

# STATISTICS & MATHS NOTES

---

**Mazher Khan - IIT (BHU) - B.Tech (DR-2)**

**Senior Data Analyst @Target | Ex - OLX (EU)**

**YouTube - 2.2M+ (Views) | LinkedIn 16k+**

---

Telegram Link-	<a href="https://t.me/+XTjv6r80eDc5ZWU1">https://t.me/+XTjv6r80eDc5ZWU1</a>
Practice Workbook - 100 Days Challenge	<a href="https://docs.google.com/spreadsheets/d/1eP8evU2JIsawAVJ7GH_NNd_2xNLOT7abpJTs5O9iUQI/edit#gid=775777503">https://docs.google.com/spreadsheets/d/1eP8evU2JIsawAVJ7GH_NNd_2xNLOT7abpJTs5O9iUQI/edit#gid=775777503</a>
Follow me on LinkedIn	<a href="https://www.linkedin.com/in/mazher-khan/">https://www.linkedin.com/in/mazher-khan/</a>
Follow on Instagram	<a href="https://www.instagram.com/khan.the.analyst">https://www.instagram.com/khan.the.analyst</a>
Book 1:1 Mentorship Plan - 1,3, 6 Months	<a href="https://www.preplaced.in/profile/mazher-khan">https://www.preplaced.in/profile/mazher-khan</a>
Book for Career Guidance, CV review & interview tip	<a href="https://topmate.io/mazher_khan">https://topmate.io/mazher_khan</a>
Follow on Youtube	<a href="https://youtube.com/@imzhr.?si=KdMGmWt-vTy12hxV">https://youtube.com/@imzhr.?si=KdMGmWt-vTy12hxV</a>

## **A. Introduction to Statistics**

**@khan.the.analyst**

## 1. Population and Sample:

- Population refers to the entire group of individuals, objects, or events that you want to study or make inferences about.
- A sample, on the other hand, is a subset of the population
- Sampling error  $= z \text{ score} * \text{Standard deviation} / \sqrt{\text{sample size}}$  can be reduced by
  - Increasing the sample size
  - Classifying the population into different groups

Example: Suppose you want to study the average income of all employees in a company. The population would be all the employees in the company, while a sample might be a randomly selected group of 100 employees.

Variables and Data Types:

## 2. Variable and Data Type

Example: In a survey, age can be a numerical variable (discrete if measured in whole years) and gender can be a categorical variable (nominal with categories like male and female).

Variables or Measurement levels -

- Quantitative (Nominal or Ordinal)
- Qualitative (Ratio or Interval)

## 3. Measures of Central Tendency:

- Central tendency is a statistical measure that represents the center or average value of a dataset.. If the data is symmetrically distributed then  $\text{Mean} = \text{Median} = \text{Mode}$
- If the distribution is Skewed, then Median is the best measure of central tendency  
Mean is most sensitive for skewed data.
- Mode is the best measure for categorical or discrete data
- Other measures : Weighted Mean, Geometric Mean, Harmonic Mean

## 4. Measures of Dispersion:

Dispersion or variability describes how items are distributed from each other and the centre of a distribution. There are 4 methods to measure the dispersion of the data:

- Range
- Interquartile Range
- Variance
- Standard Deviation (Best Measure)

## **B. Descriptive Statistics**

These visualization techniques help in understanding the distribution, shape, and characteristics of data, enabling better analysis and interpretation.

### **1. Frequency Distributions:**

Frequency distribution is a table that summarizes the frequency of each value in a dataset. It helps visualize the distribution of data and identify patterns.

Example: Consider the dataset [1, 1, 2, 3, 4, 4, 4, 5, 5, 5, 5]. The frequency distribution would be:

Value	Frequency
1	2
2	1
3	1
4	3
5	4

### **2. Histograms, Bar Charts, and Pie Charts:**

- Histograms display the distribution of numerical data by dividing the data into intervals (bins) and plotting the frequency or count of values in each bin as bars.
- Bar charts represent categorical data using rectangular bars whose heights indicate the frequency or count of each category.
- Pie charts represent the composition of categorical data as a circular chart, where each category is shown as a slice of the pie proportional to its frequency or count.

Example: Suppose you have survey data on the favorite colors of individuals. A bar chart could show the frequencies of different colors, while a pie chart could show the proportion of each color category.

### **3. Measures of Skewness and Kurtosis:**

- Skewness measures the asymmetry of a distribution, indicating whether the data is concentrated on one side or stretched out on both sides. (Positive or Negative Skewness)
- There are different ways to measure the skewness like
  - Pearson Mode

- Pearson Median
- Kurtosis measures the degree of peakedness or flatness of a distribution compared to the normal distribution. Outliers are detected in a data distribution using kurtosis. The higher the kurtosis, the higher the number of outliers in the data.

Example: A positively skewed distribution would have a long tail on the right side, indicating that most values are concentrated on the left. High kurtosis suggests a distribution with a sharp peak and heavy tails, while low kurtosis indicates a flatter distribution.

#### **4. Box Plots:**

##### **5.**

- Box plots provide a graphical representation of the distribution of numerical data through quartiles. They show the minimum, maximum, median (middle value), and the lower and upper quartiles (25th and 75th percentiles) of a dataset.

Example: In a box plot, a box is drawn from the lower quartile to the upper quartile, with a line indicating the median. Whiskers extend from the box to the minimum and maximum values, and potential outliers are shown as individual data points.

### **C. Probability**

#### **1. Basic Probability Rules (Union, Intersection, Complement):**

- Union: The probability of the union of two events A and B, denoted as  $P(A \cup B)$ , is the probability that either A or B or both occur.
- Intersection: The probability of the intersection of two events A and B, denoted as  $P(A \cap B)$ , is the probability that both A and B occur.
- Complement: The probability of the complement of an event A, denoted as  $P(A')$ , is the probability that A does not occur.

#### **2. Conditional Probability:**

Conditional probability measures the probability of an event A occurring given that event B has already occurred, denoted as  $P(A|B)$ .

It is calculated as the probability of A and B occurring together divided by the probability of B.

$$P(B|A) = P(A \cap B) / P(A).$$

### 3. Bayes' Theorem:

Bayes' theorem is a fundamental concept in probability theory that relates conditional probabilities. It calculates the probability of an event A given the occurrence of another event B, by incorporating prior probabilities and likelihoods.

$$\text{Bayes' theorem: } P(A|B) = (P(B|A) * P(A)) / P(B)$$

Example: In medical diagnostics, Bayes' theorem can be used to calculate the probability of a disease given the presence of certain symptoms, by considering the prevalence of the disease, the sensitivity and specificity of the tests, and the conditional probabilities.

### 4. Probability Distributions (Discrete and Continuous):

Probability distributions describe the likelihood of different outcomes in a random event.

- Discrete distributions are used when the variable can take on a countable number of distinct values, such as the binomial distribution for the number of successes in a fixed number of trials.
- Continuous distributions are used when the variable can take on any value within a range, such as the normal distribution for variables like height or weight.

Example: The binomial distribution describes the probability of obtaining a specific number of successes in a fixed number of independent trials with the same probability of success. The normal distribution describes the probability distribution of a continuous variable that follows a bell-shaped curve.

## D. Statistics Distribution

### 1. Normal Distribution:

- The normal distribution (or Gaussian distribution) is a continuous probability distribution that is symmetric and bell-shaped.

- It is characterized by its mean ( $\mu$ ) and standard deviation ( $\sigma$ ), which determine the location and spread of the distribution, respectively.

Example: The heights of adult males in a population often follow a normal distribution. If the mean height is 175 cm and the standard deviation is 6 cm, you can use the normal distribution to calculate the probability of finding a male with a height between certain values or above a certain height.

## **2. Binomial Distribution:**

- The binomial distribution models the probability of a binary outcome (success or failure) in a fixed number of independent Bernoulli trials, each with the same probability of success ( $p$ ).
- It is characterized by two parameters: the number of trials ( $n$ ) and the probability of success ( $p$ ).

Example: Tossing a fair coin 10 times can be modeled using a binomial distribution. The probability of getting exactly 5 heads (successes) in 10 coin flips can be calculated using the binomial distribution formula.

## **C. Poisson Distribution:**

- The Poisson distribution models the probability of the number of events occurring in a fixed interval of time or space when the events occur randomly and independently.
- It is characterized by a single parameter,  $\lambda$  (lambda), which represents the average rate of occurrence of the events.

Example: The number of phone calls received at a call center in a given hour can be modeled using a Poisson distribution. If the average rate of calls is 4 per hour, you can use the Poisson distribution to calculate the probability of receiving a certain number of calls in a specific time interval.

## **D. Exponential Distribution:**

The exponential distribution models the time between consecutive events in a Poisson process, where events occur randomly and independently at a constant average rate.

It is characterized by a single parameter,  $\lambda$  (lambda), >average rate of occurrence of the events.

Example: The time between arrivals of customers at a service counter can be modeled using an exponential distribution. If the average arrival rate is 5 customers per hour, you can use the exponential distribution to calculate the probability of waiting a certain amount of time before the next customer arrives.

## **E. Statistical Inference**

### **1. Hypothesis Testing (Null and Alternative Hypotheses):**

Hypothesis testing is a statistical method used to make inferences about a population based on sample data.

Common assumptions include:

- Independence: The observations in the sample are independent of each other.
- Random sampling: The sample is representative of the population being studied.
- Normality: The data or test statistics follow a normal distribution.
- Homogeneity of variance: The variance of the variables is equal across groups being compared.

### **2. Type I and Type II Errors:**

- Type I error occurs when you reject the null hypothesis ( $H_0$ ) when it is actually true. It is the probability of falsely claiming an effect or difference.
- Type II error occurs when you fail to reject the null hypothesis ( $H_0$ ) when it is actually false. It is the probability of missing a true effect or difference.

Example: In a clinical trial, a Type I error would mean falsely concluding that the drug is effective when it actually has no effect. A Type II error would mean failing to identify the drug's effectiveness when it does have an effect.

### 3. Confidence Intervals:

- A confidence interval is a range of values within which the true population parameter is estimated to lie with a certain level of confidence.
- It provides a measure of uncertainty around a sample estimate and is often used to make inferences about population parameters.

Example: Suppose you want to estimate the average height of a population. A 95% confidence interval would provide a range of values within which you can be 95% confident that the true population mean lies.

### 4. P-values:

- A p-value is a probability that measures the strength of evidence against the null hypothesis ( $H_0$ ) in a hypothesis test.
- It quantifies the likelihood of observing the data or more extreme data if the null hypothesis were true.
- The p-value ranges between 0 and 1, where a smaller p-value indicates stronger evidence against the null hypothesis.

Example: In a hypothesis test, if the p-value is less than a pre-defined significance level (e.g., 0.05), it is considered statistically significant, and you reject the null hypothesis ( $H_0$ ) in favor of the alternative hypothesis ( $H_a$ ).

## F. Regression Analysis

### 1. Simple Linear Regression:

- Simple linear regression is a statistical method used to model the relationship between a dependent variable and a single independent variable.



## **Assumptions**

- Linear relationship between the independent and the dependent variable.
- The independent variables should not be highly correlated with each other.
- No correlation between the residuals or errors of the model.
- The variance of the residuals should be constant
- Normality: The residuals should follow a normal distribution.

### **2. Multiple Linear Regression:**

- Multiple linear regression extends simple linear regression by considering multiple independent variables to predict a dependent variable.

### **3. Logistic Regression:**

- Logistic regression is used when the dependent variable is categorical or binary, and the aim is to predict the probability of an event occurring.

Example: Suppose you want to predict whether a customer will churn or not

### **4. Model Evaluation Metrics (R-squared, AIC, BIC):**

- R-squared : R-squared measures the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. It ranges from 0 to 1, with higher values indicating a better fit.
- Adjusted R -squared penalizes the addition of unnecessary variables, preventing an inflation of R-squared due to model complexity.
- 

## **G. Experimental Design and A/B Testing**

### **1. Randomized Controlled Trials**

- Participants are randomly assigned to either a treatment group or a control group. This random assignment helps ensure that any observed differences between the groups can be attributed to the treatment.

Example: Suppose a pharmaceutical company wants to test the effectiveness of a new drug. They conduct an RCT by randomly assigning participants to two groups: the treatment group receives the new drug, while the control group receives a placebo. By comparing the outcomes between the two groups, the company can determine if the drug has a significant impact.

## **2. Control and Treatment Groups**

- In experimental studies, a control group serves as a baseline for comparison against the treatment group.
- The control group does not receive the treatment or intervention being studied, while the treatment group does.

Example: In a study investigating the impact of a new teaching method on student performance, one group of students receives the new teaching method (treatment group), while another group continues with the traditional teaching method (control group).

## **3. Hypothesis Testing in Experiments**

- Hypothesis testing in experiments involves setting up null and alternative hypotheses to evaluate the statistical significance of observed differences.
- Statistical tests, such as t-tests or chi-square tests, are then performed to determine whether the observed differences are likely due to chance or if they provide evidence to reject the null hypothesis.

## **4. A/B Testing Methodology**

- A/B testing is a methodology used in product development to compare two versions to determine which one performs better.
- A randomly selected sample of users is divided into two groups:
- Control vs Treatment
- Key metrics, such as conversion rates or click-through rates, are then compared between the two groups

## **H. Statistical Learning**

Understanding these concepts and techniques allows you to apply different algorithms, evaluate their performance, and make informed decisions regarding model selection and evaluation.

### **1. Random Forests:**

- Random forests are an ensemble learning method that combines multiple decision trees to make more accurate predictions.
- It works by creating a multitude of decision trees on different subsets of the training data and averaging their predictions.

Example: Consider a scenario where you want to predict housing prices based on various features such as area, number of bedrooms, and location. A random forest can be built by training multiple decision trees on different subsets of the data and aggregating their predictions to create a more reliable and accurate price prediction model.

### **2. Support Vector Machines (SVM):**

- Support Vector Machines are a supervised machine learning algorithm used for classification and regression tasks.

### **3. Model Evaluation and Validation Techniques (Cross-Validation, Bias-Variance Tradeoff):**

- Cross-Validation: Cross-validation is a technique used to assess the performance of a model by splitting the data into training and validation sets. It helps estimate the model's generalization capability and identify potential overfitting or underfitting issues.

## **I. Statistical Sampling**

### **1. Simple Random Sampling:**

- It involves randomly selecting individuals from the population without any specific characteristics or grouping considerations.

Example: Suppose you want to estimate the average height of students in a school. You could assign a unique number to each student and use a random number generator to select a sample of students. Each student has an equal chance of being chosen, regardless of their grade or other attributes.

### **2. Stratified Sampling:**

- Stratified sampling involves dividing the population into homogeneous groups called strata, and then taking a random sample from each stratum.

Example: Consider a company with employees from different departments (e.g., HR, Finance, Operations). To conduct an employee satisfaction survey, you can use stratified sampling by selecting a random sample of employees from each department. This ensures representation from all departments in the survey.

### **3. Cluster Sampling:**

- Cluster sampling involves dividing the population into clusters or groups, and then randomly selecting some of these clusters for inclusion in the sample.

Example: Suppose you want to estimate the average income in different neighborhoods of a city. Instead of sampling individual residents, you could divide the city into clusters

(e.g., by zip code) and randomly select a few clusters. Then, you can collect data on income from all individuals within the selected clusters.

4. **Systematical:** Picks up every 'n' member in the data

#### 5. **Sampling Distributions:**

- A sampling distribution is a probability distribution of a sample statistic (e.g., mean, proportion) obtained from multiple samples of the same size from a population.

Example: Consider taking multiple random samples of 100 individuals from a population and calculating the mean height of each sample. The distribution of these sample means would form a sampling distribution. From this distribution, you can estimate the population mean and assess the variability or precision of the estimate

### **J. Correlation and Covariance**

#### 1. **Correlation and Covariance:**

- Correlation and covariance are statistical measures that describe the relationship between two variables.
- Covariance measures how two variables vary together, indicating the direction (positive or negative) and strength of their linear relationship.
- Correlation measures the linear relationship between two variables on a standardized scale, ranging from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.

Example: Suppose you want to examine the relationship between the age of a car (in years) and its resale value (in dollars). If the covariance between age and resale value is

positive, it suggests that as the age of the car increases, its resale value tends to decrease. A positive correlation coefficient confirms this relationship.

## **2. Pearson Correlation Coefficient:**

- The Pearson correlation coefficient (also known as Pearson's  $r$ ) measures the linear relationship between two continuous variables.
- It assesses the strength and direction of the relationship and is bound between -1 and 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.

Example: Consider two variables, such as the hours studied and the exam scores of a group of students. A positive Pearson correlation coefficient indicates that as the number of hours studied increases, the exam scores tend to be higher.

## **3. Covariance Matrix:**

- A covariance matrix summarizes the covariances between multiple variables.
- It is a square matrix where each element represents the covariance between two variables.

The diagonal elements of the matrix represent the variances of the individual variables.

Example: In a dataset with multiple variables like age, income, and education level, the covariance matrix would provide insights into how these variables relate to each other. Positive off-diagonal elements indicate a positive relationship, negative off-diagonal elements indicate a negative relationship, and diagonal elements indicate the variability of each variable.

## **4. Correlation vs. Causation:**

- Correlation refers to a statistical relationship between two variables where changes in one variable tend to be associated with changes in the other variable.
- It measures the strength and direction of the linear relationship between variables but does not imply a cause-and-effect relationship.

Example: Let's consider the relationship between ice cream sales and temperature.

There is a positive correlation between these variables because as temperature increases, ice cream sales tend to increase. However, this correlation does not imply that temperature causes people to buy more ice cream. Other factors like seasonality and consumer behavior might contribute to the observed correlation.

- Causation refers to a cause-and-effect relationship where a change in one variable directly influences or causes a change in another variable.
- It suggests a directional relationship between variables, indicating that one variable is responsible for the changes observed in the other.

Example: Let's consider the relationship between smoking and lung cancer. Numerous studies have established a causal link between smoking and the development of lung cancer. The evidence shows that smoking causes an increased risk of developing lung cancer

## **K. Analysis of Variance (ANOVA)**

Understanding ANOV allows you to compare group means and assess the significance of differences. Post-hoc tests help provide more detailed information about specific group comparisons.

### **1. ANOVA (Analysis of Variance):**

- ANOVA is a statistical method used to compare the means of two or more groups to determine if there are significant differences among them.

Example: Suppose you want to compare the effectiveness of three different diets (A, B, and C) in terms of weight loss. ANOVA can be used to determine if there is a significant difference in mean weight loss among the three diets.

#### Assumptions:

- Observations are independent
- Populations being compared follow a normal distribution
- Variances within each group are equal (homoscedasticity)

Interpretations: If the null hypothesis is rejected in ANOVA, it indicates that there are significant differences among the groups. Post-hoc tests can help identify the specific groups that differ significantly.

One-way ANOVA is used when there is only one independent variable with >2 groups and tests whether means of the groups are significantly different from each other while Two-way Anova involves two independent variables

One-way (example) Consider a study examining the effect of different teaching methods (A, B, C, and D) on students' test scores. One-way ANOVA can be used to determine if there is a significant difference in mean test scores among the four teaching methods.



Two-way (example) Suppose you want to investigate the effects of both gender (male, female) and exercise intensity (low, medium, high) on heart rate. Two-way ANOVA can be used to determine if there are significant main effects of gender and exercise intensity, as well as a significant interaction effect between them.

## **2. Post-hoc tests:**

- Post-hoc tests are used after performing ANOVA to determine which specific groups or combinations of groups differ significantly from each other.

Example: Following a significant result in a one-way ANOVA comparing the means of three different treatments, post-hoc tests such as Tukey's HSD (Honestly Significant Difference) or Bonferroni's test can be conducted to determine which specific treatments differ significantly from each other.

## **3. Parametric vs Non-Parametric Testing**

- The choice between parametric and non-parametric methods depends on the nature of the data and the assumptions you are willing to make.
- If the data meets the assumptions of parametric tests, they can provide more powerful and precise results. However, when dealing with non-normal or skewed data, or when assumptions cannot be met, non-parametric tests offer a valid alternative

### **Parametric Statistics and Tests:**

- Parametric statistics assume that the data follows a specific distribution, typically the normal (Gaussian) distribution.
- These methods make assumptions about the population parameters, such as mean and variance.
- Parametric tests include t-tests, analysis of variance (ANOVA), and linear regression.

**Parametric tests are generally more powerful** (i.e., have higher statistical power) when the underlying assumptions are met. However, violating the assumptions can lead to inaccurate results and conclusions.

### **Non-parametric Statistics and Tests:**

- Non-parametric statistics do not make any specific assumptions about the underlying population distribution. Non-parametric tests are used when the data is not normally distributed or when there are concerns about the assumptions of parametric tests.
- Non-parametric tests include the Wilcoxon signed-rank test, Mann-Whitney U test, and Kruskal-Wallis test.

Non-parametric tests are generally less powerful than their parametric counterparts but provide more robustness against violations of assumptions.

### **Interview Questions**

1. Explain Population vs Sample
2. Explain Descriptive vs Inferential Statistics
  - Descriptive statistics describe some sample or population.
  - Inferential statistics attempts to infer from some sample to the larger population.
3. What is Standard Deviation?
  - Standard deviation measures the dispersion of a dataset relative to its mean. It tells you, on average, how far each value lies from the mean.
  - A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.
4. Give an example where the median is a better measure than the mean  
The median is a better measure of central tendency than the mean when the distribution of data values is skewed or when there are clear outliers.
5. How do you calculate the needed sample size?

$$\text{Sample size} = \frac{(Z \text{ score})^2 \times \sigma \times (1 - \sigma)}{(\text{margin of error})^2}$$

- Population Size
- Margin of error - Also known as Confidence of Interval
- Confidence of Interval - 95%, 99%
- Standard Deviation

6. What do you understand by the term Normal Distribution?

- The normal distribution (or Gaussian distribution) is a continuous probability distribution that is symmetric and bell-shaped.
- Unimodal: normal distribution has only one peak. (i.e., one mode)
- Symmetric: a normal distribution is perfectly symmetrical around its center
- The Mean, Mode, and Median are all located in the center
- Asymptotic: normal distributions are continuous and have tails that are asymptotic. The curve approaches the x-axis, but it never touches.
- Standard Normal Distribution– Mean =0 and Standard deviation=1

7. What are Outliers?

An outlier is a data point that differs significantly from other data points in a dataset. An outlier in the data is due to:

- Variability in the data- Genuine – 100 33 score
- Experimental Error -
- Heavy skewness in data - Income data (assymtery)
- Missing values

8. Mention methods to screen for outliers in a dataset

- a. Sort the data from high to low or low to high
- b. InterQuartile Range
- c. Box Plot - This chart highlights statistical information like minimum and maximum values (the range), the median, and the interquartile range for the data. When reviewing a box plot, an outlier is a data point outside the box plot's whiskers.
- d. Use Z-score ( $z = (x - \text{mean}) / \text{standard deviation}$ )
  - If the z-score is positive, the data point is above average.
  - If the z-score is negative, the data point is below average.
  - If the z-score is close to zero, the data point is close to average.
  - If the z-score is above or below 3 , it is an outlier

Cases when outliers are kept

- Results are critical
- Outliers add meaning to the data
- The data is highly skewed

## Methods to handle missing data

- Prediction of the missing values
- Assignment of individual (unique) values
- Deletion of rows, which have the missing data
- Mean imputation or median imputation
- Using random forests, which support the missing values

## 9. What is the meaning of an inlier?





- An inlier is a data value that lies within the general distribution of other observed values but is an error. Inliers are difficult to distinguish from good data values, therefore, they are sometimes difficult to find and correct.

An example of an inlier might be a value recorded in the wrong units.

## 10. What is the difference between one-tailed and two-tail hypothesis testing?

- One-tailed tests allow for the possibility of an effect in one direction. Here, the critical region lies only on one tail while Two-tailed tests test for the possibility of an effect in two directions—positive and negative. Here, the critical region is one of both tails.

## 11. What is the difference between type I vs. type II errors?

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	 Type I Error (False positive)	 Correct Outcome! (True positive)
Fail to reject null hypothesis	 Correct Outcome! (True negative)	 Type II Error (False negative)

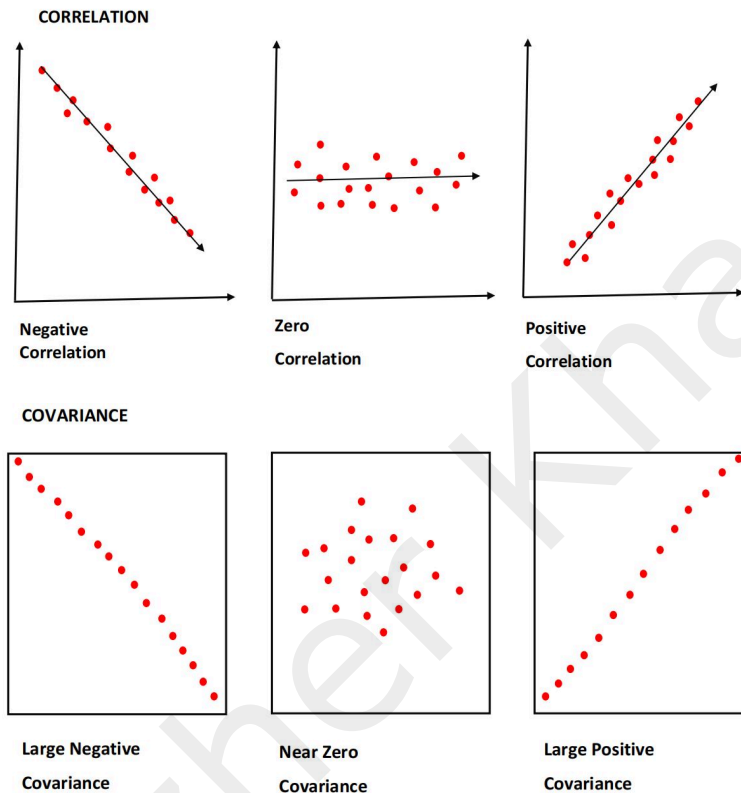
## 12. What is the Central Limit Theorem?

The central limit theorem states that the sampling distribution of the mean will always follow a normal distribution under the following conditions:

- The sample size is sufficiently large (i.e., the sample size is  $n \geq 30$ ).
- The samples are independent and identically distributed random variables.
- The population's distribution has finite variance.

13. What are correlation and covariance in statistics?

- Correlation indicates how strongly two variables are related. The value of correlation between two variables ranges from -1 to +1.
- Covariance is a measure that indicates the extent to which a pair of random variables vary with each other. A higher number denotes a higher dependency.



14. What is the difference between Point Estimate and Confidence Interval Estimate?

A point estimate gives a single value as an estimate of a population parameter. A confidence interval estimate gives a range of values likely to contain the population parameter.

15. What is the left-skewed distribution and the right-skewed distribution?

- Left-skewed : the left tail is longer than the right side.  $\text{Mean} < \text{median} < \text{mode}$
- Right-skewed: the right tail is longer. It is also known as positive-skew distribution. ( $\text{Mode} < \text{median} < \text{mean}$ )

16. How to convert normal distribution to standard normal distribution?

- Any point (x) from the normal distribution can be converted into standard normal distribution (Z) using this formula –  $Z(\text{standardized}) = (x - \mu) / \sigma$

17. What is Bessel's correction?

Bessel's correction advocates the use of n-1 instead of n in the formula of standard deviation.

18. What is the difference between the first quartile, the second quartile, and the third quartile?

First quartile (Q1) - Median of the lower half of the data set.

Second quartile (Q2) - Median of the data set.

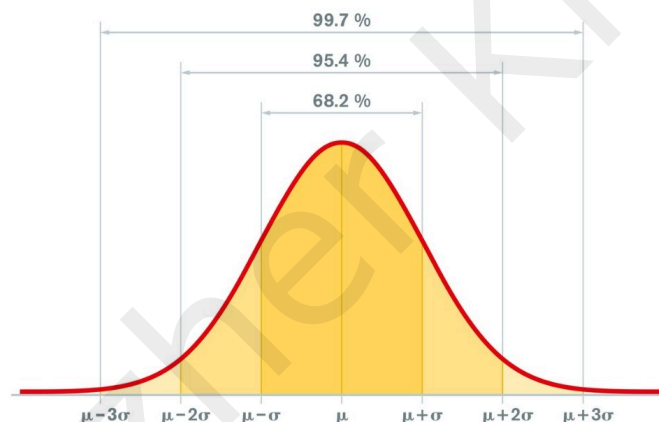
Third quartile (Q3) - Median of the upper half of the data set.

18. What is the difference between Probability and Likelihood?

- Probability attaches to possible results (chances) while Likelihood attaches to the hypothesis. Probability: Only two possibilities >Choose to Bat or not
- Likelihood: Choosing to bat first will depend on Weather Conditions etc.

19. What is the empirical rule?

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



20. What Chi-square test?

- A statistical method is used to find the difference or correlation between the observed and expected categorical variables in the dataset.
- Example: A food delivery company wants to find the relationship between gender, location and food choices of people in India.

21. What is the meaning of selection bias?

Selection bias is a phenomenon that involves the selection of individual or grouped data in a way that is not considered to be random.

22. What is the power of a statistical test?

The power of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. It is equal to 1 minus the probability of a Type II error ( $1 - \beta$ )

22. What are alternatives to traditional A/B testing?

Sequential testing, Bayesian methods

23. What is a one-tailed Z-test or two-tailed Z-test?

- A one-tailed Z-test is used when the alternative hypothesis specifies the direction of the difference between the sample mean and the population mean
- A one-tailed Z-test is used when the alternative hypothesis specifies the direction of the difference between the sample mean and the population mean

24. How can one reduce the p-value in a hypothesis test?

- Increasing sample size
- Strengthening the effect

25. What are the main advantages of K-means clustering?

- Simplicity: K-means is easy to understand and implement.
- Scalability: It can handle large datasets efficiently.
- Speed: It converges relatively quickly compared to other clustering algorithms.
- Interpretability: The cluster assignments and centroids provide meaningful insights.

26.: What are the limitations of K-means clustering?

- Dependency on K: The algorithm requires specifying the number of clusters in advance.
- Sensitive to initial centroids
- Assumes spherical clusters
- Sensitive to outliers

27: How can you evaluate the quality of a K-means clustering solution?

- Inertia: Measure the sum of squared distances between each data point and its centroid. Lower inertia indicates better clustering.
- F1-score

28: What are some alternatives to K-means clustering?

- Hierarchical clustering
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Gaussian Mixture Models (GMM)
- Spectral clustering
- Mean-Shift clustering