**DATA**

Data refers to raw facts, information, or values that are collected, stored, and processed. It can take various forms, such as numbers, text, images, audio, and more. Data becomes meaningful when it is organized, analysed, and interpreted.

Example:

Consider a simple spreadsheet containing information about students in a class. The spreadsheet might have columns such as "Student ID," "Name," "Age," and "Grade." Each row represents a different student, and the cells within each row contain specific pieces of information.

| **Student ID** | **Name** | **Age** | **Grade** |
| --- | --- | --- | --- |
| 001 | John Doe | 18 | A |
| 002 | Jane Smith | 19 | B |
| 003 | Bob Brown | 17 | C |

**Structured Data:**

Data that is highly organized and follows a predefined model or schema.

Characteristics:

* Organized into tables with rows and columns.
* Each data element is defined and has a specific data type.
* Examples include relational databases, spreadsheets, and CSV files.

Example: -

| **ID** | **Name** | **Age** | **Grade** |
| --- | --- | --- | --- |
| 001 | John Doe | 18 | A |
| 002 | Jane Smith | 19 | B |
| 003 | Bob Brown | 17 | C |

**Unstructured Data:**

Data that does not have a predefined data model or is not easily organized into a structured format.

Example:

A paragraph of text, an image file, or a video recording without a specific format.

**Semi-Structured Data:**

Data that has some structure but does not fit neatly into a relational database or a strict schema.

Examples include XML (eXtensible Markup Language) and JSON (JavaScript Object Notation)

{

  "id": 1,

  "name": "John Doe",

  "age": 25,

  "grades": {

    "math": "A",

    "science": "B",

    "history": "A"

  }

}

**Numeric Data:**

* Discrete: Consists of separate values. For example, the number of students in a class (1, 2, 3, ...).
* Continuous: Represents a range of values. For example, the height or weight of a person.

**Text Data (Categorical):**

* Also known as qualitative or nominal data.
* Represents categories or labels. For example, color, gender, or types of fruits.

**Variable**

**Quantitative Variable:**

* Represents quantities or measurable amounts.
* Associated with continuous or discrete variables.
* Example: Income, Age

**Discrete Variable:**

* Represents countable, distinct values.
* Often whole numbers.
* Example: Number of siblings, Number of cars in a parking lot.

**Continuous Variable:**

* Represents measurable quantities that can take any value within a range.
* Can have an infinite number of possible values.
* Example: Height, Weight, Temperature.

**Qualitative Variable:**

* Represents qualities or characteristics.
* Often associated with categorical variables.
* Example: Hair color, Marital status.

**Nominal Variable:**

* A type of categorical variable with categories that do not have a specific order.
* Example: Colours, Types of fruit.

**Ordinal Variable:**

* A type of categorical variable with categories that have a meaningful order or ranking.
* Example: Education level (e.g., High School, Bachelor's, Master's).

**Population:**

The entire group that is the subject of a study and about which the researcher wants to draw conclusions.

Characteristics:

* Represents the complete set of individuals, items, or data points that meet certain criteria.
* Often denoted by the symbol "N."

Example:

If a researcher is interested in studying the average income of all households in a particular city, the population would consist of every household in that city.

**Sample:**

A subset of the population that is selected for the actual study.

Characteristics:

* Typically, more manageable and practical to study than the entire population.
* Findings from the sample are then generalized to make inferences about the population.
* Often denoted by the symbol "n."

Example:

In the study mentioned earlier, it might be impractical and time-consuming to collect data from every household in the city. Instead, the researcher might select a sample of, say, 500 households, gather data from them, and use the findings to make inferences about the average income of all households in the city.

**Some key reasons for using sampling**

**Necessity: -**

When dealing with a large population, it may be impractical or impossible to study every individual or element. Sampling allows researchers to study a representative subset, making it more manageable.

**Practicality: -**

Conducting research on an entire population may be time-consuming, require extensive resources, and be logistically challenging. Sampling allows researchers to gather meaningful insights while efficiently using resources.

**Cost-Effective: -**

Conducting a study on an entire population can be costly. Sampling is often more cost-effective as it reduces the expenses associated with data collection, analysis, and other logistical aspects.

**Management: -**

Managing and coordinating a study involving the entire population may be challenging. Sampling facilitates easier management of the research process.

**Data Cleansing (Data Cleaning):**

**Tasks Involved:**

- Handling missing values.

- Correcting inaccuracies and errors.

- Standardizing formats.

- Removing duplicates.

- Dealing with outliers.

**Missing Data**

Missing data refers to the absence of values or information for certain observations or variables in a dataset.

**Causes of Missing Data:**

**Non-response:**

Participants or entities may choose not to provide certain information when responding to surveys or questionnaires.

**Data Entry Errors:**

Mistakes during data entry or recording can lead to missing values.

**Systematic Issues:**

Issues with data collection systems or instruments can result in missing data.

**Incomplete Surveys:**

Some participants might not complete all parts of a survey, leading to missing responses.

**Measurement Issues:**

Certain variables may be challenging to measure, resulting in missing data for those variables.

**Dealing with Missing Data:**

**Accepting Missing Data:**

In some cases, researchers may choose to accept missing data if the proportion of missing values is small and the messiness is deemed random or inconsequential.

**Deleting Missing Data**

In this approach, observations with any missing values are excluded from the analysis.

**Imputation of Missing Data:**

* + Imputation involves estimating or predicting missing values based on observed data.
  + Mean/Median Imputation: Replace missing values with the mean or median of the observed values for that variable.

**Outlier**

An outlier is an observation or data point that significantly differs from the rest of the data in a dataset. Outliers can skew statistical analyses and distort the interpretation of results, making it important to identify and understand them.

**Key Characteristics of Outliers:**

**Deviation from the Norm:**

Outliers exhibit values that deviate substantially from the typical or expected range of values in a dataset.

**Impact on Statistical Measures:**

Outliers can heavily influence summary statistics such as the mean and standard deviation, leading to misleading representations of central tendency and dispersion.

**Identification:**

Outliers are often identified through statistical methods or visual inspection of graphs; such as box plots or scatter plots.

**Causes of Outliers:**

Outliers can arise from measurement errors, data entry mistakes, natural variability, or genuine extreme observations in the population.

**Ways to Identify Outliers:**

**Visual Inspection:**

Plotting the data using graphs like box plots, scatter plots, or histograms can reveal observations that stand out from the majority.

**Statistical Methods:**

* Z-Score: Identifying data points with z-scores beyond a certain threshold (e.g., |z| > 3) as potential outliers.

*Z = (x-µ)/σ*

* Interquartile Range (IQR): Using the IQR to identify observations outside a defined range.

*IQR = Q3 - Q1*

*LF = Q1 - (1.5\*IQR)*

*UF = Q3 + (1.5\*IQR)*

*Ex:*

*l = [2,5,7,10,12,15,18,20,22,30]*

*Q1 = 7*

*Q3 = 20*

*IQR = 20 - 7 => 13*

*LF = 7-(1.5\*13) => -12.5*

*UF = 20-(1.5\*13) => 39.5*

**Dealing with Outliers:**

**Retaining Outliers:**

* In some cases, it may be appropriate to retain outliers, especially if they represent genuine extreme values in the data.
* Retaining outliers allows for an inclusive analysis, considering the full range of variability in the dataset.

**Removing Outliers:**

* + Removing outliers involves excluding extreme values from the dataset before analysis.
  + Common methods include using statistical criteria (e.g., Z-scores, IQR) to identify and exclude observations beyond a certain threshold.
  + Reduces the impact of extreme values on summary statistics and model results
  + Loss of information: Excluding outliers may discard meaningful data points.

**Transformation:**

* + Transformation involves applying mathematical functions to the data to modify its distribution and reduce the impact of outliers.
  + Common transformations include logarithmic, square root, or Cube root transformations.

**What are quartiles?**

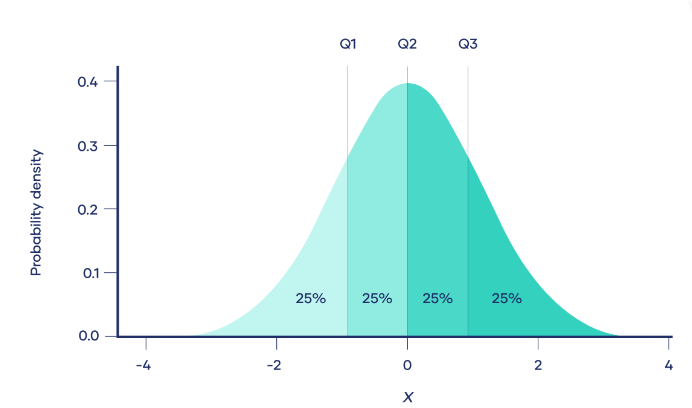
Quartiles are a set of descriptive statistics. They summarize the central tendency and variability of a dataset or distribution.

Quartiles are a type of percentile. A percentile is a value with a certain percentage of the data falling below it. In general terms, *k*% of the data falls below the *k*th percentile.

* The first quartile (Q1, or the lowest quartile) is the 25th percentile, meaning that 25% of the data falls below the first quartile.
* The second quartile (Q2, or the median) is the 50th percentile, meaning that 50% of the data falls below the second quartile.
* The third quartile (Q3, or the upper quartile) is the 75th percentile, meaning that 75% of the data falls below the third quartile.

By splitting the data at the 25th, 50th, and 75th percentiles, the quartiles divide the data into four equal parts.

* In a sample or dataset, the quartiles divide the data into four groups with equal numbers of observations.
* In a probability distribution, the quartiles divide the distribution’s range into four intervals with equal probability.

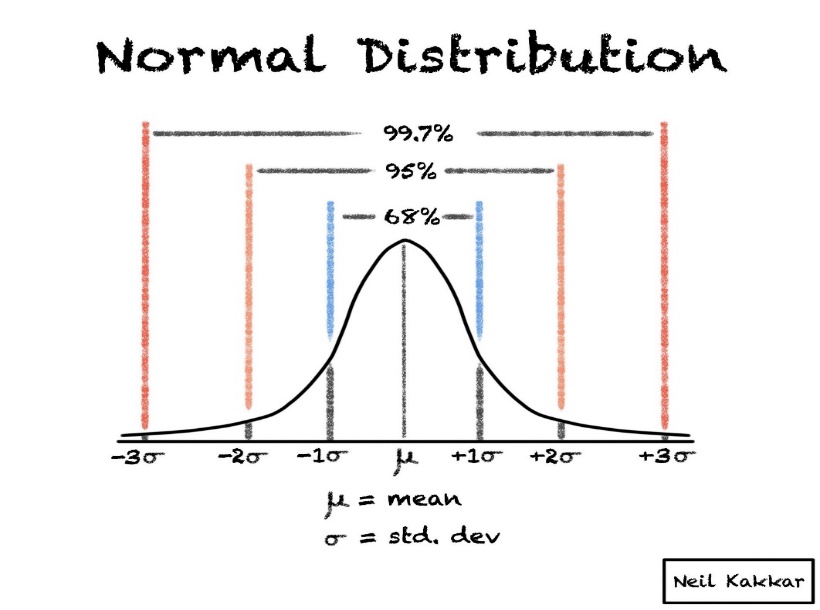
****

**Distribution**

a distribution refers to the pattern or shape that the values in a dataset take. It describes how the different values are spread or distributed across the range of possible values. Understanding the distribution of a dataset is fundamental for statistical analysis

**Normal Distribution:**

* + The normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is symmetric and bell-shaped.
  + The normal distribution is symmetric around its mean (average), which is located at the centre of the distribution. This means that the probabilities of observing values on the left and right sides of the mean are equal.
  + The normal distribution is completely described by two parameters: the mean (μ) and the standard deviation (σ). The mean determines the centre of the distribution, and the standard deviation controls the spread or variability of the distribution.
  + Approximately 68% of the data falls within one standard deviation of the mean (±μ±σ).
  + Approximately 95% falls within two standard deviations of the mean (±2μ±2σ).
  + Approximately 99.7% falls within three standard deviations of the mean (±3μ±3σ).



**Standard normal distribution**

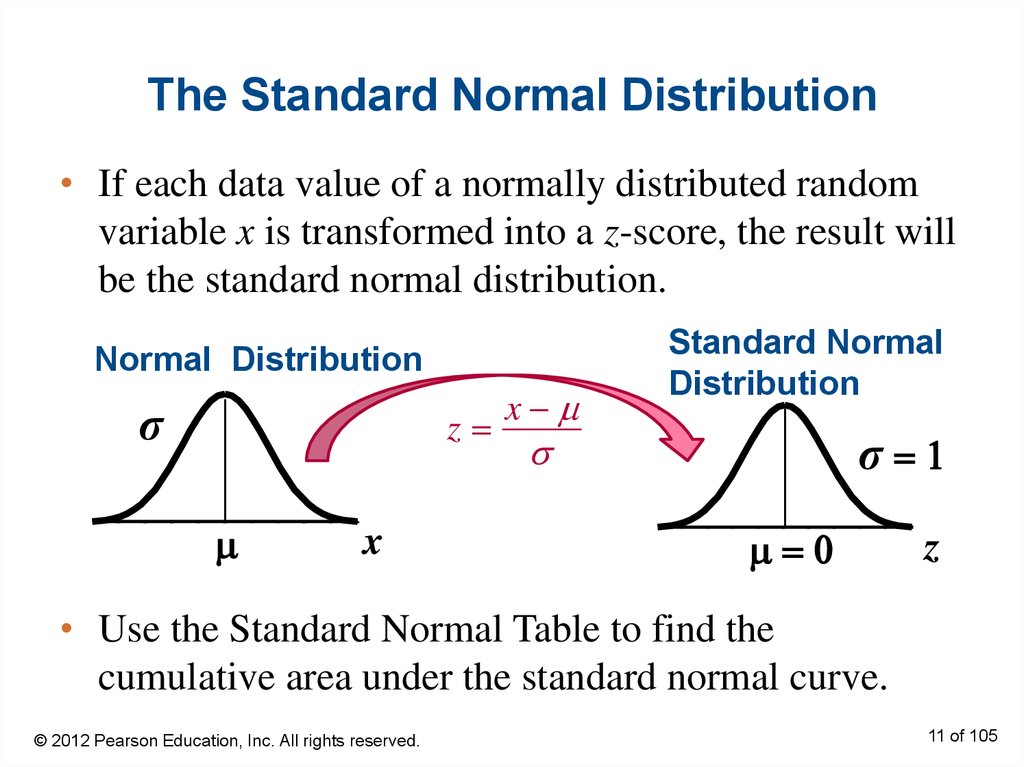
The standard normal distribution, also known as the z-distribution or the standard Gaussian distribution, is a special case of the normal distribution where the mean (μ) is 0 and the standard deviation (σ) is 1

Properties of the standard normal distribution include:

- Mean (μ): 0

- Standard Deviation (σ): 1

- Symmetry: Like the general normal distribution, the standard normal distribution is symmetric around 0.



**Probability Distribution:**

 probability distribution is an idealized frequency distribution.

A **frequency distribution**describes a specific sample or dataset. It’s the number of times each possible value of a variable occurs in the dataset.

Infinitely large samples are impossible in real life, so probability distributions are theoretical. They’re idealized versions of frequency distributions that aim to describe the population the sample was drawn from.

Probability distributions are used to describe the populations of real-life variables, like coin tosses or the weight of chicken eggs. They’re also used in hypothesis testing to determine *p*values.

Ex: -

Consider the probability distribution of the number of heads (successes) when flipping a fair coin five times.

**Probability mass functions**

A probability mass function (PMF) is a mathematical function that describes a discrete probability distribution. It gives the probability of every possible value of a variable.

**Probability density function**

The Probability Density Function (PDF) is a concept used in probability theory and statistics to describe the likelihood of a continuous random variable falling within a particular range of values. The PDF is associated with continuous probability distributions

Probability within an interval: The probability that X falls within a specific interval [ a, b] is given by the integral of the PDF over that interval.

**Statistics**

Statistics is a branch of mathematics that involves collecting, analysing, interpreting, presenting, and organizing data. It helps in making informed decisions and drawing conclusions from numerical information.

For example, imagine you want to study the average height of students in a class. You collect the heights of each student and calculate the mean (average) height. This mean height represents a central tendency, providing a summary statistic that gives insight into the overall height distribution within the class.

**Types of Statistics**

**Descriptive Statistics:**

- Descriptive statistics are used to summarize, organize, and describe the main features of a dataset.

- Measures of central tendency (mean, median, mode)

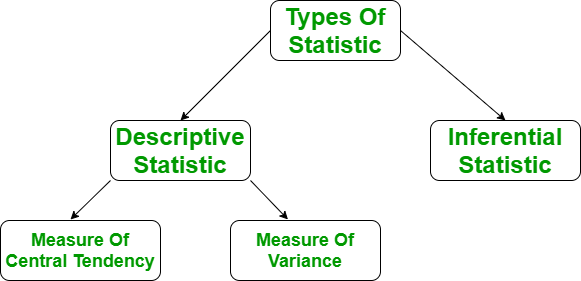
- Measures of dispersion (range, variance, standard deviation)

- Frequency distributions and histograms

- Percentiles and quartiles

**Inferential Statistics:**

inferential statistics involve making inferences or predictions about a population based on a sample of data. It extends findings from a sample to the entire population.



**Understanding the Descriptive Statistics**

In layman’s terms, descriptive statistics generally means describing the data with the help of some representative methods like charts, tables, Excel files, etc. The data is described in such a way that it can express some meaningful information that can also be used to find some future trends.

**There are two types of Descriptive Statistics:**

* 1. The measure of central tendency
  2. Measure of variability

**Measure of Central Tendency**

The measure of central tendency is a single value that attempts to describe the whole set of data. There are three main features of central tendency:

* + Mean
  + Median
  + Mode

**Mean**

It is the sum of observations divided by the total number of observations. It is also defined as average which is the sum divided by count.

The mean () function returns the mean or average of the data passed in its arguments. If the passed argument is empty, Statistics Error is raised.

Mean = Sum/N

Ex:

L = [15,3,12,0,24,3]

mean = (15+3+12+0+24+3)/6 =>9.5

**Median**

It is the middle value of the data set. It splits the data into two halves. If the number of elements in the data set is odd, then the centre element is the median and if it is even then the median would be the average of two central elements. it first sorts the data i=and then performs the median operation

Ex:

L = [15,3,12,0,24,3]

median = (12+0)/2 => 6

**Mode**

It is the value that has the highest frequency in the given data set. The data set may have no mode if the frequency of all data points is the same. Also, we can have more than one mode if we encounter two or more data points having the same frequency.

Ex:

L = [15,3,12,0,24,3]

median = 3

**Measure of Variability**

Till now, we have studied the measure of central tendency but this alone is not sufficient to describe the data. To overcome this, we need the measure of variability. The measure of variability is known as the spread of data or how well our data is distributed. The most common variability measures are:

* + Range
  + Variance
  + Standard deviation

**Range**

The difference between the largest and smallest data point in our data set is known as the range. The range is directly proportional to the spread of data which means the bigger the range, the more the spread of data and vice versa.

Range = Largest data value – smallest data value

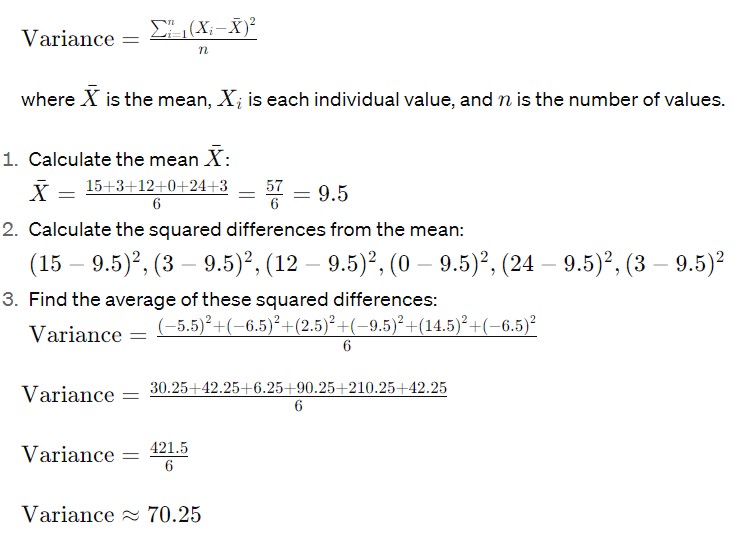
Ex:

L = [15,3,12,0,24,3]

Range=max(L)−min(L)=24−0=24

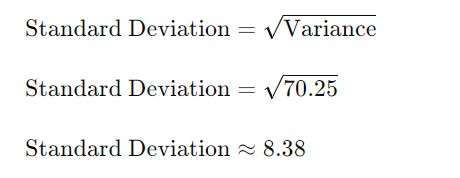
**Variance**

It is defined as an average squared deviation from the mean. It is calculated by finding the difference between every data point and the average which is also known as the mean, squaring them, adding all of them, and then dividing by the number of data points present in our data set.



**Standard Deviation**

It is defined as the square root of the variance. It is calculated by finding the Mean, then subtracting each number from the Mean which is also known as the average, and squaring the result. Adding all the values and then dividing by the no of terms followed by the square root.

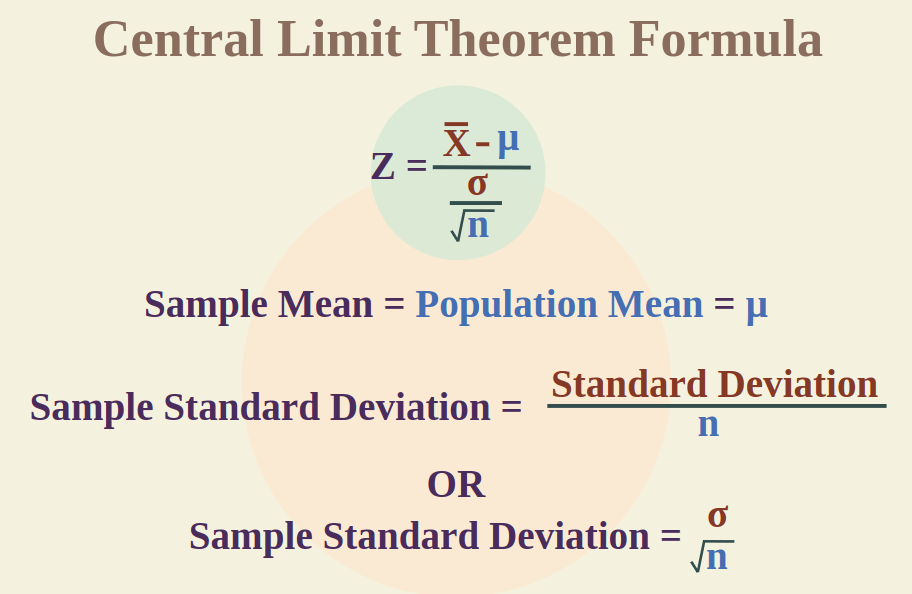
****

**Inferential Statistics**

**Central Limit Theorem**

Central Limit Theorem explains that the sample distribution of the sample mean resembles the normal distribution irrespective of the fact that whether the variables themselves are distributed normally or not. Central Limit Theorem is often called CLT in abbreviated form.

Central Limit Theorem states that when large samples usually greater than thirty are taken into consideration then the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random variables were originally distributed normally or not.



Ex: -

Imagine you have a school with a large number of students. The heights of individual students may not follow a perfect normal distribution. Some students might be shorter, some taller, and the distribution of heights may not look like a bell curve.

Now, let's apply the Central Limit Theorem:

Original Distribution:

The individual heights of students may have a varied distribution; it might not follow a normal pattern.

Sample Means:

Now, randomly select groups of students (samples) from the school and calculate the average height in each group. For example, you might take a sample of 30 students and find the average height of those 30.

Distribution of Sample Means:

According to the Central Limit Theorem, as you collect more and more of these average heights from different samples, the distribution of these sample means will start to resemble a bell curve, even if the individual heights themselves don't follow a normal distribution.

**Test statistic**

A test statistic describes how closely the distribution of your data matches the distribution predicted under the null hypothesis of the statistical test you are using.

The distribution of data is how often each observation occurs, and can be described by its central tendency and variation around that central tendency. Different statistical tests predict different types of distributions, so it’s important to choose the right statistical test for your hypothesis.

**Types of Statistic test**

1. Chi-Square Test: This test is used for categorical data to determine if there is a significant association between two variables. It’s commonly used for feature selection and assessing independence.
2. T-Test: The T-test is used to compare the means of two groups (e.g., comparing the performance of two different algorithms). It’s often used when the sample size is small.
3. Z-Test: Similar to the T-test, the Z-test compares means, but it’s more suitable for larger sample sizes. It assumes that the data follows a normal distribution.
4. Analysis of Variance (ANOVA): ANOVA is used to compare means across multiple groups (more than two). It’s useful for assessing the impact of different features on the model’s performance.

**Hypothesis Testing**

Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.

There are 5 main steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis and alternate hypothesis (Ho) and (Ha or H1).
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test.
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section.

**Ex:**

You want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, you restate it as:

* H0: Men are, on average, not taller than women.
* Ha: Men are, on average, taller than women.

**P-values**

**Definition**: The p-value (short for “probability value”) represents the probability of observing a test statistic (or something more extreme) under the assumption that the null hypothesis is true.

**Null Hypothesis (H₀)**: In hypothesis testing, we start with a null hypothesis (H₀) that there is no effect or no difference (e.g., no difference between two groups, no impact of a treatment, etc.). The alternative hypothesis (H₁) contradicts the null hypothesis.

**Test Statistic**: When we perform a statistical test (such as a t-test, chi-square test, etc.), we calculate a test statistic based on our data. This test statistic measures how far our observed data deviates from what we would expect under the null hypothesis.

**P-Value Interpretation**:

* If the p-value is small (typically less than a significance level, often denoted as α), we reject the null hypothesis. This suggests that our observed data is unlikely to occur by chance alone if the null hypothesis were true.
* If the p-value is large, we fail to reject the null hypothesis. This means that our data is consistent with the null hypothesis.

**Significance Level (α)**: The significance level (α) is a predetermined threshold (often set at 0.05 or 0.01) that helps us decide whether to reject the null hypothesis. If the p-value is less than α, we reject H₀; otherwise, we do not.

**Type I Error (False Positive):**

A Type I error occurs when you incorrectly reject the null hypothesis (H₀) even though it is actually true.

In other words, you make a false positive conclusion.

Example: Imagine testing a new drug for an autoimmune disease. If you conclude that the drug is effective (rejecting the null hypothesis) when it actually has no effect, that’s a Type I error.

**Type II Error (False Negative):**

A Type II error happens when you fail to reject the null hypothesis (H₀) even though it is false (i.e., there is a real effect).

In this case, you miss detecting a genuine effect.

Example: Continuing with the drug test scenario, if you fail to conclude that the drug is effective (not rejecting the null hypothesis) when it actually is, that’s a Type II error.

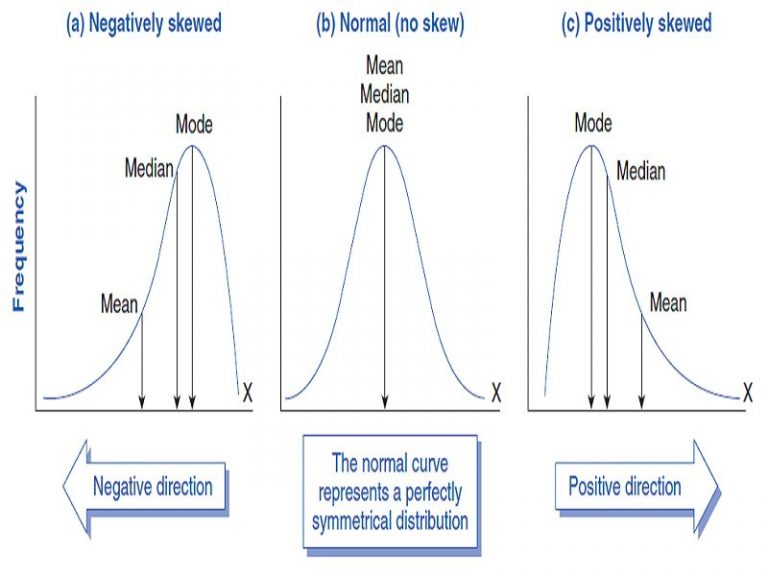
**Assumptions of Statistical test**

**Skew**

skewness is a measure of the asymmetry of a probability distribution. It quantifies the extent and direction of skew (departure from horizontal symmetry) in a dataset.

There are three main types of skewness:

1. **Positive skewness (right-skewed):** The right tail (larger values) of the distribution is longer or fatter than the left tail. In other words, the majority of the data points are concentrated on the left side, and the right side has some extreme values.
2. **Negative skewness (left-skewed):** The left tail (smaller values) of the distribution is longer or fatter than the right tail. This indicates that the majority of the data points are concentrated on the right side, and the left side has some extreme values.
3. **Zero skewness:** A perfectly symmetric distribution has a skewness of zero. This means that the right and left sides of the distribution are mirror images of each other.



Skewness is often measured using different formulas, with one common method being the Pearson's skewness coefficient. It is calculated as:

Skewness=3(*x*ˉ−*μ*)/σ​

where:

* ˉ*x*ˉ is the sample mean,
* *μ* is the population mean,
* *σ* is the standard deviation.

Skewness provides valuable information about the shape of a distribution and can help identify potential outliers

**Kurtosis**

Kurtosis is a statistical measure that describes the distribution of data in terms of the tails and the peakedness (or flatness) relative to a normal distribution. It provides information about the shape and thickness of the tails of a probability distribution.

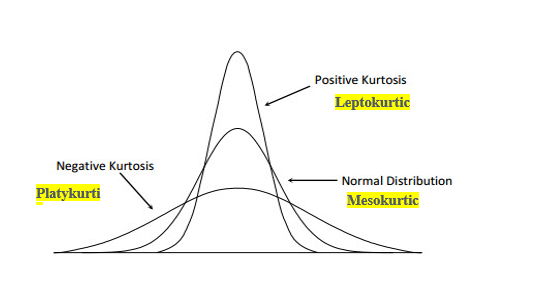
There are three main types of kurtosis:

1. **Mesokurtic (Normal kurtosis):** A distribution with kurtosis equal to 3 (or 0 excess kurtosis) is considered mesokurtic. This means that the distribution has similar tail behavior to a normal distribution.
2. **Leptokurtic (Excess kurtosis > 0):** A distribution with positive excess kurtosis has heavier tails and a more peaked central portion compared to a normal distribution. This indicates the presence of more extreme values.
3. **Platykurtic (Excess kurtosis < 0):** A distribution with negative excess kurtosis has lighter tails and a flatter central portion compared to a normal distribution. This suggests fewer extreme values than a normal distribution.

The excess kurtosis is often used in practice, and it is calculated as:

Excess Kurtosis=Kurtosis−3

Kurtosis is affected by the presence of outliers in a dataset. High kurtosis can indicate the presence of heavy tails and potentially more extreme values, while low kurtosis can suggest lighter tails and a lack of extreme values.



**Correlation**

Correlation is the statistical analysis of the relationship or dependency between two variables. Correlation allows us to study both the strength and direction of the relationship between two sets of variables.

Change in variable Couse change in another variable

**Types**

**+ ve Correlation**

It is a relation b/w two variable where if one variable increases, the other variable also increases & vice-verse

Ex: - No of tree cut down increases, the probability of erosion Increases

**- ve Correlation**

Inverse relationship b/w Variable, where one variable decreases other variable increases.

ex: - Car speed decreases, time taken to reach destination increases.

It always takes on a value between -1 and 1 where:

* + -1 indicates a perfectly negative linear correlation between two variables
  + 0 indicates no linear correlation between two variables
  + 1 indicates a perfectly positive linear correlation between two variables

**Correlation coefficients**

The correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two variables.

The most commonly used correlation coefficient is Pearson's correlation coefficient, denoted by ***r***.

**Coefficient of determination**

The coefficient of determination, often denoted as *R*2 (pronounced "R-squared"), is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s)

he *R*2 value ranges between 0 and 1, or it can be expressed as a percentage between 0% and 100%. The interpretation is as follows:

* *R*2=0: The model does not explain any of the variability in the dependent variable.
* *R*2=1: The model explains all of the variability in the dependent variable.
* 0<*R*2<1: The model explains a proportion of the variability in the dependent variable.

it is calculated as the ratio of the explained sum of squares to the total sum of squares. The formula for *R*2 is:

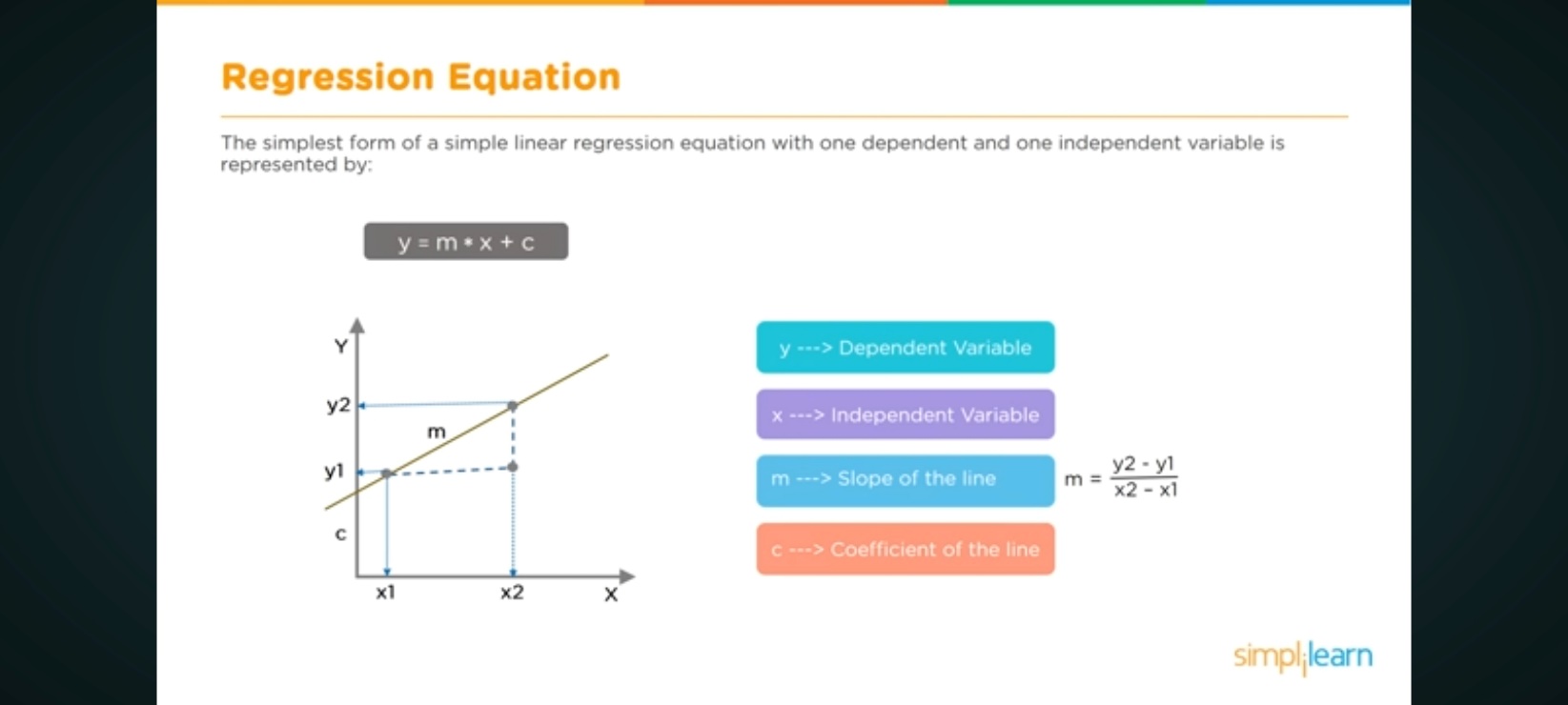
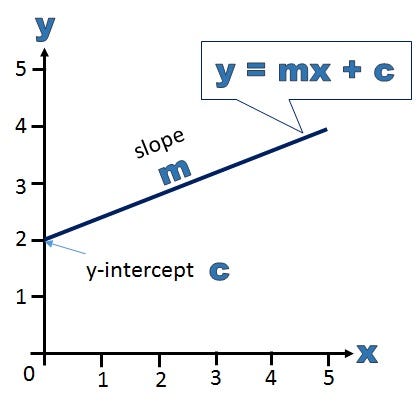
*R2 = Explained Sum of Squares (ESS)/ Total Sum of Squares (TSS)*

**Regression**

It involves modelling the relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (also known as predictors or features). The primary goal of regression is to understand and quantify the relationship between variables and to make predictions based on that understanding.

**Simple Linear Regression**

* Simple Linear Regression is a statistical method used to model the relationship between a single independent variable (predictor) and a dependent variable (response).
* The relationship is assumed to be linear, meaning it can be represented by a straight line.
* The purpose of simple linear regression is to find the best-fitting line that minimizes the difference between the predicted and observed values of the dependent variable.



The simple linear regression model is represented by the equation:

y=mx + c

where:

* *y* is the dependent variable.
* *x* is the independent variable.
* *m* is the slope of the line, representing the rate of change in *y* for a one-unit change in *x*.
* *c* is the y-intercept, which is the value of *y* when *x* is 0.

The goal in simple linear regression is to estimate the values of *m* and *c* that minimize the sum of squared differences between the observed values (*y*) and the values predicted by the line (*mx*+*c*).

The formula for the slope (*m*) is given by:

m=n(∑x\*y) −(∑x) (∑y) ​/n(∑x2) −(∑x)2

where:

* *n* is the number of data points.
* ∑ denotes summation.
* *x* and *y* are the respective means of the independent and dependent variables.

The formula for the y-intercept (*c*) is given by: C = (∑x) (∑y) ​/ n