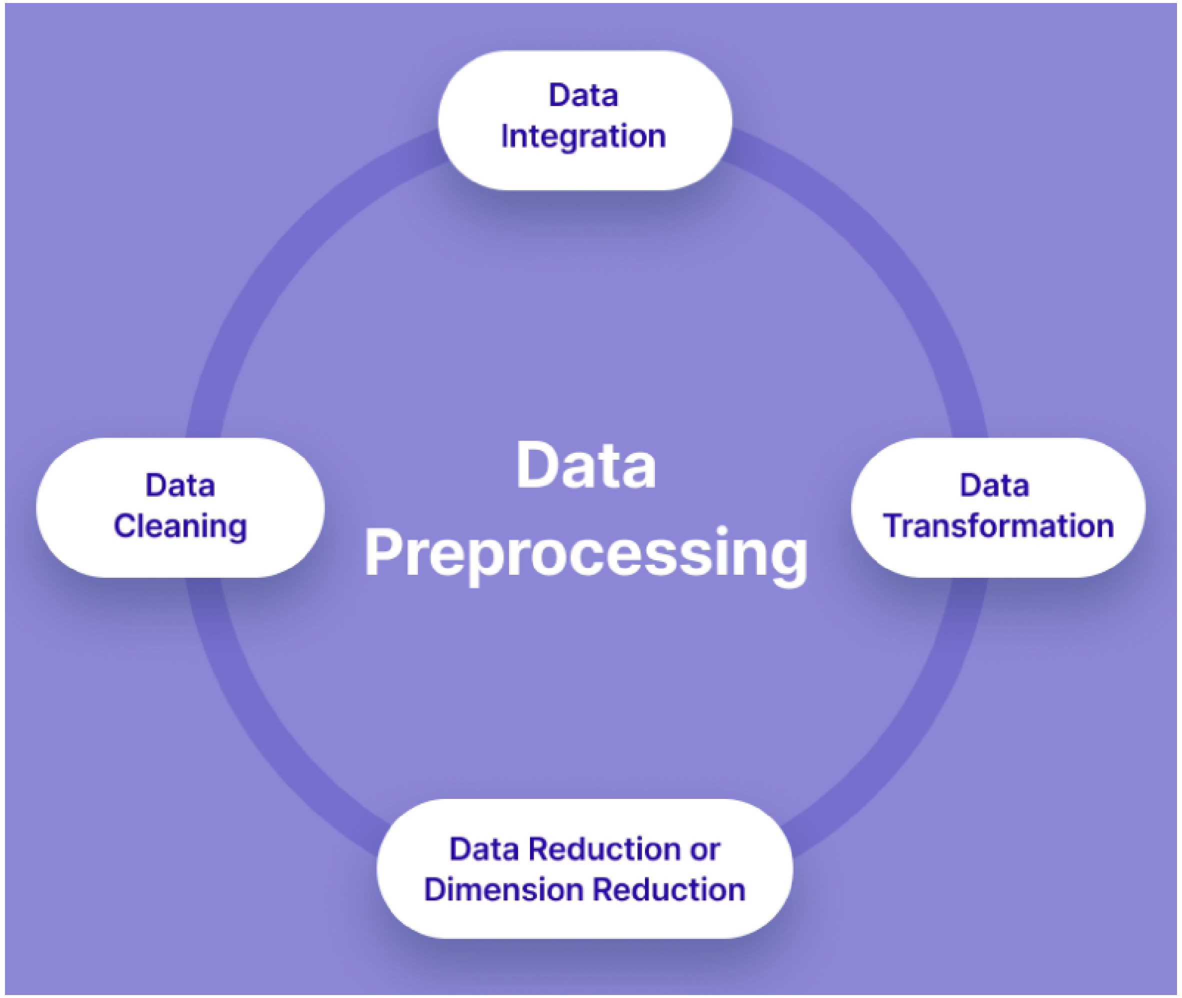


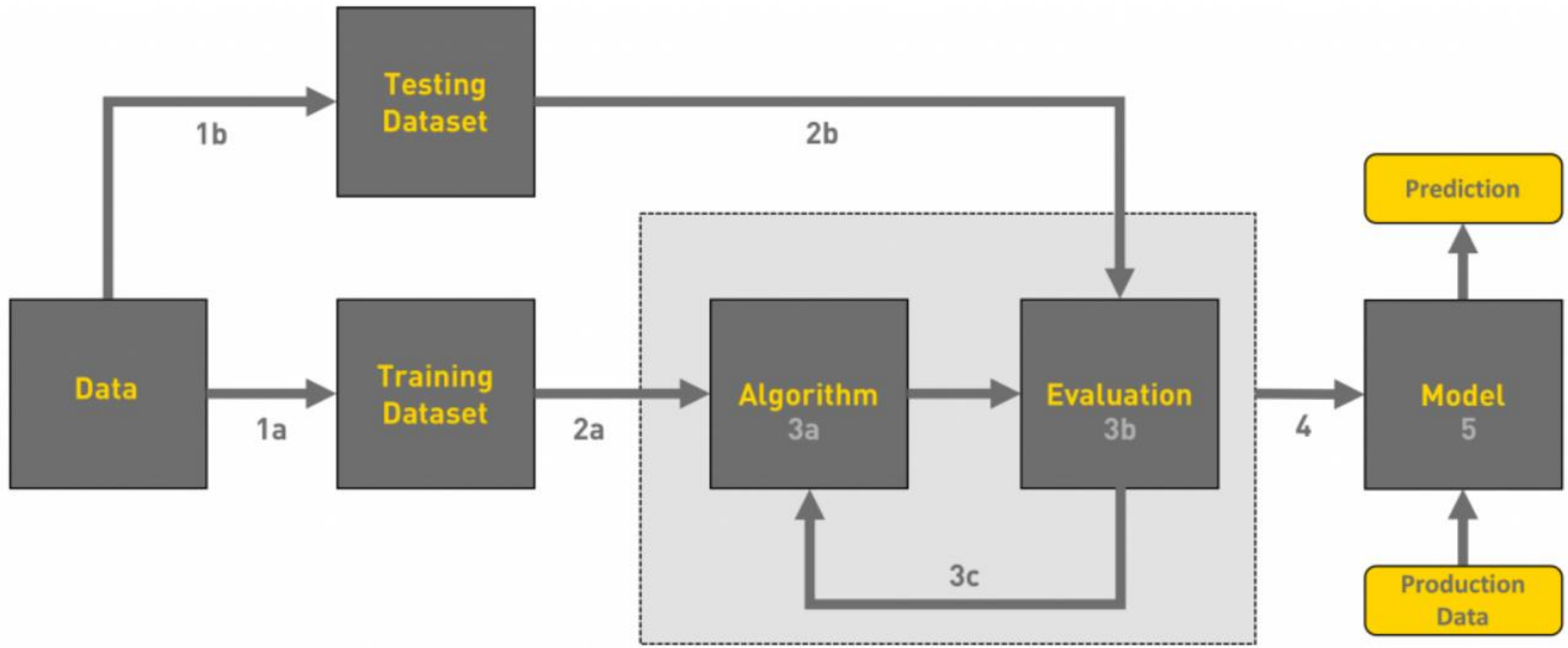
$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

- A relation matrix is a two-dimensional array that represents the relationship between two sets using an algebraic method.

What is Data Preprocessing?

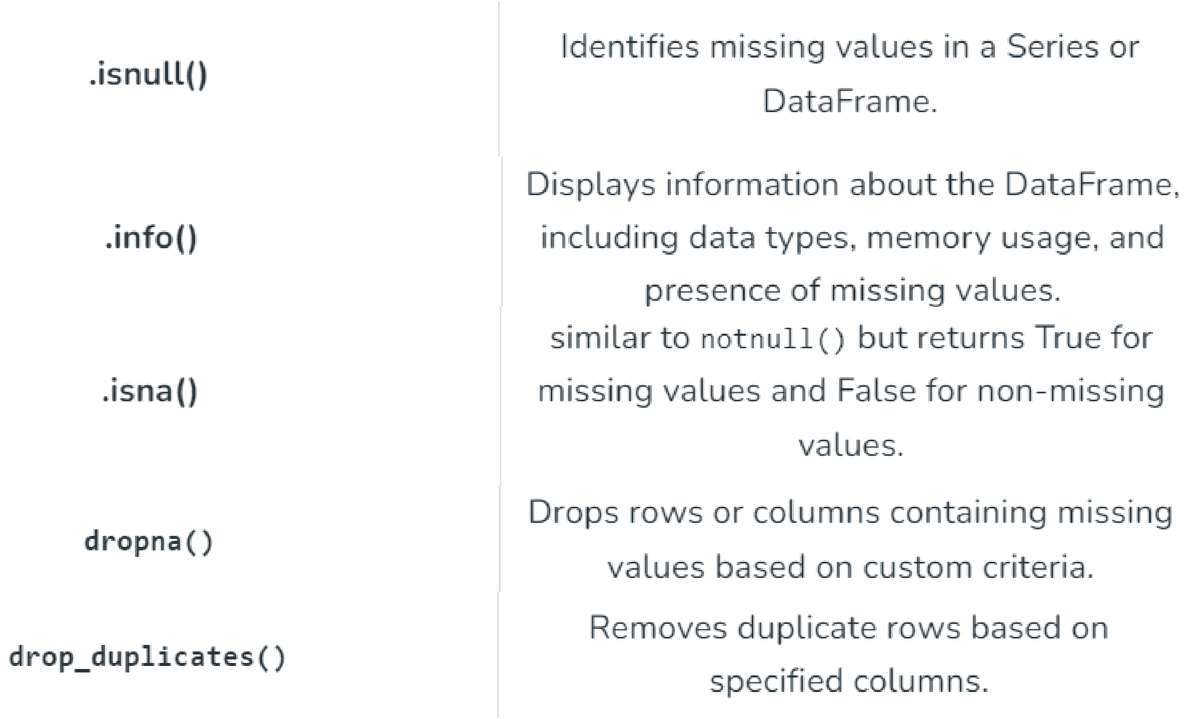
- Data preprocessing is a critical step in the data analysis and machine learning pipeline, involving a series of operations aimed at transforming raw data into a clean, structured format suitable for analysis.
- Key steps in data preprocessing include data cleaning, where missing values are handled, and inconsistencies or errors are corrected.
- Data integration follows, combining data from multiple sources to provide a unified view.
- Data transformation is another crucial step, involving normalization or scaling to ensure that the data fits within a specific range, making it suitable for algorithms that are sensitive to the scale of input data.
- Feature extraction and selection reduce the dimensionality of the data, focusing on the most relevant attributes, which helps in improving model efficiency and interpretability.


Data Integration
Data Cleaning
Data Preprocessing
Data Transformation
Data Reduction or Dimension Reduction


Testing Dataset
1b
2b
Prediction
Data
1a
Training Dataset
2a
Algorithm
3a
Evaluation
3b
4
Model
5
3c
Production Data

Handling Missing values:

- Handling missing values is a critical step in data preprocessing that can significantly impact the performance and accuracy of machine learning models.
- Missing data can arise from various sources, such as human error during data entry, equipment malfunctions, or incomplete data collection processes.
- Data can be missing for many reasons like technical issues, human errors, privacy concerns, data processing issues, or the nature of the variable itself.



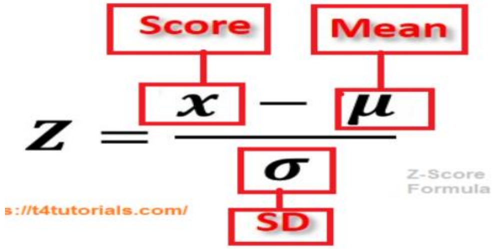
.isnull()	Identifies missing values in a Series or DataFrame.
.info()	Displays information about the DataFrame, including data types, memory usage, and presence of missing values.
.isna()	similar to `notnull()` but returns True for missing values and False for non-missing values.
`dropna()`	Drops rows or columns containing missing values based on custom criteria.
`drop_duplicates()`	Removes duplicate rows based on specified columns.

Data Normalization

- Normalization is a critical step in data preprocessing, particularly when working with machine learning algorithms.
- It involves adjusting the values of numeric data features to a common scale, without distorting differences in the ranges of values.
- Two common normalization techniques are min-max normalization and z-score normalization.
- Min-max normalization, also known as feature scaling, transforms the data to fit within a specific range, typically [0, 1]

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

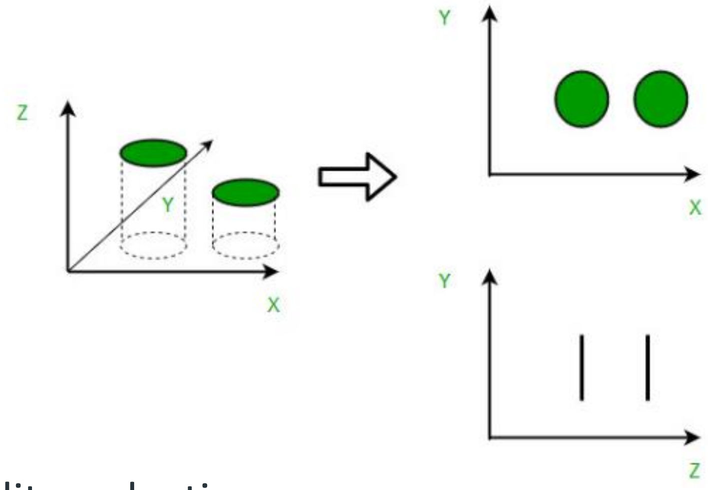
- Z-score normalization, or standardization, adjusts the data to have a mean of zero and a standard deviation of one.
- Standardization is especially beneficial for algorithms that assume data is normally distributed, such as linear regression and principal component analysis.



Dimensionality Reduction

- Dimensionality reduction is a crucial step in the preprocessing of data, particularly in the realm of machine learning and data analysis.
- It involves reducing the number of random variables under consideration, simplifying the dataset while retaining its essential features.
- Firstly, it helps in mitigating the "curse of dimensionality," which refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces.
- As dimensions increase, the volume of the space increases exponentially, leading to sparse data that complicates analysis and model training.
- Dimensionality reduction techniques, address this by transforming the data into a lower-dimensional space that captures most of the variability or discriminatory information.
- Dimensionality reduction enhances computational efficiency. High-dimensional data can be computationally expensive to process and can lead to increased training times for machine learning models.
- By reducing the number of features, the computational burden is significantly lowered, making the analysis more feasible and faster.

Dimensionality Reduction



- Two components of dimensionality reduction:

Feature selection: In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:

  - Filter
  - Wrapper
  - Embedded

Feature extraction: This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Feature Selection

- Feature selection is a technique in machine learning and data preprocessing that involves selecting a subset of relevant features (variables, predictors) from the original set of features in a dataset.
- The main objective of feature selection is to improve the performance of machine learning models by removing irrelevant, redundant, or noisy features.
- This process enhances the efficiency of model training, reduces overfitting, and improves the interpretability of models.

1. Filter method:

- The filter method for feature selection involves evaluating the relevance of each feature independently of any machine learning model.
- This method uses statistical techniques to assess the relationship between each feature and the target variable, selecting features based on their scores from these assessments.
- Common statistical measures used in filter methods include correlation coefficients for continuous data, chi-square tests for categorical data, and mutual information for non-linear relationships.
- By filtering out features that are less relevant or redundant before model training, filter methods improve the efficiency and performance of models without being computationally intensive.

2. Wrapper method:

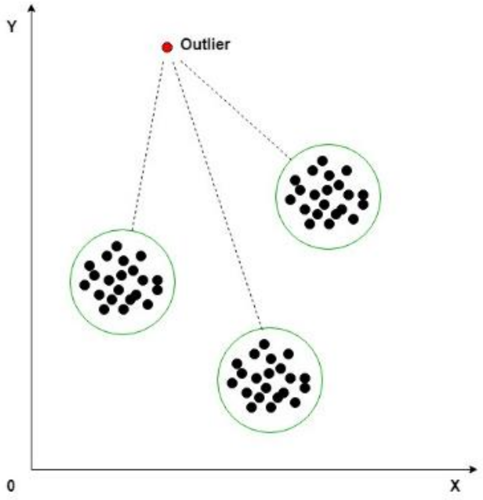
- The wrapper method for feature selection involves using a predictive model to evaluate the combination of features and iteratively adding or removing features based on their impact on model performance.
- Common techniques within wrapper methods include forward selection, where features are added one at a time, backward elimination, where features are removed one at a time, and recursive feature elimination (RFE), which repeatedly builds the model and eliminates the least important features.

3. Embedded method:

- Embedded methods are a type of feature selection technique that integrate the process of feature selection directly into the model training phase.
- This integration allows these methods to consider feature interactions and dependencies more effectively.
- Examples include regularization techniques like Lasso (L1 regularization), which penalizes the absolute size of coefficients, effectively shrinking some to zero and thus performing feature selection.

Outlier Reduction

- Outlier reduction, also known as outlier detection and removal, is a data preprocessing technique used to identify and mitigate the impact of outliers in a dataset.
- Outliers are data points that significantly deviate from the majority of the data and can arise due to measurement errors, data entry mistakes, or genuine variability in the data.
- Techniques for outlier reduction include statistical methods (e.g., Z-score, IQR), clustering methods (e.g., DBSCAN), and machine learning approaches (e.g., isolation forests).
- By detecting and handling outliers, we can ensure a more robust and reliable analysis, leading to more accurate and generalizable models.



Methods of Outlier Reduction

1. Statistical Methods:

- **Z-Score:** This method calculates the standard deviation of the data points and identifies outliers as those with Z-scores exceeding a certain threshold (typically 3 or -3).
- **Interquartile Range (IQR):** IQR identifies outliers as data points falling outside the range defined by Q1-k*(Q3-Q1) and Q3+k*(Q3-Q1), where Q1 and Q3 are the first and third quartiles, and k is a factor (typically 1.5).

2. Distance-Based Methods:

- **K-Nearest Neighbors (KNN):** KNN identifies outliers as data points whose K nearest neighbors are far away from them.
- **Local Outlier Factor (LOF):** This method calculates the local density of data points and identifies outliers as those with significantly lower density compared to their neighbors.

3. Clustering-Based Methods:

- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN):** In DBSCAN, clusters data points based on their density and identifies outliers as points not belonging to any cluster.
- **Hierarchical clustering:** Hierarchical clustering involves building a hierarchy of clusters by iteratively merging or splitting clusters based on their similarity. Outliers can be identified as clusters containing only a single data point or clusters significantly smaller than others.

Importance of outlier detection

1. **Biased models:** Outliers can bias a machine learning model towards the outlier values, leading to poor performance on the rest of the data. This can be particularly problematic for algorithms that are sensitive to outliers, such as linear regression.
2. **Reduced accuracy:** Outliers can introduce noise into the data, making it difficult for a machine learning model to learn the true underlying patterns. This can lead to reduced accuracy and performance.
3. **Reduced interpretability:** Outliers can make it difficult to understand what a machine learning model has learned from the data. This can make it difficult to trust the model's predictions and can hamper efforts to improve its performance.

Statistics Library (R):

- The R Language stands out as a powerful tool in the modern era of statistical computing and data analysis.
- Here are several reasons why professionals across various fields prefer R:

1. Comprehensive Statistical Analysis:

- R language is specifically designed for statistical analysis and provides a vast array of statistical techniques and tests, making it ideal for data-driven research.

2. Extensive Packages and Libraries:

- The R Language boasts a rich ecosystem of packages and libraries that extend its capabilities, allowing users to perform advanced data manipulation, visualization, and machine learning tasks with ease.

3. Strong Data Visualization Capabilities:

- R language excels in data visualization, offering powerful tools like ggplot2 and plotly, which enable the creation of detailed and aesthetically pleasing graphs and plots.

Numpy

- NumPy is a general-purpose array-processing package.
- It provides a high-performance multidimensional array object and tools for working with these arrays.
- It is the fundamental package for scientific computing with Python. It is open-source software.

Some of these important features include:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Arrays in NumPy

- NumPy’s main object is the homogeneous multidimensional array.
- It is a table of elements (usually numbers), all of the same type, indexed by a tuple of positive integers.
- In NumPy, dimensions are called axes. The number of axes is rank.
- NumPy’s array class is called ndarray. It is also known by the alias array.

Pandas

- Pandas is a powerful and open-source Python library.
- The Pandas library is used for data manipulation and analysis.
- Pandas consist of data structures and functions to perform efficient operations on data.
- Pandas is well-suited for working with **tabular data**, such as **spreadsheets** or **SQL tables**.

Here is a list of things that we can do using Pandas.

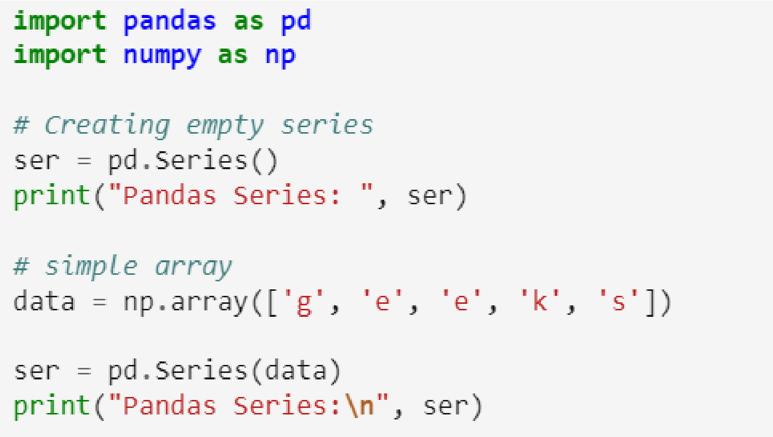
- Data set cleaning, merging, and joining.
- Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data.
- Columns can be inserted and deleted from DataFrame and higher-dimensional objects.
- Powerful group by functionality for performing split-apply-combine operations on data sets.
- Data Visualization.

Pandas generally provide two data structures for manipulating data. They are:

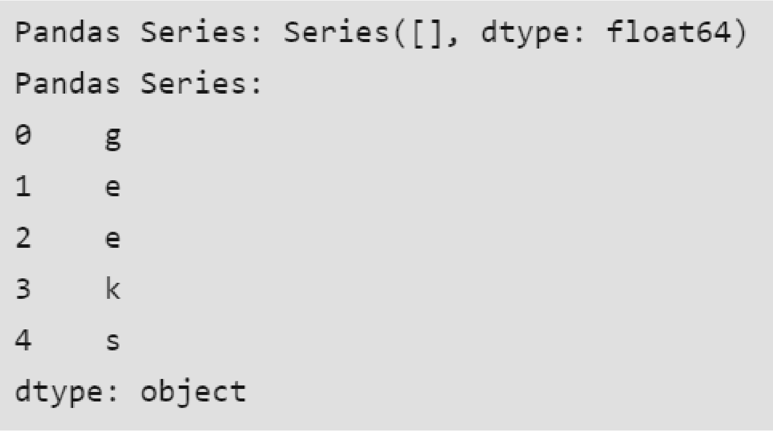
- Series
- DataFrame

Pandas Series

- A Pandas Series is a one-dimensional labeled array capable of holding data of any type (integer, string, float, Python objects, etc.).
- The axis labels are collectively called **indexes**.
- The Pandas Series is nothing but a column in an Excel sheet.
- Pandas Series can be created from lists, dictionaries, scalar values, etc.



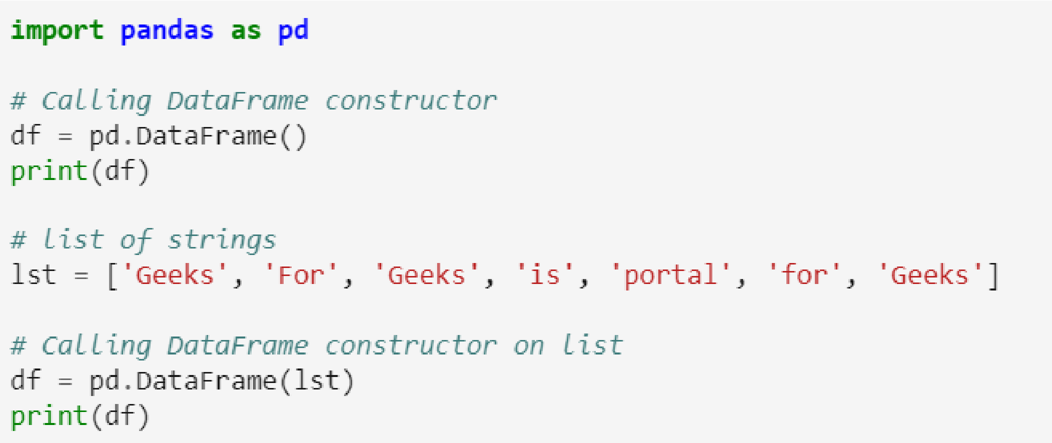
```
import pandas as pd
import numpy as np

# Creating empty series
ser = pd.Series()
print("Pandas Series: ", ser)

# simple array
data = np.array(['g', 'e', 'e', 'k', 's'])

ser = pd.Series(data)
print("Pandas Series:\n", ser)
```



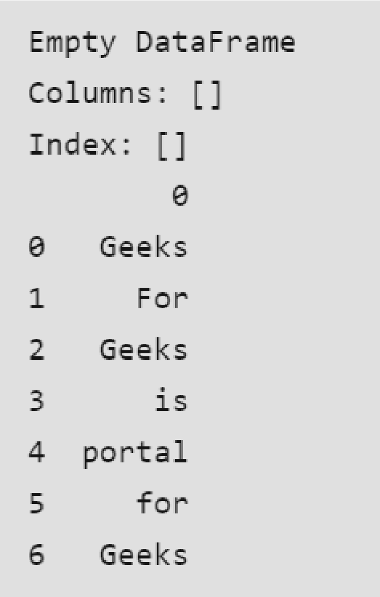
```
Pandas Series: Series([], dtype: float64)
Pandas Series:
0    g
1    e
2    e
3    k
4    s
dtype: object
```

Pandas DataFrame

- Pandas DataFrame is a two-dimensional data structure with labeled axes (rows and columns).
- Pandas DataFrame is created by loading the datasets from existing storage (which can be a SQL database, a CSV file, or an Excel file).
- Pandas DataFrame can be created from lists, dictionaries, a list of dictionaries, etc.



```python
import pandas as pd

# Calling DataFrame constructor
df = pd.DataFrame()
print(df)

# list of strings
lst = ['Geeks', 'For', 'Geeks', 'is', 'portal', 'for', 'Geeks']

# Calling DataFrame constructor on list
df = pd.DataFrame(lst)
print(df)
```



```
Empty DataFrame
Columns: []
Index: []
        0
0   Geeks
1     For
2   Geeks
3      is
4  portal
5     for
6   Geeks
```

Function in Pandas

- .max()
- .min()
- .sum()

e.g. (df['Maths'].sum())

- .count() :- will display the total number of values for each column or row of a DataFrame. To count the rows we need to use the argument axis=1.
- .mean()
- .median()
- .mode()
- .quantile() is used to get the quartiles. It will output the quartile of each column or row of the DataFrame in four parts i.e. the first quartile is 25% (parameter q = .25), the second quartile is 50% (Median), the third quartile is 75% (parameter q = .75).

e.g. df.quantile(q=.75)

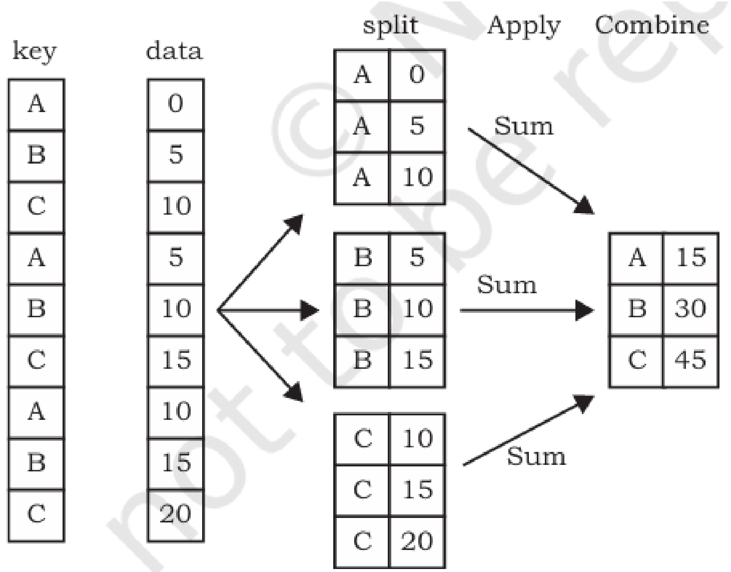
- .var() is used to display the variance. It is the average of squared differences from the mean.
- .std() returns the standard deviation of the values. Standard deviation is calculated as the square root of the variance.
- .describe() function displays the descriptive statistical values in a single command.
- Aggregation means to transform the dataset and produce a single numeric value from an array. Aggregation can be applied to one or more columns together. Aggregate functions are max(),min(), sum(), count(), std(), var().

Functions of Pandas

- Sorting refers to the arrangement of data elements in a specified order, which can either be ascending or descending. Pandas provide sort_values() function to sort the data values of a DataFrame.

E.g. df.sort_values(by=['Name'], axis=0, ascending=True)

- .GROUP BY()



key	data
A	0
B	5
C	10
A	5
B	10
C	15
A	10
B	15
C	20

- For reshaping data, two basic functions are available in Pandas, pivot and pivot_table. The pivot function is used to reshape and create a new DataFrame from the original one.
- Pivot_table syntax:

```
pandas.pivot_table(data, values=None,
index=None, columns=None, aggfunc='mean')
```

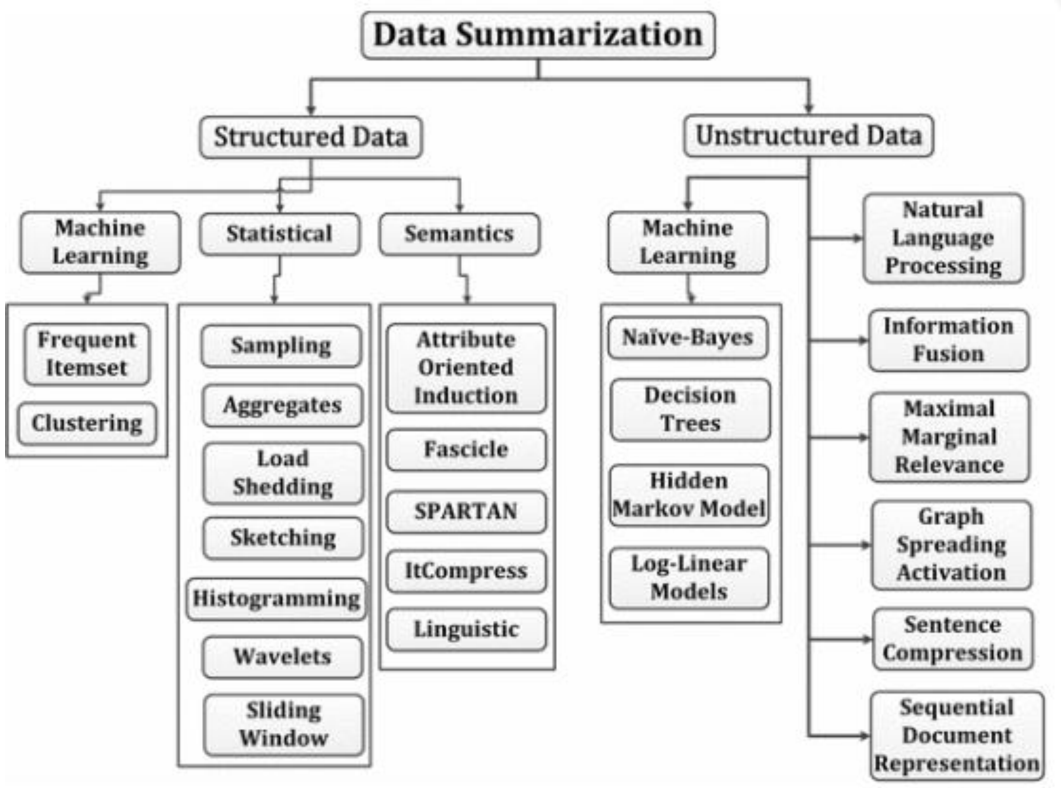
Scipy

- SciPy is a scientific computation library that uses NumPy underneath.
- SciPy stands for Scientific Python.
- It provides more utility functions for optimization, stats and signal processing.
- SciPy has optimized and added functions that are frequently used in NumPy and Data Science.
- It is designed on the top of Numpy library that gives more extension of finding scientific mathematical formulae like Matrix Rank, Inverse, polynomial equations, LU Decomposition, etc. Using its high-level functions will significantly reduce the complexity of the code and helps better in analyzing the data.
- Use descriptive statistics from SciPy’s stats module to gain insights into the dataset.
- Calculate measures such as mean, median, standard deviation, skewness, kurtosis, etc.

Chapter 2

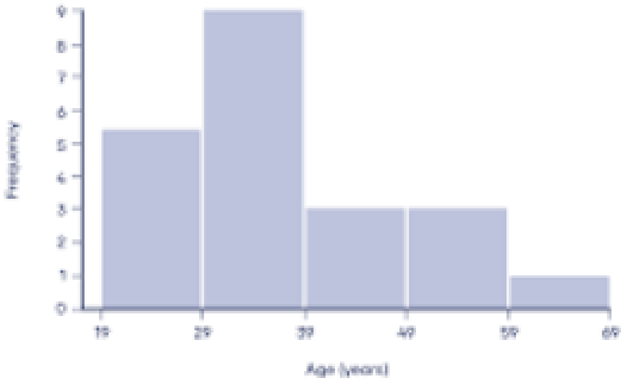
Data Summarization

- Data summarization is the science and art of conveying information more effectively and efficiently. Data summarization is typically numerical, visual or a combination of the two.
- It is a key skill in data analysis - we use it to provide insights both to others and to ourselves. Data summarization is also an integral part of exploratory data analysis.



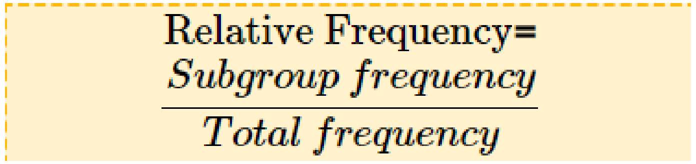
Frequency Distribution

- A frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval.
- The frequency is how often a value occurs in an interval, while the distribution is the pattern of frequency of the variable.
- Frequency distribution is used to organize the collected data in table form. The data could be marks scored by students, temperatures of different towns, points scored in a volleyball match, etc.
- After data collection, we have to show data in a meaningful manner for better understanding. Organize the data in such a way that all its features are summarized in a table. This is known as frequency distribution.

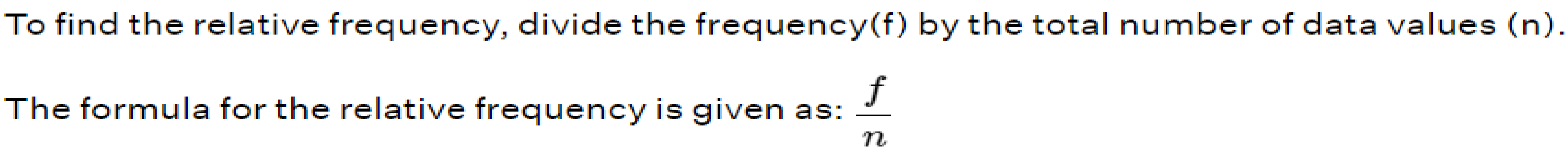


Relative Frequency

- Relative frequency can be defined as the number of times an event occurs divided by the total number of events occurring in a given scenario.



$$\text{Relative Frequency}= \frac{Subgroup\ frequency}{Total\ frequency}$$



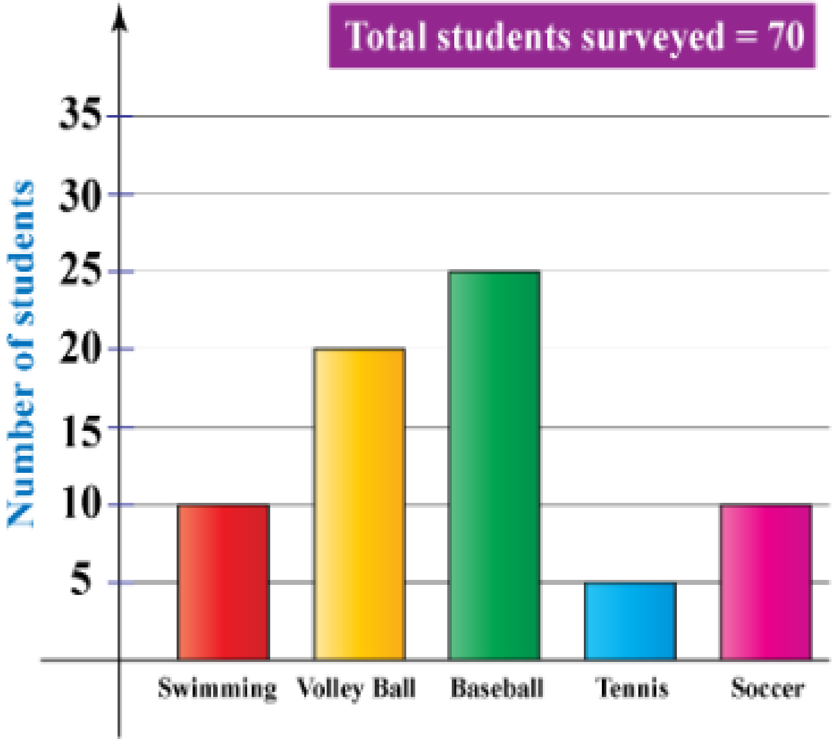
To find the relative frequency, divide the frequency(f) by the total number of data values (n).

The formula for the relative frequency is given as: $\frac{f}{n}$

- A relative frequency distribution is a statistical representation that shows the frequency of each unique value or group of values in a dataset as a proportion of the total number of data points.
- This distribution is particularly useful for understanding the distribution of data across different categories or intervals, especially when comparing datasets of different sizes.

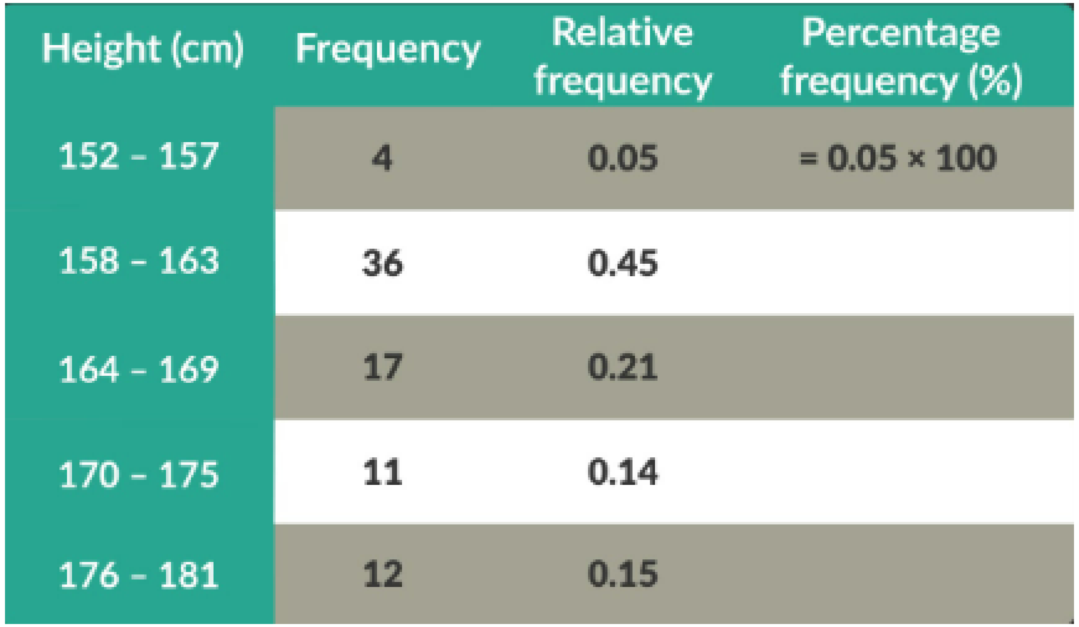
Example:

Ellie surveys a group of students in her school to learn about their favorite sport. The data collected has been presented below. What will be the relative frequency for volleyball?



Percent Frequency

- Percentage Distribution is a frequency distribution in which the individual class frequencies are expressed as a percentage of the total frequency equated to 100. Also known as relative frequency distribution; relative frequency table.



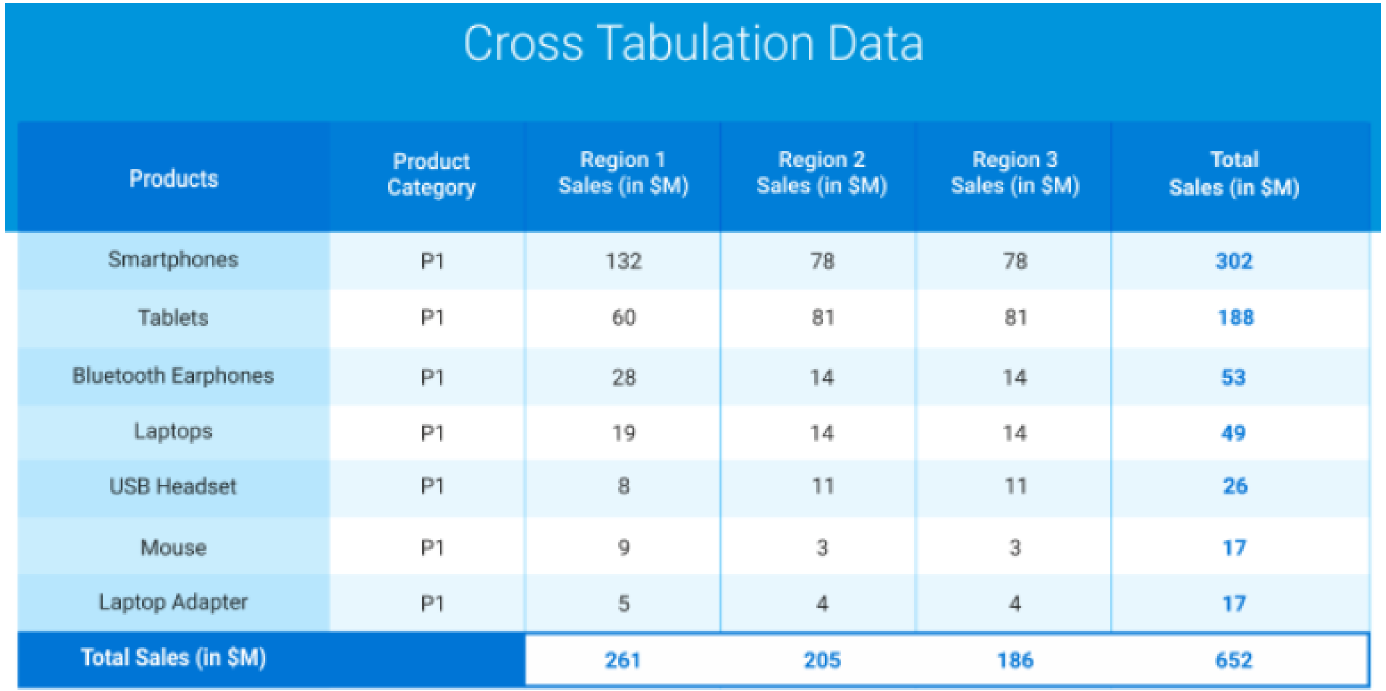
Height (cm)	Frequency	Relative frequency	Percentage frequency (%)
152 – 157	4	0.05	= 0.05 × 100
158 – 163	36	0.45	
164 – 169	17	0.21	
170 – 175	11	0.14	
176 – 181	12	0.15	

Cross - tabulations

- Cross-tabulation is a statistical tool for categorizing data and making sense of it. It involves data values that are mutually exclusive from each other.
- This data is collected in numbers but has no value unless it means something. Like 1, 2, and 3 are mere numbers, but 1 trousers, 2 books, and 3 pencils are meaningful data points.
- Cross-tabulation, or Cross-tabulation analysis, helps you make informed decisions from raw data by identifying patterns, trends, and a correlation between parameters.
- Advantages.

1. Data Simplification: It gives a clear snapshot of how variables relate, making it easier to spot patterns without sifting through raw data.
2. Visualization: Cross-tabs quickly display relationships in a table, and when paired with graphics, like bar graphs, the data becomes even clearer.
3. Testing Ideas: If you have a guess, like “more women prefer this product,” cross-tabs can quickly confirm or refute it.

Cross - tabulations



Cross Tabulation Data

Products	Product Category	Region 1 Sales (in $M)	Region 2 Sales (in $M)	Region 3 Sales (in $M)	Total Sales (in $M)
Smartphones	P1	132	78	78	**302**
Tablets	P1	60	81	81	**188**
Bluetooth Earphones	P1	28	14	14	**53**
Laptops	P1	19	14	14	**49**
USB Headset	P1	8	11	11	**26**
Mouse	P1	9	3	3	**17**
Laptop Adapter	P1	5	4	4	**17**
Total Sales (in $M)		**261**	**205**	**186**	**652**

Graphical Methods:

- **Bar Chart**: Represents categorical data with rectangular bars. The length of each bar is proportional to the value it represents.
- **Scatter Plot**: Displays data points on a 2D plane, showing the relationship between two variables.
- **Line Chart**: Shows trends over time or ordered categories, with data points connected by lines.
- **Area Chart**: Similar to a line chart, but the area under the line is filled, showing cumulative values over time.
- **Pie Chart**: Represents proportions of a whole, with each slice corresponding to a category's percentage.
- **Stem-and-Leaf Display**: Represents numerical data in a textual format, showing the distribution of data.
- **Dot Plot**: Displays individual data points along an axis, often used for small datasets.
- **Histogram**: Displays the frequency distribution of continuous data by dividing it into bins.
- **Cumulative Distribution (Ogive)**: A line graph showing the cumulative frequency or cumulative percentage of data.

Numerical Methods:

- The **mean** (or average) is the sum of all the data points divided by the number of data points. It gives a central value of the dataset.
- The **median** is the middle value of a dataset when it is arranged in ascending or descending order. If the dataset has an odd number of elements, the median is the middle number. If it has an even number of elements, the median is the average of the two middle numbers.
- The **mode** is the value that appears most frequently in a dataset. A dataset can have one mode, more than one mode, or no mode at all.
- A **percentile** indicates the relative standing of a value within a dataset. The nth percentile is the value below which n% of the data falls.
- **Quartiles** divide the dataset into four equal parts:

where, **Q1 (First Quartile)**: 25th percentile (middle of the first half).

Q2 (Second Quartile): 50th percentile (the median).

Q3 (Third Quartile): 75th percentile (middle of the second half).

Numerical Methods:

- Range: The range is the simplest measure of statistical dispersion. It represents the difference between the maximum and minimum values in a data set.

  Range= Maximum value – Minimum value

- Interquatile Range: The Interquartile Range (IQR) measures the spread of the middle 50% of the data. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1).

  IQR = Q3 - Q1

- Variance: It measures the average squared deviation of each data point from the mean. It provides a sense of how much the data varies from the mean.

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- The standard deviation is the square root of the variance. It measures the average distance of each data point from the mean, providing a sense of the spread of the data in the same units as the data itself.

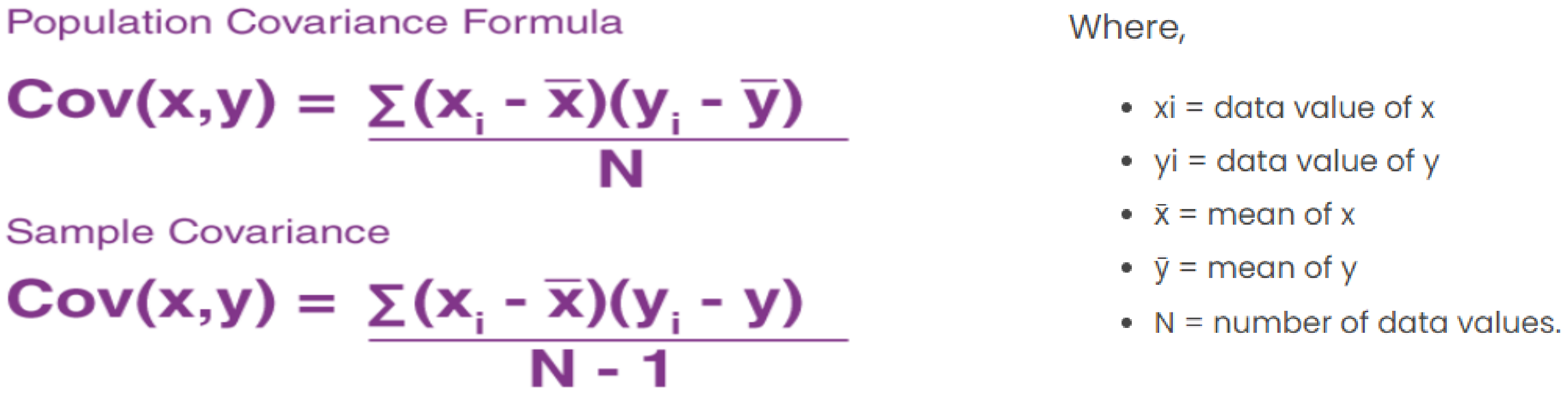
$$\text{Standard Deviation} = \sqrt{Variance}$$

- The coefficient of variation is the ratio of the standard deviation to the mean, expressed as a percentage. It is a normalized measure of dispersion, allowing comparison between data sets with different units or means.

$$\text{CV} = \frac{\sigma}{\mu} \times 100$$

Covariance

- **Covariance** is a statistical measure that indicates the extent to which two random variables change together. It shows the direction of the linear relationship between variables, meaning whether an increase in one variable tends to correspond to an increase (positive covariance) or decrease (negative covariance) in the other.



Population Covariance Formula

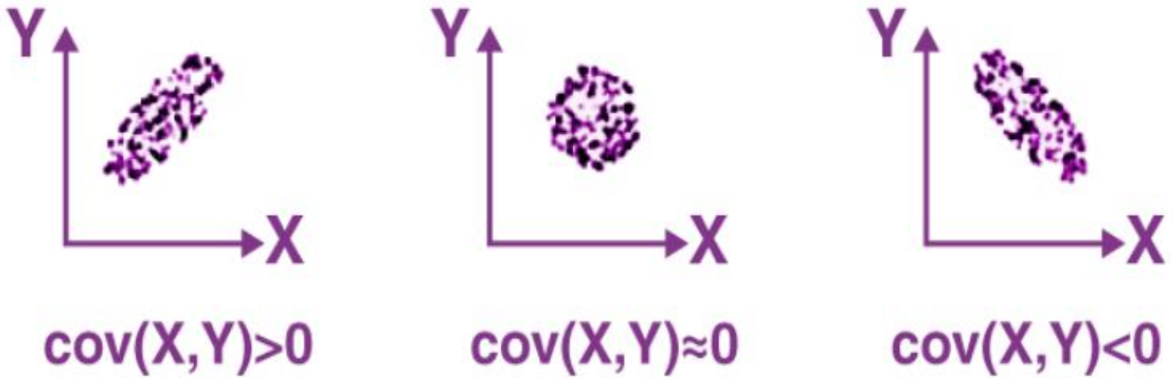
$$\mathrm{Cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$\mathrm{Cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - y)}{N - 1}$$

Where,

- xi = data value of x
- yi = data value of y
- x̄ = mean of x
- ȳ = mean of y
- N = number of data values.



Correlation

- The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is a dimensionless value that ranges from -1 to 1.
- **+1** indicates a perfect positive linear relationship: as one variable increases, the other also increases proportionally.
- **-1** indicates a perfect negative linear relationship: as one variable increases, the other decreases proportionally.
- The most commonly used correlation coefficient is the Pearson correlation coefficient, which measures the linear association between two continuous variables.
- A p-value is a statistical measurement used to validate a hypothesis against observed data.
- If the p-value is below 0.05, this correlation is statistically significant, suggesting that the observed relationship is unlikely to be due to random chance.
- Positive Correlation – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
- Negative Correlation – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
- No Correlation – when there is no linear dependence or no relation between the two variables.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

n = Data quantity or number of data available

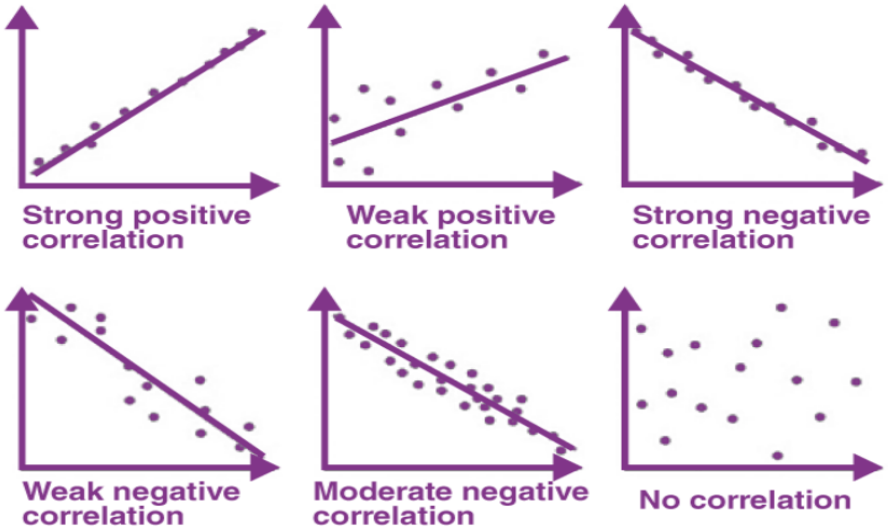
Σx = Total of the First Variable Value

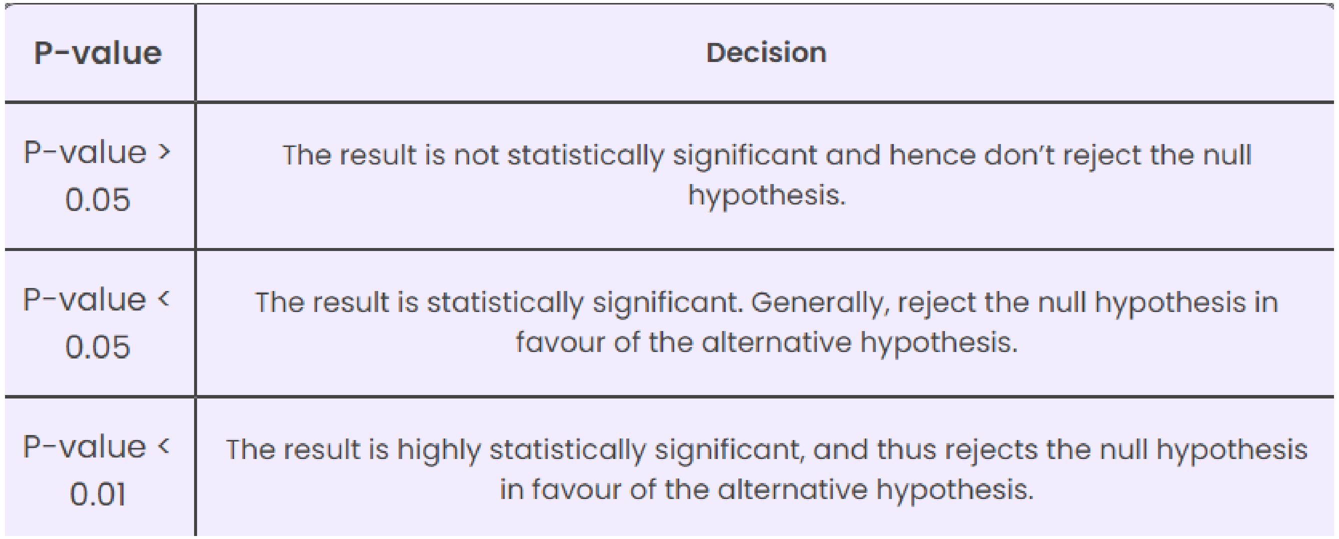
Σy = Total of the Second Variable Value

Σxy = Sum of the Product of First & Second Value

Σx^2 = Sum of the Squares of the First Value

Σy^2 = Sum of the Squares of the Second Value


Strong positive correlation
Weak positive correlation
Strong negative correlation
Weak negative correlation
Moderate negative correlation
No correlation



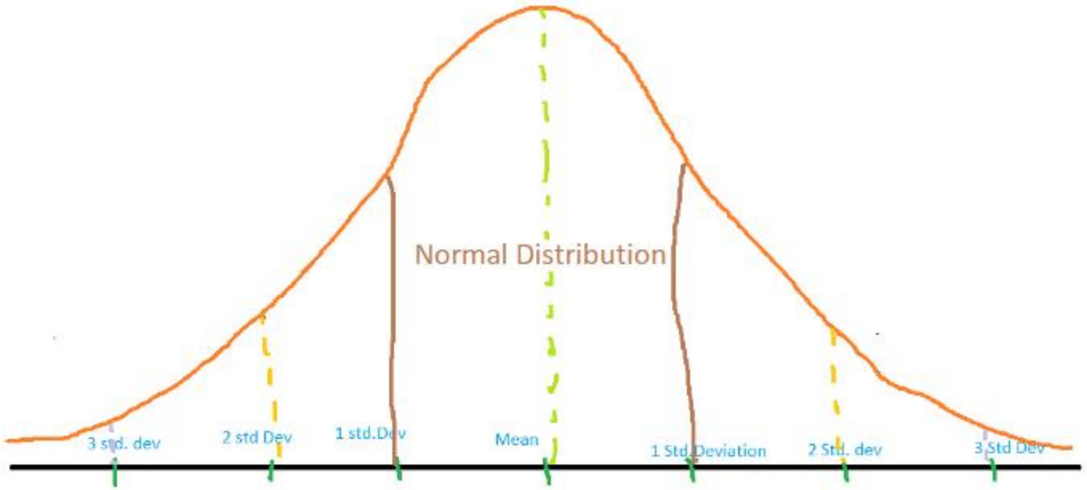
P-value	Decision
P-value > 0.05	The result is not statistically significant and hence don’t reject the null hypothesis.
P-value < 0.05	The result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis.
P-value < 0.01	The result is highly statistically significant, and thus rejects the null hypothesis in favour of the alternative hypothesis.

Detecting Outliers Using Z Scores

- Outliers are one of the most interesting and often challenging topics in statistics. They are data points that lie far away from the other data points in a dataset, and they can have a significant impact on statistical analyses.
- Detecting outliers is crucial in many fields, including finance, healthcare, and social sciences. There are many statistical approaches to detecting outliers, and one of them is using z-scores.
- Z-scores are a measure of how many standard deviations a data point is away from the mean of a dataset. A z-score of 3 means that the data point is three standard deviations above the mean, while a z-score of -2 means that the data point is two standard deviations below the mean. In general, a z-score of 2 or more is considered an outlier.
- The advantage of using z-scores to detect outliers is that they are independent of the scale of the dataset. This means that z-scores can be used to compare data points from different datasets that have different units of measurement.
- Z-Scores, also known as standard scores, measure the distance between a data point and the mean of the data set in terms of standard deviations. The resulting value is an indication of how many standard deviations a data point is away from the mean.

$$Z = \frac{(X-\mu)}{\sigma}$$

where X is the data point, μ is the mean, and σ is the standard deviation.


Normal Distribution
3 std. dev
2 std Dev
1 std.Dev
Mean
1 Std Deviation
2 Std. dev
3 Std Dev

Regression model

1. Simple linear regression model.

- **Simple Linear Regression** is a statistical method used to model the relationship between two continuous variables: one independent variable (predictor) and one dependent variable (response).
- The goal is to fit a straight line (linear equation) through the data points that best predicts the dependent variable from the independent variable. The general form of the simple linear regression equation is:

$$y=\beta0+\beta1x+\epsilon$$

- **y**: Dependent variable (response).
- **x**: Independent variable (predictor).
- **β0**: Intercept (value of y when x=0).
- **β1**: Slope (change in y for a one-unit change in x).
- **ϵ**: Error term (captures the randomness or noise not explained by the model).

I. Intercept (β0): Represents the point where the regression line crosses the y-axis. It is the expected value of y when x is zero.

II. Slope (β1): Indicates the rate of change in the dependent variable for each unit change in the independent variable. A positive slope indicates a positive relationship, while a negative slope indicates a negative relationship.

III. Residuals: The differences between the observed values and the predicted values from the model. Residuals are used to measure the accuracy of the model.

IV. **Assumptions**:

- The relationship between the independent and dependent variables is linear.
- The residuals are normally distributed.
- Independence: Observations are independent of each other.

2. Multiple regression model.

- **Multiple Linear Regression (MLR)** is an extension of simple linear regression that models the relationship between a dependent variable and two or more independent variables.
- It is used to understand how the dependent variable (response) changes as the independent variables (predictors) change.
- The primary goal of MLR is to determine the best-fitting plane or hyperplane that predicts the dependent variable from multiple predictors.
- The general equation for a multiple linear regression model is:

$$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\epsilon$$

where,

y: Dependent variable (response).

x1,x2,…,xn: Independent variables (predictors).

β0: Intercept (value of y when all x values are zero).

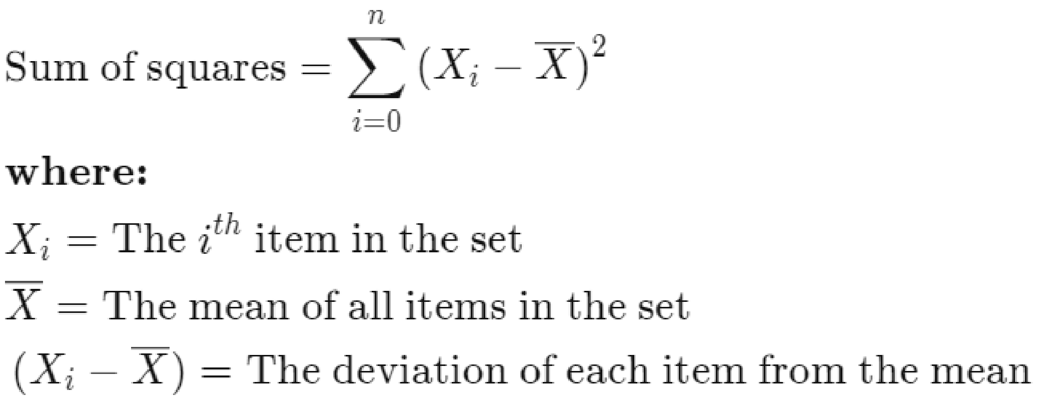
β1,β2,…,βn: Coefficients representing the change in y for a one-unit change in each x.

ϵ: Error term (random noise not explained by the model).

1. **Coefficients (β)**: Each coefficient measures the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.
2. **Intercept (β0)**: Represents the expected value of the dependent variable when all predictors are zero.
3. **Residuals**: The differences between observed and predicted values of the dependent variable. Analysis of residuals helps in validating model assumptions.

Least Squares Method

- The least squares method is a form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points.
- Each point of data represents the relationship between a known independent variable and an unknown dependent variable. This method is commonly used by statisticians and traders who want to identify trading opportunities and trends.
- The sum of squares measures the deviation of data points away from the mean value.
- A higher sum of squares indicates higher variability while a lower result indicates low variability from the mean.
- To calculate the sum of squares, subtract the mean from the data points, square the differences, and add them together.
- There are three types of sum of squares: total, residual, and regression.



$$\text{Sum of squares} = \sum_{i=0}^{n} (X_i - \overline{X})^2$$

where:

$X_i =$ The i^{th} item in the set

$\overline{X} =$ The mean of all items in the set

$(X_i - \overline{X}) =$ The deviation of each item from the mean

Residual sum of squares:

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where:

y_i = Observed value

$\hat{y}_i$ = Value estimated by regression line

Regression sum of squares:

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

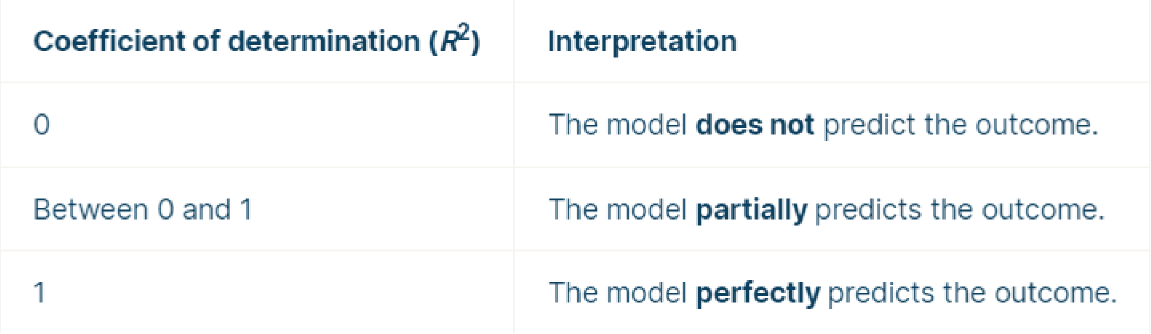
where:

$\hat{y}_i$ = Value estimated by regression line

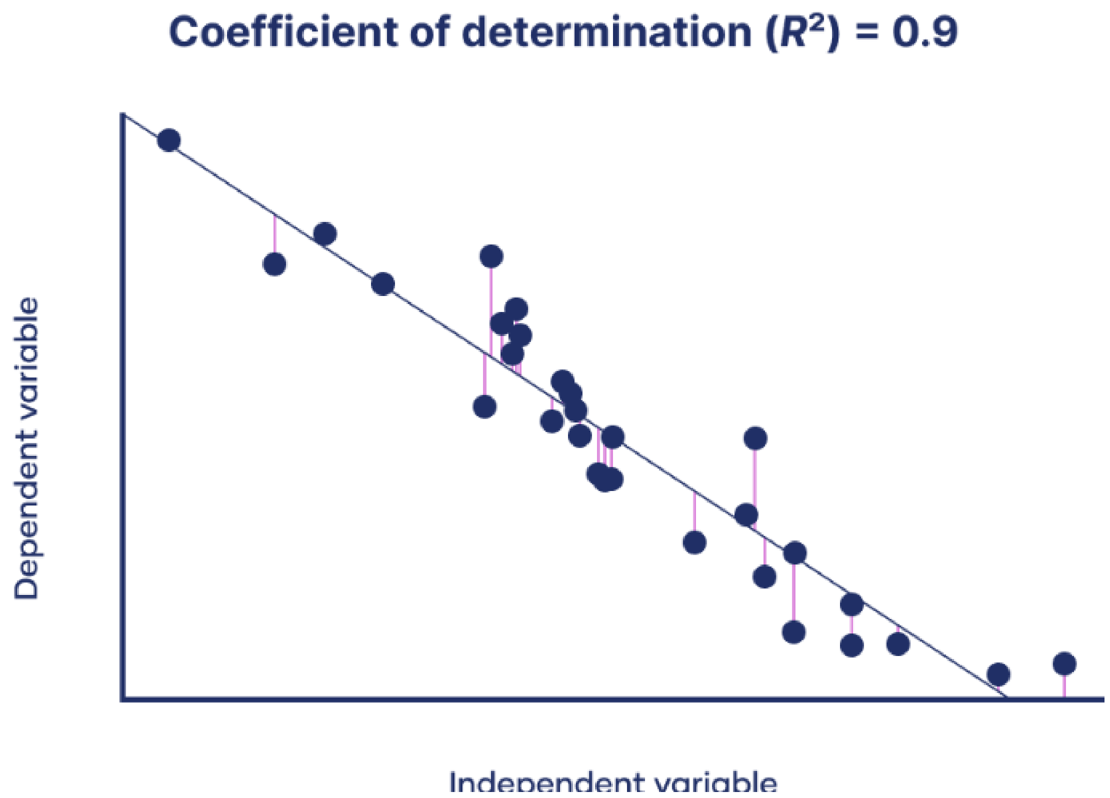
$\bar{y}$ = Mean value of a sample

Coefficient of determination

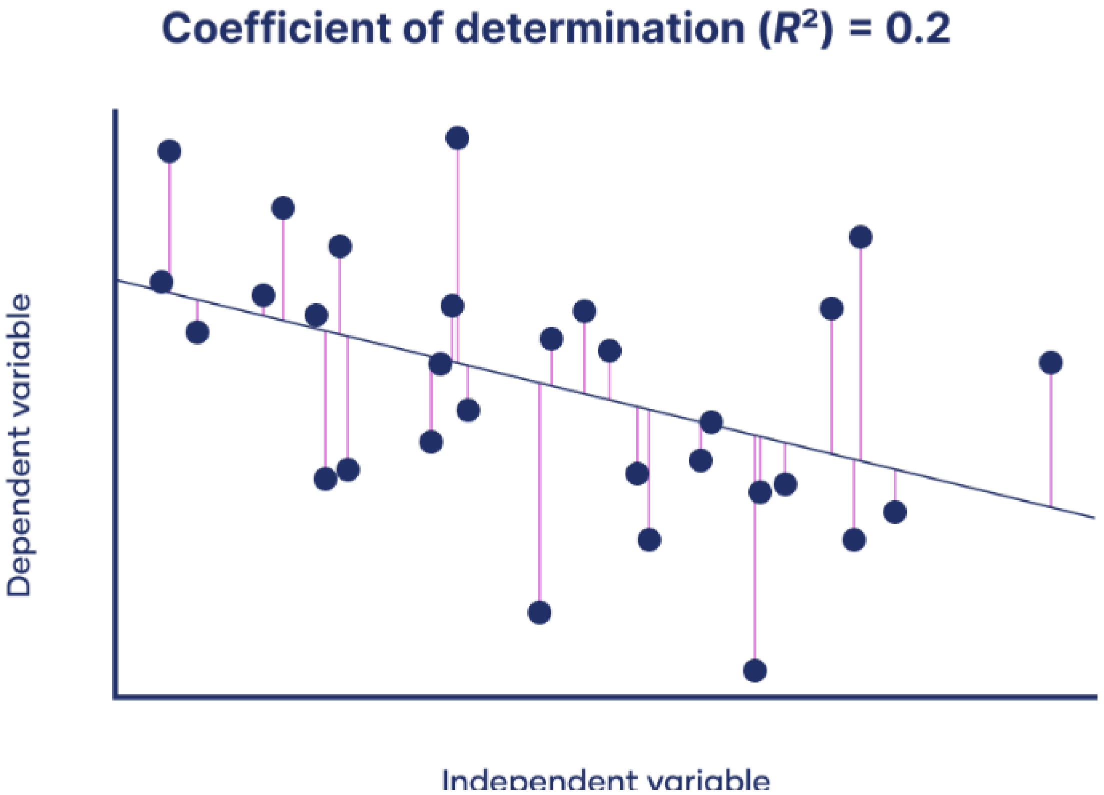
- The coefficient of determination (R2) is a statistical measurement that shows how well a model predicts outcomes based on a given set of variables.
- It's a number between 0 and 1 that can be interpreted as the proportion of variance in the dependent variable that's explained by the model.



Coefficient of determination (R^2)	**Interpretation**
0	The model **does not** predict the outcome.
Between 0 and 1	The model **partially** predicts the outcome.
1	The model **perfectly** predicts the outcome.



- The distance between the observations and their predicted values (the residuals) are shown as purple lines.
- The coefficient of determination is always positive, even when the correlation is negative.


Coefficient of determination (R^2) = 0.2
Dependent variable
Independent variable

- Formula 1: Using the correlation coefficient

$$R^2 = (r)^2$$

where, r = pearson correlation coefficient

Example: r = 0.28

Solution: R^2 = 0.08

- Formula 2: Using the regression outputs

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

Example: RSS = 629.22   TSS = 2187.04

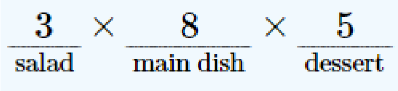
Solution: 0.71

Counting Rules

- There are times when the sample space is very large and is not feasible to write out. In that case, it helps to have mathematical tools for counting the size of the sample space. These tools are known as counting techniques or counting rules.
- **Fundamental Counting Rule:** If event 1 can be done m_1 ways, event 2 can be done m_2 ways, and so forth to event n being done m_n ways, then the number of ways to do event 1, followed by event 2,…, followed by event n together would be to multiply the number of ways for each event: $m_1 \cdot m_2 \cdots m_n$.

Example 1: A menu offers a choice of 3 salads, 8 main dishes, and 5 desserts. How many different meals consisting of one salad, one main dish, and one dessert are possible?

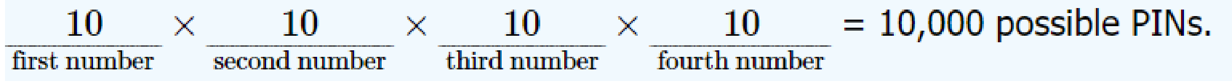
Solution: There are three events: choosing a salad, a main dish, and a dessert. There are 3 choices for salad, 8 choices for the main dish, and 5 choices for dessert. The ways to choose a salad, main dish, and dessert are:



$\underset{\text{salad}}{\underline{3}} \times \underset{\text{main dish}}{\underline{8}} \times \underset{\text{dessert}}{\underline{5}}$ =120 different ways

Example 2: How many 4-digit debit card personal identification numbers (PIN) can be made?

Solution: There are four events in this example. The events are picking the first number, then the second number, then the third number, and then the fourth number. The first event can be done 10 ways since the choices are the numbers 1 through 9 and 0. We can use the same numbers over again (repeats are allowed) for the second number, so it can also be done 10 ways. The same with the third and fourth numbers, which also have 10 choices.



$\underset{\text{first number}}{\underline{10}} \times \underset{\text{second number}}{\underline{10}} \times \underset{\text{third number}}{\underline{10}} \times \underset{\text{fourth number}}{\underline{10}}$ = 10,000 possible PINs.

Question 3: How many ways can the three letters *a*, *b*, and *c* be arranged with no letters repeating?

Answer: 6 ways

Factorials, Permutations and Combinations

- If we have 10 different letters for, say, a password, the tree diagram would be very time-consuming to make because of the length of options and tasks, so we have some shortcut formulas that help count these arrangements.
- Many counting problems involve multiplying a list of decreasing numbers, which is called a **factorial**. The factorial is represented mathematically by the starting number followed by an exclamation point, in this case $3! = 3\cdot2\cdot1 = 6$. There is a special symbol for this and a special button on your calculator for the factorial.
- **Factorial Rule:** The number of different ways to arrange *n* distinct objects is

$n!=n.(n-1).(n-2)\ldots.3.2.1$, where repetitions not allowed

0 factorial is defined to be $0! = 1$ and 1 factorial is defined to be $1! = 1$

Example 1: How many ways can you arrange 5 people standing in line?

Solution: No repeats are allowed since you cannot reuse a person twice in the line. Order is important since the first person is first in line and will be selected first. This meets the requirements for the factorial rule. There are $5! = 5\cdot4\cdot3\cdot2\cdot1 = 120$ ways to arrange 5 people standing in line.

- Sometimes we do not want to select the entire group but only select *r* objects from *n* total objects. The number of ways to do this depends on if the order you choose the *r* objects matters or if it does not matter.
- As an example, if you are trying to call a person on the phone, you have to have the digits of their number in the correct order. In this case, the order of the numbers matters. If you were picking random numbers for the lottery, it does not matter which number you pick first since they always arrange the numbers from the smallest to largest once the numbers are drawn. As long as you have the same numbers that the lottery officials pick, you win. In this case, the order does not matter.
- A **permutation** is an arrangement of items with a specific order. You use permutations to count items when the order matters.
- **Permutation Rule:** The number of different ways of picking *r* objects from *n* distinct total objects when repeats are not allowed and order matters is:

$$_nP_r = P(n,r) = \frac{n!}{(n-r)!}.$$

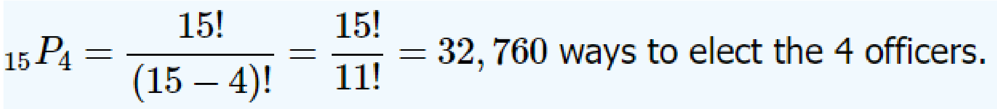
- When the order does not matter, you use combinations. A **combination** is an arrangement of items when order is not important.
- When you do a counting problem, the first thing you should ask yourself is “Are repeats allowed?”, then ask yourself “Does order matter?”
- **Combination Rule:** The number of ways to select *r* objects from *n* distinct total objects when repeats are not allowed and order does not matter is:

$$_nC_r = C(n,r) = \frac{n!}{r!(n-r)!}.$$

Example 1: Circle K International, a college community service club, has 15 members this year. How many ways can a board of officers consisting of a president, vice-president, secretary and treasurer be elected?

Solution: In this case, repeats are not allowed since the same member cannot hold more than one position. The order matters because if you elect person 1 for president and person 2 for vice-president, there would be different members in those positions than if you elect person 2 for president, person 1 for vice-president. Thus, this is a permutation problem with *n* = 15 and *r* = 4.

There are

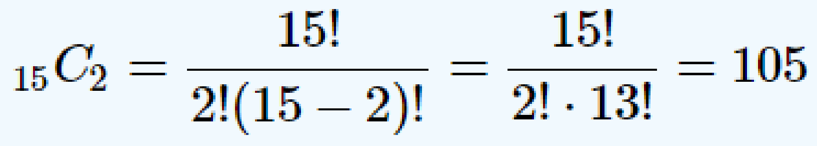


$${}_{15}P_4 = \frac{15!}{(15-4)!} = \frac{15!}{11!} = 32,760 \text{ ways to elect the 4 officers.}$$

Example 2: Circle K International, a college community service club, has 15 members this year. They need to select 2 members to be representatives for the school's Inter-Club Council. How many ways can the 2 members be chosen?

Solution: In this case, repeats are not allowed, since there must be 2 different members as representatives. The order in which the representatives are selected does not matter since they have the same position. Thus, this is a combination problem with n = 15 and r = 2.

There are



$${}_{15}C_2 = \frac{15!}{2!(15-2)!} = \frac{15!}{2! \cdot 13!} = 105$$

ways to select 2 representatives.

Events in probability

- Events in probability can be defined as a set of outcomes of a random experiment.
- The sample space indicates all possible outcomes of an experiment.
- Thus, events in probability can also be described as subsets of the sample space.
- There are many different types of events in probability.
- Each type of event has its own individual properties.
- This classification of events in probability helps to simplify mathematical calculations.

What are Event in Probability?

- Events in probability are outcomes of random experiments. Any subset of the sample space will form events in probability.
- The likelihood of occurrence of events in probability can be calculated by dividing the number of favorable outcomes by the total number of outcomes of that experiment.

Definitions:

- Events in probability can be defined as certain likely outcomes of an experiment that form a subset of a finite sample space. The probability of occurrence of any event will always lie between 0 and 1. There could be many events associated with one sample space.

Example:

- Suppose a fair die is rolled. The total number of possible outcomes will form the sample space and are given by {1, 2, 3, 4, 5, 6}. Let an event, E, be defined as getting an even number on the die. Then E = {2, 4, 6}. Thus, it can be seen that E is a subset of the sample space and is an outcome of the rolling of a die.

Types of Event in Probability

- There are several different types of events in probability. There can only be one sample space for a random experiment however, there can be many different types of events. Some of the important events in probability are listed below.

Independent and Dependent Events:

- Independent events in probability are those events whose outcome does not depend on some previous outcome. No matter how many times an experiment has been conducted the probability of occurrence of independent events will be the same. For example, tossing a coin is an independent event in probability.
- Dependent events in probability are events whose outcome depends on a previous outcome. This implies that the probability of occurrence of a dependent event will be affected by some previous outcome. For example, drawing two balls one after another from a bag without replacement.

Impossible and Sure Events:

- An event that can never happen is known as an impossible event. As impossible events in probability will never take place thus, the chance that they will occur is always 0. For example, the sun revolving around the earth is an impossible event.
- A sure event is one that will always happen. The probability of occurrence of a sure event will always be 1. For example, the earth revolving around the sun is a sure event.

Simple and Compound events:

- If an event consists of a single point or a single result from the sample space, it is termed a simple event. The event of getting less than 2 on rolling a fair die, denoted as E = {1}, is an example of a simple event.
- If an event consists of more than a single result from the sample space, it is called a compound event. An example of a compound event in probability is rolling a fair die and getting an odd number. E = {1, 3, 5}.

Complementary event:

- When there are two events such that one event can occur if and only if the other does not take place, then such events are known as complementary events in probability. The sum of the probability of complementary events will always be equal to 1. For example, on tossing a coin let E be defined as getting a head. Then the complement of E is E' which will be the event of getting a tail. Thus, E and E' together make up complementary events. Such events are mutually exclusive and exhaustive.

Mutually Exclusive events:

- Events that cannot occur at the same time are known as mutually exclusive events. Thus, mutually exclusive events in probability do not have any common outcomes. For example, S = {10, 9, 8, 7, 6, 5, 4}, A = {4, 6, 7} and B = {10, 9, 8}. As there is nothing common between sets A and B thus, they are mutually exclusive events.

Exhaustive events:

- Exhaustive events in probability are those events when taken together from the sample space of a random experiment. In other words, a set of events out of which at least one is sure to occur when the experiment is performed are exhaustive events. For example, the outcome of an exam is either passing or failing.

Example 1: A random card is drawn from a deck of 52 cards. What is the probability that it is an ace?

Solution: E = event of drawing an ace.
Total number of outcomes = 52
The favorable number of outcomes = 4 (there are 4 ace cards in a deck of cards. One belonging to each suit).
P(E) = 4 / 52 = 1 / 13

Example 2: On rolling a fair dice, A is the event of getting a number less than 5, B is the event of getting an odd number and C is the event of getting a multiple of 3. Find the AND event.

Solution: Sample space of a dice roll = {1, 2, 3, 4, 5, 6}
A = {1, 2, 3, 4}
B = {1, 3, 5}
C = {3, 6}
$A \cap B \cap C = \{3\}$

Example 3: If a coin is tossed 3 times what would be the event of getting at most two heads?

Solution: The sample space for tossing coin thrice is {(H, H, H), (T, H, H), (H, T, H), (H, H, T), (T, T, H), (H, T, T), (T, H, T), (T, T, T)}
E = {(T, H, H), (H, T, H), (H, H, T), (T, T, H), (H, T, T), (T, H, T), (T, T, T)}

Example 4: In the game of snakes and ladders, a fair die is thrown. If event E_1 represents all the events of getting a natural number less than 4, event E_2 consists of all the events of getting an even number and E_3 denotes all the events of getting an odd number. List the sets representing the following:

i)E_1 or E_2 or E_3

ii)E_1 and E_2 and E_3

iii)E_1 but not E_3

Solution: The sample space is given as S = {1 , 2 , 3 , 4 , 5 , 6}

$E_1 = \{1,2,3\}$

$E_2 = \{2,4,6\}$

$E_3 = \{1,3,5\}$

i)E_1 or E_2 or E_3= E_1 E_2 E_3= {1, 2, 3, 4, 5, 6}

ii)E_1 and E_2 and E_3 = E_1 E_2 E_3 = ∅

iii)E_1 but not E_3 = {2}

Example 5: A die is rolled, let’s define two events, event A is getting the number 2 and Event B is getting an even number. Are these events mutually exclusive?

Solution: Sample space for die roll will be,

S = {1, 2, 3, 4, 5, 6}

For the event A,

A = {2}

For the event B,

B = {2, 4, 6}

For two events to be mutually exclusive, their intersection must be an empty set

A ∩ B = {2} ∩ {2, 4, 6}

⇒ A ∩ B = {2}

- Since it is not an empty set, these events are not mutually exclusive.

Binomial Distribution

Example: A fair coin is tossed 10 times, what are the probability of getting exactly 6 heads and at least six heads.

Solution: Let x denote the number of heads in an experiment.

Here, the number of times the coin tossed is 10. Hence, n=10.

The probability of getting head, p ½

The probability of getting a tail, q = 1-p = 1-(½) = ½.

The binomial distribution is given by the formula:

$P(X= x) = {}^{n}C_{x}p^{x}q^{n-x}$, where = 0, 1, 2, 3, …

Therefore, $P(X = x) = {}^{10}C_{x}(½)^{x}(½)^{10-x}$

(i) The probability of getting exactly 6 heads is:

$P(X=6) = {}^{10}C_{6}(½)^{6}(½)^{10-6}$

$P(X= 6) = {}^{10}C_{6}(½)^{10}$

$P(X = 6) = 105/512.$

Hence, the probability of getting exactly 6 heads is 105/512.

(ii) The probability of getting at least 6 heads is $P(X \geq 6)$

$P(X \geq 6) = P(X=6) + P(X=7) + P(X= 8) + P(X = 9) + P(X=10)$

$P(X \geq 6) = {}^{10}C_{6}(½)^{10} + {}^{10}C_{7}(½)^{10} + {}^{10}C_{8}(½)^{10} + {}^{10}C_{9}(½)^{10} + {}^{10}C_{10}(½)^{10}$

$P(X \geq 6) = 193/512.$

Poisson distribution

- It is used for calculating the possibilities for an event with the average rate of value. Poisson distribution is a discrete probability distribution.
- The Poisson distribution is a discrete probability function that means the variable can only take specific values in a given list of numbers, probably infinite. A Poisson distribution measures how many times an event is likely to occur within "x" period of time.
- A Poisson random variable "x" defines the number of successes in the experiment. This distribution occurs when there are events that do not occur as the outcomes of a definite number of outcomes. Poisson distribution is used under certain conditions. They are:

  The number of trials "n" tends to infinity

  Probability of success "p" tends to zero

  np = 1 is finite
- The formula for the Poisson distribution function is given by:

  $$\mathbf{f(x) = (e^{-\lambda} \lambda^{x})/x!}$$

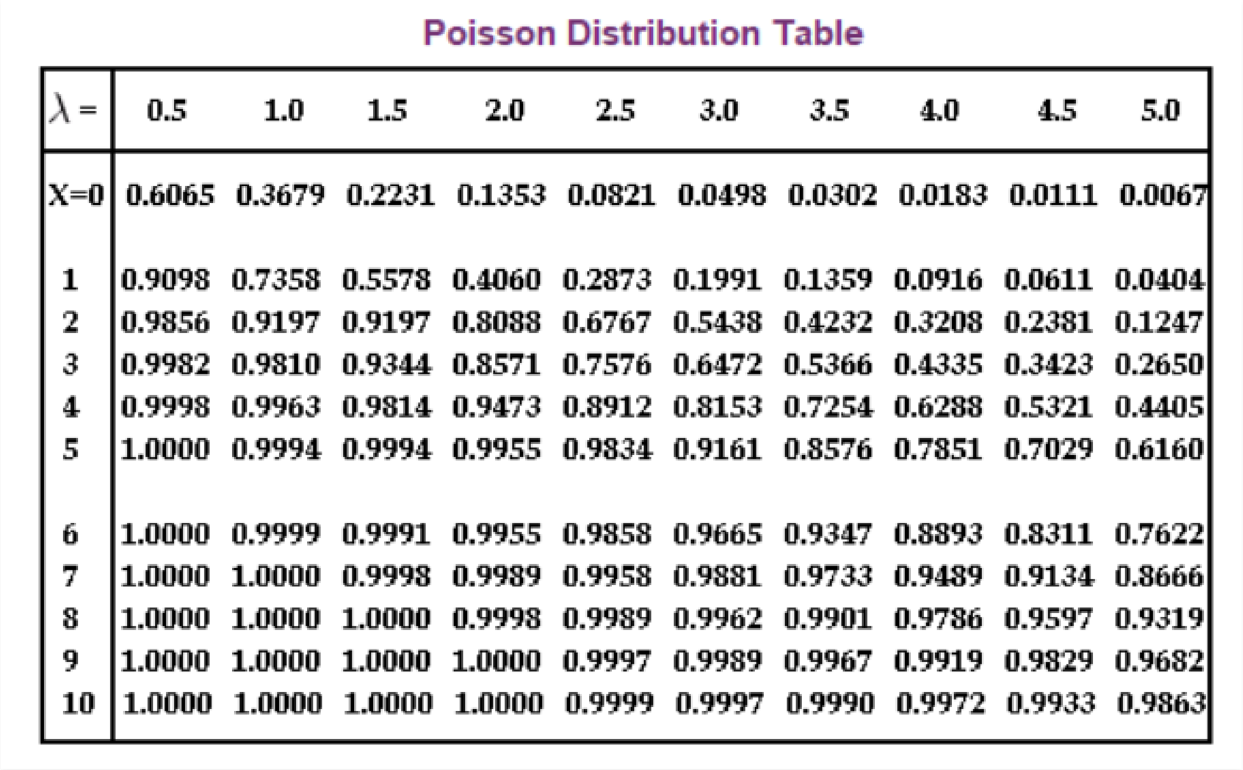
  Where,

  e is the base of the logarithm

  x is a Poisson random variable

  λ is an average rate of value

- The table is showing the values of f(x) = P(X ≥ x), where X has a Poisson distribution with parameter λ. Refer the values from the table and substitute it in the Poisson distribution formula to get the probability value.



Poisson Distribution Table

λ =	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
X=0	0.6065	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067
1	0.9098	0.7358	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404
2	0.9856	0.9197	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1247
3	0.9982	0.9810	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650
4	0.9998	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405
5	1.0000	0.9994	0.9994	0.9955	0.9834	0.9161	0.8576	0.7851	0.7029	0.6160
6	1.0000	0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622
7	1.0000	1.0000	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666
8	1.0000	1.0000	1.0000	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319
9	1.0000	1.0000	1.0000	1.0000	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682
10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9972	0.9933	0.9863

Example 1: A random variable X has a Poisson distribution with parameter λ such that P (X = 1) = (0.2) P (X = 2). Find P (X = 0).

Solution: For the Poisson distribution, the probability function is defined as:

$P (X =x) = (e^{-\lambda} \lambda^{x})/x!$, where λ is a parameter.

Given that, $P (x = 1) = (0.2) P (X = 2)$

$(e^{-\lambda} \lambda^{1})/1! = (0.2)(e^{-\lambda} \lambda^{2})/2!$

$\Rightarrow \lambda = \lambda^{2}/ 10$

$\Rightarrow \lambda = 10$

Now, substitute λ = 10, in the formula, we get:

$P (X =0) = (e^{-\lambda} \lambda^{0})/0!$

$P (X =0) = e^{-10} = 0.0000454$

Thus, P (X= 0) = 0.0000454

Example 2: Telephone calls arrive at an exchange according to the Poisson process at a rate λ= 2/min. Calculate the probability that exactly two calls will be received during each of the first 5 minutes of the hour.

Solution: Assume that "N" be the number of calls received during a 1 minute period.

Therefore,

$P(N= 2) = (e^{-2}. 2^2)/2!$

$P(N=2) = 2e^{-2}$.

Now, "M" be the number of minutes among 5 minutes considered, during which exactly 2 calls will be received. Thus "M" follows a binomial distribution with parameters n=5 and p= $2e^{-2}$.

$P(M=5) = 32 \text{ x } e^{-10}$

$P(M =5) = 0.00145$, where "e" is a constant, which is approximately equal to 2.718.

Normal Distribution

- The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

  $f(x) \geq 0 \ \forall \ x \in (-\infty,+\infty)$

  And ${}_{-\infty}\int^{+\infty} f(x) = 1$

- The probability density function of normal or gaussian distribution is given by;

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where,

x is the variable

μ is the mean

σ is the standard deviation

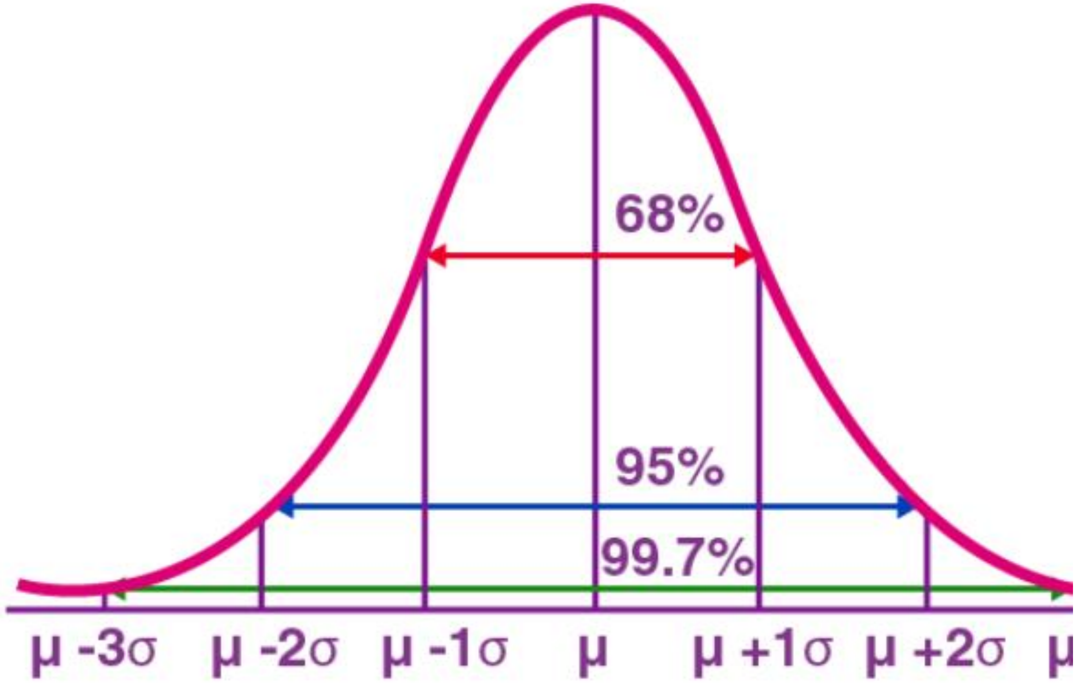
Normal Distribution Curve

- The random variables following the normal distribution are those whose values can find any unknown value in a given range.
- For example, finding the height of the students in the school. Here, the distribution can consider any value, but it will be bounded in the range say, 0 to 6ft.
- Whereas, the normal distribution doesn’t even bother about the range. The range can also extend to –∞ to + ∞ and still we can find a smooth curve.
- These random variables are called Continuous Variables, and the Normal Distribution then provides here probability of the value lying in a particular range for a given experiment.

Normal distribution standard deviation

- Generally, the normal distribution has any positive standard deviation.
- We know that the mean helps to determine the line of symmetry of a graph, whereas the standard deviation helps to know how far the data are spread out.
- If the standard deviation is smaller, the data are somewhat close to each other and the graph becomes narrower.
- If the standard deviation is larger, the data are dispersed more, and the graph becomes wider. The standard deviations are used to subdivide the area under the normal curve.
- Each subdivided section defines the percentage of data, which falls into the specific region of a graph.

- Using 1 standard deviation, the Empirical Rule states that,
- Approximately 68% of the data falls within one standard deviation of the mean. (i.e., Between Mean- one Standard Deviation and Mean + one standard deviation)
- Approximately 95% of the data falls within two standard deviations of the mean. (i.e., Between Mean- two Standard Deviation and Mean + two standard deviations)
- Approximately 99.7% of the data fall within three standard deviations of the mean. (i.e., Between Mean- three Standard Deviation and Mean + three standard deviations)


68%
95%
99.7%
μ -3σ
μ -2σ
μ -1σ
μ
μ +1σ
μ +2σ
μ

Question 1: Calculate the probability density function of normal distribution using the following data. x = 3, μ = 4 and σ = 2.

Solution: Given, variable, x = 3

Mean = 4 and

Standard deviation = 2

By the formula of the probability density of normal distribution, we can write;

$$f(3,4,2) = \frac{1}{2\sqrt{2\pi}} e^{\frac{-(3-2)^2}{2\times 2^2}}$$

Hence, f(3,4,2) = 1.106.

Question 2: If the value of random variable is 2, mean is 5 and the standard deviation is 4, then find the probability density function of the gaussian distribution.

Solution: Given,

Variable, x = 2

Mean = 5 and

Standard deviation = 4

By the formula of the probability density of normal distribution, we can write;

$$f(2,2,4) = \frac{1}{4\sqrt{2\pi}} e^{\frac{-(2-2)^2}{2\times 4^2}}$$

$f(2,2,4) = 1/(4\sqrt{2}\pi)\ e^0$

$f(2,2,4) = 0.0997$

- There are two main parameters of normal distribution in statistics namely mean and standard deviation. The location and scale parameters of the given normal distribution can be estimated using these two parameters.

Normal distribution properties

- In a normal distribution, the mean, median and mode are equal.(i.e., Mean = Median= Mode).
- The total area under the curve should be equal to 1.
- The normally distributed curve should be symmetric at the center.
- There should be exactly half of the values are to the right of the center and exactly half of the values are to the left of the center.
- The normal distribution should be defined by the mean and standard deviation.
- The normal distribution curve must have only one peak. (i.e., Unimodal)
- The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.