STATISTICS INTERVIEW QUESTIONS

AMRIT PRITAM SANGRAMSINGH

https://www.linkedin.com/in/amritpritam/

https://github.com/Project-Amrit

Q 1. What are the most important topics in sta s cs?

Some of the important topics in sta s cs are:

- Measure of central tendency
- Measure of dispersion
- Covariance and Correla on
- Probability Distribu on Func on
- Standardiza on and normaliza on
- Central limit theorem
- Popula on and sample
- Hypothesis Tes ng

Q 2.What is EDA (Exploratory Data Analysis)?

- EDA involves the process of visually and sta s cally analysing data to understand its underlying pa erns distribu ons and rela onships.
- The goal of EDA is to gain insights into the data, iden fy poten al issues, and guide for the data processing steps.

Q 3. What are quan ta ve data and qualita ve data?

Data can be categorized into two main types: quan ta ve data and qualita ve data.

Quan ta ve Data--(numeric)	Qualita ve Data--(categorical)
It is numbers-based, countable, or measur	It is interpreta on-based, descrip ve, and rela ng to language.
It is analyzed using sta s cal analysis.	It is analyzed by grouping the data into categories and themes.
Quan ta ve Data Types: Discrete Data and Con nuous Data.	Qualita ve Data Types: Nominal Data and Ordinal Data.
Ex: Age, Height, Weight, Income, Groupe size, Test score.	Ex: Gender, Marital status, Na ve language, Qualifica ons, Colours.

Q 4. What is the meaning of KPI in sta s cs?

KPI stands for "Key Performance Indicator." KPIs are specific metrics or measures that are used to evaluate and assess the performance of a process, system, or organiza on.

They are used in various fields, including business, finance, healthcare, educa on, and more. The choice of KPIs depends on the goals and objec ves of the organiza on or process being assessed.

By regularly monitoring and analyzing KPIs, organiza ons can iden fy areas of improvement, make data-driven decisions, and measure progress toward their strategic goals.

Q 5. What is the difference between Univariate, Bivariate, and Mul variate Analysis?

Univariate Analysis	Bivariate Analysis	Mul variate Analysis
Involves the examina on of a single variable.	Involves examining the rela onship between two variables.	Involves analyzing mple variables simultaneously.
Analyzing the distribu ons, summary sta scs, and characteriscs.	Focuses on how changes in one variable are associated with changes in another variable.	We can observe, how mple variables interact and influence each other.
Ex: Histograms, Box plots, Mean, Median, Standard devia on.	Ex: Sca er plots, Correla on coefficients, cross-tabula ons.	Ex: Pairplot, Principal Component Analysis (PCA), Factor Analysis.

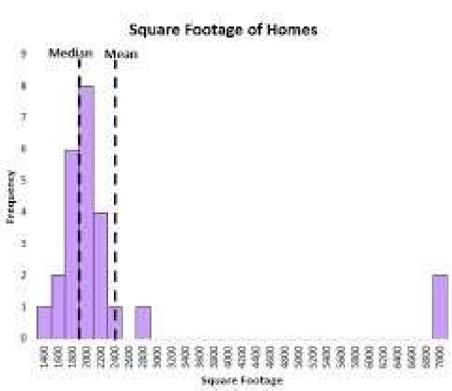
Q 6. How would you approach a dataset that's missing more than 30% of its values?

Choose an appropriate imputa on method based on the nature of the missing data:

- Mean/Median Imputa on:
  Impute missing values with the mean or median of the variable. This is a simple method but may not be suitable for variables with non-normal distribu ons.
- Mode Imputa on:
  Impute missing values with the mode (most frequent value) of the variable for categorical data.
- K-Nearest Neighbors (KNN) Imputa on:
  Impute missing values by finding the nearest neighbors based on other variables.

Q 7. Give an example where the median is a be er measure than the mean.

- The choice between using the median or the mean as a measure of central tendency depends on the distribu on of the data and the specific characteristics of the dataset.
- One common situa on where the median is a be er measure than the mean is when dealing with data that has extreme outliers or a highly skewed distribu on.



- Example:
  Suppose you have the following incomes for ten residents in the town (in thousands of dollars): {25,28,30,32,35,38,40,42,45,5000}
  Now, let's calculate both the mean and the median:
  a. Mean (Average):

$$Mean = \frac{25 + 28 + 30 + 32 + 35 + 38 + 40 + 42 + 45 + 5000}{10} = 488.7$$

  The mean income (488.7) is heavily influenced by the extreme outlier (5000), making it much higher than the typical income of the residents in the town.

  b. Median:
  To find the median, first, arrange the incomes in ascending order:
  {25,28,30,32,35,38,40,42,45,5000}

$$Median = \frac{35 + 38}{2} = 36.5$$

  The median income (36.5) is a be er measure of central tendency in this scenario because it is not affected by extreme values.

Q 8. What is the difference between Descrip ve and Inferen al Sta s cs?

Descrip ve sta s cs and inferen al sta s cs are two fundamental branches of sta s cs that serve different purposes in data analysis. Here's an overview of the key differences between them:

Descrip ve Sta s cs	Inferen al Sta s cs
Used to summarize and describe the main feat or characteris cs of a dataset. They aim to provic clear and concise overview of the data.	Used to make inferences or draw conclusions about a larger popula on based on a sample of data. They involve generalizing from a sample to a popula on.
Typically used at the ini al stage of data analysis to understand the dataset and iden fy pa erns, trends, and important features.	Typically used a er the ini al data explora on (descrip ve sta s cs) when researchers want to make predic ons, test hypotheses, or make statements about a popula on.
Are generally applied to both popula ons and samples. They can be used to summarize data t a complete popula on or from a sample drawn the popula on.	Are focused on making statements or inferences about a popula on based on data from a sample. They involve es ma ng popula on parameters and assessing the uncertainty associated with those es mates.
Examples: Common descrip ve sta s cs include measures of central tendency (e.g., mean, median, mode), measures of dispersion (e.g., range, variance, standard devia on), frequency distribu ons, histograms, and summary tables.	Examples: Common inferen al sta s cal techniques include hypothesis tes ng, confidence intervals, regression analysis, analysis of variance (ANOVA), chi-square tests, and various forms of mul variate analysis.

Q 9. Can you state the method of dispersion of the data in sta s cs?

In sta s cs, measures of dispersion, also known as measures of variability or spread, are used to describe how data points in a dataset are spread out or dispersed. These measures provide valuable insights into the extent to which data values deviate from the central tendency (e.g., the mean) and how variable or homogeneous the dataset is.

Here are some common methods of measuring dispersion:

- Range:
  The range is the simplest measure of dispersion and is calculated as the difference between the maximum and minimum values in a dataset. It provides an idea of the spread of data but is sensive to outliers.
- Variance:
  Variance quan fies the average squared difference between each data point and the mean. It is calculated by taking the average of the squared devia ons from the mean.
- Standard Devia on:
  The standard devia on is the square root of the variance. It provides a measure of dispersion in the same units as the original data, making it easier to interpret.

Q 10. How can we calculate the range of the data?

The range is a measure of the spread or dispersion of data, and it is simply the difference between the maximum and minimum values in the dataset. It represents the span or spread of values from the lowest to the highest within your data.

$$\boxed{Range = Max - Min}$$

- Example:
  Suppose you have a dataset of exam scores for a class of students:
  Exam Scores: [60, 72, 78, 85, 92, 95]

  Range = Max - Min = 95 - 60 = 35
  So, the range of the exam scores in this dataset is 35. This means that the scores vary from a minimum of 60 to a maximum of 95, covering a range of 35 points.

Q 11. Is range sensi ve to outliers?

Yes, the range is sensi ve to outliers. Since it depends solely on the extreme values in the dataset (the maximum and minimum)., outliers, which are extreme values that fall far from the central tendency of the data, can have a significant impact on the range.

Q 12. What are the scenarios where outliers are kept in the data?

Outliers may be kept in data when they represent important and meaningful informa on, unusual events, or rare occurrences that are relevant to the analysis, such as detec ng anomalies, understanding extreme behavior, or studying unique cases.

Q 13. What is the meaning of standard devia on?

- The standard devia on is a sta s cal measure that quan fies the amount of varia on or dispersion in a set of data values.
- It provides insight into how spread out or clustered the data points are around the mean (average) value.
- In other words, the standard devia on helps us understand the extent to which individual data points deviate from the mean.
- The standard devia on is calculated as the square root of variance by determining each data point's devia on rela ve to the mean.

$$\sigma = \sqrt{\frac{\sum(x - \bar{x}\ )}{N}}$$

Q 14. What is Bessel's correc on?

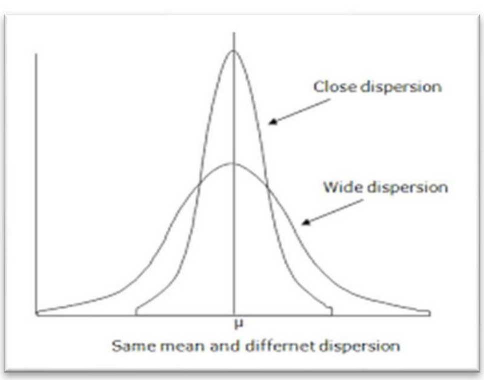
Bessel's correon is a sta scal adjustment made to the formula for calcula ng the sample variance and sample standard devia on. It is used to provide a more accurate es mate of the popula on variance and standard devia on when working with a sample from a larger popula on.

The key idea behind Bessel's correc on is that when you calculate the variance or standard devia on using sample data (rather than data from the en re popula on), you tend to underes mate the true popula on variance or standard devia on. This underes ma on occurs because you are basing your calcula ons on a smaller subset of the data.

Bessel's correon adjusts for this underes ma on by dividing the sum of squared differences from the mean by (n - 1), where "n" is the sample size. In contrast, when calcula ng popula on variance and standard devia on, you divide by "n" (the actual popula on size). By using (n - 1) instead of "n" in the formula, Bessel's correc on increases the calculated variance and standard devia on slightly, making them more representa ve of the popula on.

Q 15. What do you understand about a spread out and concentrated curve?

In the context of data distribu ons and sta s cs, these terms describe the degree of variability or dispersion in the data.



Same mean and differnet dispersion

Spread Out Curve (Wider Dispersion)	Concentrated Curve (Narrower Dispersion)
A spread-out curve or distribu on typically a larger spread or range of values. This me that the data points are more spread out f each other.	A concentrated curve or distribu on typically has a smaller spread or range of values. This means that the data points are closer together.
It is associated with a higher standard devia on and a larger range or interqu le range (IQR).	It is associated with a lower standard devia on and a smaller range or interqu le range (IQR).
In graphical representa ons, it o en results wider or fla er distribu on with a larger spr of data points.	In graphical representa ons, it o en results in a narrower, taller distribu on with data points clustered closely together.
Example: A dataset of income levels for a diverse popula on, where some individuals have very high incomes, and others have very low incomes, crea ng a wide spread.	Example: A dataset of test scores for a group of students who all scored very close to each other, crea ng a concentrated distribu on.

Q 16. Can you calculate the coefficient of varia on?

- The coefficient of varia on (CV) is a measure of rela ve variability and is calculated as the ra o of the standard devia on (σ) to the mean (μ) of a dataset. It is o en expressed as a percentage to make it more interpretable.
- The formula for calcula ng the coefficient of varia on is as follows:

$$\text{Coef icient of Variation}(CV) = \frac{\sigma}{\mu} \times 100$$

  Where:

  CV= Coefficient of varia on, σ= standard devia on of the dataset, μ= mean of the dataset.

- The coefficient of varia on is par cularly useful when you want to compare the rela ve variability of two or more datasets with different units of measurement or different means. It provides a standardized way to express the dispersion of data rela ve to the mean, making it easier to compare datasets of varying scales.
- Example:
  Test Scores: Consider two classes, Class A and Class B, with test scores. Here are the sta s cs for both classes:
  Class A: Mean Score = 85, Standard Devia on = 10
  Class B: Mean Score = 90, Standard Devia on = 8

  Now, let's calculate the coefficient of varia on for both classes:
  For Class A: CV= (σ/ μ) * 100 = (10 /85) * 100 ≈ 11.76%
  For Class B: CV= (σ/ μ) * 100= (8 /90) * 100% ≈ 8.89%

  In this example, Class A has a higher coefficient of varia on (11.76%) compared to Class B (8.89%). This suggests that the test scores in Class A are more variable rela ve to their mean compared to Class B.

Q 17. What is meant by mean imputa on for missing data? Why is it bad?

Mean imputa on is a method for handling missing data by replacing missing values with the mean (average) value of the available data in the same column.

Disadvantages of Mean Imputa on:

- Bias Introduc on:
  Mean imputa on can introduce bias into the dataset.
- Loss of Variability:
  Impu ng missing values with the mean reduces the variability of the data because all imputed values are the same.
- Disregards Data Pa erns:
  Mean imputa on does not take into account any underlying pa erns or rela onships in the data. It treats all missing values as if they were independent of other variables or condi ons, which may not be the case.
- Impact on Model Performance:
  In machine learning, mean imputa on can nega vely impact model performance, especially when missing values are related to the target variable or when they carry important informa on. It can lead to inaccurate predic ons and reduced model effec veness.
- Imputa on of Categorical Data:
  Mean imputa on is primarily suitable for numerical data. When dealing with categorical data, other imputa on methods like mode imputa on (replacing missing values with the mode, or most common category) are more appropriate.

Q 18. What is the benefit of using box plots?

Box plots, are valuable graphical tools in sta s cs and data analysis that provide several benefits for visualizing and summarizing data distribu ons.

Here are some of the key benefits of using box plots:

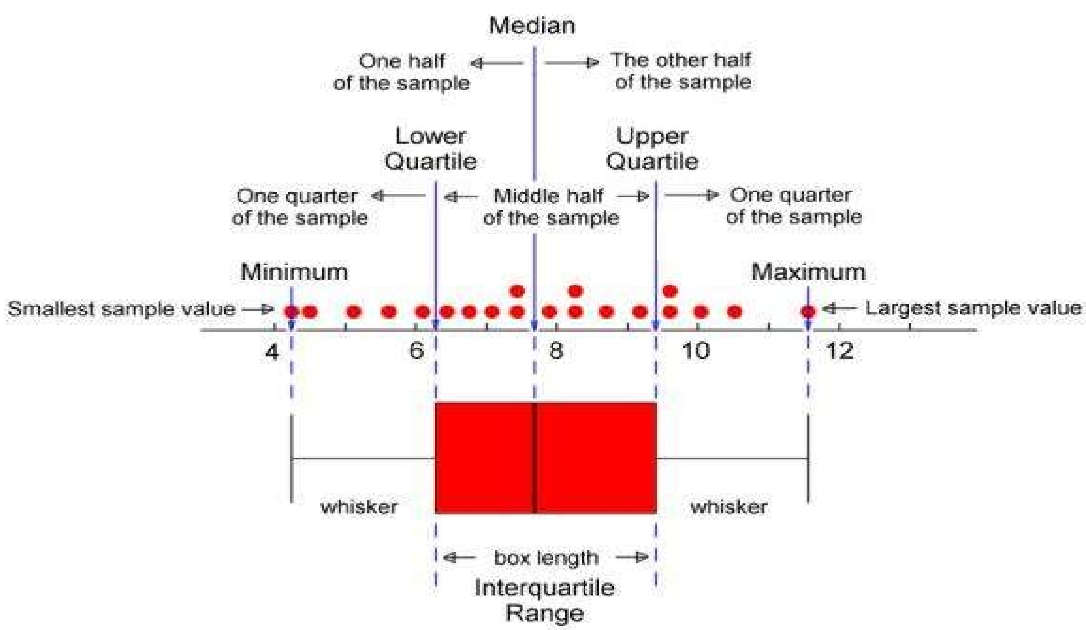
- Summary of Data Distribu on
- Iden fica on of Outliers
- Comparison of Mul ple Groups
- Detec on of Skewness
- Visualiza on of Quar les
- Robustness to Outliers
- Ease of Interpreta on
- Data Quality Assessment

Q 19. What is the meaning of the five-number summary in Sta s cs?

The five-number summary consists of five key values that help describe the central tendency, spread, and shape of a dataset.

The five values in the five-number summary are:

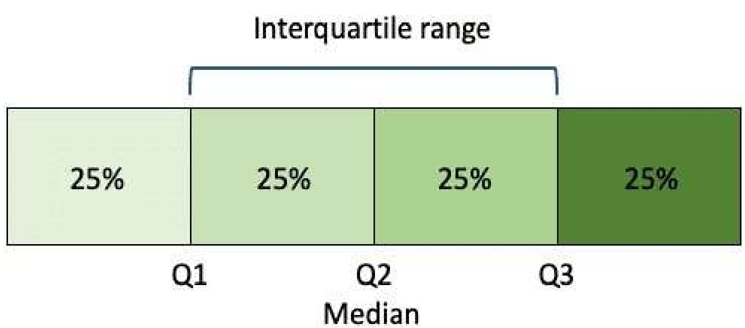
a. Minimum (Min): This is the smallest value in the dataset, represen ng the lowest data point. It gives you an idea of the floor or lower boundary of the data.
b. First Quar le (Q1): The first quar le, also known as the lower quar le, is the value below which 25% of the data falls. It marks the 25th percen le of the dataset and represents the lower boundary of the middle 50% of the data.
c. Median (Q2): The median, or the second quar le, is the middle value of the dataset when it is sorted in ascending order. It divides the data into two equal halves, with 50% of the data falling below it and 50% above it. The median represents the central tendency of the data.
d. Third Quar le (Q3): The third quar le, also known as the upper quar le, is the value below which 75% of the data falls. It marks the 75th percen le of the dataset and represents the upper boundary of the middle 50% of the data.
e. Maximum (Max): This is the largest value in the dataset, represen ng the highest data point. It gives you an idea of the ceiling or upper boundary of the data.



The five-number summary is o en used to create box plots (box-and-whisker plots), which provide a visual representa on of these five summary sta s cs. Box plots are helpful for understanding the spread, central tendency, and presence of outliers in a dataset. The box in the plot represents the interquar le range (IQR), which is the range between the first quar le (Q1) and the third quar le (Q3), while the whiskers extend to the minimum and maximum values, indica ng the range of the data.

Q 20. What is the difference between the 1st quar le, the 2nd quar le and the 3rd quar le?

- The 1st quar le (Q1) is the value below which 25% of the data falls. It represents the lower boundary of the middle 50% of the data.
- The 2nd quar le (Q2), also known as the median, is the middle value of the data when it's sorted. It divides the data into two equal halves, with 50% below it and 50% above it.
- The 3rd quar le (Q3) is the value below which 75% of the data falls. It represents the upper boundary of the middle 50% of the data.



Think of quar les as dividing your data into four equal parts, with Q1 marking the 25% point, Q2 (median) marking the 50% point, and Q3 marking the 75% point. These values help you understand where the data is concentrated and how it's spread out.

Q 21. What is the difference between percent and percen le?

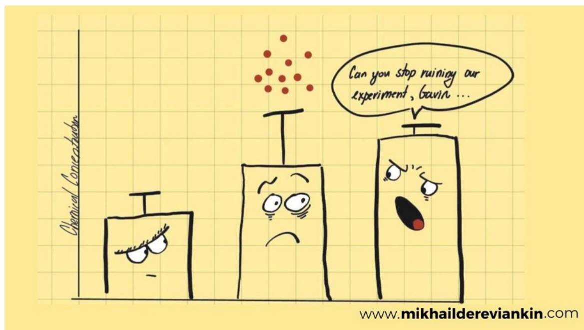
Percent and percen le are related concepts in sta s cs, but they have dis nct meanings.

Percent	Percen le
Percent is a unit of measurement denoted the symbol "%."	Percen le is a sta s cal concept used to describe a specific posi on or loca on within a dataset.
It represents a propor on or frac on of a whole, divided by 100. In other words, when you express a quan ty as a percentage, you are dividing it by 100.	It represents the value below which a given percentage of the data falls. Percen les are used to understand the distribu on of data and iden fy how a par cular data point ranks in comparison to others.
For example, 25 percent (25%) is equivalen 0.25 or 25/100. It means 25 out of every 10 or one-quarter of the whole.	For example, the 25th percen le (also known as the first quar le, Q1) is the value below which 25% of the data points in a dataset lie.

Q 22. What is an Outlier?

- An outlier is a data point that significantly deviates from the rest of the data in a dataset.
- In other words, it's an observa on that is unusually distant from other observa ons in the dataset.
- Outliers can be either excep onally high values (posi ve outliers) or excep onally low values (nega ve outliers).

Q 23. What is the impact of outliers in a dataset?



1. Nega ve Impacts:

- <u>Influence on Measures of Central Tendency</u>:
  A single extreme outlier can pull the mean in its direc on, making it unrepresenta ve of the majority of the data.
- <u>Impact on Dispersion Measures</u>:
  The presence of outliers can inflate the measures like the standard devia on and the interquar le range (IQR), making them larger than they would be without outliers.
- <u>Skewing Data Distribu ons</u>:
  Posi ve outliers can result in right-skewed distribu ons, while nega ve outliers can result in le -skewed distribu ons. This can affect the interpreta on of the data.
- <u>Misleading Summary Sta s cs</u>:
  Outliers can distort the interpreta on of summary sta s cs.
- <u>Impact on Hypothesis Tes ng</u>:
  Outliers can affect the results of hypothesis tests. They can lead to incorrect conclusions, such as detec ng significant differences when none exist or failing to detect real differences when outliers mask them.

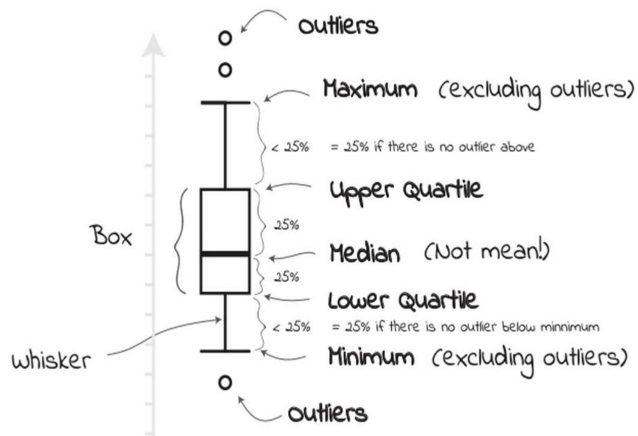
2. Posi ve Impacts:

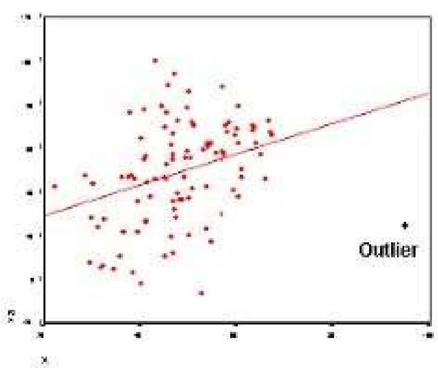
- <u>Detec on of Anomalies</u>:
  Outliers can signal the presence of anomalies or rare events in a dataset. Iden fying these anomalies can be valuable in various fields, including fraud detec on, quality control, and outlier detec on in scien fic experiments.
- <u>Robust Modeling</u>:
  In some cases, outliers can be genuine observa ons that are important to model. For example, in financial modeling, extreme stock price movements may contain valuable informa on for predic ng market trends.

Q 24. Men on methods to screen for outliers in a dataset.

There are several methods to screen for outliers in a dataset, ranging from graphical techniques to sta s cal tests. Here are some commonly used methods:

- Box Plots (Box-and-Whisker Plots
  Box plots provide a visual representa on of the distribu on o data, including the iden fica on of poten al outliers. In a box plot, outliers are typically shown as individual data points beyond the whiskers of the plot.



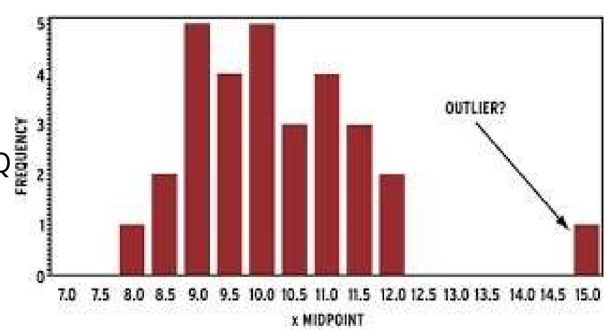


- Sca erplots:
  Sca erplots are parcularly useful for iden fying outliers in bivariate or mvaliriate data. Outliers can appear as data points that are far from the main cluster of points in the sca erplot.

- Z-Scores:
  Z-scores (standard scores) measure how many standard devia ons a data point is away from the mean. Data points with high absolute Z-scores (typically greater than 2 or 3) are o en considered poten al outliers.

- IQR (Interquale Range) Method:
  The IQR method involves calcula ng the interquarange (IQR = Q3 - Q1) and then iden fying values that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR as poten al outliers.

- Visual Inspecon:
  Some mes, simple visual inspecon of the data through histograms, Q plots (quan le-quan le plots), or other visualiza on techniques can reveal the presence of outliers.



It's important to note that the choice of outlier detec on method should be guided by the characteriscs of your data and the specific goals of your analysis.

Q 25. How you can handle outliers in the datasets.

Handling outliers in datasets is an important step in data preprocessing to ensure that they do not unduly influence the results of your analysis or modeling. The approach you choose for handling outliers depends on the nature of the data, the context of the analysis, and your specific objec ves. Here are several methods for handling outliers:

- Data Trunca on or Removal:
  One common approach is to simply remove outliers from the dataset. This should be done cau ously, especially if the outliers represent valid and important observa ons. Removing outliers is appropriate when they are likely the result of data entry errors or measurement errors.
- Data Transforma on:
  Transforming the data can be a useful way to mi gate the impact of outliers. Common transforma ons include logarithmic, square root, or inverse transforma ons. These transforma ons tend to compress the range of extreme values.
- Winsoriza on:
  Winsoriza on involves capping or limi ng extreme values by replacing them with a specified percen le value. For example, you might replace values above the 95th percen le with the value at the 95th percen le.
- Imputa on:
  For missing values that are not extreme outliers, you can impute them using various methods, such as mean imputa on, median imputa on, or more advanced techniques like regression imputa on.
- Robust Sta s cs:
  Using robust sta s cal methods that are less sensi ve to outliers can be an effec ve approach. For example, replacing the mean with the median and using the interquar le range (IQR) instead of the standard devia on can make sta s cal analysis more robust.
- Model-Based Approaches:
  In predic ve modeling, consider using algorithms that are less sensi ve to outliers, such as robust regression methods or ensemble methods like random forests, which can handle outliers be er than linear regression.
- Domain Knowledge:
  Rely on domain knowledge to understand the context of the outliers. Some mes, what appears as an outlier may be a valid and important data point. Consult with domain experts to determine the appropriateness of handling outliers.
- Repor ng and Transparency:
  Regardless of the approach chosen, it's crucial to transparently document how outliers were handled in the analysis to ensure the reproducibility and interpretability of your results.

Q 26. How to calculate range and interquar le range?

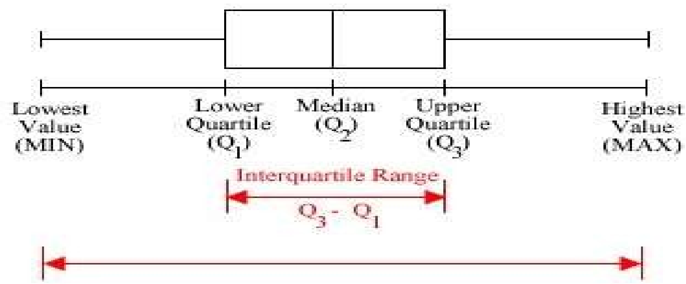
Calcula ng the range and interquar le range (IQR) is a straigh orward process involving the use of basic sta s cal formulas. Here's how to calculate both the range and the IQR:

- Range:
  The range is the simplest measure of spread in a dataset. It is the difference between the maximum and minimum values in the dataset.

$$\boxed{Range = Maximum\,Value - Minimum\,Value}$$

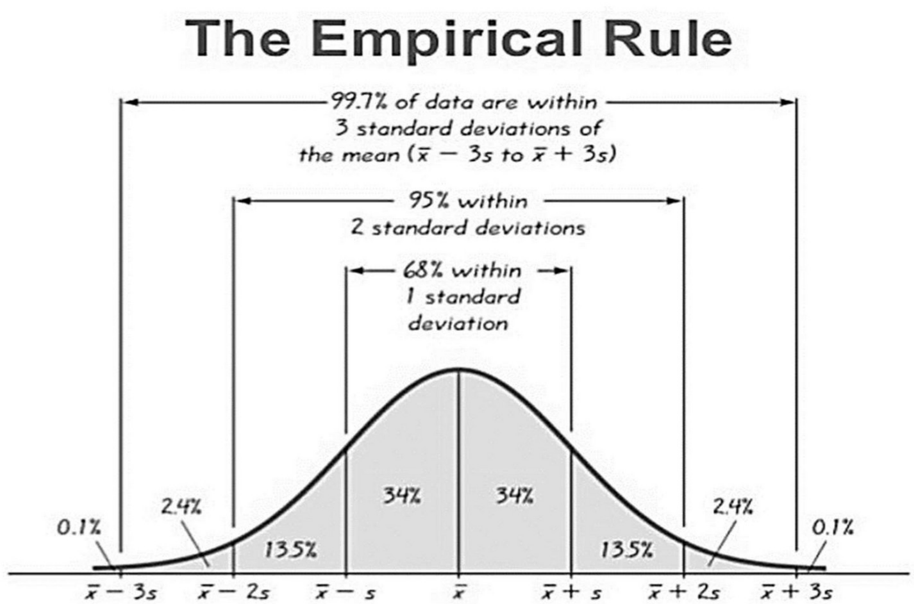
- Interquar le Range (IQR):
  The interquar le range (IQR) is a measure of the spread or variability of the middle 50% of the data. It is calculated as the difference between the third quar le (Q3) and the first quar le (Q1) of the dataset.

$$\boxed{IQR = Q3 - Q1}$$



Range

Q 27. What is the empirical rule?



The empirical rule, also known as the 68-95-99.7 rule or the three-sigma rule, is a sta s cal guideline used to describe the approximate distribu on of data in a normal distribu on (bell-shaped) curve. It provides insights into how data values are distributed around the mean (average) in a normally distributed dataset.

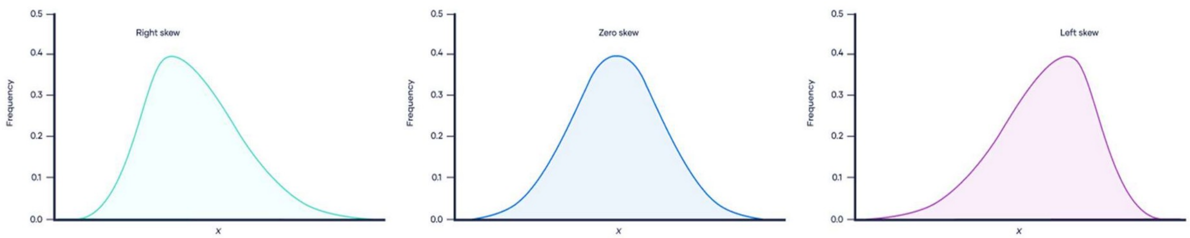
The empirical rule states that:

- Approximately 68% of the data falls within one standard devia on of the mean.
- Approximately 95% of the data falls within two standard devia ons of the mean.
- Approximately 99.7% of the data falls within three standard devia ons of the mean.

Q 28. What is skewness?

Skewness is a measure of the asymmetry of a distribu on. A distribu on is asymmetrical when its le and right side are not mirror images.

A distribu on can have right (or posi ve), le (or nega ve), or zero skewness. A right-skewed distribu on is longer on the right side of its peak, and a le -skewed distribu on is longer on the le side of its peak:



Q29. What are the different measures of Skewness?

There are different measures of skewness used to quan fy this property. The three most common measures of skewness are:

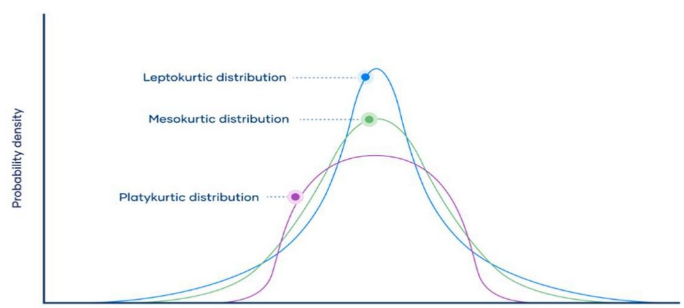
- Pearson's First Coefficient of Skewness (or Moment Skewness)
- Fisher-Pearson Standardized Moment Coefficient of Skewness (or Sample Skewness)
- Bowley's Coefficient of Skewness (or Quar le Skewness)

Q 30. What is kurtosis?

Kurtosis is a sta s cal measure that quan fies the "tailedness" or "peakedness" of the probability distribu on of a real-valued random variable. In other words, it tells you how the data is distributed with respect to the tails (extreme values) and the central peak of the distribu on.

Kurtosis classifica ons based on the shape of the data distribu on:

- Mesokurc
- Leptokurc
- Platykurc



Q 31. Where are long-tailed distribu ons used?

Long-tailed distribu ons are used in various fields and applica ons where the presence of rare but significant events, extreme values, or outliers is of par cular interest or importance. Here are some areas where long-tailed distribu ons are commonly used:

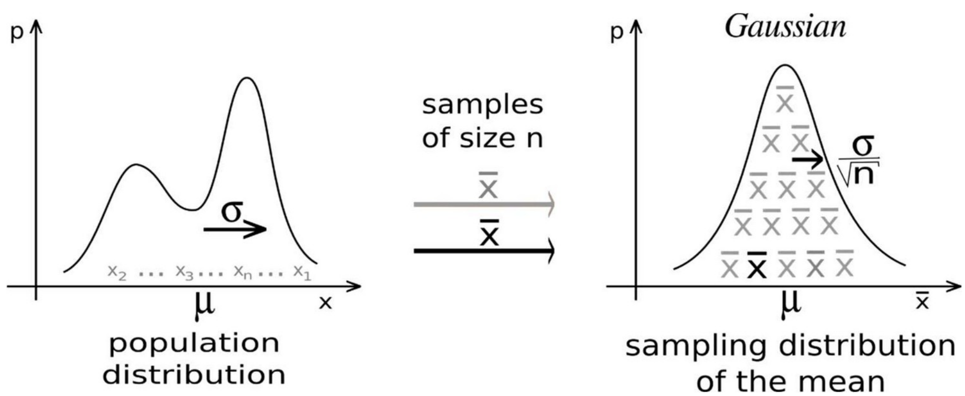
- Finance and Risk Management:
  Long-tailed distribu ons are frequently used to model asset returns, market vola lity, and financial risk.

They are employed in risk assessment and por olio management to account for extreme events like market crashes or large investment gains.

- Insurance:
  Insurance companies use long-tailed distribu ons to model insurance claims. These distribu ons account for rare but costly events, such as natural disasters or large medical claims.
- Environmental Science:
  In studies related to natural disasters, such as hurricanes, earthquakes, and floods, long-tailed distribu ons are used to es mate the likelihood of extreme events occurring.
- Epidemiology:
  Epidemiologists may use long-tailed distribu ons to model the spread of infec ous diseases, as they account for sporadic outbreaks or superspreading events.

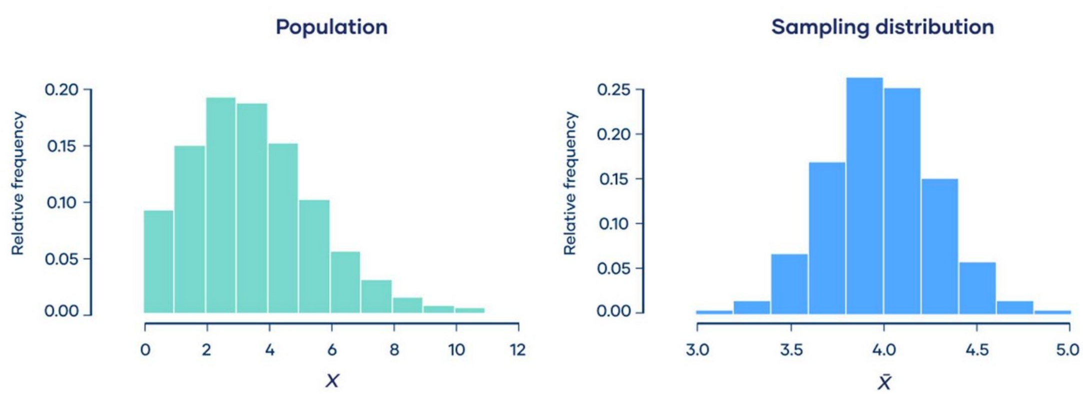
Q 32. What is the central limit theorem?

In probability theory, the central limit theorem (CLT) states that the distribu on of a sample variable approximates a normal distribu on (i.e., a “bell curve”) as the sample size becomes larger i.e., n>=30, assuming that all samples are iden cal in size, and regardless of the popula on's actual distribu on shape.



Q 33. Can you give an example to denote the working of the central limit theorem?

A popula on follows a Poisson distribu on (le image). If we take 10,000 samples from the popula on, each with a sample size of 50, the sample means follow a normal distribu on, as predicted by the central limit theorem (right image).



Q 34. What general condi ons must be sa sfied for the central limit theorem to hold?

For the Central Limit Theorem (CLT) to hold:

- Random Sampling:
  Data must be randomly selected from the popula on.
- Independence:
  Data points must be independent of each other.
- Sufficient Sample Size:
  The sample size should generally be greater than or equal to 30.
- Finite Variance:
  The popula on should have a finite variance.
- Iden cal Distribu on:
  Ideally, data should come from a popula on with the same distribu on.

The CLT states that as sample size increases, sample means approach a normal distribu on.

Q 35. What is the meaning of selec on bias?

Selec on bias is the bias that occurs during the sampling of data. This kind of bias occurs when a sample is not representa ve of the popula on, which is going to be analyzed in a clinical study.

Q 36. What are the types of selec on bias in sta s cs?

There are many types of selec on bias as shown below:

- Observer selec on
- A ri on
- Protopathic bias
- Time intervals
- Sampling bias

Q 37. What is the probability of throwing two fair dice when the sum is 8?

- To find the probability of throwing two fair dice and ge ng a sum of 8, we need to determine how many favorable outcomes (sums of 8) there are and divide that by the total number of possible outcomes when rolling two dice.
- Each die has 6 sides, numbered from 1 to 6. When you roll two dice, there are 6 x 6 = 36 possible outcomes because each die has 6 possible outcomes, and they are independent.
- Now, let's calculate the favorable outcomes where the sum is 8:
  (2, 6), (3, 5), (4, 4), (5, 3), (6, 2) ---- There are 5 favorable outcomes.
- So, the probability of ge ng a sum of 8 when rolling two fair dice is:

$$Probability = \frac{(\qquad)}{(\qquad)} = -$$

Therefore, the probability is 5/36.

Q 38. What are the different types of Probability Distribu on used in Data Science?

Probability distribu ons are mathema cal func ons that describe the likelihood of different outcomes or events in a random process. There are several types of probability distribu ons, each with its own characteris cs and applica ons.

There are two main types of probability distribu ons: Discrete and Con nuous.

1. Discrete Probability Distribu ons:
   In a discrete probability distribu on, the random variable can only take on dis nct, separate values, o en integers. Common examples of discrete probability distribu ons include:
   a. Bernoulli Distribu on
   b. Binomial Distribu on
   c. Poisson Distribu on
2. Con nuous Probability Distribu ons:
   In a con nuous probability distribu on, the random variable can take on any value within a specified range. Common examples of con nuous probability distribu ons include:
   a. Normal Distribu on (Gaussian Distribu on)
   b. Uniform Distribu on
   c. Log-Normal Distribu on
   d. Power Law
   e. Pareto Distribu on

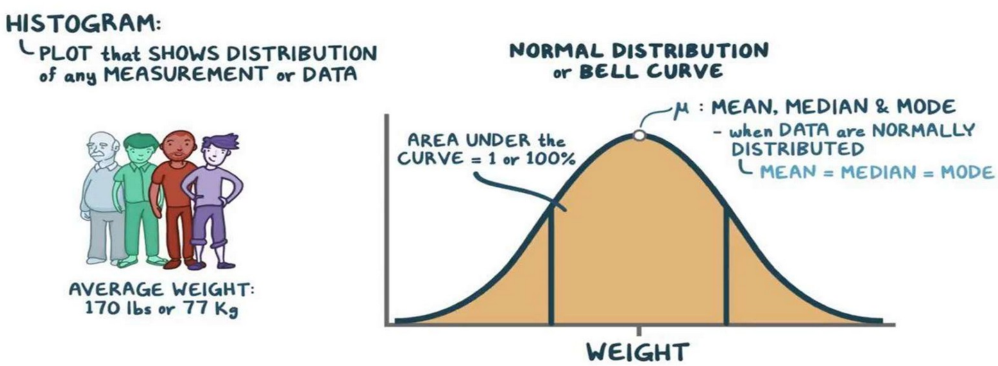
Q 39. What do you understand by the term Normal/Gaussian/bell-curve distribu on?

A normal distribu on, also known as a Gaussian distribu on or a bell curve, is a fundamental sta s cal concept in probability theory and sta s cs. It is a con nuous probability distribu on that is characterized by a specific shape of its probability density func on (PDF), which has the following key proper es:

- Symmetry: The normal distribu on is symmetric, meaning that it is centred around a single peak, and the le and right tails are mirror images of each other. The mean, median, and mode of a normal distribu on are all equal and located at the centre of the distribu on.
- Bell-shaped: The PDF of a normal distribu on has a bell-shaped curve, with the highest point (peak) at the mean value and gradually decreasing probabili es as you move away from the mean in either direc on.
- Mean and Standard Devia on: The normal distribu on is fully characterized by two parameters: the mean (μ) and the standard devia on (σ). The mean represents the centre of the distribu on, while the standard devia on controls the spread or dispersion of the data. Larger standard devia ons result in wider distribu ons.
- Empirical Rule: The normal distribu on follows the empirical rule (also known as the 68-95-99.7 rule), which states that approximately:
   a. About 68% of the data falls within one standard devia on of the mean.
   b. About 95% of the data falls within two standard devia ons of the mean.
   c. About 99.7% of the data falls within three standard devia ons of the mean.

- Con nuous: The normal distribu on is a con nuous probability distribu on, which means that it can take on an infinite number of values within its range. There are no gaps or discon nui es in the distribu on.

Many natural phenomena, such as weights, heights and IQ scores, approximate a normal distribu on. It is also fundamental in hypothesis tes ng and sta s cal modeling.



Q 40. Can you state the formula for normal distribu on?

This formula represents the bell-shaped curve of the normal distribu on, which is symmetric around the mean (μ) and characterized by its mean and standard devia on. It describes the probability of observing a specific value (x) in a normally distributed dataset.

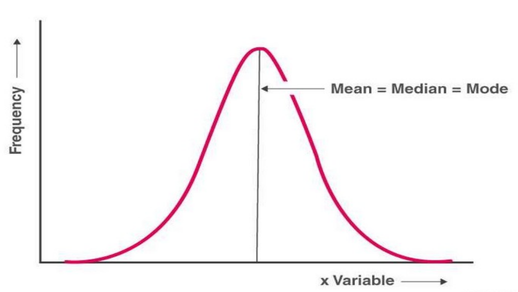
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(\quad)}{}}$$

Where:

- $f(x)$ is the probability density func on at a given value of
- μ is the mean of the normal distribu on.
- σ is the standard devia on of the normal distribu on.
- π is the mathema cal constant pi (approximately 3.14159).
- e is the base of the natural logarithm (approximately 2.71828).

Q 41. What is the rela onship between mean and median in a normal distribu on?

In a normal distribu on, the mean and median are equal and coincide at the centre of the distribu on.



Q 42. What are some of the proper es of a normal distribu on?

A normal distribu on, also known as a Gaussian distribu on or bell curve, has several key proper es:

- Bell-Shaped Curve: The distribu on looks like a symmetrical bell, with a peak in the middle and tails that taper off gradually on both sides.
- Symmetry: It's perfectly symmetric, meaning if you fold the curve in half, one side is a mirror image of the other.
- Central Peak: The highest point (peak) of the curve is at the mean, which is also the middle of the data.
- Mean = Median = Mode: The mean (average), median (middle value), and mode (most common value) are all at the same point in the middle of the distribu on.
- Tails Extend to Infinity: The tails of the curve stretch infinitely in both direc ons, but they get closer and closer to the horizontal axis as they go farther from the mean.
- Standard Devia on Controls Spread: The width of the bell curve is determined by the standard devia on. A larger standard devia on makes the curve wider, and a smaller one makes it narrower.
- Empirical Rule: This rule helps you es mate where data points are likely to be within the distribu on. It's based on the 68-95-99.7 rule. Approximately 68% of the data falls within one standard devia on of the mean, about 95% falls within two standard devia ons, and roughly 99.7% falls within three standard devia ons.
- Used in Many Real-Life Situa ons: The normal distribu on is commonly seen in nature and in human-made systems, including things like height measurements, IQ scores, and errors in manufacturing.
- Easy for Sta s cal Analysis: Because of its well-defined proper es, the normal distribu on is o en used in sta s cs for modeling and making predic ons about data.

Q 43. What is the assump on of normality?

The assump on of normality in sta s cs is the idea that data or residuals in a sta s cal analysis should follow a bell-shaped, symmetric, and con nuous probability distribu on called the normal distribu on.

Q 44. How to convert normal distribu on to standard normal distribu on?

Conver ng a normal distribu on to a standard normal distribu on involves a process called "standardiza on" or "normaliza on". This process transforms the values from the original normal distribu on into equivalent values that follow a standard normal distribu on with a mean of 0 and a standard devia on of 1.

Here are the steps to convert a value from a normal distribu on to a standard normal distribu on:

- Determine the Mean and Standard Devia on of the Original Normal Distribu on:
  Iden fy the mean (μ) and standard devia on (σ) of the original normal distribu on.
- Calculate the Z-Score:
  The Z-score (also known as the standard score) measures how many standard devia ons a par cular value is from the mean in the original distribu on.

Calculate the Z-score using the formula:

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

where:
Z is the Z-score.
X is the value from the original distribu on that you want to convert.
μ is the mean of the original distribu on.
σ is the standard devia on of the original distribu on.

- The Resul ng Z-Score Represents the Standard Normal Distribu on:
  The Z-score you calculate in step 2 represents the equivalent value in a standard normal distribu on.

By following these steps, you can convert any value from a normal distribu on into a corresponding value in the standard normal distribu on. This conversion is useful for performing standard normal distribu on-based calcula ons and making comparisons between data from different normal distribu ons.

Q 45. Can you tell me the range of the values in standard normal distribu on?

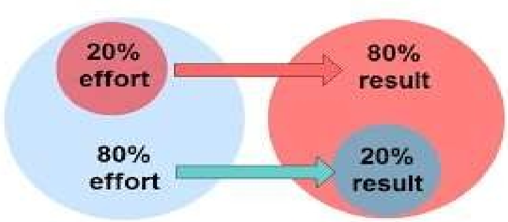
In a standard normal distribu on, also known as the standard normal or Z-distribu on, the range of possible values extends from nega ve infinity (−∞) to posi ve infinity (+∞).

However, it's important to note that while the range of possible values is theore cally infinite, the vast majority of values in a standard normal distribu on are concentrated within a rela vely narrow range around the mean, which is 0. The distribu on is bell-shaped, and as you move away from the mean in either direc on, the probability density of values decreases. The tails of the distribu on extend to infinity, but they become increasingly rare as you move farther from the mean.

Sta s cally, most of the values in a standard normal distribu on fall within a few standard devia ons of the mean. Approximately:

This means that the values within the range of roughly -3 to +3 standard devia ons from the mean cover the vast majority of observa ons in a standard normal distribu on. Beyond this range, the probability of observing a value becomes extremely low.

Q 46. What is the Pareto principle?



- The Pareto Principle, also known as the 80/20 Rule or the Law of the Vital Few, is principle named a er the Italian econom Vilfredo Pareto.
- It suggests that, in many situa ons, a sm percentage of causes or inputs is responsible for a large percentage of th results or outputs.
- In its simplest form, the Pareto Principle states that roughly 80% of the effects come from 20% of the causes.

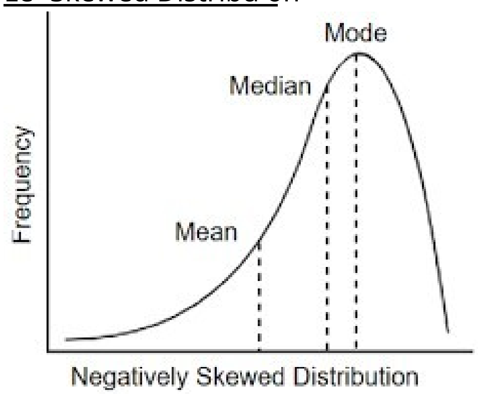
Q 47. What are le -skewed and right-skewed distribu ons?

Le -skewed and right-skewed distribu ons, also known as nega vely skewed and posi vely skewed distribu ons, are types of asymmetric distribu ons in sta s cs. They describe the shape of the distribu on of data points in a dataset.

1. Le -Skewed (Nega vely Skewed) Distribu on:

- Le -skewed distribu ons have a longer tail on the le (or nega ve) side of the distribu on.
- The peak of the distribu on (mode) is typically located to the right of the centre.
- The mean (average) is typically less than the median.
- In a le -skewed distribu on, the data is concentrated on the right side and tails off to the le .

Le -Skewed Distribu on

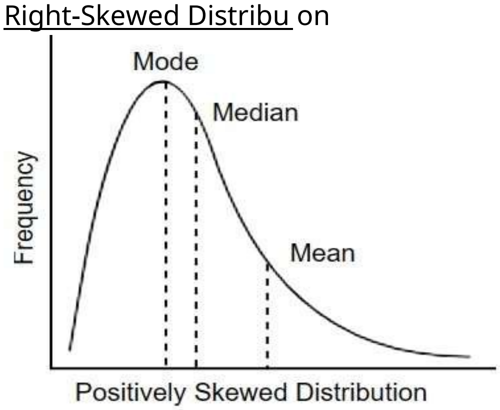


Example: The distribu on of ages at re rement may be le -skewed, as most people re re around a certain age, but very few re re at a younger age.

2. Right-Skewed (Posi vely Skewed) Distribu on:

- Right-skewed distribu ons have a longer tail on the right (or posi ve) side of the distribu on.
- The peak of the distribu on (mode) is typically located to the le of the centre.
- The mean (average) is typically greater than the median.
- In a right-skewed distribu on, the data is concentrated on the le side and tails off to the right.

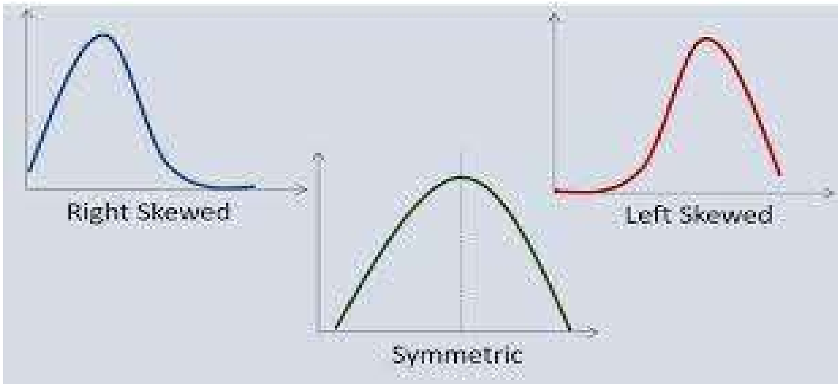
Right-Skewed Distribu on



Example: The distribu on of income in a popula on may be right-skewed, as most people earn average incomes, but a few earn very high incomes.

Skewness is a measure used to quan fy the degree of asymmetry in a distribu on.

- A posi ve skewness value indicates right-skewness.
- A nega ve skewness value indicates le -skewness.
- A skewness of 0 indicates a perfectly symmetrical distribu on.



Understanding the skewness of a dataset is essen al in sta s cs because it can affect the choice of appropriate sta s cal analyses and modeling techniques. Le -skewed and right-skewed distribu ons o en require different approaches for analysis and interpreta on.

Q 48. If a distribu on is skewed to the right and has a median of 20, will the mean be greater than or less than 20?

If a distribu on is skewed to the right (posi vely skewed) and has a median of 20, then the mean will typically be greater than 20.

<u>In a posi vely skewed distribu on</u>:

- The tail of the distribu on extends to the right, meaning there are some rela vely large values that pull the mean in that direc on.
- The median, being the middle value, is less affected by extreme values in the tail, so it is typically lower than the mean in a posi vely skewed distribu on.

Q 49. Given a le -skewed distribu on that has a median of 60, what conclusions can we draw about the mean and the mode of the data?

In a le -skewed (nega vely skewed) distribu on with a median of 60:

<u>Mean, Median and Mode Rela onship</u>:

- Since the distribu on is le -skewed, it means that the tail of the distribu on is on the le side, and there are some rela vely small values that are pulling the mean in that direc on.
- The median, being the middle value, is less affected by extreme values in the tail. In a le -skewed distribu on, the median is typically greater than the mean.
- In a le -skewed distribu on, the mode is typically greater than the median and the mean. It is o en closer to the peak of the distribu on, which is located to the right of the centre.

In summary, you can conclude that in a le -skewed distribu on with a median of 60, the mean is likely less than 60, and the mode is likely greater than 60.

Q 50. Imagine that Jeremy took part in an examina on. The test has a mean score of 160, and it has a standard devia on of 15. If Jeremy’s z-score is 1.20, what would be his score on the test?

To find Jeremy's score on the test given his z-score, you can use the formula for calcula ng a score from a z-score in a normal distribu on:

$$\boxed{Z = \frac{(\quad)}{\quad} \quad \leftrightarrow \quad Z \times \sigma = X - \mu \quad \leftrightarrow \quad X = (Z \times \sigma) + \mu}$$

<u>In this case</u>:

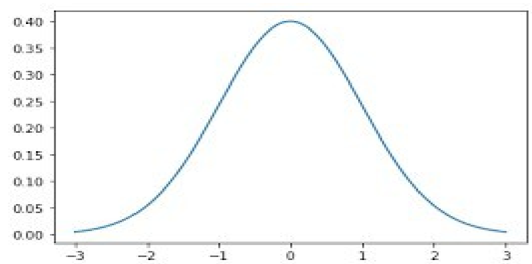
Z=1.20 (Jeremy's z-score), σ=15 (standard devia on), μ=160 (mean)

$X = (1.20 \times 15) + 160 = 178$

So, Jeremy's score on the test would be 178.

Q 51. The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False?

True. The standard normal curve, also know the standard normal distribu on or the Z-distribu on, is a specific type of normal distribu on with a mean (average) of 0 and a standard devia on of 1.



Q 52. What is the meaning of covariance?

Covariance is a measure of the rela onship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable.

Covariance can help you understand whether two variables tend to move in the same direc on (posi ve covariance) or in opposite direc ons (nega ve covariance).

Q 53. Can you tell me the difference between unimodal bimodal and bell-shaped curves?

Unimodal, bimodal, and bell-shaped curves are terms used to describe different characteris cs of the shape of a data distribu on:

1. Unimodal Curve:
- Defini on: A unimodal curve represents a data distribu on with a single peak or mode, meaning that there is one value around which the data cluster the most.
- Shape: Unimodal distribu ons are typically symmetric or asymmetric but have only one primary peak.

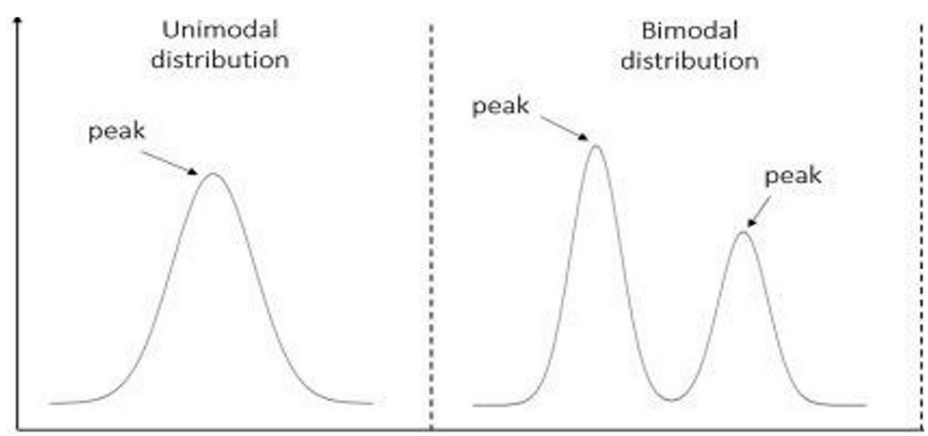
Examples: A normal distribu on, where data is symmetrically distributed around the mean, is a classic example of a unimodal curve. Other unimodal distribu ons can be skewed to the le (nega vely skewed) or to the right (posi vely skewed).

2. Bimodal Curve:
- Defini on: A bimodal curve represents a data distribu on with two dis nct peaks or modes, indica ng that there are two values around which the data cluster the most.
- Shape: Bimodal distribu ons have two primary peaks separated by a trough or dip in the distribu on.

Examples: The distribu on of test scores in a classroom with two dis nct groups of high achievers and low achievers might be bimodal. Similarly, a distribu on of daily temperatures in a year might have two peaks, one for summer and one for winter.

3. Bell-Shaped Curve:
- Defini on: A bell-shaped curve represents a data distribu on that has a symmetric, smooth, and roughly symmetrical shape resembling a bell.
- Shape: Bell-shaped distribu ons have a single peak (unimodal) and are symmetric, with the tails of the distribu on tapering off gradually as you move away from the peak.

Examples: The classic example of a bell-shaped curve is a normal distribu on, where data is symmetrically distributed around the mean. However, other distribu ons with a similar bell-shaped appearance can also exist.



Q 54. Does symmetric distribu on need to be unimodal?

No, a symmetric distribu on does not necessarily need to be unimodal. A symmetric distribu on simply means that the data is distributed in a way that is mirror-image symmetric, with values being equally likely on both sides of the distribu on's centre point (usually the mean or median).

So, while symmetry and unimodality o en go together, symmetry does not inherently require unimodality, and a symmetric distribu on can have mul ple modes.

Q 55. What are some examples of data sets with non-Gaussian distribu ons?

Many real-world datasets exhibit non-Gaussian or non-normal distribu ons due to various underlying factors. Here are some examples of data sets with non-Gaussian distribu ons:

1. Income Distribu on: Income data is o en right-skewed, with most people earning average incomes and a few earning very high incomes. This leads to a distribu on that does not follow a normal curve.
2. Stock Returns: Daily stock returns can have fat tails and exhibit vola lity clustering, making their distribu on non-normal. Events like stock market crashes can cause significant devia ons from normality.
3. Website Traffic: The number of visitors to a website on any given day o en follows a distribu on with a long tail. A few days with extremely high traffic can result in a skewed distribu on.
4. Ages at Re rement: The distribu on of ages at which people re re can be le -skewed, with many re ring around a certain age and very few re ring at younger ages.
5. Number of Customer Arrivals: The number of customers arriving at a store or service centre follows a Poisson distribu on, which is discrete and not normal.
6. Test Scores: Test scores, par cularly in educa onal se ngs, o en have a distribu on with modes due to various subpopula ons of students, leading to a mul modal distribu on.
7. City Popula on Sizes: The distribu on of city popula on sizes worldwide is o en right-skewed, with a few megaci es having very high popula ons and the majority of ci es having smaller popula ons.
8. Wait Times: The distribu on of wait mes in queues or lines can o en be right-skewed, with a few people experiencing very long waits and most people experiencing shorter waits.
9. Social Media Engagement: The number of likes, shares, or comments on social media posts can exhibit a highly skewed distribu on, with a few posts going viral and receiving a dispropor onate number of interac ons.
10. Height and Weight: While human height and weight o en follow roughly normal distribu ons, they can also be influenced by factors like nutri on and gene cs, leading to devia ons from normality in some popula ons.

These examples illustrate that real-world data can take on various shapes and characteris cs, and not all datasets follow the idealized Gaussian or normal distribu on. Understanding the distribu on of data is essen al for making accurate sta s cal inferences and modeling.

Q 56. What is the Binomial Distribu on Formula?

The binomial distribu on formula is used to calculate the probability of a specific number of successes (usually denoted as "k") in a fixed number of independent Bernoulli trials, where each trial has two possible outcomes: success (usually denoted as "p") and failure (usually denoted as "q," where q = 1 - p).

The probability mass func on (PMF) of the binomial distribu on is given by the formula:

$$\boxed{P(X = k) = \binom{n}{k} * p * q}$$

where,

- $P(X = k)$ is the probability of exactly k successes.
- n is the total number of trials.
- k is the number of successes you want to find the probability for.
- p is the probability of success on a single trial.
- q is the probability of failure on a single trial ($q = 1 – p$).
- represents the binomial coefficient, which is o en calculated as $= \frac{!}{!(\quad)!}$, where "!" denotes factorial.

Q 57. What are the criteria that Binomial distribu ons must meet?

The binomial distribu on is a probability distribu on that models a specific type of random experiment. To use the binomial distribu on, certain criteria or assump ons must be met:

- Fixed Number of Trials (n):
  The experiment consists of a fixed number of iden cal, independent trials, denoted as "n." Each trial can result in one of two possible outcomes: success or failure.
- Independence:
  The outcome of one trial does not affect the outcome of any other trial. In other words, the trials are independent of each other.
- Constant Probability of Success (p):
  The probability of success (o en denoted as "p") remains constant from trial to trial. This means that the probability of success is the same for each trial.
- Binary Outcomes:
  Each trial has only two possible outcomes: success and failure. These outcomes are mutually exclusive, meaning that a trial cannot result in both success and failure simultaneously.
- Bernoulli Trials:
  The individual trials are Bernoulli trials, which are experiments with two possible outcomes (success and failure) that meet the criteria men oned above (fixed n, independence, constant p, and binary outcomes).

Q 58. What are the examples of symmetric distribu on?

Symmetric distribu ons are characterized by their mirror-image symmetry, where the data is equally likely to occur on both sides of the centre point. Some examples of symmetric distribu ons include:

- Normal Distribu on (Gaussian Distribu on)
  1. The most well-known symmetric distribu on.
  2. Bell-shaped and characterized by its mean and standard devia on.
  3. Many natural phenomena and measurements, such as height and weight in a popula on, closely follow a normal distribu on.
- Uniform Distribu on
  1. In a con nuous uniform distribu on, all values within an interval have equal probability.
  2. In a discrete uniform distribu on, all outcomes have equal probability.
  3. For example, rolling a fair six-sided die follows a discrete uniform distribu on.
- Logis c Distribu on
  1. S-shaped curve similar to the normal distribu on but with heavier tails.
  2. O en used in logis c regression and modeling growth processes.

Q 59. Briefly explain the procedure to measure the length of all sharks in the world.

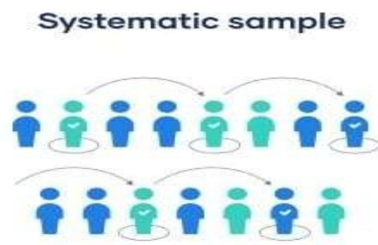
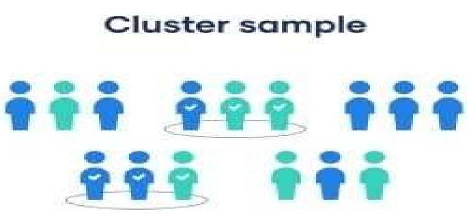
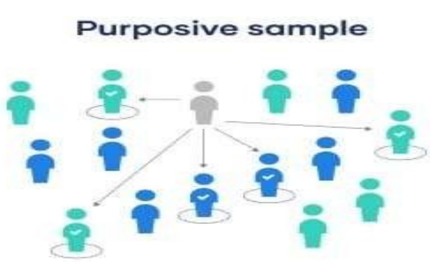
- Define the confidence level (most common is 95%)
- Take a sample of sharks from the sea (to get be er results the number of fishes > 30)
- Calculate the mean length and standard devia on of the lengths
- Calculate t-sta s cs
- Get the confidence interval in which the mean length of all the sharks should be.

Q 60. What are the types of sampling in Sta s cs?

In sta s cs, sampling is the process of selec ng a subset of individuals or items from a larger popula on to make inferences about the en re popula on. There are several types of sampling methods, each with its own advantages and use cases.

Here are some of the most common types of sampling:

1. Simple Random Sampling: • Involves randomly selec ng individuals or items from the popula on without any specific pa ern or criteria. • Every member of the popula on has an equal chance of being selected. • Can be done with or without replacement (i.e., the same individual/item can be selected more than once or not).	
 	2. Stra fied Sampling: • Divides the popula on into non-overlapping subgroups or strata based on certain characteris cs (e.g., age, gender, loca on). • Random samples are then taken from each stratum. • Ensures that each subgroup is represented in the sample, making it useful when there are significant differences between subgroups.

3. Systema c Sampling: - Involves selec ng every nth individual/item from a list or sequence. - Typically, a random star ng point is chosen, and then every nth individual/item is selected. - Useful when there's a natural order or sequence in the popula on.	
	4. Cluster Sampling: - Divides the popula on into clusters or groups, o en based on geographic proximity or another criterion. - A random sample of clusters is selected, and all individuals/items within the selected clusters are included in the sample. - Efficient for large and geographically dispersed popula ons.
5. Convenience Sampling: - Involves selec ng individuals or items that are readily available and convenient to sample. - O en used in exploratory or preliminary research but can introduce bias because it may not be representa ve of the en re popula on.	
	6. Purposive Sampling (Judgmental Sampling): - Involves selec ng individuals/items based on the researcher's judgment and specific criteria. - Useful when the researcher wants to focus on a par cular subgroup or characteris c. - Can be biased if not done carefully.

The choice of sampling method depends on the researcher's objec ves, available resources, and the characteris cs of the popula on being studied. Each method has its own strengths and limita ons, and researchers must consider these factors when designing and conduc ng a study.

Q 61. Why is sampling required?

Sampling is required for several simple and prac cal reasons:

1. Efficiency: Sampling is faster and more cost-effec ve than collec ng data from an en re popula on, especially when the popula on is large.
2. Resource Conserva on: It saves me, money, and resources, making research more feasible and prac cal.
3. Timeliness: Allows for quicker data collec on and analysis, which can be crucial in me-sensi ve situa ons.
4. Accessibility: Some popula ons are difficult to access, making sampling the only prac cal op on.
5. Accuracy: When done correctly, sampling provides accurate es mates of popula on characteris cs.
6. Risk Reduc on: Reduces the poten al for errors in data collec on and analysis.
7. Inference: Provides a basis for making conclusions about the en re popula on based on the characteris cs of the sample.
8. Privacy and Ethics: Respects privacy and ethical considera ons, especially in sensi ve research areas.
9. Analysis: Simplifies data analysis, par cularly for large datasets.

Sampling is a prac cal and essen al tool for researchers to gather valuable informa on while managing constraints and prac cal limita ons.

Q 62. How do you calculate the needed sample size?

To calculate the needed sample size:

- Define your research objec ves and ques ons.
- Choose a significance level (α) and desired margin of error (E).
- Es mate popula on variability (σ) or use conserva ve es mates.
- Determine the popula on size (N).
- Select the type of sampling (random or stra fied).
- Choose the sta s cal test or analysis.
- Use a sample size formula or so ware tool to calculate the sample size.
- Consider prac cal constraints and adjust for non-response.
- Conduct the study, analyze data, and interpret results.

Sample size calcula ons ensure your study has enough data to draw meaningful conclusions while controlling for errors and precision.

Q 63. Can you give the difference between stra fied sampling and clustering sampling?

The key dis nc on between stra fied sampling and cluster sampling lies in how the popula on is divided and sampled:

- Stra fied sampling divides the popula on into homogeneous subgroups (strata) and selects samples from each stratum independently to ensure representa on from all subgroups.
- Cluster sampling divides the popula on into clusters and randomly selects clusters to sample, then collects data from all individuals/items within the selected clusters.

Q 64. Where is inferen al sta cs used?

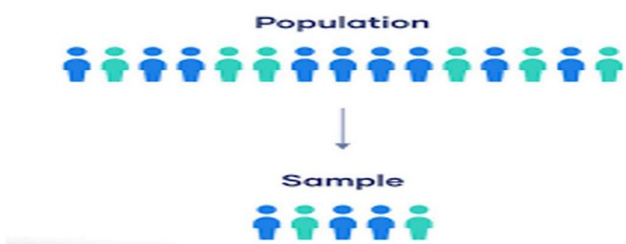
Inferen al sta scs is used in various fields and contexts to make predic ons, draw conclusions, and make inferences about popula ons based on sample data.

Here are some common areas and applica ons where inferen al sta s cs is used:

1. Scien fic Research:
   Inferen al sta scs is fundamental in scien fic research across disciplines such as biology, physics, chemistry, and environmental science. Researchers use sta s cal tests to analyze data and draw conclusions about hypotheses.
2. Business and Economics:
   Businesses use inferen al sta cs for market research, sales forecas ng, quality control, and decision-making. Econometric models are employed to analyze economic data and make policy recommenda ons.
3. Healthcare and Medicine:
   Medical researchers and healthcare professionals use inferen al sta s cs to study the effec veness of treatments, analyze pa ent data, and draw conclusions about disease prevalence. Clinical trials rely heavily on inferen al sta s cs
4. Educa on:
   In the field of educa on, inferen al sta cs are used to assess the effec veness of teaching methods, evaluate standardized test scores, and make policy decisions about educa onal programs.
5. Market Research and Data Analysis:
   Market researchers use inferen al sta s cs to make predic ons about consumer preferences, market trends, and the impact of marke ng campaigns.
6. Finance and Investment:
   In finance, inferen al sta cs are used to assess investment risk, analyze stock market data, and es mate future asset prices. Por olio op miza on and risk management rely on sta s cal modeling.
7. Criminal Jus ce and Criminology:
   Researchers and law enforcement agencies use inferen al sta s cs to analyze crime data, study crime pa erns, and evaluate the effec veness of criminal jus ce programs.
8. Sports and Athle cs:
   In sports analy cs, inferen al sta cs are used to analyze player performance, predict game outcomes, and make strategic decisions in sports management.

Q 65. What are popula on and sample in Inferen al Sta s cs, and how are they different?

In inferen al sta scs, the concepts of "popula on" and "sample" are fundamental and play dis nct roles.

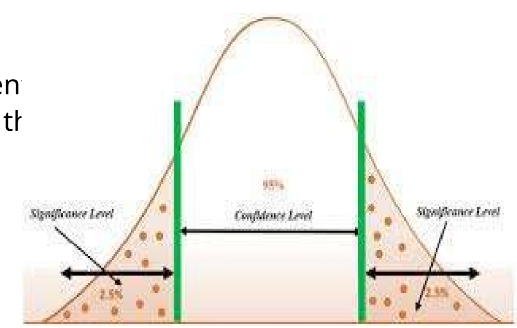


Popula on	Sample
Defini on: The popula on refers to the en re group or collec on of individuals, items, or data poin about which you want to draw conclusions. represents the larger, o en theore cal, set t you're interested in studying.	Defini on: A sample is a subset or a smaller, carefully selected group of individuals, items, or data points taken from the larger popula on. It is a representa ve por on of the popula on used for data collec on and analysis.
Characteris cs: -The popula on can be finite (e.g., all students in a school) or infinite (e.g., all poten al customers in a market). -It includes every possible individual or element that falls within the scope of your research ques on.	Characteris cs: -The sample is a finite and manageable subset of the popula on. -It is chosen through a systema c process, such as random sampling, stra fied sampling, or cluster sampling. -The sample should be representa ve of the popula on, meaning that it should reflect the diversity and characteris cs of the popula on.
Purpose: -In inferen al sta s cs, the popula on is the ul mate target for making conclusions and generaliza ons. However, it is o en imprac cal or impossible to collect data from the en re popula on.	Purpose: -The primary purpose of taking a sample is prac cality. It's o en more feasible, cost-effec ve, and efficient to collect data from a sample rather than the en re popula on. -Inferen al sta s cs use data from the sample to make inferences, predic ons, or generaliza ons about the larger popula on.

Q 66. What is the rela onship between the confidence level and the significance level in sta s cs?

The rela onship between the confidence level and the significance level in sta s cs is inverse and complementary. These two concepts are essen al in hypothesis tes ng and sta s cal inference.

Rela onship:



1.The rela onship between the two is complemen meaning that if you increase one, you decrease th other, and vice versa.

2.Higher confidence levels correspond to lower significance levels, and lower confidence levels correspond to higher significance levels.

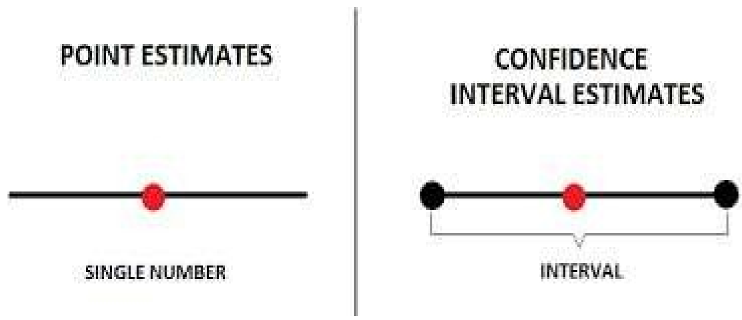
For example:

- If you set a confidence level of 95% ($1-\alpha=0.95$), the significance level would be 0.05($\alpha=0.05$).
- If you set a confidence level of 99% ($1-\alpha=0.99$), the significance level would be 0.01($\alpha=0.01$).

Confidence level	Significance level
The confidence level (often denoted as $1-\alpha$) represents the probability that a confidence interval calculated from sample data contains the true population parameter.	The significance level (denoted as α) is the probability of making a Type I error in hypothesis testing. It is also known as the "alpha level" or "level of significance."
It is a measure of how confident you are that the interval you calculated captures the parameter you're estimating.	A Type I error occurs when you incorrectly reject a true null hypothesis. In other words, it represents the probability of finding a significant result (rejecting the null hypothesis) when there is no real effect or difference in the population.
Commonly used confidence levels include 90%, 95%, and 99%.	Commonly used significance levels are 0.05 (5%), 0.01 (1%), and 0.10 (10%).

Q 67. What is the difference between Point Estimate and Confidence Interval Estimate?

Point Estimate	Confidence Interval Estimate
A point estimate is a single value that is used to estimate an unknown population parameter, such as the population mean (μ) or population proportion (p).	A confidence interval estimate is a range or interval of values that is used to estimate a population parameter.
It provides a "best guess" or a single numerical value for the parameter.	It provides a range of plausible values for the parameter along with a level of confidence (e.g., 95% confidence interval). The confidence interval reflects the uncertainty associated with the estimate and quantifies how confident you are that the true parameter falls within the interval.
For example, if you calculate the sample mean ($\bar{x}$) from a sample of data, $\bar{x}$ itself is a point estimate of the population mean (μ).	For example, a 95% confidence interval for the population mean (μ) might be (60,70), indicating that you are 95% confident that the true population mean falls between 60 and 70.



Key Difference:

- The main difference between a point es mate and a confidence interval es mate is that a point es mate provides a single value, while a confidence interval es mate provides a range of values.
- Point es mates are useful for providing a single es mate of a parameter when you need a single, specific value.
- Confidence interval es mates are useful when you want to convey the uncertainty associated with your es mate and provide a range of values within which the parameter is likely to fall.

Q 68. What do you understand about biased and unbiased terms?

In sta s cs, the terms "biased" and "unbiased" are used to describe the accuracy of an es mator in es ma ng a popula on parameter. These terms relate to how close the expected value of the es mator is to the true (or popula on) value of the parameter being es mated.

Biased	Unbiased
A sta s cal es mator is said to be "biased" if on average, it systema cally overes mates or underes mates the true popula on paramet	A sta s cal es mator is considered "unbiased" if, on average, it provides es mates that are equal to the true popula on parameter.
In other words, a biased es mator tends to consistently deviate from the true value in a specific direc on (either consistently too high or too low).	In mathema cal terms, the expected value (mean) of an unbiased es mator is equal to the true value of the parameter being es mated.
Biased es mators can result from flaws in th es ma on method or sampling procedure.	Unbiased es mators are desirable because, over repeated sampling, they provide accurate and unbiased es mates of the popula on parameter.
When using a biased es mator, it's important to be aware of the direc on and magnitude of the bias to adjust for it in data analysis or decision making.	While unbiased es mators are preferred, they are not always achievable, and in some cases, biased es mators may be the best available op on.

Q 69. How does the width of the confidence interval change with length?

The width of a confidence interval changes inversely with the level of confidence and the precision of the es mate. In other words, as you increase the level of confidence or decrease the precision (increase the margin of error), the width of the confidence interval increases, and vice versa.

Q 70. What is the meaning of standard error?

The width of a confidence interval changes inversely with the level of confidence and the precision of the es mate. In other words, as you increase the level of confidence or decrease the precision (increase the margin of error), the width of the confidence interval increases, and vice versa.

- Standard Error of the Sample Mean (SE(***x***)):

1. The standard error of the sample mean represents the standard devia on of the distribu on of sample means.
2. It measures how much individual sample means are expected to deviate from the true popula on mean (μ) on average.
3. The formula for the standard error of the sample mean depends on the popula on standard devia on (σ) and the sample size (n) and is given by:

$$\boxed{SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}}$$

4. As the sample size (n) increases, the standard error decreases. This means that larger samples tend to produce sample means that are closer to the true popula on mean.

- The standard error is a cri cal concept in inferen al sta s cs because it is used to calculate confidence intervals and conduct hypothesis tests. Here's how it is typically used:

1. <u>Confidence Interva</u>ls: The standard error is used to calculate the margin of error for a confidence interval. A confidence interval represents a range of values within which you are confident that the true popula on parameter lies.
2. <u>Hypothesis Tes ng</u>: In hypothesis tes ng, the standard error is used to calculate test sta s cs, such as the t-sta s c or z-sta s c, which are then compared to cri cal values to assess the significance of an observed effect or difference.

Q 71. What is a Sampling Error and how can it be reduced?

Sampling error is a type of error that occurs when a sample is used to es mate popula on parameters, and the es mate differs from the true popula on value. It's the difference between the sample sta s c (e.g., sample mean or propor on) and the true popula on parameter. It happens because we can't study everyone in the popula on, so we use a sample (a smaller group) to make predic ons.

Here's how sampling error can be reduced or minimized:

- Use a Larger Sample: The bigger the sample, the closer the es mate is to reality.
- Randomly Choose the Sample: Ensure that everyone in the popula on has an equal chance of being in the sample.
- Be Careful with Surveys: Encourage more people to respond to surveys to make sure they represent the whole popula on.
- Use Proper Methods: Follow good sta s cal methods to analyze the data from your sample.

Reducing sampling error helps us make more accurate predictions about the population based on our sample.

Q 72. How do the standard error and the margin of error relate?

In simple words, think of the standard error (SE) as a measure of how much sample data can vary from the true population value. It's like a measure of how shaky or uncertain our estimate is.

The margin of error (MOE) is directly related to the standard error. It tells us how much we should add to and subtract from our sample estimate to create a range that likely includes the true population value. It's like a safety buffer around our estimates.

So, the standard error tells us about the uncertainty in our estimates, and the margin of error tells us the size of the safety buffer we need to account for that uncertainty. If you want a narrower margin of error, you need a more precise estimate, which usually means a larger sample size or a lower level of confidence.

Q 73. What is hypothesis testing?

Hypothesis testing is a fundamental statistical technique used to make inferences and draw conclusions about populations based on sample data. It involves a structured process of formulating and testing hypotheses (statements or claims) about population parameters, such as means, proportions, or variances.

Here are the key components and steps involved in hypothesis testing:

<u>Components of Hypothesis Testing</u>

- Null Hypothesis (H0)
- Alternative Hypothesis (Ha or H1)
- Test Statistic
- Significance Level (α)
- Critical Region or Rejection Region
- P-Value

<u>Steps in Hypothesis Testing:</u>

- Formulate Hypotheses
- Collect Data
- Calculate Test Statistic
- Determine Critical Region
- Compare Test Statistic and Critical Region
- Calculate P-Value
- Make a Decision
- Draw Conclusions

Q 74. What is an alternative hypothesis?

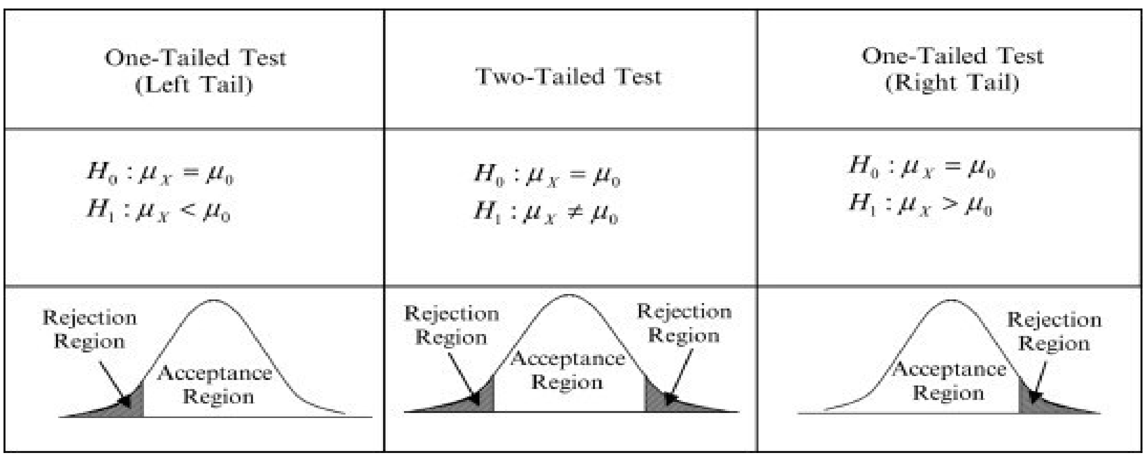
The alternative hypothesis contradicts the null hypothesis. It typically states what you expect to find in the population based on your research question or hypothesis. It is denoted as Ha or H1.

Q 75. What is the difference between one-tailed and two-tail hypothesis tes ng?

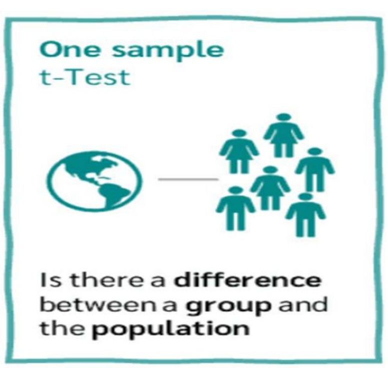
One-tailed and two-tailed hypothesis tes ng are two different approaches used in sta s cal hypothesis tes ng to inves gate research ques ons or hypotheses. They differ in terms of the direc onality of the research ques on and the way they assess evidence from sample data.

Here's a comparison of the two:

One-Tailed Hypothesis Tes ng	Two-Tailed Hypothesis Tes ng
One tail test is a sta s cal hypothesis test in which the alterna ve hypothesis onl has one end.	Two-tail test refers to a significance test in which the alterna ve hypothesis has two ends.
Region of rejec on is either le or right.	Region of rejec on is both le and right.
Determines rela onship between varia in single direc on.	Determines rela onship between variables in either direc on.
Results are greater or less than certain value.	Results are greater or less than certain range of values.
Direc onal: > or <	Non-direc onal: ≠



Q 76. What is one sample t-test?



A one-sample t-test is a sta s cal hypothesis test used to determine whether the mean of a single sample of data is sta s cally different from a known or hypothesized popula on mean.

It's par cularly useful when you have a sample and you want to assess whether it represents a popula on with a specific mean.

Q 77. What is the meaning of degrees of freedom (DF) in sta s cs?

In sta s cs, degrees of freedom (DF) refer to the number of values in the final calcula on of a sta s c that are free to vary. Degrees of freedom are a fundamental concept in hypothesis tes ng, confidence intervals, and various sta s cal analyses. They are used in various sta s cal tests, such as t-tests, chi-square tests, and analysis of variance (ANOVA).

The concept of degrees of freedom can be a bit abstract, but it's essen al to understand because it affects the behaviour of sta s cal tests and the interpreta on of their results. Here's a basic explana on:

- T-Tests:
  In a t-test, degrees of freedom are related to the sample size. If you have a sample of size "n" then,
  1. <u>One-sample t-test</u>: $\text{Degrees of freedom} = n - 1$

  2. <u>Two-sample t-test</u>: $\text{Degrees of freedom} = n1 + n2 - 2$

  where "n1" and "n2" are the sample sizes of the two groups being compared. This "n1 + n2 - 2" represents the number of data points that are free to vary a er es ma ng the means of the two groups.

- Chi-Square Tests:
  In chi-square tests, degrees of freedom are related to the number of categories being compared.
  For a chi-square test of independence, the degrees of freedom are calculated as,

  $$\boxed{\text{Degrees of freedom} = (\text{rows} - 1) * (\text{columns} - 1)}$$

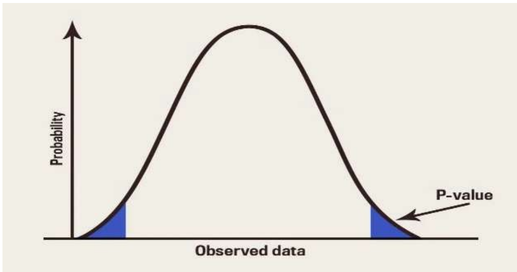
  where "rows" and "columns" represent the number of categories in the rows and columns of the con ngency table. This calcula on reflects the number of categories that can vary freely.

- ANOVA:
  In analysis of variance (ANOVA), degrees of freedom are associated with the number of groups being compared.
  There are two types of degrees of freedom in ANOVA:
  1. Between-group degrees of freedom:
     The between-group degrees of freedom are related to the number of groups minus one.
  2. Within-group degrees of freedom.
     The within-group degrees of freedom are related to the total sample size minus the number of groups.

  These degrees of freedom help determine whether there are significant differences between group means.

In essence, degrees of freedom represent the flexibility or "freedom" in the data or the sta s cal model. Understanding degrees of freedom is crucial because they affect the distribu on of test sta s cs and, consequently, the interpreta on of p-values and the conclusions drawn from sta s cal analyses. Different sta s cal tests have different formulas for calcula ng degrees of freedom, and they are chosen to ensure the validity of the sta s cal test being performed.

Q 78. What is the p-value in hypothesis tes ng?



The p-value, short for "probability value," is a crucial concept in hypothesis tes ng in sta s cs. It measures the strength of evidence against a null hypothesis.

Q 79. How can you calculate the p-value?

In general, calcula ng a p-value involves the following steps:

- Formulate Hypotheses:
  Start by defining your null hypothesis (H0) and alterna ve hypothesis (Ha). H0 typically represents a statement of no effect or no difference, while Ha suggests there is an effect or difference.
- Choose a Sta s cal Test:
  Select the appropriate sta s cal test based on your research ques on and the type of data you have. The choice of test depends on whether you're comparing means, tes ng propor ons, examining associa ons, etc.
- Collect Data:
  Collect relevant data for your analysis. The data should match the assump ons and requirements of the chosen sta s cal test.
- Calculate the Test Sta s c:
  Calculate the test sta s c that corresponds to your chosen test. This involves using mathema cal formulas specific to the test.
- Determine the Sampling Distribu on:
  Determine the theore cal sampling distribu on of the test sta s c under the assump on that the null hypothesis is true. This distribu on depends on the test you're conduc ng (e.g., t-distribu on, chi-square distribu on, F-distribu on, normal distribu on).
- Find the Observed Test Sta s c
  Calculate the observed test sta s c using your data.
- Calculate the p-value:
  The p-value is calculated based on the observed test sta s c and its distribu on under the null hypothesis.
  1. For one-tailed tests (where you are only interested in one direc on of an effect), the p-value is the probability of observing a test sta s c as extreme or more extreme than the observed value in that direc on.
  2. For two-tailed tests (where you are interested in both direc ons of an effect), the p-value is the probability of observing a test sta s c as extreme or more extreme than the observed value in either direc on.
- Compare the p-value to the Significance Level (α):
  Decide on a significance level (α), which is typically set at 0.05 but can vary depending on the study.

1. If the p-value is less than or equal to α, you reject the null hypothesis (conclude there is evidence for the alterna ve hypothesis).
2. If the p-value is greater than α, you fail to reject the null hypothesis (insufficient evidence to support the alterna ve hypothesis).

It's important to note that the specific calcula ons for the test sta s c and p-value depend on the chosen sta s cal test. Different tests have different formulas and assump ons. In prac ce, sta s cal so ware or calculators are o en used to perform these calcula ons automa cally, as they can be complex for many tests. Addi onally, when conduc ng hypothesis tests, make sure to consider the assump ons and limita ons of the chosen test to ensure the validity of your results.

Q 80. If there is a 30 percent probability that you will see a supercar in any 20-minute me interval, what is the probability that you see at least one supercar in the period of an hour (60 minutes)?

- The probability of not seeing a supercar in 20 minutes is:

  $= 1 - P(Seeing\ one\ supercar) = 1 - 0.3 = 0.7$

- Probability of not seeing any supercar in the period of 60 minutes is:

  $= (0.7) \wedge 3 = 0.343$

- Hence, the probability of seeing at least one supercar in 60 minutes is:

  $= 1 - P(Not\ seeing\ any\ supercar) = 1 - 0.343 = 0.657$

Q 81. How would you describe a 'p-value'?

p-values help you make decisions about whether the results of a sta s cal analysis are sta s cally significant. They don't tell you whether the null hypothesis is true or false; instead, they inform you about the likelihood of observing the data if the null hypothesis were true.

Q 82. What is the difference between type I vs type II errors?

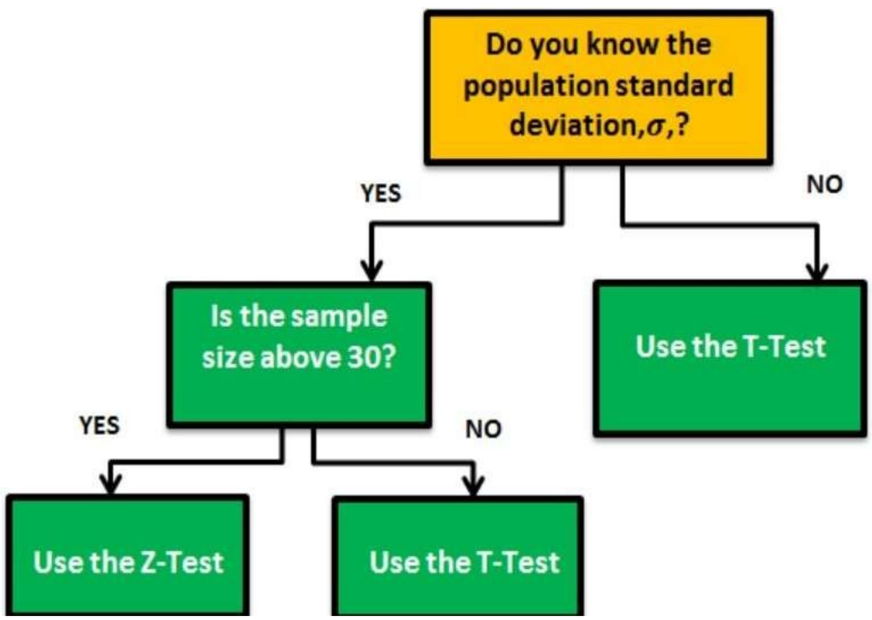
A type I error (false-posi ve) occurs if an inves gator rejects a null hypothesis that is actually true in the popula on; a type II error (false-nega ve) occurs if the inves gator fails to reject a null hypothesis that is actually false in the popula on.

Type I	Type II
1. The chance or probability that you will reject a null hypothesis that should not have been rejected. 2. This will result in you deciding two groups are different or two variables are related when they really are not. 3. The probability of a Type I error is called alpha (α).	1. The chance or probability that you will not reject a null hypothesis when it should have been rejected. 2. This will result in you deciding two groups are different or two variables are not related when they really are. 3. The probability of a Type II error is called beta (β).

Type I and Type II Error		
Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = β

Q 83. When should you use a t-test vs a z-test?

A z-test is used to test a Null Hypothesis if the popula on variance is known, or if the sample size is larger than 30, for an unknown popula on variance. A t-test is used when the sample size is less than 30 and the popula on variance is unknown.



Q 84. What is the difference between the f test and anova test?

The F-test and ANOVA (Analysis of Variance) are related statistical tests, but they serve different purposes and are used in different contexts.

f test	anova test
Purpose: The F-test is a sta s cal test used to compare the variances of two or more popula ons or samples.	Purpose: ANOVA, on the other hand, is used to compare means of three or more groups to determine if there are sta s cally significant differences among the group means.
Number of Groups: The F-test is primarily used for comparing the variances of two groups. It's commonly employed in the context of comparing the variances of two groups when tes ng for the equality of popula on variances (e.g., in the context of two-sample hypothesis tes ng).	Number of Groups: ANOVA is specifically designed for comparing the means of three or more groups. It is used when you have mul ple groups, and you want to test if there are any significant differences among them.
Test Sta s c: The test sta s c for the F-test follows an F-distribu on, which is a right-skewed distribu on. The F-sta s c is calculated by dividing the variance of one group by the variance of another group.	Test Sta s c: ANOVA uses an F-sta s c as well, but the calcula on is different from the F-test. It assesses the ra o of varia on between group means to the varia on within groups.
Use Cases: Common use cases for the F-test include comparing the variances of two groups (F-test for equality of variances), assessing the goodness of fit of a sta s cal model, and performing regression analysis (F-test for overall model fit).	Use Cases: ANOVA is commonly used in experimental designs where you have several treatments or condi ons and you want to determine if there is a sta s cally significant difference in the means of these groups. It is o en followed by post-hoc tests to iden fy which specific group means differ from each other.

Q 85. What is Resampling and what are the common methods of resampling?

Resampling is a series of techniques used in sta s cs to gather more informa on about a sample. This can include retaking a sample or es ma ng its accuracy. With these addi onal techniques, resampling o en improves the overall accuracy and es mates any uncertainty within a popula on.

Common methods of resampling include:

1. Bootstrapping:
   Bootstrap Sampling: In bootstrap resampling, you randomly select data points from your dataset with replacement to create mul ple "bootstrap samples" of the same size as the original dataset.

Purpose: Bootstrapping is o en used to es mate the sampling distribu on of a sta s c (e.g., mean, median, standard devia on) or to construct confidence intervals.

2. Cross-Valida on:
   K-Fold Cross-Valida on: In cross-valida on, you par on your dataset into "k" subsets (folds). You itera vely use k-1 folds for training and the remaining fold for tes ng, repea ng this process k mes.
   Purpose: Cross-valida on is widely used in machine learning to assess model performance, tune hyperparameters, and detect overfi ng.

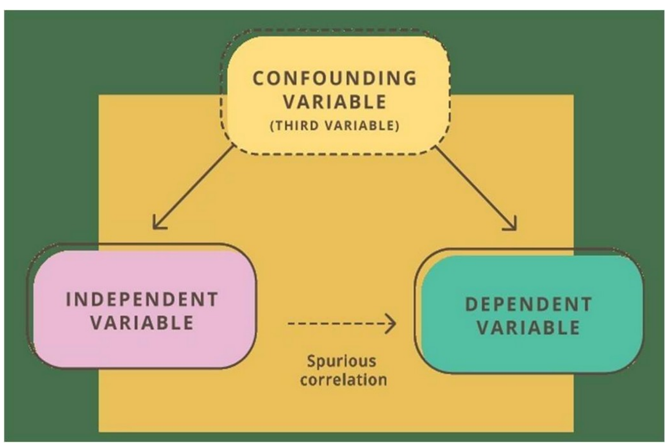
Q 86. What is the propor on of confidence intervals that will not contain the popula on parameter?

The propor on of confidence intervals that will not contain the popula on parameter (o en denoted as 1 - confidence level) is equal to the significance level (α) chosen for construc ng the confidence intervals.

In other words, if you construct a large number of confidence intervals using the same method and the same confidence level (e.g., 95% confidence level), and if you repeat this process many mes, then approximately 5% of these intervals will not contain the true popula on parameter.

Q 87. What is a confounding variable?

A confounding variable, also known as a confounder or confounding factor, is a variable in a research study that is related to both the independent variable (the variable being studied or manipulated) and the dependent variable (the outcome or response of interest). The presence of a confounding variable can lead to a misleading or incorrect interpreta on of the rela onship between the independent and dependent variables.



In simpler terms, a confounding variable is an extra factor that can distort the observed rela onship between two other variables by either masking or falsely sugges ng a connec on between them.

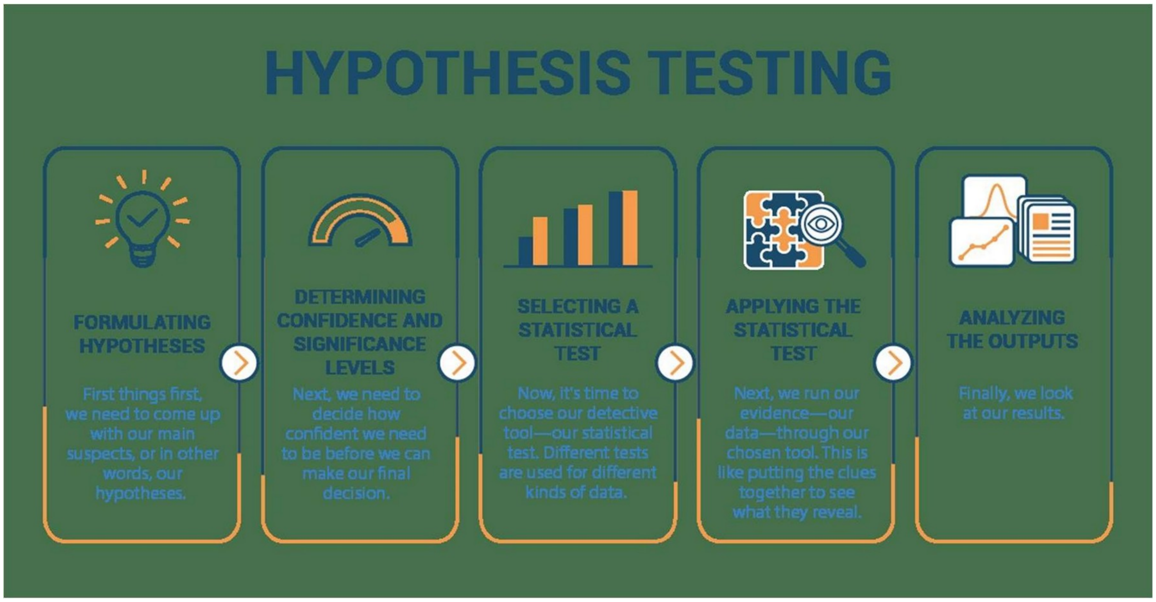
Example: Suppose you are studying the rela onship between coffee consump on (independent variable) and the risk of heart disease (dependent variable). Age is a confounding variable because it is related to both coffee consump on (as people of different ages may drink different amounts of coffee) and the risk of heart disease (as older individuals tend to have a higher risk). Without considering age as a confounder, you may mistakenly conclude that coffee consump on directly affects heart disease risk.

Q 88. What are the steps we should take in hypothesis tes ng?

Hypothesis tes ng is a structured process used in sta s cs to make inferences about popula on parameters based on sample data. Here are the steps typically involved in hypothesis tes ng:

1. Formulate Hypotheses:
   State the null hypothesis (H0): This is a statement of no effect or no difference. It represents the default assump on you want to test.
   State the alterna ve hypothesis (Ha): This is the hypothesis you want to provide evidence for, sugges ng that there is an effect, difference, or rela onship in the popula on.
2. Choose a Significance Level (α):
   Select the significance level (α), which represents the probability of making a Type I error (rejec ng the null hypothesis when it is true). Common choices include 0.05 (5%) and 0.01 (1%).
3. Collect and Analyse Data:
   Collect sample data that are relevant to your research ques on.
   Perform appropriate sta s cal analysis based on the type of data and research design. This analysis depends on the specific hypothesis test you're conduc ng (e.g., t-test, chi-square test, ANOVA).
4. Calculate the Test Sta s c:
   Calculate the test sta s c based on your sample data and the null hypothesis. The test sta s c quan fies how different your sample data are from what you would expect under the null hypothesis.
5. Determine the Cri cal Region:
   Iden fy the cri cal region or rejec on region in the probability distribu on of the test sta s c. This is the range of values that would lead to reject the null hypothesis if the test sta s c falls within it.
6. Compare the Test Sta s c to Cri cal Values:
   Compare the calculated test sta s c to the cri cal values (cut-off values) corresponding to the chosen significance level. If the test sta s c falls in the cri cal region, you reject the null hypothesis. Otherwise, you fail to reject it.
7. Calculate the P-Value:
   Alterna vely, you can calculate the p-value, which is the probability of observing a test sta s c as extreme as, or more extreme than, the one calculated, assuming the null hypothesis is true.
   - If the p-value is less than or equal to the chosen significance level (α), you reject the null hypothesis.
   - If the p-value is greater than α, you fail to reject it.
8. Make a Decision:
   Based on the comparison of the test sta s c (or p-value) to the cri cal values (or α), make a decision:
   - If you reject the null hypothesis, conclude that there is evidence for the alterna ve hypothesis.
   - If you fail to reject the null hypothesis, conclude that there is insufficient evidence to support the alterna ve hypothesis.
9. Interpret Results:
   Interpret the results in the context of your research ques on. Explain the prac cal significance of your findings and their implica ons.
10. Report Findings:

Clearly communicate your results, including the test sta s c, p-value (if used), conclusion, and any relevant effect size measures, in a clear and concise manner.



Q 89. How would you describe what a 'p-value' is to a non-technical person or in a layman term?

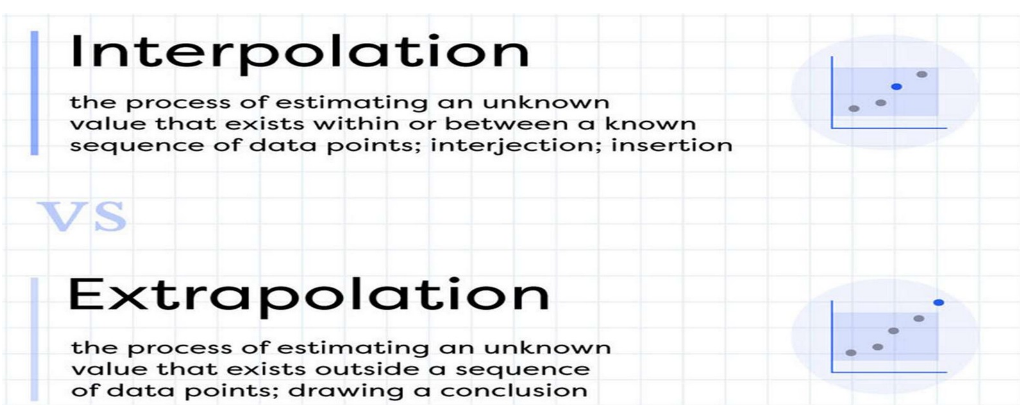
Explaining a p-value to a non-technical person or in layman's terms:

Imagine you're a detec ve inves ga ng a case. You have a suspect on trial, and you want to know if there's enough evidence to say they are guilty.

The p-value is like a measure of how strong your evidence is against the suspect. It tells you the likelihood of ge ng the evidence you have if the suspect is innocent.

Q 90. What does interpola on and extrapola on mean? Which is generally more accurate?

Interpola on and extrapola on are two mathema cal techniques used to es mate values within or outside a given range of known data points. They serve different purposes and have different degrees of accuracy:



Which Is Generally More Accurate?

Interpola on is generally more accurate than extrapola on. Here's why:

Interpola on es mates values within the range of known data, where you have observed the actual pa ern or rela onship between data points. As long as this rela onship is rela vely consistent, interpola on tends to provide reasonably accurate es mates.

Extrapola on, on the other hand, involves predic ng values beyond the range of known data, which is inherently uncertain. Extrapola on assumes that the same pa ern or trend will con nue, and this assump on may not always hold true, especially when data are subject to changing condi ons or unobserved factors.

Q 91. What is an inlier?

An inlier is a data point in a dataset that conforms to the general pa ern or behavior of the majority of the data points. In other words, an inlier is a point that is considered typical or consistent with the overall characteris cs of the dataset. Inliers are contrasted with outliers, which are data points that deviate significantly from the expected or typical behaviour of the dataset.

Q 92. You roll a biased coin (p(head)=0.8) five mes. What's the probability of ge ng three or more heads?

To start off the ques on, we need 3, 4, or 5 heads to sa sfy the cases.

- 5 heads: All heads, so

  – = ——.

- 4 heads: All heads but 1. There are 5 ways to organize this, and then a

  – ∗ – = 256/3125.

  Since there are 5 cases, we have 1280/3125.

- 3 heads: All heads but 2. There are 10 ways to organize this, and then a

  – ∗ – = 64/3125.

  Since there are 10 cases, we have 640/3125.

We sum all these cases up to get (1024 + 1280 + 640)/3125=2944/3125.

We have a 2944/3125 or 0.94208 probability to get 3 or more heads.

Q 93. Infec on rates at a hospital above a 1 infec on per 100 person-days at risk are considered high. A hospital had 10 infec ons over the last 1787 person-days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard.

To find the p-value for the one-sided test of whether the hospital infec on rate is below the standard of 1 infec on per 100 person-days at risk, you can use the Poisson distribu on. The Poisson distribu on is appropriate for modeling the number of rare events, such as infec ons in a hospital, over a known interval of me.

Here's how to calculate the p-value for this test:

1. Calculate the expected number of infections under the standard rate: Standard infection rate = 1 infection per 100 person-days.

$$Expected\ infections = (1787)\ \frac{1}{100} = 17.87$$

2. Use the Poisson distribution to find the probability of observing 10 or fewer infections when the expected number is 17.87. The Poisson probability mass function is:

$$P(X = x) = \frac{e \quad * \lambda}{x!}$$

3. Calculate the cumulative probability of observing 10 or fewer infections:

$$P(X \leq 10) = \frac{e \quad \cdot \quad * 17.87}{x!}$$

4. Find the p-value, which is the probability of observing 10 or fewer infections:

$$P(X \leq 10) = 0.033$$

So, the p-value for the one-sided test of whether the hospital is below the standard infection rate of 1 infection per 100 person-days at risk is 35/44 approximately 0.033. This p-value indicates strong evidence that the hospital's infection rate is below the standard, as it is smaller than a typical significance level a such as 0.05

Q 94. In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?

To calculate a 95% Student's t-confidence interval for the mean brain volume in the population, you can use the following formula:

$$Confidence\ Interval = \bar{x} \pm \left(t * \frac{s}{\sqrt{n}}\right)$$

Where:

$\bar{x}$ is the sample mean (1,100cc in this case).

t is the critical t-value for a 95% confidence interval with (n - 1) degrees of freedom.

s is the sample standard deviation (30cc in this case).

n is the sample size (9 in this case).

First, let's find the critical t-value for a 95% confidence interval with 8 degrees of freedom (9 - 1 = 8). You can use a t-table or a calculator to find this value. For a 95% confidence level and 8 degrees of freedom, the critical t-value is approximately 2.306.

Now, plug in the values into the formula:

$$Confidence\ Interval = 1100 \pm (2.306 * \frac{30}{\sqrt{9}})$$

$$Confidence\ Interval = 1100 \pm (2.306 * 10)$$

Now, calculate the lower and upper bounds of the confidence interval:

Lower Bound = 1,100 - (2.306 * 10) = 1,100 - 23.06 = 1,076.94 cc

Upper Bound = 1,100 + (2.306 * 10) = 1,100 + 23.06 = 1,123.06 cc

So, the 95% confidence interval for the mean brain volume in this new popula on is approximately 1,076.94 cc to 1,123.06 cc. This means that we are 95% confident that the true mean brain volume in the popula on falls within this range.

Q 95. What Chi-square test?

A chi-square test is a sta s cal test used to determine if there is a significant associa on or rela onship between categorical variables. It is par cularly useful for analyzing data that can be organized into a con ngency table, which is a tabular representa on of data where rows and columns correspond to different categories or groups.

Q 96. What is the ANOVA test?

ANOVA, or Analysis of Variance, is a sta s cal test used to analyse the differences among group means in a sample. It's a powerful and widely used technique for comparing means from mul ple groups to determine whether there are sta s cally significant differences among them.

The main idea behind ANOVA is to par on the total variance in the data into different components, which can be a ributed to different sources or factors.

Q 97. What do we mean by – making a decision based on comparing p-value with significance level?

Making a decision based on comparing a p-value with a significance level involves determining whether the evidence from a sta s cal test supports or contradicts a null hypothesis.

- If the p-value is less than or equal to the chosen significance level (α), typically 0.05, it suggests that the observed results are sta s cally significant. In this case, you reject the null hypothesis.
- If the p-value is greater than the significance level, it suggests that the observed results are not sta s cally significant. In this case, you fail to reject the null hypothesis.

In short, it's a way to decide whether the data provides enough evidence to challenge a specific hypothesis or not.

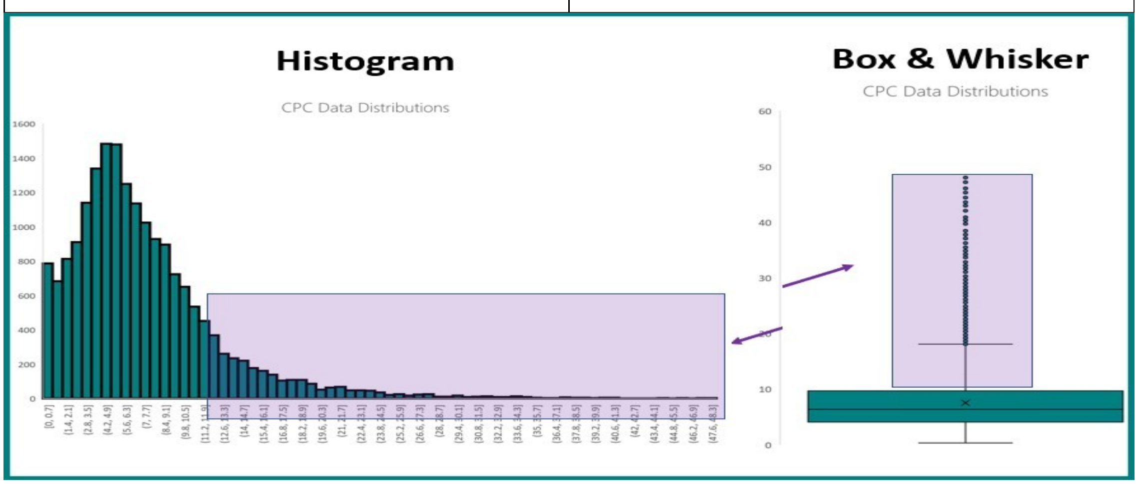
Q 98. What is the goal of A/B tes ng?

The goal of A/B tes ng is to compare different varia ons of a digital element (such as a webpage or app feature) to determine which one performs be er in terms of a specific outcome, with the aim of op mizing that element for improved user engagement, conversions, or other desired metrics.

Q 99. What is the difference between a box plot and a histogram

Box plots and histograms are both graphical representa ons used in statis cs to visualize the distribu on of data. However, they have different purposes and characteris cs:

Histogram	Box Plot
Purpose: Histograms are used to visualize the distribu on of con nuous data by dividing it into bins or intervals and displaying the frequency or count of data points within each bin. Appearance: A histogram consists of a series of adjacent bars or bins, with the width of each bin represen ng a range of values. The height of each bar represents the frequency or count of data points in that bin. Informa on: Histograms provide a detailed view of the data's shape, centre, spread, skewness, and poten al modes. Data Type: Histograms are primarily used for con nuous data, although they can be adapted for discrete data by adjus ng bin widths. Usage: Commonly used for exploring the distribu on of data, iden fying pa erns, and assessing data characteris cs.	Purpose: Box plots are used to display the distribu on, central tendency, and spread (variability) of a dataset. They are par cularly useful for iden fying outliers and comparing the distribu on of mul ple datasets. Appearance: A box plot consists of a rectangular "box" with a line inside it (the median), and "whiskers" that extend from the box. Some mes, individual data points are plo ed as dots. Informa on: A box plot provides informa on about the median, quar les (25th and 75th percen les), the interquar le range (IQR), and the presence of outliers. Data Type: Box plots are suitable for summarizing both con nuous and categorical data. Usage: Commonly used for comparing distribu ons between different groups or visualizing the spread of data.



Q 100. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 mes. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

You use Bayes Theorem to find the answer. Let's split problem into two parts:

1. What is the probability you picked the double-headed coin (now referred as D)?
2. What is the probability of g ng a head on the next toss?

PART1

We are trying to find the probability of having a double-headed coin. We know that the same coin has been flipped 10 mes, and we've go en 10 heads (in vely, you're probably thinking that there is a significant chance we have the double-headed coin). Formally, we're trying to find $P(D \mid 10\ heads)$.

Using Bayes rule:

$$P(D \mid 10\ H) = \frac{P(10\ H \mid D) * P(D)}{P(10H)}$$

- Tackling the numerator, the prior probability, $P(D) = 1/1000$.
- If we used the double headed coin, the chance of g ng 10 heads, $P(10\ H \mid D) = 1$ (we always flip heads).
- So, the $numerator = 1 / 1000 * 1 = 1 / 1000$.

- The denominator, P(10H) is just $P(10\ H \mid D) * P(D) + P(10\ H \mid Fair) * P(Fair)$.
  This makes sense because we are simply enumera ng over the two possible coins. The first part of P(10H) is the exact same as the numerator (1 / 1000).
- Then the second part: $P(Fair) = 999/1000$. $P(10\ H \mid Fair) = (1/2)^{\wedge} 10 = 1/1024$.
  Thus, $P(10\ H \mid Fair) * P(Fair) = 0.0009756$.
  The denominator then equals $0.001 + 0.0009756$.

Since we have all the components of $P(D \mid 10\ H)$, compute and you'll find the the probability of having a double headed coin is 0.506. We have finished the first ques on.

PART2

The second ques on is then easily answered: we just compute the two individual probabili es and add.

$$P(H) = P(D) * P(H \mid D) + P(Fair) * P(H \mid Fair)$$

$$= 0.506 * 1 + (1 - 0.506) * (0.5) = 0.753.$$

So, there is a 75.3% chance you will flip a head.

Q 101. What is a confidence interval and how do you interpret it?

A confidence interval is a sta s cal concept used to es mate a range of values within which a popula on parameter (such as a mean, propor on, or regression coefficient) is likely to fall with a certain level of confidence. It provides a measure of the uncertainty or variability associated with es ma ng a parameter from a sample of data.

Interpre ng a confidence interval:

<u>Example</u>: Suppose you calculate a 95% confidence interval for the average height of a popula on, and you obtain the interval [165 cm, 175 cm].

<u>Interpreta on</u>: You can interpret this confidence interval as follows:

"We are 95% confident that the true average height of the popula on falls within the range of 165 cm to 175 cm."

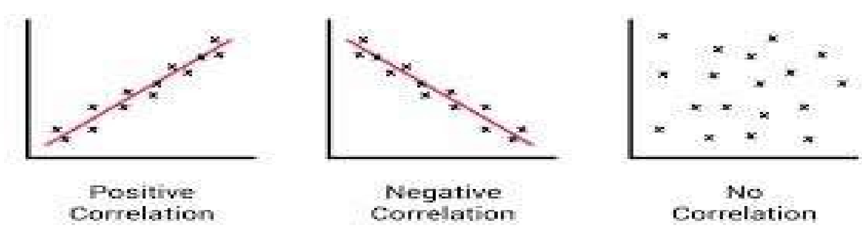
Q 102. How do you stay up-to-date with the new and upcoming concepts in sta s cs?

To stay up-to-date with new concepts in sta s cs

- Read Journals: Regularly read sta s cal journals and publica ons.
- Online Courses: Take online courses and webinars.
- Conferences: A end sta s cal conferences and workshops.
- Join Forums: Par cipate in online sta s cal forums and communi es.
- Network: Connect with sta s cians and data scien sts.
- Subscribe: Subscribe to sta s cal newsle ers and blogs.
- Follow Researchers: Follow leading sta s cians on social media.
- Con nuous Learning: Embrace a culture of con nuous learning.

Q 103. What is correla on?

Correla on is a sta s cal measure used to describe the degree to which two or more variables change together or are related to each other. In other words, it quan fies the strength and direc on of the linear rela onship between two or more variables.



<u>Key points about correla on</u>:

- Correla on Coefficient:
  The most common way to measure correla on is by calcula ng the correla on coefficient, which is represented by the symbol "r" or "ρ" (rho). The correla on coefficient is a numerical value that ranges between -1 and 1, with the following interpreta ons:

1. A posi ve correla on (r > 0) indicates that as one variable increases, the other tends to increase as well.
2. A nega ve correla on (r < 0) indicates that as one variable increases, the other tends to decrease.
3. A correla on coefficient of 0 (r = 0) suggests no linear rela onship between the variables.

- Strength of Correla on:
  The absolute value of the correla on coefficient (|r|) indicates the strength of the rela onship. Values closer to -1 or 1 represent stronger correla ons, while values closer to 0 represent weaker correla ons.
- Direc on of Correla on:
  The sign of the correla on coefficient (+ or -) indicates the direc on of the rela onship. A posi ve coefficient means the variables move in the same direc on, while a nega ve coefficient means they move in opposite direc ons.
- Sca erplots:
  Sca erplots are o en used to visually represent the rela onship between two variables. Points on the plot represent data points, and the pa ern they form can give an indica on of the correla on.

Q 104. What types of variables are used for Pearson's correla on coefficient?

Pearson's correla on coefficient, o en denoted as "r," is used to measure the strength and direc on of the linear rela onship between two con nuous variables. In other words, it is applied when both of the variables being studied are quan ta ve and numeric in nature.

Q 105.In an observa on, there is a high correla on between the me a person sleeps and the amount of produc ve work he does. What can be inferred from this?

A high correla on between the me a person sleeps and the amount of produc ve work they do suggests a significant rela onship between these two variables. However, it's important to note that correla on does not imply causa on. Here's what can be inferred and what cannot be inferred from this observa on:

<u>What Can Be Inferred</u>:

- Associa on: A high posi ve correla on implies that, on average, as the amount of me a person sleeps increases, their produc ve work also tends to increase. In other words, there appears to be a connec on between sleep and produc vity.
- Predic ve Value: The strength of the correla on can indicate the extent to which sleep me can be used to predict or es mate produc ve work. If the correla on is strong, sleep me may be a good predictor of work produc vity.
- Direc on: A posi ve correla on means that as one variable (sleep me) increases, the other variable (produc ve work) tends to increase as well. This suggests that ge ng more sleep is associated with higher produc vity, which aligns with common understanding.

Q 106. What does autocorrela on mean?

Autocorrela on, also known as serial correla on, refers to the correla on or rela onship between a variable and its past values in a me series or sequence of data points. In simpler terms, autocorrela on assesses how a data point at a given me is related to the data points that occurred at previous me points within the same series.

Q 107. How will you determine the test for the con nuous data?

Common tests for analyzing con nuous data in sta s cs include:

- T-Test: Used to compare means between two groups.
- Analysis of Variance (ANOVA): Compares means among three or more groups.
- Correla on Tests: Assess rela onships between con nuous variables, e.g., Pearson correla on or Spearman rank correla on.
- Regression Analysis: Predicts one con nuous variable based on one or more predictors.
- Chi-Squared Test for Independence: Examines associa ons between categorical and con nuous variables.
- ANOVA with Repeated Measures: ANOVA extension for within-subject or repeated measures designs.
- Mul variate Analysis of Variance (MANOVA): Extends ANOVA to analyze mul ple dependent variables simultaneously.

The choice of test depends on your research ques on, data distribu on, and experimental design.

Q 108. What can be the reason for non-normality of the data?

Non-normality of data, meaning that the data does not follow a normal distribu on (also known as a Gaussian distribu on), can occur for various reasons. It's important to iden fy the underlying causes of non-normality because the choice of sta s cal analysis and the interpreta on of results may depend on the distribu on of the data.

Here are some common reasons for non-normality:

- Skewness: Data may be skewed to the le (nega vely skewed) or right (posi vely skewed), leading to non-normality.
- Outliers: Extreme values or outliers in the dataset can distort the normal distribu on.
- Sampling Bias: Non-random sampling or selec on bias may result in data that does not reflect the popula on's true distribu on.
- Non-linear Rela onships: Data influenced by non-linear rela onships or complex interac ons may deviate from normality.
- Data Transforma on: Some data, such as counts or propor ons, inherently follow non-normal distribu ons.
- Natural Varia on: In some cases, data may naturally follow a non-normal distribu on due to the underlying process being studied.
- Measurement Errors: Errors in data collec on or measurement can introduce non-normality.
- Censoring or Floor/Ceiling Effects: Data may be bounded, leading to devia ons from normality.

Understanding the cause of non-normality is essen al for appropriate data analysis and choosing the right sta s cal techniques or transforma ons.

Q 109. Why is there no such thing like 3 samples t- test? why t-test failed with 3 samples?

There is no dedicated "3 samples t-test" because tradi onal t-tests are designed for comparing means between two groups, not three. When you have three or more groups to compare, you typically use analysis of variance (ANOVA) or its varia ons, which can determine whether there are sta s cally significant differences among the mul ple groups. T-tests can be applied to compare pairs of groups within an ANOVA framework, but they are not used to directly compare three groups simultaneously.