

STATISTICS INTERVIEW QUESTIONS

GOUTHAMI K

in <https://www.linkedin.com/in/gouthami-krishnamurthy/>

 <https://github.com/Gouthami-K>

Q 1. What are the most important topics in statistics?

Some of the important topics in statistics are:

- Measure of central tendency
- Measure of dispersion
- Covariance and Correlation
- Probability Distribution Function
- Standardization and normalization
- Central limit theorem
- Population and sample
- Hypothesis Testing

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

- EDA involves the process of visually and statistically analysing data to understand its underlying patterns distributions and relationships.
- The goal of EDA is to gain insights into the data, identify potential issues, and guide for the data processing steps.

Q 3. What are quantitative data and qualitative data?

Data can be categorized into two main types: quantitative data and qualitative data.

Quantitative Data-- (numeric)	Qualitative Data-- (categorical)
It is numbers-based, countable, or measurable.	It is interpretation-based, descriptive, and relating to language.
It is analyzed using statistical analysis.	It is analyzed by grouping the data into categories and themes.
Quantitative Data Types: Discrete Data and Continuous Data.	Qualitative Data Types: Nominal Data and Ordinal Data.
Ex: Age, Height, Weight, Income, Groupe size, Test score.	Ex: Gender, Marital status, Native language, Qualifications, Colours.

Q 4. What is the meaning of KPI in statistics?

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 organization.

They are used in various fields, including business, finance, healthcare, education, and more. The choice of KPIs depends on the goals and objectives of the organization or process being assessed.

By regularly monitoring and analyzing KPIs, organizations can identify areas of improvement, make data-driven decisions, and measure progress toward their strategic goals.

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

Univariate Analysis	Bivariate Analysis	Multivariate Analysis
Involves the examination of a single variable.	Involves examining the relationship between two variables.	Involves analyzing multiple variables simultaneously.
Analyzing the distributions, summary statistics, and characteristics.	Focuses on how changes in one variable are associated with changes in another variable.	We can observe, how multiple variables interact and influence each other.
Ex: Histograms, Box plots, Mean, Median, Standard deviation.	Ex: Scatter plots, Correlation coefficients, cross-tabulations.	Ex: Pairplot, Principal Component Analysis (PCA), Factor Analysis.

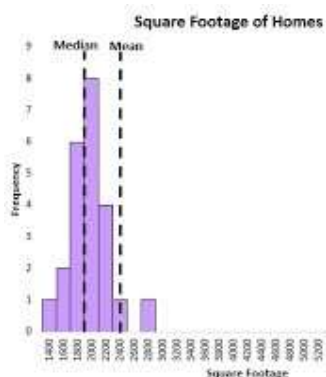
Q 6. How would you approach a dataset that's missing more than 30% of its values?

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

- **Mean/Median Imputation:**
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 distributions.
- **Mode Imputation:**
Impute missing values with the mode (most frequent value) of the variable for categorical data.
- **K-Nearest Neighbors (KNN) Imputation:**
Impute missing values by finding the nearest neighbors based on other variables.

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

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



- **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):

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

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}

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

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

Q 8. What is the difference between Descriptive and Inferential Statistics?

Descriptive statistics and inferential statistics are two fundamental branches of statistics that serve different purposes in data analysis. Here's an overview of the key differences between them:

Descriptive Statistics	Inferential Statistics
Used to summarize and describe the main features or characteristics of a dataset. They aim to provide a clear and concise overview of the data.	Used to make inferences or draw conclusions about a larger population based on a sample of data. They involve generalizing from a sample to a population.
Typically used at the initial stage of data analysis to understand the dataset and identify patterns, trends, and important features.	Typically used after the initial data exploration (descriptive statistics) when researchers want to make predictions, test hypotheses, or make statements about a population.
Are generally applied to both populations and samples. They can be used to summarize data from a complete population or from a sample drawn from the population.	Are focused on making statements or inferences about a population based on data from a sample. They involve estimating population parameters and assessing the uncertainty associated with those estimates.
Examples: Common descriptive statistics include measures of central tendency (e.g., mean, median, mode), measures of dispersion (e.g., range, variance, standard deviation), frequency distributions, histograms, and summary tables.	Examples: Common inferential statistical techniques include hypothesis testing, confidence intervals, regression analysis, analysis of variance (ANOVA), chi-square tests, and various forms of multivariate analysis.

Q 9. Can you state the method of dispersion of the data in statistics?

In statistics, 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 sensitive to outliers.
- **Variance:**
Variance quantifies the average squared difference between each data point and the mean. It is calculated by taking the average of the squared deviations from the mean.
- **Standard Deviation:**
The standard deviation 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.

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

- **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 sensitive to outliers?

Yes, the range is sensitive 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 information, unusual events, or rare occurrences that are relevant to the analysis, such as detecting anomalies, understanding extreme behavior, or studying unique cases.

Q 13. What is the meaning of standard deviation?

- The standard deviation is a statistical measure that quantifies the amount of variation 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 deviation helps us understand the extent to which individual data points deviate from the mean.
- The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

Q 14. What is Bessel's correction?

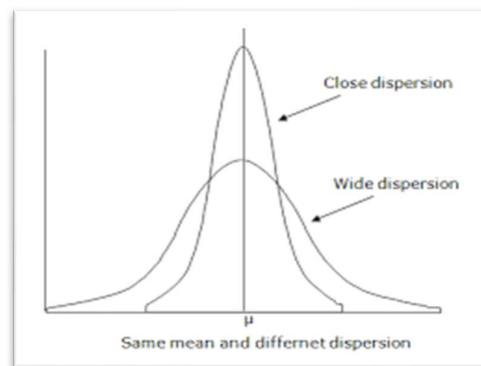
Bessel's correction is a statistical adjustment made to the formula for calculating the sample variance and sample standard deviation. It is used to provide a more accurate estimate of the population variance and standard deviation when working with a sample from a larger population.

The key idea behind Bessel's correction is that when you calculate the variance or standard deviation using sample data (rather than data from the entire population), you tend to underestimate the true population variance or standard deviation. This underestimation occurs because you are basing your calculations on a smaller subset of the data.

Bessel's correction adjusts for this underestimation by dividing the sum of squared differences from the mean by $(n - 1)$, where "n" is the sample size. In contrast, when calculating population variance and standard deviation, you divide by "n" (the actual population size). By using $(n - 1)$ instead of "n" in the formula, Bessel's correction increases the calculated variance and standard deviation slightly, making them more representative of the population.

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

In the context of data distributions and statistics, these terms describe the degree of variability or dispersion in the data.



Spread Out Curve (Wider Dispersion)	Concentrated Curve (Narrower Dispersion)
A spread-out curve or distribution typically has a larger spread or range of values. This means that the data points are more spread out from each other.	A concentrated curve or distribution 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 deviation and a larger range or interquartile range (IQR).	It is associated with a lower standard deviation and a smaller range or interquartile range (IQR).
In graphical representations, it often results in a wider or flatter distribution with a larger spread of data points.	In graphical representations, it often results in a narrower, taller distribution with data points clustered closely together.
Example: A dataset of income levels for a diverse population, where some individuals have very high incomes, and others have very low incomes, creating a wide spread.	Example: A dataset of test scores for a group of students who all scored very close to each other, creating a concentrated distribution.

Q 16. Can you calculate the coefficient of variation?

- The coefficient of variation (CV) is a measure of relative variability and is calculated as the ratio of the standard deviation (σ) to the mean (μ) of a dataset. It is often expressed as a percentage to make it more interpretable.
- The formula for calculating the coefficient of variation is as follows:

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

Where:

CV= Coefficient of variation, σ = standard deviation of the dataset, μ = mean of the dataset.

- The coefficient of variation is particularly useful when you want to compare the relative 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 relative 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 statistics for both classes:
Class A: Mean Score = 85, Standard Deviation = 10
Class B: Mean Score = 90, Standard Deviation = 8

Now, let's calculate the coefficient of variation for both classes:

For Class A: $CV = (\sigma / \mu) * 100 = (10 / 85) * 100 \approx 11.76\%$

For Class B: $CV = (\sigma / \mu) * 100 = (8 / 90) * 100 \approx 8.89\%$

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

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

Mean imputation 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 Imputation:

- **Bias Introduction:**
Mean imputation can introduce bias into the dataset.
- **Loss of Variability:**
Imputing missing values with the mean reduces the variability of the data because all imputed values are the same.
- **Disregards Data Patterns:**
Mean imputation does not take into account any underlying patterns or relationships in the data. It treats all missing values as if they were independent of other variables or conditions, which may not be the case.
- **Impact on Model Performance:**
In machine learning, mean imputation can negatively impact model performance, especially when missing values are related to the target variable or when they carry important information. It can lead to inaccurate predictions and reduced model effectiveness.
- **Imputation of Categorical Data:**
Mean imputation is primarily suitable for numerical data. When dealing with categorical data, other imputation methods like mode imputation (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 statistics and data analysis that provide several benefits for visualizing and summarizing data distributions.

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

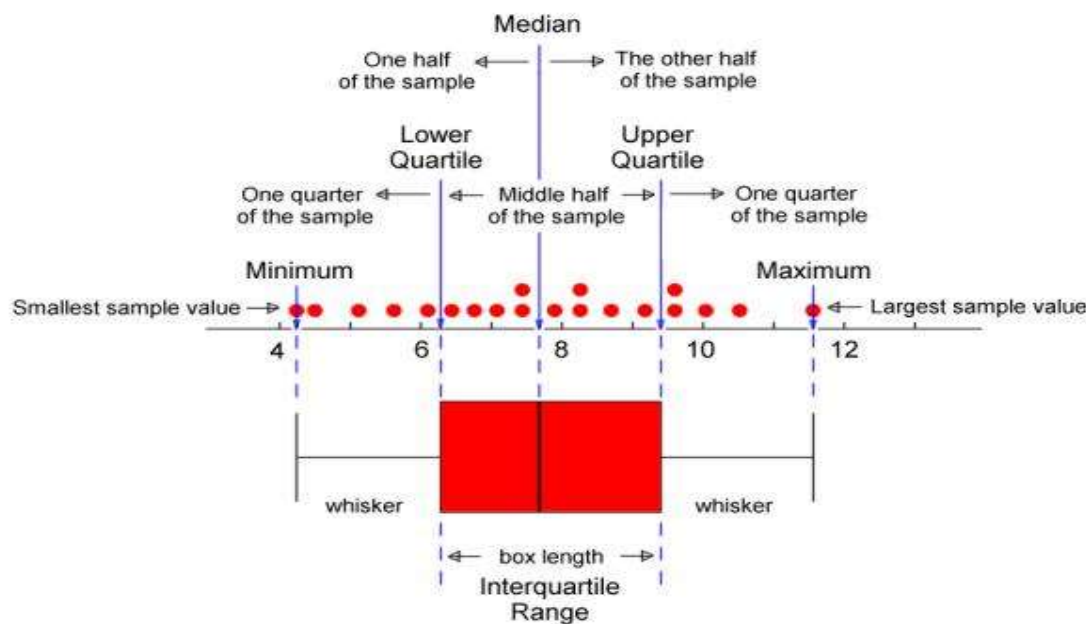
- Summary of Data Distribution
- Identification of Outliers
- Comparison of Multiple Groups
- Detection of Skewness
- Visualization of Quartiles
- Robustness to Outliers
- Ease of Interpretation
- Data Quality Assessment

Q 19. What is the meaning of the five-number summary in Statistics?

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:

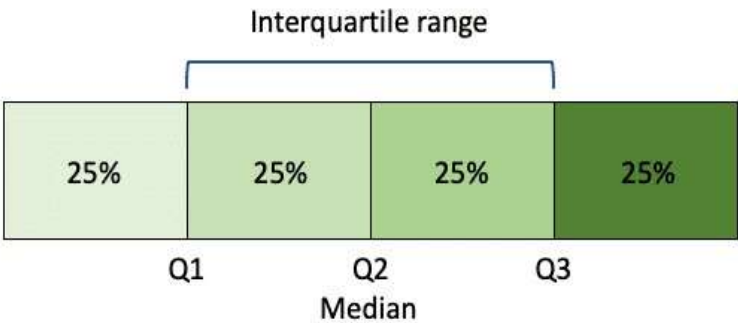
- a. Minimum (Min): This is the smallest value in the dataset, representing the lowest data point. It gives you an idea of the floor or lower boundary of the data.
- b. First Quartile (Q1): The first quartile, also known as the lower quartile, is the value below which 25% of the data falls. It marks the 25th percentile of the dataset and represents the lower boundary of the middle 50% of the data.
- c. Median (Q2): The median, or the second quartile, 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 Quartile (Q3): The third quartile, also known as the upper quartile, is the value below which 75% of the data falls. It marks the 75th percentile 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, representing the highest data point. It gives you an idea of the ceiling or upper boundary of the data.



The five-number summary is often used to create box plots (box-and-whisker plots), which provide a visual representation of these five summary statistics. Box plots are helpful for understanding the spread, central tendency, and presence of outliers in a dataset. The box in the plot represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3), while the whiskers extend to the minimum and maximum values, indicating the range of the data.

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

- The 1st quartile (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 quartile (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 quartile (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 quartiles 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 percentile?

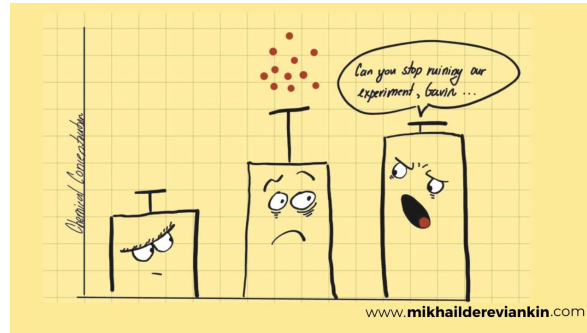
Percent and percentile are related concepts in statistics, but they have distinct meanings.

Percent	Percentile
Percent is a unit of measurement denoted by the symbol "%."	Percentile is a statistical concept used to describe a specific position or location within a dataset.
It represents a proportion or fraction of a whole, divided by 100. In other words, when you express a quantity as a percentage, you are dividing it by 100.	It represents the value below which a given percentage of the data falls. Percentiles are used to understand the distribution of data and identify how a particular data point ranks in comparison to others.
For example, 25 percent (25%) is equivalent to 0.25 or 25/100. It means 25 out of every 100, or one-quarter of the whole.	For example, the 25th percentile (also known as the first quartile, 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 observation that is unusually distant from other observations in the dataset.
- Outliers can be either exceptionally high values (positive outliers) or exceptionally low values (negative outliers).

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



1. Negative Impacts:

- Influence on Measures of Central Tendency:
A single extreme outlier can pull the mean in its direction, making it unrepresentative of the majority of the data.
- Impact on Dispersion Measures:
The presence of outliers can inflate the measures like the standard deviation and the interquartile range (IQR), making them larger than they would be without outliers.
- Skewing Data Distributions:
Positive outliers can result in right-skewed distributions, while negative outliers can result in left-skewed distributions. This can affect the interpretation of the data.
- Misleading Summary Statistics:
Outliers can distort the interpretation of summary statistics.
- Impact on Hypothesis Testing:
Outliers can affect the results of hypothesis tests. They can lead to incorrect conclusions, such as detecting significant differences when none exist or failing to detect real differences when outliers mask them.

2. Positive Impacts:

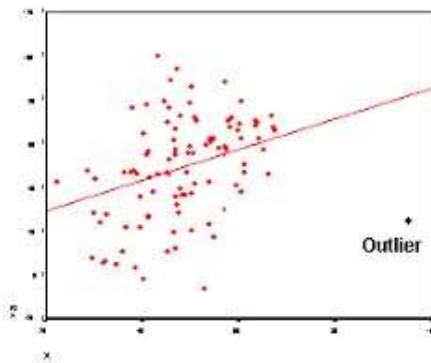
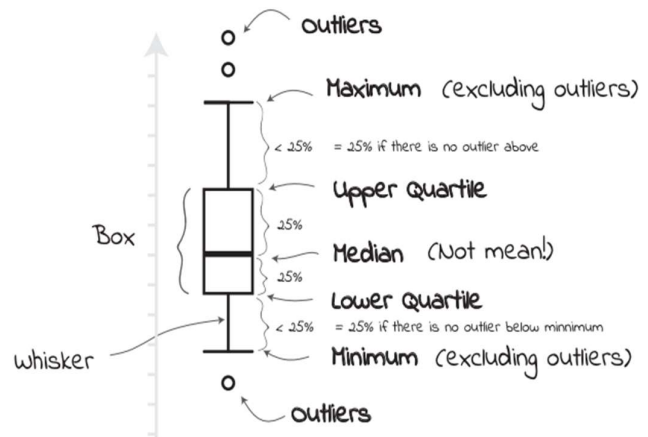
- Detection of Anomalies:
Outliers can signal the presence of anomalies or rare events in a dataset. Identifying these anomalies can be valuable in various fields, including fraud detection, quality control, and outlier detection in scientific experiments.
- Robust Modeling:
In some cases, outliers can be genuine observations that are important to model. For example, in financial modeling, extreme stock price movements may contain valuable information for predicting market trends.

Q 24. Mention methods to screen for outliers in a dataset.

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

- Box Plots (Box-and-Whisker Plots):

Box plots provide a visual representation of the distribution of data, including the identification of potential outliers. In a box plot, outliers are typically shown as individual data points beyond the whiskers of the plot.



- Scatterplots:

Scatterplots are particularly useful for identifying outliers in bivariate or multivariate data. Outliers can appear as data points that are far from the main cluster of points in the scatterplot.

- Z-Scores:

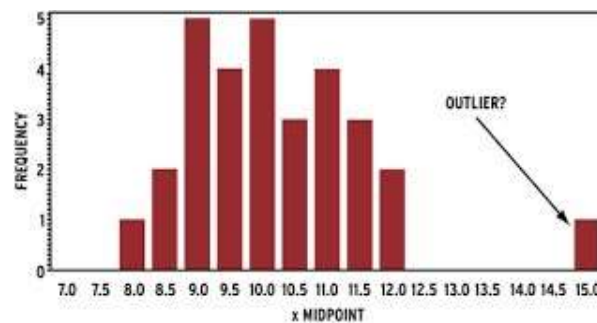
Z-scores (standard scores) measure how many standard deviations a data point is away from the mean. Data points with high absolute Z-scores (typically greater than 2 or 3) are often considered potential outliers.

- IQR (Interquartile Range) Method:

The IQR method involves calculating the interquartile range ($IQR = Q3 - Q1$) and then identifying values that fall below $Q1 - 1.5 * IQR$ or above $Q3 + 1.5 * IQR$ as potential outliers.

- Visual Inspection:

Sometimes, simple visual inspection of the data through histograms, QQ-plots (quantile-quantile plots), or other visualization techniques can reveal the presence of outliers.



It's important to note that the choice of outlier detection method should be guided by the characteristics 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 objectives. Here are several methods for handling outliers:

- Data Truncation or Removal:
One common approach is to simply remove outliers from the dataset. This should be done cautiously, especially if the outliers represent valid and important observations. Removing outliers is appropriate when they are likely the result of data entry errors or measurement errors.
- Data Transformation:
Transforming the data can be a useful way to mitigate the impact of outliers. Common transformations include logarithmic, square root, or inverse transformations. These transformations tend to compress the range of extreme values.
- Winsorization:
Winsorization involves capping or limiting extreme values by replacing them with a specified percentile value. For example, you might replace values above the 95th percentile with the value at the 95th percentile.
- Imputation:
For missing values that are not extreme outliers, you can impute them using various methods, such as mean imputation, median imputation, or more advanced techniques like regression imputation.
- Robust Statistics:
Using robust statistical methods that are less sensitive to outliers can be an effective approach. For example, replacing the mean with the median and using the interquartile range (IQR) instead of the standard deviation can make statistical analysis more robust.
- Model-Based Approaches:
In predictive modeling, consider using algorithms that are less sensitive to outliers, such as robust regression methods or ensemble methods like random forests, which can handle outliers better than linear regression.
- Domain Knowledge:
Rely on domain knowledge to understand the context of the outliers. Sometimes, what appears as an outlier may be a valid and important data point. Consult with domain experts to determine the appropriateness of handling outliers.
- Reporting 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 interquartile range?

Calculating the range and interquartile range (IQR) is a straightforward process involving the use of basic statistical formulas. Here's how to calculate both the range and the IQR:

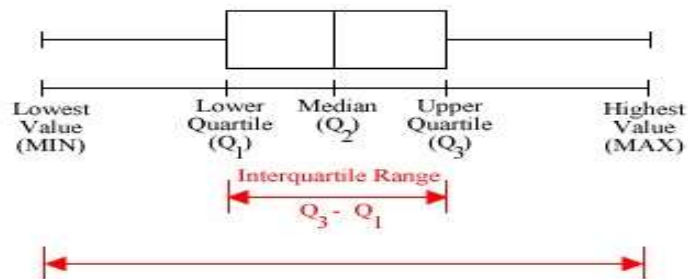
- Range:
The range is the simplest measure of spread in a dataset. It is the difference between the maximum and minimum values in the dataset.

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

- Interquartile Range (IQR):

The interquartile 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 quartile (Q3) and the first quartile (Q1) of the dataset.

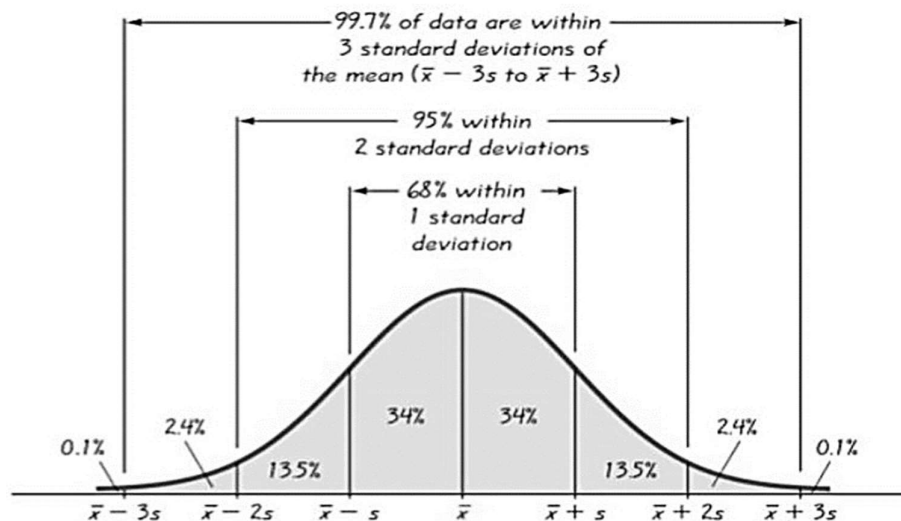
$$IQR = Q_3 - Q_1$$



Range

Q 27. What is the empirical rule?

The Empirical Rule



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

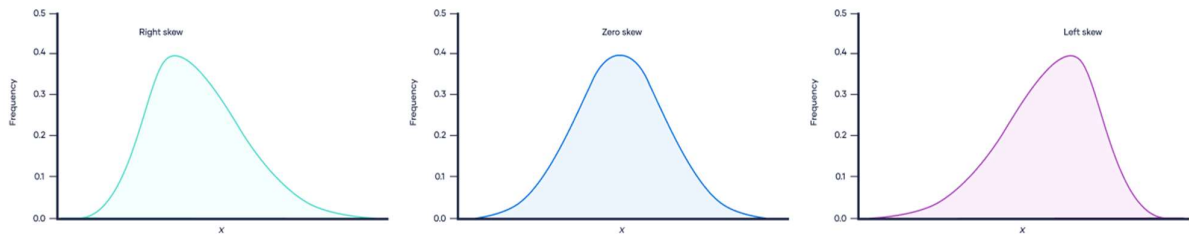
The empirical rule states that:

- Approximately 68% of the data falls within one standard deviation of the mean.
- Approximately 95% of the data falls within two standard deviations of the mean.
- Approximately 99.7% of the data falls within three standard deviations of the mean.

Q 28. What is skewness?

Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images.

A distribution can have right (or positive), left (or negative), or zero skewness. A right-skewed distribution is longer on the right side of its peak, and a left-skewed distribution is longer on the left side of its peak:



Q29. What are the different measures of Skewness?

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

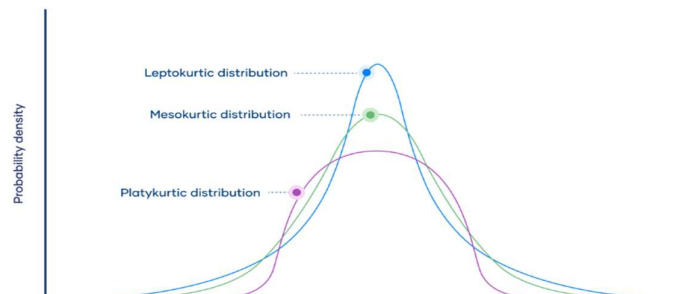
- Pearson's First Coefficient of Skewness (or Moment Skewness)
- Fisher-Pearson Standardized Moment Coefficient of Skewness (or Sample Skewness)
- Bowley's Coefficient of Skewness (or Quartile Skewness)

Q 30. What is kurtosis?

Kurtosis is a statistical measure that quantifies the "tailedness" or "peakedness" of the probability distribution 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 distribution.

Kurtosis classifications based on the shape of the data distribution:

- Mesokurtic
- Leptokurtic
- Platykurtic



Q 31. Where are long-tailed distributions used?

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

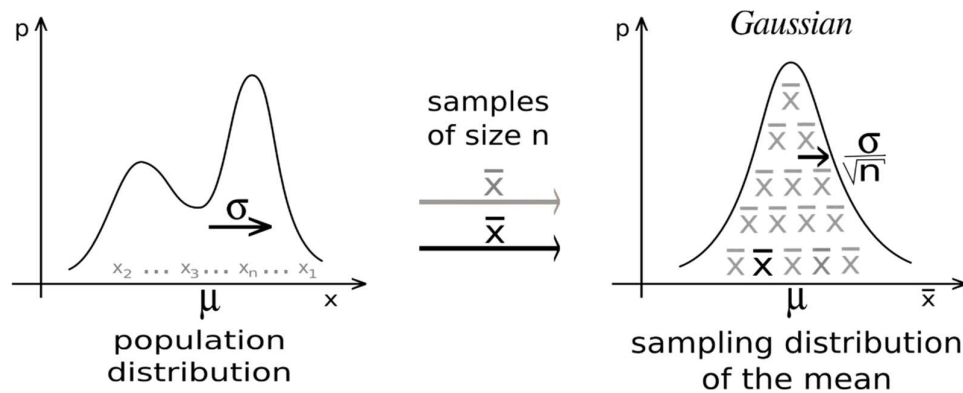
- **Finance and Risk Management:**
Long-tailed distributions are frequently used to model asset returns, market volatility, and financial risk.

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

- **Insurance:**
Insurance companies use long-tailed distributions to model insurance claims. These distributions 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 distributions are used to estimate the likelihood of extreme events occurring.
- **Epidemiology:**
Epidemiologists may use long-tailed distributions to model the spread of infectious diseases, as they account for sporadic outbreaks or superspreading events.

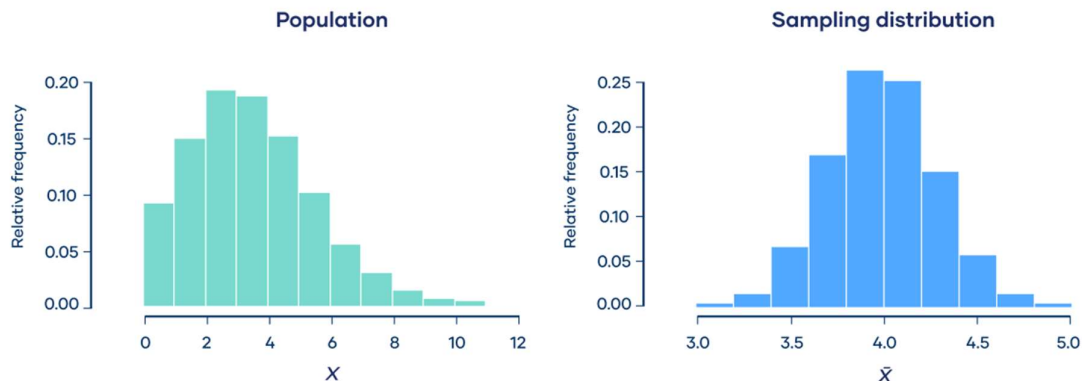
Q 32. What is the central limit theorem?

In probability theory, the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger i.e., $n \geq 30$, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.



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

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



Q 34. What general conditions must be satisfied for the central limit theorem to hold?

For the Central Limit Theorem (CLT) to hold:

- **Random Sampling:**
Data must be randomly selected from the population.
- **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 population should have a finite variance.
- **Identical Distribution:**
Ideally, data should come from a population with the same distribution.

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

Q 35. What is the meaning of selection bias?

Selection bias is the bias that occurs during the sampling of data. This kind of bias occurs when a sample is not representative of the population, which is going to be analyzed in a statistical study.

Q 36. What are the types of selection bias in statistics?

There are many types of selection bias as shown below:

- Observer selection
- Attrition
- 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 getting 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 $6 \times 6 = 36$ 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 getting a sum of 8 when rolling two fair dice is:

$$\text{Probability} = \frac{(\text{Number of Favorable Outcomes})}{(\text{Total Number of Possible Outcomes})} = \frac{5}{36}$$

Therefore, the probability is 5/36.

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

Probability distributions are mathematical functions that describe the likelihood of different outcomes or events in a random process. There are several types of probability distributions, each with its own characteristics and applications.

There are two main types of probability distributions: Discrete and Continuous.

1. Discrete Probability Distributions:

In a discrete probability distribution, the random variable can only take on distinct, separate values, often integers. Common examples of discrete probability distributions include:

- a. Bernoulli Distribution
- b. Binomial Distribution
- c. Poisson Distribution

2. Continuous Probability Distributions:

In a continuous probability distribution, the random variable can take on any value within a specified range. Common examples of continuous probability distributions include:

- a. Normal Distribution (Gaussian Distribution)
- b. Uniform Distribution
- c. Log-Normal Distribution
- d. Power Law
- e. Pareto Distribution

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

A normal distribution, also known as a Gaussian distribution or a bell curve, is a fundamental statistical concept in probability theory and statistics. It is a continuous probability distribution that is characterized by a specific shape of its probability density function (PDF), which has the following key properties:

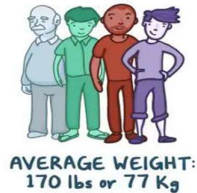
- Symmetry: The normal distribution is symmetric, meaning that it is centred around a single peak, and the left and right tails are mirror images of each other. The mean, median, and mode of a normal distribution are all equal and located at the centre of the distribution.
- Bell-shaped: The PDF of a normal distribution has a bell-shaped curve, with the highest point (peak) at the mean value and gradually decreasing probabilities as you move away from the mean in either direction.
- Mean and Standard Deviation: The normal distribution is fully characterized by two parameters: the mean (μ) and the standard deviation (σ). The mean represents the centre of the distribution, while the standard deviation controls the spread or dispersion of the data. Larger standard deviations result in wider distributions.
- Empirical Rule: The normal distribution 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 deviation of the mean.
 - b. About 95% of the data falls within two standard deviations of the mean.
 - c. About 99.7% of the data falls within three standard deviations of the mean.

- Continuous: The normal distribution is a continuous probability distribution, which means that it can take on an infinite number of values within its range. There are no gaps or discontinuities in the distribution.

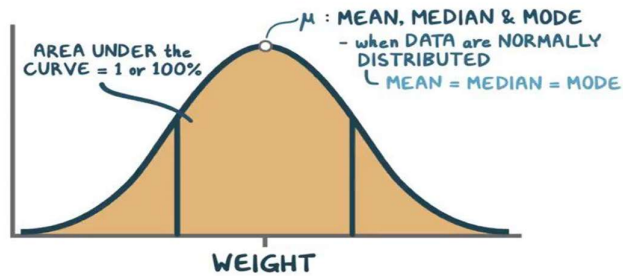
Many natural phenomena, such as weights, heights and IQ scores, approximate a normal distribution. It is also fundamental in hypothesis testing and statistical modeling.

HISTOGRAM:

↳ PLOT that SHOWS DISTRIBUTION of any MEASUREMENT or DATA



NORMAL DISTRIBUTION or BELL CURVE



Q 40. Can you state the formula for normal distribution?

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

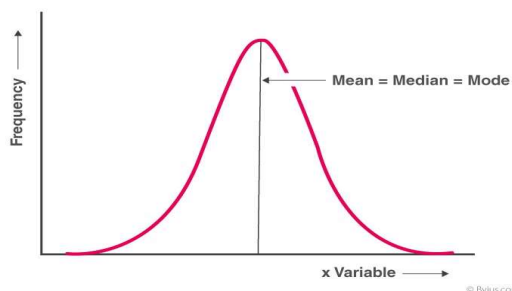
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where:

- $f(x)$ is the probability density function at a given value of
- μ is the mean of the normal distribution.
- σ is the standard deviation of the normal distribution.
- π is the mathematical constant pi (approximately 3.14159).
- e is the base of the natural logarithm (approximately 2.71828).

Q 41. What is the relationship between mean and median in a normal distribution?

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



Q 42. What are some of the properties of a normal distribution?

A normal distribution, also known as a Gaussian distribution or bell curve, has several key properties:

- Bell-Shaped Curve: The distribution 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 distribution.
- Tails Extend to Infinity: The tails of the curve stretch infinitely in both directions, but they get closer and closer to the horizontal axis as they go farther from the mean.
- Standard Deviation Controls Spread: The width of the bell curve is determined by the standard deviation. A larger standard deviation makes the curve wider, and a smaller one makes it narrower.
- Empirical Rule: This rule helps you estimate where data points are likely to be within the distribution. It's based on the 68-95-99.7 rule. Approximately 68% of the data falls within one standard deviation of the mean, about 95% falls within two standard deviations, and roughly 99.7% falls within three standard deviations.
- Used in Many Real-Life Situations: The normal distribution is commonly seen in nature and in human-made systems, including things like height measurements, IQ scores, and errors in manufacturing.
- Easy for Statistical Analysis: Because of its well-defined properties, the normal distribution is often used in statistics for modeling and making predictions about data.

Q 43. What is the assumption of normality?

The assumption of normality in statistics is the idea that data or residuals in a statistical analysis should follow a bell-shaped, symmetric, and continuous probability distribution called the normal distribution.

Q 44. How to convert normal distribution to standard normal distribution?

Converting a normal distribution to a standard normal distribution involves a process called "**standardization**" or "**normalization**". This process transforms the values from the original normal distribution into equivalent values that follow a standard normal distribution with a mean of 0 and a standard deviation of 1.

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

- **Determine the Mean and Standard Deviation of the Original Normal Distribution:**
Identify the mean (μ) and standard deviation (σ) of the original normal distribution.
- **Calculate the Z-Score:**
The Z-score (also known as the standard score) measures how many standard deviations a particular value is from the mean in the original distribution.

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 distribution that you want to convert.

μ is the mean of the original distribution.

σ is the standard deviation of the original distribution.

- **The Resulting Z-Score Represents the Standard Normal Distribution:**

The Z-score you calculate in step 2 represents the equivalent value in a standard normal distribution.

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

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

In a standard normal distribution, also known as the standard normal or Z-distribution, the range of possible values extends from negative infinity ($-\infty$) to positive infinity ($+\infty$).

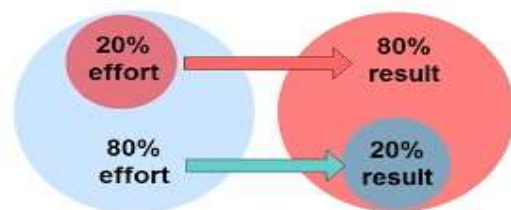
However, it's important to note that while the range of possible values is theoretically infinite, the vast majority of values in a standard normal distribution are concentrated within a relatively narrow range around the mean, which is 0. The distribution is bell-shaped, and as you move away from the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

Statistically, most of the values in a standard normal distribution fall within a few standard deviations of the mean. Approximately:

This means that the values within the range of roughly -3 to +3 standard deviations from the mean cover the vast majority of observations in a standard normal distribution. 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 a principle named after the Italian economist Vilfredo Pareto.
- It suggests that, in many situations, a small percentage of causes or inputs is responsible for a large percentage of the results or outputs.
- In its simplest form, the Pareto Principle states that roughly 80% of the effects come from 20% of the causes.



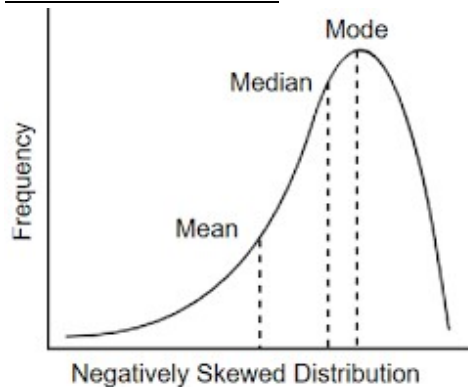
Q 47. What are left-skewed and right-skewed distributions?

Left-skewed and right-skewed distributions, also known as negatively skewed and positively skewed distributions, are types of asymmetric distributions in statistics. They describe the shape of the distribution of data points in a dataset.

1. Left-Skewed (Negatively Skewed) Distribution:

- Left-skewed distributions have a longer tail on the left (or negative) side of the distribution.
- The peak of the distribution (mode) is typically located to the right of the centre.
- The mean (average) is typically less than the median.
- In a left-skewed distribution, the data is concentrated on the right side and tails off to the left.

Left-Skewed Distribution

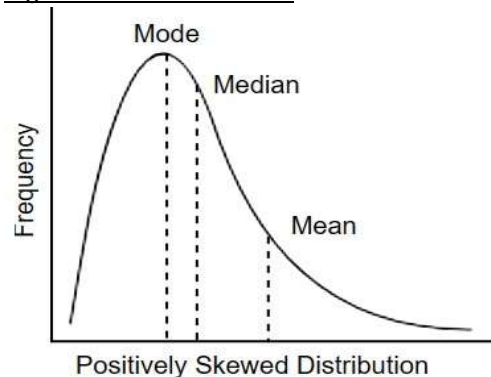


Example: The distribution of ages at retirement may be left-skewed, as most people retire around a certain age, but very few retire at a younger age.

2. Right-Skewed (Positively Skewed) Distribution:

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

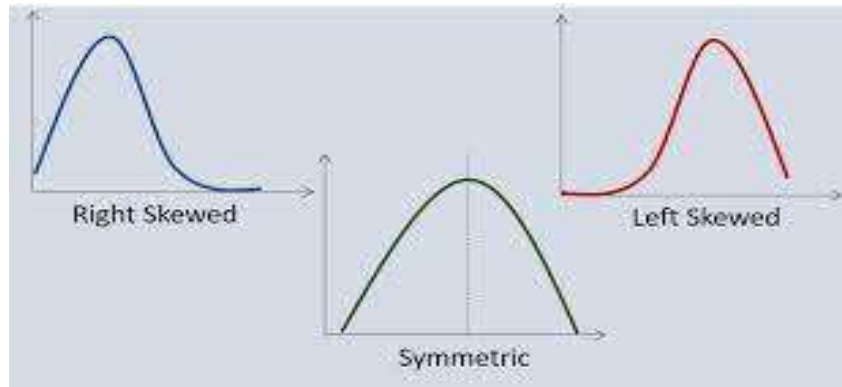
Right-Skewed Distribution



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

Skewness is a measure used to quantify the degree of asymmetry in a distribution.

- A positive skewness value indicates right-skewness.
- A negative skewness value indicates left-skewness.
- A skewness of 0 indicates a perfectly symmetrical distribution.



Understanding the skewness of a dataset is essential in statistics because it can affect the choice of appropriate statistical analyses and modeling techniques. Left-skewed and right-skewed distributions often require different approaches for analysis and interpretation.

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

If a distribution is skewed to the right (positively skewed) and has a median of 20, then the mean will typically be greater than 20.

In a positively skewed distribution:

- The tail of the distribution extends to the right, meaning there are some relatively large values that pull the mean in that direction.
- 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 positively skewed distribution.

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

In a left-skewed (negatively skewed) distribution with a median of 60:

Mean, Median and Mode Relationship:

- Since the distribution is left-skewed, it means that the tail of the distribution is on the left side, and there are some relatively small values that are pulling the mean in that direction.
- The median, being the middle value, is less affected by extreme values in the tail. In a left-skewed distribution, the median is typically greater than the mean.
- In a left-skewed distribution, the mode is typically greater than the median and the mean. It is often closer to the peak of the distribution, which is located to the right of the centre.

In summary, you can conclude that in a left-skewed distribution 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 examination. The test has a mean score of 160, and it has a standard deviation 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 calculating a score from a z-score in a normal distribution:

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

In this case:

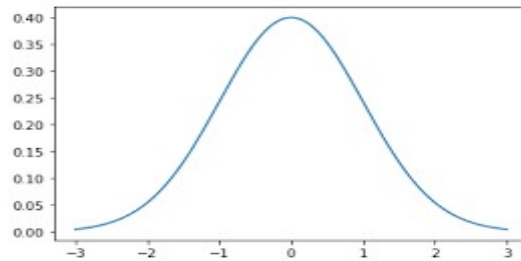
$Z=1.20$ (Jeremy's z-score), $\sigma=15$ (standard deviation), $\mu=160$ (mean)

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

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 known as the standard normal distribution or the Z-distribution, is a specific type of normal distribution with a mean (average) of 0 and a standard deviation of 1.



Q 52. What is the meaning of covariance?

Covariance is a measure of the relationship 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 direction (positive covariance) or in opposite directions (negative 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 characteristics of the shape of a data distribution:

1. Unimodal Curve:

- Definition: A unimodal curve represents a data distribution with a single distinct peak or mode, meaning that there is one value around which the data cluster the most.
- Shape: Unimodal distributions are typically symmetric or asymmetric but have only one primary peak.

Examples: A normal distribution, where data is symmetrically distributed around the mean, is a classic example of a unimodal curve. Other unimodal distributions can be skewed to the left (negatively skewed) or to the right (positively skewed).

2. Bimodal Curve:

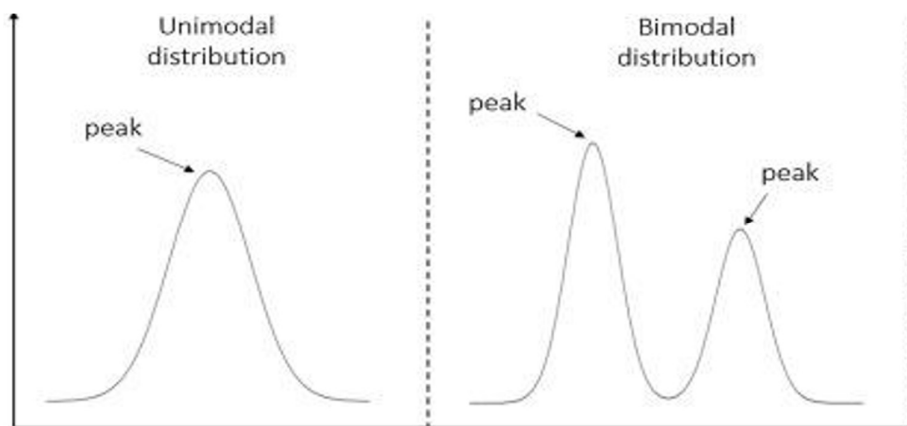
- Definition: A bimodal curve represents a data distribution with two distinct peaks or modes, indicating that there are two values around which the data cluster the most.
- Shape: Bimodal distributions have two primary peaks separated by a trough or dip in the distribution.

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

3. Bell-Shaped Curve:

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

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



Q 54. Does symmetric distribution need to be unimodal?

No, a symmetric distribution does not necessarily need to be unimodal. A symmetric distribution 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 distribution's centre point (usually the mean or median).

So, while symmetry and unimodality often go together, symmetry does not inherently require unimodality, and a symmetric distribution can have multiple modes.

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

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

1. **Income Distribution:** Income data is often right-skewed, with most people earning average incomes and a few earning very high incomes. This leads to a distribution that does not follow a normal curve.
2. **Stock Returns:** Daily stock returns can have fat tails and exhibit volatility clustering, making their distribution non-normal. Events like stock market crashes can cause significant deviations from normality.
3. **Website Traffic:** The number of visitors to a website on any given day often follows a distribution with a long tail. A few days with extremely high traffic can result in a skewed distribution.
4. **Ages at Retirement:** The distribution of ages at which people retire can be left-skewed, with many retiring around a certain age and very few retiring at younger ages.
5. **Number of Customer Arrivals:** The number of customers arriving at a store or service centre follows a Poisson distribution, which is discrete and not normal.
6. **Test Scores:** Test scores, particularly in educational settings, often have a distribution with modes due to various subpopulations of students, leading to a multimodal distribution.
7. **City Population Sizes:** The distribution of city population sizes worldwide is often right-skewed, with a few megacities having very high populations and the majority of cities having smaller populations.
8. **Wait Times:** The distribution of wait times in queues or lines can often 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 distribution, with a few posts going viral and receiving a disproportionate number of interactions.
10. **Height and Weight:** While human height and weight often follow roughly normal distributions, they can also be influenced by factors like nutrition and genetics, leading to deviations from normality in some populations.

These examples illustrate that real-world data can take on various shapes and characteristics, and not all datasets follow the idealized Gaussian or normal distribution. Understanding the distribution of data is essential for making accurate statistical inferences and modeling.

Q 56. What is the Binomial Distribution Formula?

The binomial distribution 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 function (PMF) of the binomial distribution is given by the formula:

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

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 ($q = 1 - p$).
- $\binom{n}{k}$ represents the binomial coefficient, which is often calculated as $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, where "!" denotes factorial.

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

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

- **Fixed Number of Trials (n):**
The experiment consists of a fixed number of identical, 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 (often 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 mentioned above (fixed n , independence, constant p , and binary outcomes).

Q 58. What are the examples of symmetric distribution?

Symmetric distributions 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 distributions include:

- **Normal Distribution (Gaussian Distribution)**
 1. The most well-known symmetric distribution.
 2. Bell-shaped and characterized by its mean and standard deviation.
 3. Many natural phenomena and measurements, such as height and weight in a population, closely follow a normal distribution.
- **Uniform Distribution**
 1. In a continuous uniform distribution, all values within an interval have equal probability.
 2. In a discrete uniform distribution, all outcomes have equal probability.
 3. For example, rolling a fair six-sided die follows a discrete uniform distribution.
- **Logistic Distribution**
 1. S-shaped curve similar to the normal distribution but with heavier tails.
 2. Often used in logistic regression and modeling growth processes.


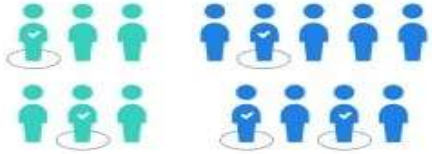
Q 59. Briefly explain the procedure to measure the length of all sharks in the world.

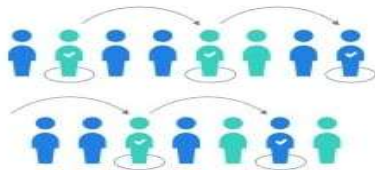
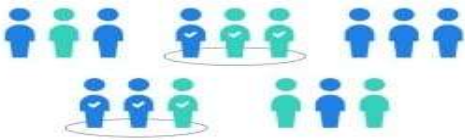

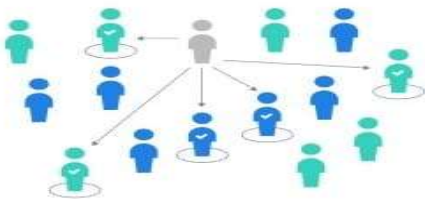
- Define the confidence level (most common is 95%)
- Take a sample of sharks from the sea (to get better results the number of fishes > 30)
- Calculate the mean length and standard deviation of the lengths
- Calculate t-statistics
- Get the confidence interval in which the mean length of all the sharks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

Q 60. What are the types of sampling in Statistics?

In statistics, sampling is the process of selecting a subset of individuals or items from a larger population to make inferences about the entire population. 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:

<p>1. Simple Random Sampling:</p> <ul style="list-style-type: none"> • Involves randomly selecting individuals or items from the population without any specific pattern or criteria. • Every member of the population 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). 	<p>Simple random sample</p>  <p>The diagram shows a population of 12 stylized human figures arranged in two rows of six. The top row has four blue figures and two green figures. The bottom row has three blue figures and three green figures. Four figures are circled, representing a random sample: the first blue figure in the top row, the second green figure in the top row, the third blue figure in the bottom row, and the fourth green figure in the bottom row.</p>
<p>Stratified sample</p>  <p>The diagram shows a population of 12 stylized human figures arranged in two groups of six. The left group consists of four green figures and two blue figures. The right group consists of four blue figures and two green figures. Within each group, two figures are circled, representing a stratified sample where individuals are selected from each subgroup.</p>	<p>2. Stratified Sampling:</p> <ul style="list-style-type: none"> • Divides the population into non-overlapping subgroups or strata based on certain characteristics (e.g., age, gender, location). • 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.

<p>3. Systematic Sampling:</p> <ul style="list-style-type: none"> • Involves selecting every nth individual/item from a list or sequence. • Typically, a random starting point is chosen, and then every nth individual/item is selected. • Useful when there's a natural order or sequence in the population. 	<p>Systematic sample</p>  <p>The diagram shows two rows of stylized human figures in blue and green. In the top row, a starting point is marked with a circle, and arrows indicate a regular interval selection of figures. In the bottom row, a similar process is shown with a different starting point and interval.</p>
<p>Cluster sample</p>  <p>The diagram shows a collection of stylized human figures arranged in small groups. One group is circled, indicating it has been selected as a cluster for sampling.</p>	<p>4. Cluster Sampling:</p> <ul style="list-style-type: none"> • Divides the population into clusters or groups, often 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 populations.
<p>5. Convenience Sampling:</p> <ul style="list-style-type: none"> • Involves selecting individuals or items that are readily available and convenient to sample. • Often used in exploratory or preliminary research but can introduce bias because it may not be representative of the entire population. 	<p>Convenience sample</p>  <p>The diagram shows a group of stylized human figures. A central figure is highlighted with a circle, and arrows point to it from various other figures, illustrating selection based on availability or convenience.</p>
<p>Purposive sample</p>  <p>The diagram shows a group of stylized human figures. A central figure is highlighted with a circle, and arrows point to it from various other figures, illustrating selection based on specific criteria or judgment.</p>	<p>6. Purposive Sampling (Judgmental Sampling):</p> <ul style="list-style-type: none"> • Involves selecting individuals/items based on the researcher's judgment and specific criteria. • Useful when the researcher wants to focus on a particular subgroup or characteristic. • Can be biased if not done carefully.

The choice of sampling method depends on the research objectives, available resources, and the characteristics of the population being studied. Each method has its own strengths and limitations, and researchers must consider these factors when designing and conducting a study.

Q 61. Why is sampling required?

Sampling is required for several simple and practical reasons:

1. **Efficiency:** Sampling is faster and more cost-effective than collecting data from an entire population, especially when the population is large.
2. **Resource Conservation:** It saves time, money, and resources, making research more feasible and practical.
3. **Timeliness:** Allows for quicker data collection and analysis, which can be crucial in time-sensitive situations.
4. **Accessibility:** Some populations are difficult to access, making sampling the only practical option.
5. **Accuracy:** When done correctly, sampling provides accurate estimates of population characteristics.
6. **Risk Reduction:** Reduces the potential for errors in data collection and analysis.
7. **Inference:** Provides a basis for making conclusions about the entire population based on the characteristics of the sample.
8. **Privacy and Ethics:** Respects privacy and ethical considerations, especially in sensitive research areas.
9. **Analysis:** Simplifies data analysis, particularly for large datasets.

Sampling is a practical and essential tool for researchers to gather valuable information while managing constraints and practical limitations.

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

To calculate the needed sample size:

- Define your research objectives and questions.
- Choose a significance level (α) and desired margin of error (E).
- Estimate population variability (σ) or use conservative estimates.
- Determine the population size (N).
- Select the type of sampling (random or stratified).
- Choose the statistical test or analysis.
- Use a sample size formula or software tool to calculate the sample size.
- Consider practical constraints and adjust for non-response.
- Conduct the study, analyze data, and interpret results.

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

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

The key distinction between stratified sampling and cluster sampling lies in how the population is divided and sampled:

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

Q 64. Where is inferential statistics used?

Inferential statistics is used in various fields and contexts to make predictions, draw conclusions, and make inferences about populations based on sample data.

Here are some common areas and applications where inferential statistics is used:

1. Scientific Research:

Inferential statistics is fundamental in scientific research across disciplines such as biology, physics, chemistry, and environmental science. Researchers use statistical tests to analyze data and draw conclusions about hypotheses.

2. Business and Economics:

Businesses use inferential statistics for market research, sales forecasting, quality control, and decision-making. Econometric models are employed to analyze economic data and make policy recommendations.

3. Healthcare and Medicine:

Medical researchers and healthcare professionals use inferential statistics to study the effectiveness of treatments, analyze patient data, and draw conclusions about disease prevalence. Clinical trials rely heavily on inferential statistics.

4. Education:

In the field of education, inferential statistics are used to assess the effectiveness of teaching methods, evaluate standardized test scores, and make policy decisions about educational programs.

5. Market Research and Data Analysis:

Market researchers use inferential statistics to make predictions about consumer preferences, market trends, and the impact of marketing campaigns.

6. Finance and Investment:

In finance, inferential statistics are used to assess investment risk, analyze stock market data, and estimate future asset prices. Portfolio optimization and risk management rely on statistical modeling.

7. Criminal Justice and Criminology:

Researchers and law enforcement agencies use inferential statistics to analyze crime data, study crime patterns, and evaluate the effectiveness of criminal justice programs.

8. Sports and Athletics:

In sports analytics, inferential statistics are used to analyze player performance, predict game outcomes, and make strategic decisions in sports management.

Q 65. What are population and sample in Inferential Statistics, and how are they different?

In inferential statistics, the concepts of "population" and "sample" are fundamental and play distinct roles.



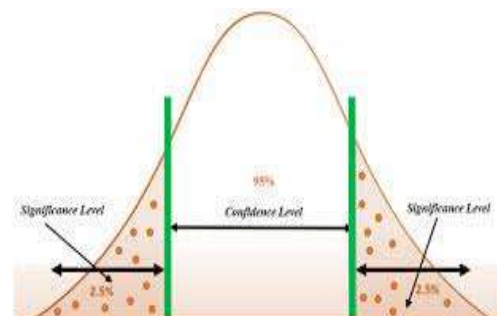
Population	Sample
Definition: The population refers to the entire group or collection of individuals, items, or data points about which you want to draw conclusions. It represents the larger, often theoretical, set that you're interested in studying.	Definition: A sample is a subset or a smaller, carefully selected group of individuals, items, or data points taken from the larger population. It is a representative portion of the population used for data collection and analysis.
Characteristics: - The population can be finite (e.g., all students in a school) or infinite (e.g., all potential customers in a market). -It includes every possible individual or element that falls within the scope of your research question.	Characteristics: -The sample is a finite and manageable subset of the population. -It is chosen through a systematic process, such as random sampling, stratified sampling, or cluster sampling. -The sample should be representative of the population, meaning that it should reflect the diversity and characteristics of the population.
Purpose: -In inferential statistics, the population is the ultimate target for making conclusions and generalizations. However, it is often impractical or impossible to collect data from the entire population.	Purpose: -The primary purpose of taking a sample is practicality. It's often more feasible, cost-effective, and efficient to collect data from a sample rather than the entire population. -Inferential statistics use data from the sample to make inferences, predictions, or generalizations about the larger population.

Q 66. What is the relationship between the confidence level and the significance level in statistics?

The relationship between the confidence level and the significance level in statistics is inverse and complementary. These two concepts are essential in hypothesis testing and statistical inference.

Relationship:

- 1.The relationship between the two is complementary, meaning that if you increase one, you decrease the other, and vice versa.
- 2.Higher confidence levels correspond to lower significance levels, and lower confidence levels correspond to higher significance levels.



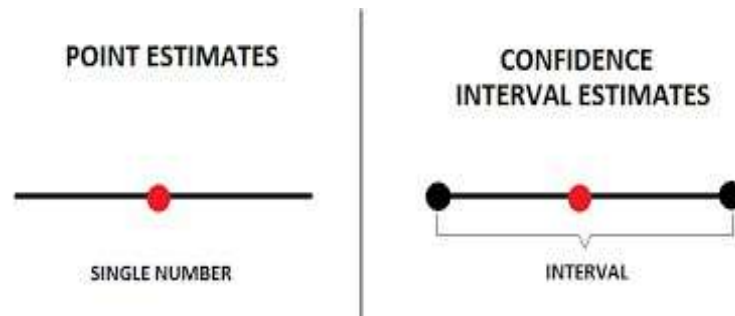
For example:

- 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 estimate and a confidence interval estimate is that a point estimate provides a single value, while a confidence interval estimate provides a range of values.
- Point estimates are useful for providing a single estimate of a parameter when you need a single, specific value.
- Confidence interval estimates are useful when you want to convey the uncertainty associated with your estimate 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 statistics, the terms "biased" and "unbiased" are used to describe the accuracy of an estimator in estimating a population parameter. These terms relate to how close the expected value of the estimator is to the true (or population) value of the parameter being estimated.

Biased	Unbiased
A statistical estimator is said to be "biased" if, on average, it systematically overestimates or underestimates the true population parameter.	A statistical estimator is considered "unbiased" if, on average, it provides estimates that are equal to the true population parameter.
In other words, a biased estimator tends to consistently deviate from the true value in a specific direction (either consistently too high or too low).	In mathematical terms, the expected value (mean) of an unbiased estimator is equal to the true value of the parameter being estimated.
Biased estimators can result from flaws in the estimation method or sampling procedure.	Unbiased estimators are desirable because, over repeated sampling, they provide accurate and unbiased estimates of the population parameter.
When using a biased estimator, it's important to be aware of the direction and magnitude of the bias to adjust for it in data analysis or decision-making.	While unbiased estimators are preferred, they are not always achievable, and in some cases, biased estimators may be the best available option.

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 estimate. 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 estimate. 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(\bar{x})):**

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

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

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 population mean.
- The standard error is a critical concept in inferential statistics because it is used to calculate confidence intervals and conduct hypothesis tests. Here's how it is typically used:
 1. Confidence Intervals: 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 population parameter lies.
 2. Hypothesis Testing: In hypothesis testing, the standard error is used to calculate test statistics, such as the t-statistic or z-statistic, which are then compared to critical 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 estimate population parameters, and the estimate differs from the true population value. It's the difference between the sample statistic (e.g., sample mean or proportion) and the true population parameter. It happens because we can't study everyone in the population, so we use a sample (a smaller group) to make predictions.

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

- **Use a Larger Sample**: The bigger the sample, the closer our estimate is to reality.
- **Randomly Choose the Sample**: Ensure that everyone in the population 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 population.
- **Use Proper Methods**: Follow good statistical 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 estimate.

So, the standard error tells us about the uncertainty in our estimate, 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:

Components of Hypothesis Testing:

- Null Hypothesis (H_0)
- Alternative Hypothesis (H_a or H_1)
- Test Statistic
- Significance Level (α)
- Critical Region or Rejection Region
- P-Value

Steps in Hypothesis Testing:

- 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?


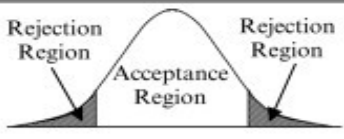
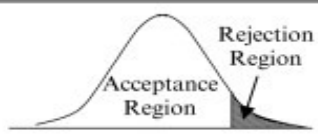
The alternative hypothesis contradicts the null hypothesis. It typically states what you expect to find in the population based on your research question or hypothesis. It is denoted as H_a or H_1 .

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

One-tailed and two-tailed hypothesis testing are two different approaches used in statistical hypothesis testing to investigate research questions or hypotheses. They differ in terms of the directionality of the research question and the way they assess evidence from sample data.


Here's a comparison of the two:

One-Tailed Hypothesis Testing	Two-Tailed Hypothesis Testing
One tail test is a statistical hypothesis test in which the alternative hypothesis only has one end.	Two-tail test refers to a significance test in which the alternative hypothesis has two ends.
Region of rejection is either left or right.	Region of rejection is both left and right.
Determines relationship between variables in single direction.	Determines relationship between variables in either direction.
Results are greater or less than certain value.	Results are greater or less than certain range of values.
Directional: > or <	Non-directional: ≠

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
		

Q 76. What is one sample t-test?

**One sample
t-Test**



Is there a **difference**
between a **group** and
the **population**

A one-sample t-test is a statistical hypothesis test used to determine whether the mean of a single sample of data is statistically different from a known or hypothesized population mean.

It's particularly useful when you have a sample and you want to assess whether it represents a population with a specific mean.

Q 77. What is the meaning of degrees of freedom (DF) in statistics?

In statistics, degrees of freedom (DF) refer to the number of values in the final calculation of a statistic that are free to vary. Degrees of freedom are a fundamental concept in hypothesis testing, confidence intervals, and various statistical analyses. They are used in various statistical 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 essential to understand because it affects the behaviour of statistical tests and the interpretation of their results. Here's a basic explanation:

- **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. One-sample t-test: $Degrees\ of\ freedom = n - 1$

2. Two-sample t-test: $Degrees\ of\ freedom = n_1 + n_2 - 2$

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 after estimating 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,

$$Degrees\ of\ freedom = (rows - 1) * (columns - 1)$$

where "rows" and "columns" represent the number of categories in the rows and columns of the contingency table. This calculation 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.

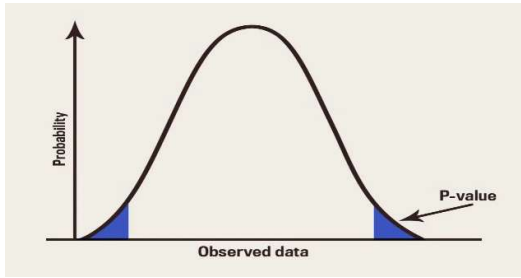
2. Within-group degrees of freedom.

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 statistical model. Understanding degrees of freedom is crucial because they affect the distribution of test statistics and, consequently, the interpretation of p-values and the conclusions drawn from statistical analyses. Different statistical tests have different formulas for calculating degrees of freedom, and they are chosen to ensure the validity of the statistical test being performed.

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



The p-value, short for "probability value," is a crucial concept in hypothesis testing in statistics. It measures the strength of evidence against a null hypothesis.

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

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

- **Formulate Hypotheses:**
Start by defining your null hypothesis (H_0) and alternative hypothesis (H_a). H_0 typically represents a statement of no effect or no difference, while H_a suggests there is an effect or difference.
- **Choose a Statistical Test:**
Select the appropriate statistical test based on your research question and the type of data you have. The choice of test depends on whether you're comparing means, testing proportions, examining associations, etc.
- **Collect Data:**
Collect relevant data for your analysis. The data should match the assumptions and requirements of the chosen statistical test.
- **Calculate the Test Statistic:**
Calculate the test statistic that corresponds to your chosen test. This involves using mathematical formulas specific to the test.
- **Determine the Sampling Distribution:**
Determine the theoretical sampling distribution of the test statistic under the assumption that the null hypothesis is true. This distribution depends on the test you're conducting (e.g., t-distribution, chi-square distribution, F-distribution, normal distribution).
- **Find the Observed Test Statistic:**
Calculate the observed test statistic using your data.
- **Calculate the p-value:**
The p-value is calculated based on the observed test statistic and its distribution under the null hypothesis.
 1. For one-tailed tests (where you are only interested in one direction of an effect), the p-value is the probability of observing a test statistic as extreme or more extreme than the observed value in that direction.
 2. For two-tailed tests (where you are interested in both directions of an effect), the p-value is the probability of observing a test statistic as extreme or more extreme than the observed value in either direction.
- **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 alternative hypothesis).
2. If the p-value is greater than α , you fail to reject the null hypothesis (insufficient evidence to support the alternative hypothesis).

It's important to note that the specific calculations for the test statistic and p-value depend on the chosen statistical test. Different tests have different formulas and assumptions. In practice, statistical software or calculators are often used to perform these calculations automatically, as they can be complex for many tests. Additionally, when conducting hypothesis tests, make sure to consider the assumptions and limitations 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 time 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:

$$= 1 - P(\text{Seeing one supercar}) = 1 - 0.3 = 0.7$$
- Probability of not seeing any supercar in the period of 60 minutes is:

$$= (0.7)^3 = 0.343$$
- Hence, the probability of seeing at least one supercar in 60 minutes is:

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

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

p-values help you make decisions about whether the results of a statistical analysis are statistically 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?

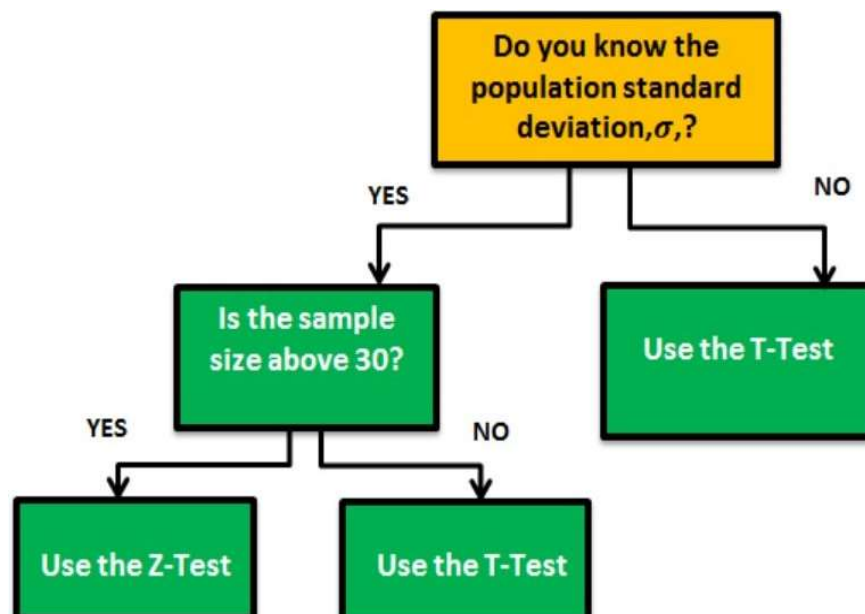
A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

Type I	Type II
1. The chance or probability that you will reject a null hypothesis that should not have been rejected.	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 related when they really are not.	2. This will result in you deciding two groups are not different or two variables are not related when they really are.
3. The probability of a Type I error is called alpha (α).	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 population variance is known, or if the sample size is larger than 30, for an unknown population variance. A t-test is used when the sample size is less than 30 and the population 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
<p>Purpose:</p> <p>The F-test is a statistical test used to compare the variances of two or more populations or samples.</p>	<p>Purpose:</p> <p>ANOVA, on the other hand, is used to compare means of three or more groups to determine if there are statistically significant differences among the group means.</p>
<p>Number of Groups:</p> <p>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 testing for the equality of population variances (e.g., in the context of two-sample hypothesis testing).</p>	<p>Number of Groups:</p> <p>ANOVA is specifically designed for comparing the means of three or more groups. It is used when you have multiple groups, and you want to test if there are any significant differences among them.</p>
<p>Test Statistic:</p> <p>The test statistic for the F-test follows an F-distribution, which is a right-skewed distribution. The F-statistic is calculated by dividing the variance of one group by the variance of another group.</p>	<p>Test Statistic:</p> <p>ANOVA uses an F-statistic as well, but the calculation is different from the F-test. It assesses the ratio of variation between group means to the variation within groups.</p>
<p>Use Cases:</p> <p>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 statistical model, and performing regression analysis (F-test for overall model fit).</p>	<p>Use Cases:</p> <p>ANOVA is commonly used in experimental designs where you have several treatments or conditions and you want to determine if there is a statistically significant difference in the means of these groups. It is often followed by post-hoc tests to identify which specific group means differ from each other.</p>

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

Resampling is a series of techniques used in statistics to gather more information about a sample. This can include retaking a sample or estimating its accuracy. With these additional techniques, resampling often improves the overall accuracy and estimates any uncertainty within a population.

Common methods of resampling include:

1. Bootstrapping:

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

Purpose: Bootstrapping is often used to estimate the sampling distribution of a statistic (e.g., mean, median, standard deviation) or to construct confidence intervals.

2. Cross-Validation:

K-Fold Cross-Validation: In cross-validation, you partition your dataset into "k" subsets (folds). You iteratively use k-1 folds for training and the remaining fold for testing, repeating this process k times.

Purpose: Cross-validation is widely used in machine learning to assess model performance, tune hyperparameters, and detect overfitting.

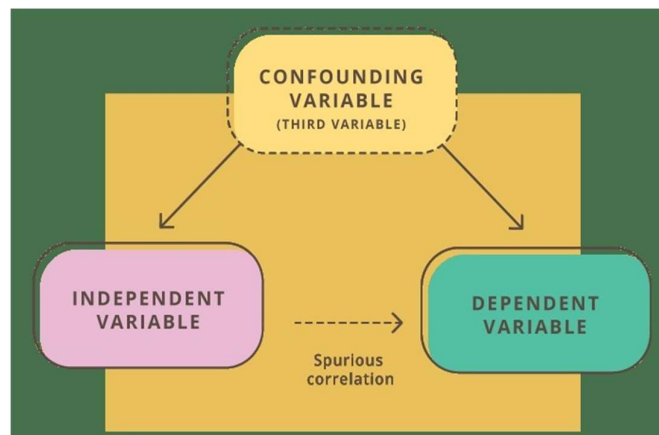
Q 86. What is the proportion of confidence intervals that will not contain the population parameter?

The proportion of confidence intervals that will not contain the population parameter (often denoted as $1 - \text{confidence level}$) is equal to the significance level (α) chosen for constructing 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 times, then approximately 5% of these intervals will not contain the true population 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 interpretation of the relationship between the independent and dependent variables.



In simpler terms, a confounding variable is an extra factor that can distort the observed relationship between two other variables by either masking or falsely suggesting a connection between them.

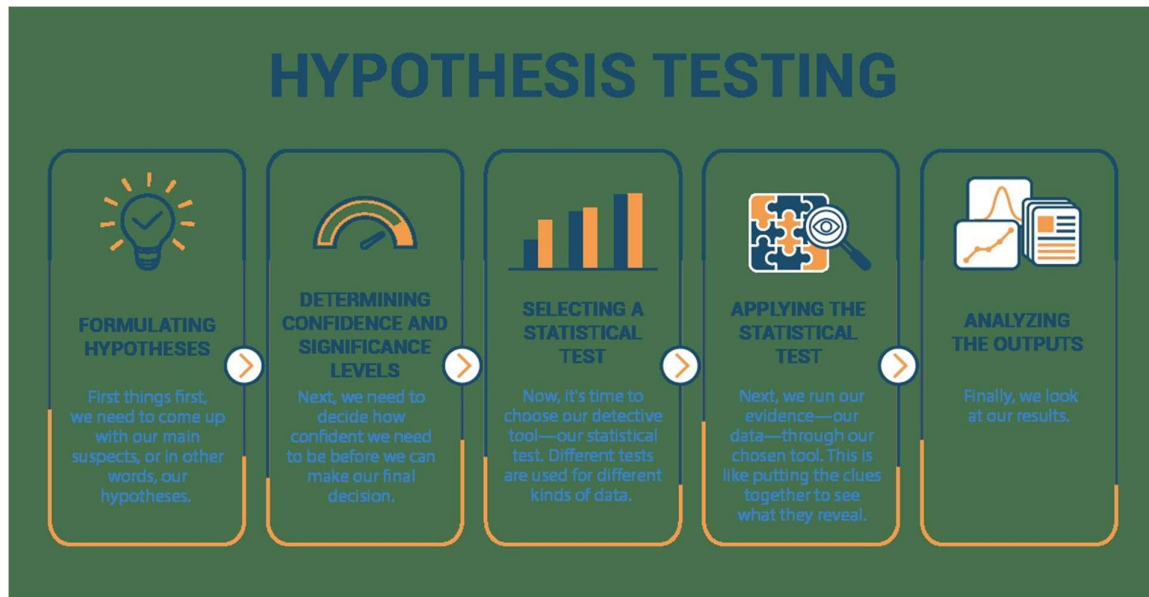
Example: Suppose you are studying the relationship between coffee consumption (independent variable) and the risk of heart disease (dependent variable). Age is a confounding variable because it is related to both coffee consumption (as people of different ages may drink different amounts of coffee) and 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 consumption directly affects heart disease risk.

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

Hypothesis testing is a structured process used in statistics to make inferences about population parameters based on sample data. Here are the steps typically involved in hypothesis testing:

1. **Formulate Hypotheses:**
State the null hypothesis (H_0): This is a statement of no effect or no difference. It represents the default assumption you want to test.
State the alternative hypothesis (H_a): This is the hypothesis you want to provide evidence for, suggesting that there is an effect, difference, or relationship in the population.
2. **Choose a Significance Level (α):**
Select the significance level (α), which represents the probability of making a Type I error (rejecting 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 question.
Perform appropriate statistical analysis based on the type of data and research design. This analysis depends on the specific hypothesis test you're conducting (e.g., t-test, chi-square test, ANOVA).
4. **Calculate the Test Statistic:**
Calculate the test statistic based on your sample data and the null hypothesis. The test statistic quantifies how different your sample data are from what you would expect under the null hypothesis.
5. **Determine the Critical Region:**
Identify the critical region or rejection region in the probability distribution of the test statistic. This is the range of values that would lead to rejecting the null hypothesis if the test statistic falls within it.
6. **Compare the Test Statistic to Critical Values:**
Compare the calculated test statistic to the critical values (cut-off values) corresponding to the chosen significance level. If the test statistic falls in the critical region, you reject the null hypothesis. Otherwise, you fail to reject it.
7. **Calculate the P-Value:**
Alternatively, you can calculate the p-value, which is the probability of observing a test statistic 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 statistic (or p-value) to the critical values (or α), make a decision:
 - If you reject the null hypothesis, conclude that there is evidence for the alternative hypothesis.
 - If you fail to reject the null hypothesis, conclude that there is insufficient evidence to support the alternative hypothesis.
9. **Interpret Results:**
Interpret the results in the context of your research question. Explain the practical significance of your findings and their implications.
10. **Report Findings:**

Clearly communicate your results, including the test statistic, 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?

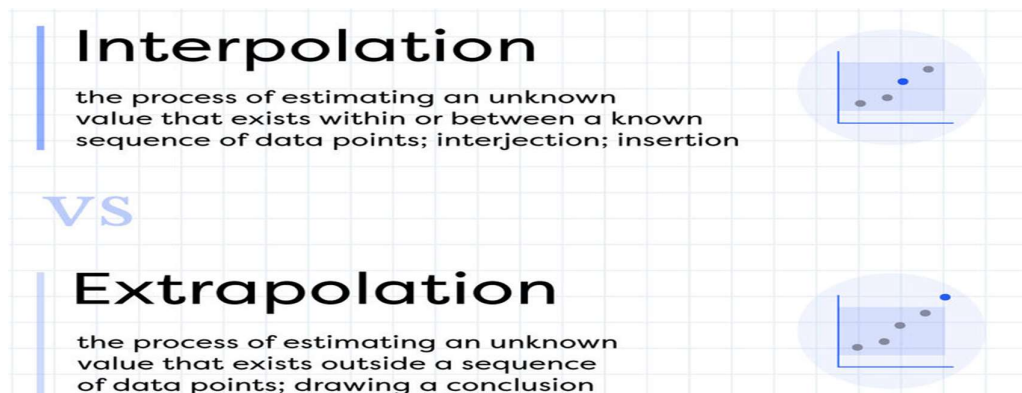
Explaining a p-value to a non-technical person or in layman's terms:

Imagine you're a detective investigating 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 getting the evidence you have if the suspect is innocent.

Q 90. What does interpolation and extrapolation mean? Which is generally more accurate?

Interpolation and extrapolation are two mathematical techniques used to estimate 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?

Interpolation is generally more accurate than extrapolation. Here's why:

Interpolation estimates values within the range of known data, where you have observed the actual pattern or relationship between data points. As long as this relationship is relatively consistent, interpolation tends to provide reasonably accurate estimates.

Extrapolation, on the other hand, involves predicting values beyond the range of known data, which is inherently uncertain. Extrapolation assumes that the same pattern or trend will continue, and this assumption may not always hold true, especially when data are subject to changing conditions or unobserved factors.

Q 91. What is an inlier?

An inlier is a data point in a dataset that conforms to the general pattern 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 characteristics 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 ($p(\text{head})=0.8$) five times. What's the probability of getting three or more heads?

To start off the question, we need 3, 4, or 5 heads to satisfy the cases.

- 5 heads: All heads, so
$$\left(\frac{4}{5}\right)^5 = \frac{1024}{3125}.$$
- 4 heads: All heads but 1. There are 5 ways to organize this, and then a
$$\left(\frac{4}{5}\right)^4 * \left(\frac{1}{5}\right)^1 = 256/3125.$$
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
$$\left(\frac{4}{5}\right)^3 * \left(\frac{1}{5}\right)^2 = 64/3125.$$
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. Infection rates at a hospital above a 1 infection per 100 person-days at risk are considered high. A hospital had 10 infections 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 infection rate is below the standard of 1 infection per 100 person-days at risk, you can use the Poisson distribution. The Poisson distribution is appropriate for modeling the number of rare events, such as infections in a hospital, over a known interval of time.

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.

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

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:

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

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

$$P(X \leq 10) = \sum_{x=0}^{10} \frac{e^{-17.87} * 17.87^x}{x!}$$

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 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:

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

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:

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

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

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

$$\text{Lower Bound} = 1,100 - (2.306 * 10) = 1,100 - 23.06 = 1,076.94 \text{ cc}$$

$$\text{Upper Bound} = 1,100 + (2.306 * 10) = 1,100 + 23.06 = 1,123.06 \text{ cc}$$

So, the 95% confidence interval for the mean brain volume in this new population 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 population falls within this range.

Q 95. What Chi-square test?

A chi-square test is a statistical test used to determine if there is a significant association or relationship between categorical variables. It is particularly useful for analyzing data that can be organized into a contingency table, which is a tabular representation 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 statistical test used to analyse the differences among group means in a sample. It's a powerful and widely used technique for comparing means from multiple groups to determine whether there are statistically significant differences among them.

The main idea behind ANOVA is to partition the total variance in the data into different components, which can be attributed 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 statistical 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 statistically 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 statistically 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.

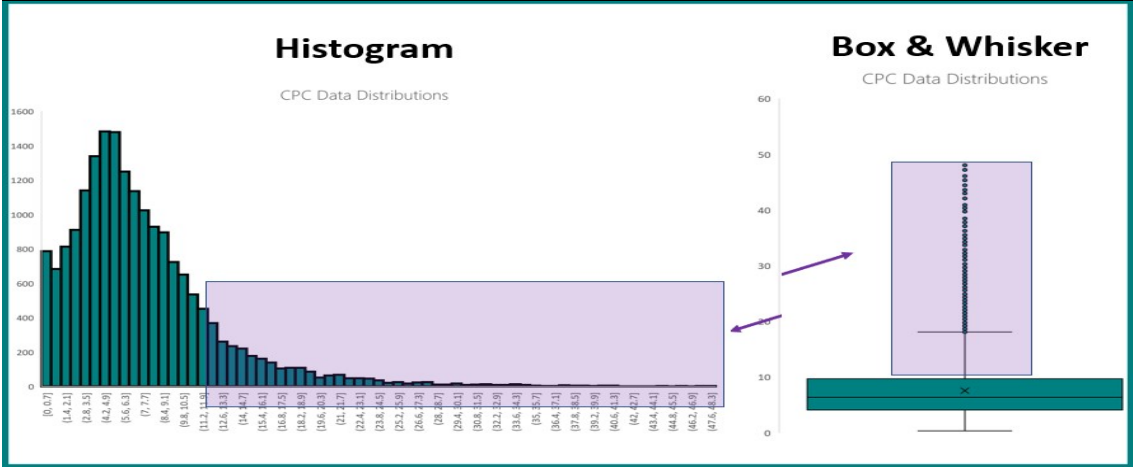
Q 98. What is the goal of A/B testing?

The goal of A/B testing is to compare different variations of a digital element (such as a webpage or app feature) to determine which one performs better in terms of a specific outcome, with the aim of optimizing 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 representations used in statistics to visualize the distribution of data. However, they have different purposes and characteristics:

Histogram	Box Plot
<p>Purpose: Histograms are used to visualize the distribution of continuous data by dividing it into bins or intervals and displaying the frequency or count of data points within each bin.</p> <p>Appearance: A histogram consists of a series of adjacent bars or bins, with the width of each bin representing a range of values. The height of each bar represents the frequency or count of data points in that bin.</p> <p>Information: Histograms provide a detailed view of the data's shape, centre, spread, skewness, and potential modes.</p> <p>Data Type: Histograms are primarily used for continuous data, although they can be adapted for discrete data by adjusting bin widths.</p> <p>Usage: Commonly used for exploring the distribution of data, identifying patterns, and assessing data characteristics.</p>	<p>Purpose: Box plots are used to display the distribution, central tendency, and spread (variability) of a dataset. They are particularly useful for identifying outliers and comparing the distribution of multiple datasets.</p> <p>Appearance: A box plot consists of a rectangular "box" with a line inside it (the median), and "whiskers" that extend from the box. Sometimes, individual data points are plotted as dots.</p> <p>Information: A box plot provides information about the median, quartiles (25th and 75th percentiles), the interquartile range (IQR), and the presence of outliers.</p> <p>Data Type: Box plots are suitable for summarizing both continuous and categorical data.</p> <p>Usage: Commonly used for comparing distributions between different groups or visualizing the spread of data.</p>



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 times. 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 getting 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 times, and we've gotten 10 heads (intuitively, you're probably thinking that there is a significant chance we have the double-headed coin). Formally, we're trying to find $P(D | 10 \text{ heads})$.

Using Bayes rule:

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

- Tackling the numerator, the prior probability, $P(D) = 1/1000$.
- If we used the double headed coin, the chance of getting 10 heads, $P(10 H | D) = 1$ (we always flip heads).
- So, the
 $numerator = 1 / 1000 * 1 = 1 / 1000$.
- The denominator, $P(10H)$ is just
 $P(10 H | D) * P(D) + P(10 H | Fair) * P(Fair)$.
This makes sense because we are simply enumerating over the two possible coins. The first part of $P(10H)$ is the exact same as the numerator ($1 / 1000$).
- Then the second part:
 $P(Fair) = 999/1000$. $P(10 H | Fair) = (1/2)^{10} = 1/1024$.
Thus,
 $P(10 H | Fair) * P(Fair) = 0.0009756$.
The denominator then equals $0.001 + 0.0009756$.

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

PART2

The second question is then easily answered: we just compute the two individual possibilities and add.

$$\begin{aligned} P(H) &= P(D) * P(H | D) + P(Fair) * P(H | Fair) \\ &= 0.506 * 1 + (1 - 0.506) * (0.5) = 0.753. \end{aligned}$$

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 statistical concept used to estimate a range of values within which a population parameter (such as a mean, proportion, or regression coefficient) is likely to fall with a certain level of confidence. It provides a measure of the uncertainty or variability associated with estimating a parameter from a sample of data.

Interpreting a confidence interval:

Example: Suppose you calculate a 95% confidence interval for the average height of a population, and you obtain the interval [165 cm, 175 cm].

Interpretation: You can interpret this confidence interval as follows:

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

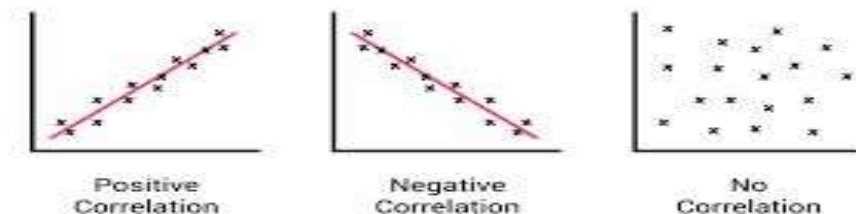
Q 102. How do you stay up-to-date with the new and upcoming concepts in statistics?

To stay up-to-date with new concepts in statistics:

- **Read Journals:** Regularly read statistical journals and publications.
- **Online Courses:** Take online courses and webinars.
- **Conferences:** Attend statistical conferences and workshops.
- **Join Forums:** Participate in online statistical forums and communities.
- **Network:** Connect with statisticians and data scientists.
- **Subscribe:** Subscribe to statistical newsletters and blogs.
- **Follow Researchers:** Follow leading statisticians on social media.
- **Continuous Learning:** Embrace a culture of continuous learning.

Q 103. What is correlation?

Correlation is a statistical measure used to describe the degree to which two or more variables change together or are related to each other. In other words, it quantifies the strength and direction of the linear relationship between two or more variables.



Key points about correlation:

- **Correlation Coefficient:**
The most common way to measure correlation is by calculating the correlation coefficient, which is represented by the symbol "r" or " ρ " (rho). The correlation coefficient is a numerical value that ranges between -1 and 1, with the following interpretations:

1. A positive correlation ($r > 0$) indicates that as one variable increases, the other tends to increase as well.
 2. A negative correlation ($r < 0$) indicates that as one variable increases, the other tends to decrease.
 3. A correlation coefficient of 0 ($r = 0$) suggests no linear relationship between the variables.
- **Strength of Correlation:**
The absolute value of the correlation coefficient ($|r|$) indicates the strength of the relationship. Values closer to -1 or 1 represent stronger correlations, while values closer to 0 represent weaker correlations.
 - **Direction of Correlation:**
The sign of the correlation coefficient (+ or -) indicates the direction of the relationship. A positive coefficient means the variables move in the same direction, while a negative coefficient means they move in opposite directions.
 - **Scatterplots:**
Scatterplots are often used to visually represent the relationship between two variables. Points on the plot represent data points, and the pattern they form can give an indication of the correlation.

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

Pearson's correlation coefficient, often denoted as "r," is used to measure the strength and direction of the linear relationship between two continuous variables. In other words, it is applied when both of the variables being studied are quantitative and numeric in nature.

Q 105. In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?

A high correlation between the time a person sleeps and the amount of productive work they do suggests a significant relationship between these two variables. However, it's important to note that correlation does not imply causation. Here's what can be inferred and what cannot be inferred from this observation:

What Can Be Inferred:

- **Association:** A high positive correlation implies that, on average, as the amount of time a person sleeps increases, their productive work also tends to increase. In other words, there appears to be a connection between sleep and productivity.
- **Predictive Value:** The strength of the correlation can indicate the extent to which sleep time can be used to predict or estimate productive work. If the correlation is strong, sleep time may be a good predictor of work productivity.
- **Direction:** A positive correlation means that as one variable (sleep time) increases, the other variable (productive work) tends to increase as well. This suggests that getting more sleep is associated with higher productivity, which aligns with common understanding.

Q 106. What does autocorrelation mean?

Autocorrelation, also known as serial correlation, refers to the correlation or relationship between a variable and its past values in a time series or sequence of data points. In simpler terms, autocorrelation assesses how a data point at a given time is related to the data points that occurred at previous time points within the same series.

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

Common tests for analyzing continuous data in statistics include:

- **T-Test:** Used to compare means between two groups.
- **Analysis of Variance (ANOVA):** Compares means among three or more groups.
- **Correlation Tests:** Assess relationships between continuous variables, e.g., Pearson correlation or Spearman rank correlation.
- **Regression Analysis:** Predicts one continuous variable based on one or more predictors.
- **Chi-Squared Test for Independence:** Examines associations between categorical and continuous variables.
- **ANOVA with Repeated Measures:** ANOVA extension for within-subject or repeated measures designs.
- **Multivariate Analysis of Variance (MANOVA):** Extends ANOVA to analyze multiple dependent variables simultaneously.

The choice of test depends on your research question, data distribution, 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 distribution (also known as a Gaussian distribution), can occur for various reasons. It's important to identify the underlying causes of non-normality because the choice of statistical analysis and the interpretation of results may depend on the distribution of the data.

Here are some common reasons for non-normality:

- **Skewness:** Data may be skewed to the left (negatively skewed) or right (positively skewed), leading to non-normality.
- **Outliers:** Extreme values or outliers in the dataset can distort the normal distribution.
- **Sampling Bias:** Non-random sampling or selection bias may result in data that does not reflect the population's true distribution.
- **Non-linear Relationships:** Data influenced by non-linear relationships or complex interactions may deviate from normality.
- **Data Transformation:** Some data, such as counts or proportions, inherently follow non-normal distributions.
- **Natural Variation:** In some cases, data may naturally follow a non-normal distribution due to the underlying process being studied.
- **Measurement Errors:** Errors in data collection or measurement can introduce non-normality.
- **Censoring or Floor/Ceiling Effects:** Data may be bounded, leading to deviations from normality at the bounds.

Understanding the cause of non-normality is essential for appropriate data analysis and choosing the right statistical techniques or transformations.

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 traditional 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 variations, which can determine whether there are statistically significant differences among multiple 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.