Name : Devika Prashant Pagare

Statistics_Interview_Questions

1. What are the most important topics in statistics?

Ans:

- Descriptive statistics: Descriptive statistics is used to summarize and describe data. It includes measures of central tendency (mean, median, mode), variability (standard deviation, variance, range), and shape (skewness, kurtosis).
- Probability: Probability is the study of chance and uncertainty. It is used to calculate the likelihood of different events occurring.
- Probability distributions: Probability distributions describe how likely different values of a variable are to occur. Some common probability distributions include the normal distribution, binomial distribution, and Poisson distribution.
- Statistical inference: Statistical inference is used to draw conclusions about a population based on a sample. It includes techniques such as hypothesis testing and confidence intervals.
- Regression: Regression is a statistical technique used to model the relationship between two or more variables.

2. What is exploratory data analysis?

Ans: Exploratory data analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

- Univariate analysis: Univariate analysis is used to examine the distribution of a single variable. This can be done using graphical methods, such as histograms and boxplots, or by calculating summary statistics, such as the mean, median, and standard deviation.
- Bivariate analysis: Bivariate analysis is used to examine the relationship between two variables. This can be done using graphical methods, such

as scatter plots and correlation matrices, or by calculating correlation coefficients.

- Multivariate analysis: Multivariate analysis is used to examine the relationship between three or more variables. This can be done using graphical methods, such as 3D scatter plots and parallel coordinate plots, or by using statistical techniques such as principal component analysis and factor analysis.

3. What are quantitative data and qualitative data?

Ans :
Quantitative data is numerical data that can be counted or measured. It is objective and can be analyzed using statistical methods. Examples of quantitative data include:

- The number of students in a class
- The height of a person
- The temperature of a room
- The sales of a product
- The results of a survey

Qualitative data is descriptive data that cannot be easily counted or measured. It is subjective and is often collected through interviews, focus groups, or observations. Examples of qualitative data include:

- The opinions of customers about a product
- The reasons why people choose to buy a certain brand
- The experiences of patients with a particular disease
- The culture of a company

4. What is the meaning of KPI in statistics?

Ans: KPI stands for Key Performance Indicator. In statistics, KPIs are used to measure the performance of a process, system, or organization against specific goals or objectives. KPIs can be used to track progress over time, identify areas for improvement, and make informed decisions about how to allocate resources.

Here are some examples of KPIs in statistics:

- Customer satisfaction: This KPI can be measured using surveys or feedback forms to collect customer feedback.
- Product quality: This KPI can be measured by tracking the number of defective products or the number of customer complaints.
- Employee productivity: This KPI can be measured by tracking the number of units produced by each employee or the amount of time it takes to complete a task.
- Financial performance: This KPI can be measured by tracking revenue, profit, or other financial metrics.

5. What Is the Difference Between Univariate, Bivariate, and Multivariate Analysis?

Ans : Univariate statistics summarize only one variable at a time.
Bivariate statistics compare two variables.
Multivariate statistics compare more than two variables.

6. How Would You Approach a Dataset That's Missing More Than 30 Percent of Its Values?

Ans : Approaching a dataset that's missing more than 30% of its values requires careful consideration. The best approach will depend on the specific dataset, the nature of the missing values, and the desired outcome.

Here are some general steps to follow:

- Assess the missing values.
- Identify the cause of the missing values.
- Choose a method for handling the missing values.

  Deletion: This involves removing rows or columns with missing values. This is a simple approach, but it can lead to a loss of data.

Imputation: This involves filling in the missing values with estimated values. There are a number of different imputation methods available, such as mean imputation, median imputation, and mode imputation.

Evaluate the results. Once you have handled the missing values, it is important to evaluate the results.

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

Ans: A classic example of where the median is a better measure than the mean is household income. In most countries, the distribution of household income is skewed, with a small number of very high incomes and a large number of lower incomes. The mean income will be pulled up by the few very high incomes, giving a misleading impression of the typical household income. The median income, on the other hand, is less affected by outliers and gives a more accurate representation of the income of the majority of households.

Here is a concrete example:

Suppose we have the following dataset of household incomes:

[1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 100000]
The mean income of this dataset is $10,000. However, the median income, which is the income of the middle household, is only $5,500. This is because the very high income of $100,000 is pulling up the mean.

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

Ans. Descriptive statistics describe some sample or population.

Inferential statistics attempts to infer from some sample to the larger population.

9. What are descriptive statistics?

Ans : Distribution – refers to the frequencies of responses.

Central Tendency – gives a measure or the average of each response.

Variability – shows the dispersion of a data set.

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

Ans:

There are two main types of dispersion of data in statistics: absolute and relative.

Absolute measures of dispersion quantify the spread of the data around the central tendency, regardless of the scale of the data.

Relative measures of dispersion compare the spread of the data to the central tendency and are therefore scale-invariant. This means that they can be used to compare the spread of two datasets, even if the datasets have different units.

11. How can we calculate the range of the data?

Ans : For example, suppose we have the following dataset of numbers:

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

To calculate the range, we would first order the data from least to greatest:

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

Then, we would subtract the smallest value from the largest value:

10 - 1 = 9

Therefore, the range of the dataset is 9.

12. Is the range sensitive to outliers?

Ans: Yes, the range is sensitive to outliers. This is because the range is calculated by subtracting the smallest value from the largest value in the dataset. If there are a few very large or very small values in the dataset, the range will be inflated.

13.What is the meaning of standard deviation?

Ans. Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. It is the average amount of variability in your dataset. It tells you, on average, how far each value lies from the mean.

A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.

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

14.What are the scenarios where outliers are kept in the data?

- Ans : To investigate the cause of the outliers: Outliers can sometimes be caused by errors in data collection or measurement. However, they can also be caused by genuine phenomena that are rare or unusual. By keeping the outliers in the data, researchers can investigate their cause and learn more about the underlying process.
- To understand the full distribution of the data: Outliers can provide valuable information about the tail ends of the distribution. For example, by studying outliers in income data, researchers can learn more about the distribution of wealth in a society.

- To avoid introducing bias into the data: Removing outliers can sometimes introduce bias into the data, especially if the outliers are not removed systematically. For example, if a researcher removes all outliers from a dataset of student test scores, the resulting dataset may not be representative of the student population as a whole.

15. What is Bessel's correction?

Ans:

In statistics, Bessel's correction is the use of n-1 instead of n in several formulas, including the sample variance and standard deviation, where n is the number of observations in a sample. This method corrects the bias in the estimation of the population variance. It also partially corrects the bias in the estimation of the population standard deviation, thereby, providing more accurate results.

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

Ans:

1. A spread out curve is a curve with a large spread of values. This means that the values are spread out over a wide range, with no central tendency. For example, a histogram with a large standard deviation would be considered a spread out curve.

2. A concentrated curve is a curve with a small spread of values. This means that the values are concentrated close to a central tendency. For example, a histogram with a small standard deviation would be considered a concentrated curve.

17. Can you calculate the coefficient of variation?

Ans:

Yes, I can calculate the coefficient of variation (CV). To do this, I need the mean and standard deviation of the data. The CV is calculated as follows:

CV = (standard deviation / mean) * 100

18. State the case where the median is a better measure when compared to the mean.

Ans: In general, the median is a better measure than the mean when we are interested in the typical value of the data, rather than the overall average.

Here are some specific examples of cases where the median is a better measure than the mean:

- Measuring household income
- Measuring student test scores
- Measuring housing prices
- Measuring the time it takes to complete a task
- Measuring the number of visitors to a website

19. How is missing data handled in statistics?

Ans : There are two main ways to handle missing data in statistics: deletion and imputation.

1. Deletion involves removing rows or columns with missing values. This is a simple approach, but it can lead to a loss of data.

2. Imputation involves filling in the missing values with estimated values. There are a number of different imputation methods available, such as mean imputation, median imputation, and mode imputation. The best imputation method will depend on the type of data and the distribution of the missing values.

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

Ans: Mean imputation for missing data is a method of handling missing values by replacing them with the mean value of the variable. It is a simple and straightforward method, but it has a number of drawbacks.

One of the main drawbacks of mean imputation is that it can introduce bias into the data. This is especially true if the missing values are not distributed randomly. For example, if we are imputing missing values for income, and the people with missing values are more likely to have lower incomes, then mean imputation will artificially increase the mean income of the population.

21. What is the benefit of using box plots?

- Ans: Easy to interpret: Box plots are very easy to interpret, even for people who are not familiar with statistics. The central tendency, dispersion, and skewness of the data can all be quickly assessed from a box plot.
- Versatility: Box plots can be used to summarize and compare data from a variety of sources, including surveys, experiments, and observational studies. They can also be used to visualize data from different categories, such as different age groups or different genders.
- Efficiency: Box plots can be used to efficiently summarize large amounts of data. A single box plot can convey a lot of information about a data set, without the need to display all of the individual data points.
- Identification of outliers and skewness: Box plots can be used to identify outliers and skewness in the data. Outliers are data points that are significantly different from the rest of the data, while skewness is a measure of the asymmetry of the data distribution. Identifying outliers and skewness can be important for understanding the data and for making accurate inferences.

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

Ans: The five-number summary in statistics is a set of five values that provides a concise summary of the distribution of a data set. It consists of the following:

- Minimum: The smallest value in the data set.
- First quartile (Q1): The middle value of the lower half of the data set.
- Median (Q2): The middle value of the entire data set.
- Third quartile (Q3): The middle value of the upper half of the data set.
- Maximum: The largest value in the data set.

23. What is the difference between the First quartile, the IInd quartile, and the IIIrd quartile?

Ans: The first quartile (Q1), the second quartile (Q2), and the third quartile (Q3) are all measures of the central tendency of a data set. However, they differ in the way they are calculated and in the information they provide.

First quartile (Q1)

The first quartile is the middle value of the lower half of the data set. This means that 25% of the values in the data set are less than or equal to the first quartile, and 75% of the values in the data set are greater than or equal to the first quartile.

Second quartile (Q2)

The second quartile is the middle value of the entire data set. This is also known as the median. The median is the most common measure of central tendency and is often used to represent the typical value in a data set.

Third quartile (Q3)

The third quartile is the middle value of the upper half of the data set. This means that 75% of the values in the data set are less than or equal to the third quartile, and 25% of the values in the data set are greater than or equal to the third quartile.

24. What is the difference between percent and percentile?

Ans:

Percent is a measure of relative frequency, calculated as the number of occurrences of an event divided by the total number of possible occurrences, and multiplied by 100. For example, if 10 out of 100 students pass a test, then the pass rate is 10%.

Percentile is a measure of relative position, calculated as the percentage of values in a data set that are below a certain value. For example, if a student's test score is in the 90th percentile, then this means that 90% of the students in the class scored lower than them on the test.

25. What is an Outlier?

Ans:

An outlier is a data point that is significantly different from the other data points in a dataset. Outliers can be caused by a variety of factors, such as errors in data collection or measurement, or by genuine phenomena that are rare or unusual.

Outliers can have a significant impact on statistical analyses. For example, outliers can skew the mean and median of a dataset, and they can also make it difficult to identify patterns and relationships in the data.

26. What is the impact of outliers in a dataset?

Ans:

Outliers can have a significant impact on a dataset, both in terms of the accuracy of statistical analyses and the interpretability of the results.

One of the most common impacts of outliers is that they can skew the mean and median of the dataset. This can lead to inaccurate conclusions about the central tendency of the data. For example, if we have a dataset of student test scores and there is one outlier who scored very high, the mean and median of the dataset will be pulled up, giving the impression that the students performed better than they actually did.

27. Mention methods to screen for outliers in a dataset.

Ans:

There are a number of methods that can be used to screen for outliers in a dataset. Some of the most common methods include:

1. Visual inspection: One of the simplest ways to screen for outliers is to visually inspect the data. This can be done by plotting the data in a histogram or box plot. Outliers will often appear as points that are far removed from the rest of the data.
2. Statistical tests: There are a number of statistical tests that can be used to identify outliers. Some of the most common tests include the interquartile range (IQR) test, the Grubbs' test, and the Dixon's test.
3. Machine learning algorithms: Machine learning algorithms can also be used to identify outliers. Some of the most commonly used algorithms include anomaly detection and outlier detection algorithms.

28. How you can handle outliers in the datasets.

Ans:

- Remove the outliers: This is the simplest approach, but it is important to only remove outliers if they are due to errors in data collection or measurement. If the outliers are caused by genuine phenomena, then removing them can introduce bias into the data.
- Transform the outliers: This involves replacing the outliers with values that are more consistent with the rest of the data. There are a number of different ways to transform outliers, such as winsorization and capping.
- Use robust statistical methods: These methods are less sensitive to outliers than traditional statistical methods. For example, instead of using the mean as a measure of central tendency, we could use the median, which is more robust to outliers.
- Model the outliers: In some cases, it may be possible to model the outliers. This can be done using statistical methods such as robust regression or outlier detection algorithms.

29. What is the empirical rule?

Ans: The empirical rule, also known as the 68-95-99.7 rule, is a statistical rule that states that approximately 68% of the values in a normal distribution will fall within one standard deviation of the mean, 95% will fall within two standard deviations of the mean, and 99.7% will fall within three standard deviations of the mean.

30. How to calculate range and interquartile range?

Ans: To calculate the range:

1. Order the data from least to greatest.
2. Subtract the smallest value from the largest value.

For example, if the data is 1, 2, 3, 4, 5, the range would be 5 - 1 = 4

To calculate the interquartile range (IQR):

1. Order the data from least to greatest.

2. Find the median, which is the middle value of the dataset.
3. Divide the dataset into two halves, with the median in the middle.
4. Find the median of each half of the dataset. These values are called the first quartile (Q1) and third quartile (Q3).
5. Subtract the first quartile from the third quartile. This is the interquartile range.

For example, if the data is 1, 2, 3, 4, 5, the median would be 3. The first quartile would be 1.5 and the third quartile would be 4.5. The interquartile range would be 4.5 - 1.5 = 3.

The interquartile range is a measure of the spread of the data. It is less sensitive to outliers than the range, which is why it is often preferred.

31.What is skewness?

Ans:
Skewness is a statistical measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images. If the distribution is skewed to the right, it means that there are more values on the left side of the distribution than on the right side. If the distribution is skewed to the left, it means that there are more values on the right side of the distribution than on the left side.

32. What are the different measures of Skewness?
Ans:
There are a number of different measures of skewness, each with its own advantages and disadvantages. The most common measure of skewness is the Pearson skewness coefficient, which is defined as follows:

Pearson skewness coefficient = (mean - median) / standard deviation

The Pearson skewness coefficient is a measure of the symmetry of the distribution around the mean. A value of 0 indicates that the distribution is perfectly symmetrical, while a positive value indicates that the distribution is skewed to the right and a negative value indicates that the distribution is skewed to the left.

33. What is kurtosis?

Measures if the distribution is peaked or flat
There is 3 types of kurtosis

1) Leptokurtic
2) Mesokurtic
3) platykurtic

34. Where are long-tailed distributions used?

Ans:

- Business: Long-tailed distributions are used in business to model the demand for products and services. For example, a bookstore may sell a small number of popular books, but it may also sell a large number of books that are less popular. The distribution of sales for these books would be long-tailed.
- Economics: Long-tailed distributions are used in economics to model the distribution of income and wealth. For example, the distribution of income in the United States is long-tailed, meaning that there are a few people with very high incomes and a large number of people with lower incomes.
- Finance: Long-tailed distributions are used in finance to model the risk of assets and portfolios. For example, the distribution of returns for stocks is long-tailed, meaning that there is a small probability of very high returns and a large probability of lower returns.
- Science: Long-tailed distributions are used in science to model a variety of phenomena, such as the distribution of word frequencies in a language, the distribution of earthquake magnitudes, and the distribution of species abundance.

35. What is the central limit theorem?

Ans:
The central limit theorem (CLT) is a statistical theorem that states that the distribution of the mean of a sample drawn from a population will be approximately normally distributed, regardless of the distribution of the population, as long as the sample size is sufficiently large.

In other words, the CLT states that if we take a large number of random samples from a population, the distribution of the sample means will be approximately normal, regardless of the distribution of the population from which the samples were drawn.

The CLT has a number of applications in statistics, including:

- Hypothesis testing: The CLT is used to develop hypothesis tests that can be used to compare the means of two or more populations.
- Confidence intervals: The CLT is used to construct confidence intervals for the means of populations.
- Estimation: The CLT is used to estimate the means of populations.

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

Ans: For example, we can construct a 95% confidence interval for the population mean height. The 95% confidence interval would be 68.5 to 69.5 inches. This means that we can be 95% confident that the population mean height is between 68.5 and 69.5 inches.

The central limit theorem is a powerful tool that can be used to make inferences about populations based on samples. It is one of the most important theorems in statistics and is used in a wide variety of fields.

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

The data must be sampled randomly

The sample values must be independent of each other

The sample size must be sufficiently large, generally it should be greater or equal than 30

38. What is the meaning of selection bias?

Ans: Selection bias is a type of bias that occurs when the sample of data that is collected is not representative of the population that the researcher is trying to study. This can happen for a variety of reasons, such as:

Sampling error: This occurs when the sample is not large enough to be representative of the population.
Self-selection: This occurs when participants choose to participate in a study, which can lead to a sample that is not representative of the population.
Non-response bias: This occurs when some members of the population are more likely to respond to a survey or questionnaire than others.

39. What are the types of selection bias in statistics?

Ans:

- Sampling bias: This is the most common type of selection bias, and it occurs when the sample of data that is collected is not representative of the population that the researcher is trying to study. Sampling bias can be caused by a variety of factors, such as using a convenience sample, not using a random sampling method, or having a response rate that is too low.
- Self-selection bias: This occurs when participants choose to participate in a study, which can lead to a sample that is not representative of the

population. For example, if a study is conducted on the effects of a new drug, the participants are likely to be people who are interested in trying the drug or who have a condition that the drug is supposed to treat.

- Non-response bias: This occurs when some members of the population are more likely to respond to a survey or questionnaire than others. For example, if a survey is conducted on political attitudes, people who are strongly interested in politics are more likely to respond than people who are not interested in politics.
- Attrition bias: This occurs when participants drop out of a study before it is completed. Attrition bias can be caused by a variety of factors, such as the study being too long or too difficult, or participants losing interest in the study.
- Observer bias: This occurs when the researcher's expectations or beliefs influence the way they collect or analyze data. For example, if a researcher believes that a new drug is effective, they may be more likely to interpret the results of the study in a way that supports their belief.

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

Ans:
The probability of throwing two fair dice when the sum is 8 is 5/36.

There are 36 possible outcomes when rolling two dice, each with equal probability. Out of these 36 outcomes, there are 5 ways to get a sum of 8:

(2,6)
(3,5)
(4,4)
(5,3)
(6,2)
Therefore, the probability of rolling a sum of 8 is 5/36.

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

Ans:

- Normal distribution: The normal distribution is also known as the Gaussian distribution, and it is the most commonly used probability distribution in data science. The normal distribution is a bell-shaped curve that is symmetrical around the mean. It is used to model a wide variety of phenomena, such as human height, test scores, and measurement errors.

- Binomial distribution: The binomial distribution is used to model the probability of a binary event occurring a certain number of times in a fixed number of trials. For example, the binomial distribution could be used to model the probability of flipping a coin and getting heads 5 times in 10 flips.
- Poisson distribution: The Poisson distribution is used to model the probability of a certain number of events occurring in a fixed interval of time or space. For example, the Poisson distribution could be used to model the probability of receiving 10 phone calls in an hour.
- Exponential distribution: The exponential distribution is used to model the time between events that occur independently and at random. For example, the exponential distribution could be used to model the time between customers arriving at a store.
- Uniform distribution: The uniform distribution is used to model the probability of a value occurring within a certain range. For example, the uniform distribution could be used to model the probability of a student's score on a test being between 80 and 90.

42. What do you understand by the term Normal Distribution or What is a bell-curve distribution?

Ans: The normal distribution, also known as the Gaussian distribution or bell-curve distribution, is a probability distribution that is symmetrical around the mean and has a bell-shaped curve. The mean, median, and mode of a normal distribution are all equal.

The normal distribution is the most commonly used probability distribution in statistics and data science. It is used to model a wide variety of phenomena, such as human height, test scores, and measurement errors.

The normal distribution is characterized by two parameters: the mean and the standard deviation. The mean is the average value of the distribution, and the standard deviation is a measure of how spread out the distribution is.

The normal distribution can be used to calculate the probability of a value occurring within a certain range. For example, we could use the normal distribution to calculate the probability of a student's score on a test being between 80 and 90.

The normal distribution can also be used to test hypotheses. For example, we could use the normal distribution to test the hypothesis that the average height of men is different from the average height of women.

43.Can you state the formula for normal distribution?

Ans: Formula:

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

where:

f(x) is the probability density function (PDF) of the normal distribution
x is the value of the variable being modelled
μ is the mean of the normal distribution
σ is the standard deviation of the normal distribution

44. What type of data does not have a normal distribution or a Gaussian distribution?

Ans:

- Count data: Count data is data that counts the number of times an event occurs. For example, the number of customers who visit a store in a day or the number of goals scored in a soccer game are both count data. Count data is often skewed to the right, meaning that there are more low values than high values.
- Categorical data: Categorical data is data that can be classified into different categories. For example, the color of a car or the type of animal are both categorical data. Categorical data does not have a natural numerical order, so it cannot be modeled with a normal distribution.
- Ordinal data: Ordinal data is data that can be ranked, but does not have a natural interval scale. For example, the results of a customer satisfaction survey (e.g., very satisfied, satisfied, neutral, dissatisfied, very

dissatisfied) are ordinal data. Ordinal data is often skewed to one side or the other, and it cannot be modeled with a normal distribution.

- Extreme value data: Extreme value data is data that consists of very high or very low values. For example, the highest temperature recorded in each year or the lowest stock price for a particular company are both extreme value data. Extreme value data is often skewed to one side or the other, and it cannot be modeled with a normal distribution.

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

Ans:
The mean and median are equal in a normal distribution. This is because the normal distribution is symmetrical around the mean, meaning that there are equal numbers of values above and below the mean.

The mean is calculated by adding up all of the values in the distribution and dividing by the number of values. The median is the middle value in the distribution when the values are ordered from least to greatest.

In a normal distribution, the mean and median are both located at the center of the bell-shaped curve. This means that the mean and median are both representative of the typical value in the distribution.

46. What are some of the properties of a normal distribution?

Ans:

- Symmetrical: The normal distribution is symmetrical around the mean, meaning that there are equal numbers of values above and below the mean.
- Unimodal: The normal distribution has a single mode, which is the most frequent value in the distribution.
- Bell-shaped: The normal distribution has a bell-shaped curve, with the mean, median, and mode all located at the center of the curve.
- Asymptotic: The normal distribution tails off to zero as we move further and further away from the mean.

- Continuous: The normal distribution is a continuous distribution, meaning that it can take on any value within a certain range.

47. What is the assumption of normality?

Ans:

The assumption of normality is a statistical assumption that states that the distribution of a random variable is normally distributed. This means that the data can be modeled by a bell-shaped curve with the mean, median, and mode all located at the center of the curve.

The assumption of normality is important for many statistical tests, such as the t-test and the chi-squared test. These tests are designed to test for differences between two groups or to test whether a sample is representative of a population. If the data does not follow a normal distribution, the results of these tests may not be reliable.

There are a number of ways to test for normality. One common method is to use a normality test, such as the Shapiro-Wilk test or the Kolmogorov-Smirnov test. These tests compare the distribution of the data to a normal distribution and return a p-value. If the p-value is less than a certain threshold, then the test rejects the null hypothesis that the data is normally distributed.

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

Ans: Formula: $z = (x - \mu) / \sigma$

where:

z is the standard normal score
x is the value in the normal distribution
μ is the mean of the normal distribution
σ is the standard deviation of the normal distribution

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

Ans: The range of values in the standard normal distribution is from negative infinity to positive infinity. However, the vast majority of values in the standard normal distribution are between -3 and 3. This is because the standard normal distribution is bell-shaped, with the mean, median, and mode all located at 0.

50. What is the Pareto principle?

Ans: The Pareto principle, also known as the 80/20 rule, is a principle that states that for many outcomes, roughly 80% of consequences come from 20% of causes. This principle serves as a general reminder that the relationship between inputs and outputs is not balanced. The Pareto principle is also known as the Pareto Rule or the 20/80 Rule.

The Pareto principle is named after Italian economist Vilfredo Pareto, who observed that 80% of the land in Italy was owned by 20% of the population. Pareto also observed that this pattern was repeated in other areas, such as income distribution and wealth distribution.

51. What are left-skewed and right-skewed distributions?

Ans: Skewness is a way to describe the symmetry of a distribution.

A left-skewed (Negative Skew) distribution is one in which the left tail is longer than that of the right tail. For this distribution, *mean* < *median* < *mode*.

Similarly, right-skewed (Positively Skew) distribution is one in which the right tail is longer than the left one. For this distribution, *mean* > *median* > *mode.*

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

Ans: If a distribution is skewed to the right and has a median of 20, the mean will be greater than 20.

A skewed distribution is a distribution that is not symmetrical. In a skewed distribution, the tail on one side of the distribution is longer than the tail on the other side. A right-skewed distribution has a longer tail on the right side, meaning that there are more values above the median than there are values below the median.

The mean of a distribution is calculated by adding up all of the values in the distribution and dividing by the number of values. The median of a distribution is the middle value when the values are ordered from least to greatest.

53. 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?

Ans:
If a distribution is left-skewed and has a median of 60, we can draw the following conclusions about the mean and the mode of the data:

- The mean will be less than the median.
- The mode will be the most frequent value in the distribution, and it will be less than the median.

In a left-skewed distribution, there are more values to the right of the median than there are values to the left of the median. This means that the mean will be pulled to the left, and it will be less than the median.

The mode is the most frequent value in the distribution. In a left-skewed distribution, the mode will be less than the median, because the most frequent values are on the left side of the distribution.

54. 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?

Ans: Formula: score = mean + (z-score * standard deviation) where:

- score is Jeremy's score on the test
- mean is the mean score of the test
- z-score is Jeremy's z-score
- standard deviation is the standard deviation of the test

Substituting in the given values, we get:

score = 160 + (1.20 * 15) = 178
Therefore, Jeremy's score on the test is 178.

We can also check our answer by using a z-score table. A z-score table shows the probability of getting a certain z-score or lower under a standard normal distribution.

To use a z-score table, we first need to find the z-score that corresponds to Jeremy's score. We can do this by looking up his score in the z-score table. Jeremy's score of 178 corresponds to a z-score of 1.20.

Once we have found Jeremy's z-score, we can look up the corresponding probability in the z-score table. This probability tells us the percentage of people who would score lower than 178 on the test.

According to the z-score table, the probability of scoring lower than 178 on the test is 88.49%. This means that 88.49% of the people who took the test scored lower than 178.

Therefore, we can be confident that Jeremy's score of 178 is above average.

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

Ans: True.

The standard normal curve is a bell-shaped curve with a mean of 0 and a standard deviation of 1. It is symmetric around zero, meaning that the area under the curve to the left of zero is equal to the area under the curve to the right of zero.

The total area under the standard normal curve is equal to 1. This means that the probability of getting a value from the standard normal distribution is 1.

We can use this information to calculate the probability of getting a value within a certain range from the standard normal distribution. For example, we can use it to calculate the probability of getting a value between 0 and 1, or the probability of getting a value greater than 2.

The standard normal curve is a very useful tool for understanding and analyzing data. It is used in a wide variety of fields, including statistics, data science, engineering, and medicine.

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

Ans: The procedure to measure the length of all sharks in the world would be a very challenging and expensive task, but it could be done in the following steps:

- Identify all shark species in the world. This could be done by consulting with marine biologists and other experts.
- Develop a sampling plan. This plan would need to determine how many sharks of each species would need to be measured in order to get an accurate estimate of the average length of all sharks. The sampling plan would also need to consider the geographic distribution of sharks and the different types of habitats they live in.
- Collect the data. This could be done by using a variety of methods, such as tagging sharks, using underwater cameras, and conducting fishing surveys.
- Analyze the data. This would involve calculating the average length of each shark species and then using this information to estimate the average length of all sharks.

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

Ans:

A unimodal distribution is a distribution that has a single mode, which is the most frequent value in the distribution. A bimodal distribution is a distribution that has two modes, and a bell-shaped distribution is a distribution that has a symmetrical shape with a single mode in the middle.

Here are some examples of unimodal, bimodal, and bell-shaped curves:

- Unimodal: The distribution of human heights is unimodal, with the most frequent height being around 5'8" for men and 5'4" for women.
- Bimodal: The distribution of test scores on a difficult exam may be bimodal, with one peak for high-scoring students and another peak for low-scoring students.
- Bell-shaped: The distribution of IQ scores is bell-shaped, with most people having IQ scores close to the average IQ score of 100.

58. Does symmetric distribution need to be unimodal?

Ans : No, a symmetric distribution does not need to be unimodal. A unimodal distribution is a distribution with a single mode, or the most frequent value. A symmetric distribution is a distribution that is symmetrical around the mean, meaning that the left side of the distribution mirrors the right side.

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

Ans. When data follows a non-normal distribution, it is frequently non-Gaussian. A non-Gaussian distribution is often seen in many statistics processes. This occurs when data is naturally clustered on one side or the other on a graph. For instance, bacterial growth follows an exponential or non-Gaussian distribution, which is non-normal.

60. What is the Binomial Distribution Formula?

Ans: The formula is as follows:

P(x) = nCr * p^x * (1 - p)^(n - x)

where:

- x is the number of successes
- n is the total number of trials
- p is the probability of success on a single trial
- q is the probability of failure on a single trial (1 - p)

61. What are the criteria that Binomial distributions must meet?

Ans: There are four criteria that binomial distributions must meet:

- Fixed number of trials: The number of trials in a binomial experiment must be fixed. This means that you know how many times you will perform the experiment before you start.
- Independent trials: The trials in a binomial experiment must be independent. This means that the outcome of one trial does not affect the outcome of any other trial.
- Two possible outcomes: There must be only two possible outcomes for each trial in a binomial experiment. These outcomes are typically called success and failure.
- Constant probability of success: The probability of success must be the same for each trial in a binomial experiment. This means that the chance of getting a success on one trial is the same as the chance of getting a success on any other trial.

62. What are the examples of symmetric distribution?

Ans: Here are some examples of symmetric distributions:

- Normal distribution: The normal distribution is the most common example of a symmetric distribution. It is bell-shaped and symmetrical around the mean.
- Uniform distribution: The uniform distribution is a distribution in which all values are equally likely. It is represented by a horizontal line on a graph.

- Binomial distribution: The binomial distribution is a distribution that models the number of successes in a series of independent trials. It is symmetrical when the probability of success is equal to the probability of failure.
- Cauchy distribution: The Cauchy distribution is a distribution that is characterized by its long tails. It is symmetrical around the mean.
- Logistic distribution: The logistic distribution is a distribution that is characterized by its sigmoid shape. It is symmetrical around the mean.

63. How to find the mean length of all fishes in the sea?

Ans: Define the confidence level (most common is 95%)

Take a sample of fishes 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 fishes should be.

64. What are the types of sampling in Statistics?

Ans: There are two main types of sampling in statistics: probability sampling and non-probability sampling.

Probability sampling is a sampling method in which every member of the population has a known chance of being selected in the sample. This means that the sample is representative of the population, and the results of the sample can be generalized to the population.

There are four main types of probability sampling:

- Simple random sampling: Simple random sampling is a sampling method in which every member of the population has an equal chance of being selected in the sample. This can be done by using a random number generator or by drawing names out of a hat.
- Stratified sampling: Stratified sampling is a sampling method in which the population is divided into groups, or strata, and then a random sample is taken from each stratum. This method is used to ensure that all groups in the population are represented in the sample.
- Cluster sampling: Cluster sampling is a sampling method in which the population is divided into groups, or clusters, and then a random sample of clusters is selected. This method is often used when it is difficult or expensive to sample the entire population.
- Systematic sampling: Systematic sampling is a sampling method in which every kth member of the population is selected in the sample. This method is simple to implement, but it can be biased if the population is not ordered in a random way.

Non-probability sampling is a sampling method in which not every member of the population has an equal chance of being selected in the sample. This means that the sample may not be representative of the population, and the results of the sample cannot be generalized to the population.

There are four main types of non-probability sampling:

- Convenience sampling: Convenience sampling is a sampling method in which the researcher selects the sample from the population that is most convenient to access. This method is often used in qualitative research.
- Purposive sampling: Purposive sampling is a sampling method in which the researcher selects the sample based on specific criteria. This method is often used in qualitative research.
- Snowball sampling: Snowball sampling is a sampling method in which the researcher selects the sample from the population and then asks the participants to refer to other participants. This method is often used in hard-to-reach populations.
- Quota sampling: Quota sampling is a sampling method in which the researcher sets quotas for different groups in the population and then

selects the sample to meet those quotas. This method is often used in market research.

65. Why is sampling required?

Ans:

- To reduce costs and time. It is often too expensive and time-consuming to collect data from the entire population. Sampling allows researchers to collect data from a smaller subset of the population and still get reliable results.
- To improve the accuracy of the results. In some cases, it is not possible to collect data from the entire population. For example, it would be impossible to collect data from all of the people in the world. Sampling allows researchers to collect data from a representative subset of the population and generalize the results to the population.
- To reduce bias. Bias can occur when the sample is not representative of the population. Sampling can help to reduce bias by ensuring that the sample is representative of the population.
- To make predictions. Sampling can be used to make predictions about the population. For example, a researcher could use a sample of voters to predict the outcome of an election.
- To test hypotheses. Sampling can be used to test hypotheses about the population. For example, a researcher could use a sample of students to test the hypothesis that a new teaching method is more effective than the traditional teaching method.

66. How do you calculate the needed sample size?

Ans: There are a number of different ways to calculate the needed sample size. The most common method is to use the following formula:

n = (z^2 * p * q) / (e^2)

where:

n is the required sample size

z is the z-score corresponding to the desired confidence level

p is the estimated proportion of the population that has the characteristic of interest

q is the complement of p (1 - p)

e is the desired margin of error

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

Ans:

Stratified and cluster sampling are both probability sampling methods, meaning that every member of the population has a known chance of being selected in the sample. However, there are some key differences between the two methods.

Stratified sampling involves dividing the population into groups, or strata, based on a shared characteristic, such as age, gender, or income level. A random sample is then taken from each stratum. This ensures that all groups in the population are represented in the sample.

Cluster sampling involves dividing the population into groups, or clusters, and then randomly selecting a sample of clusters. All members of the selected clusters are then included in the sample. This method is often used when it is difficult or expensive to sample the entire population.

68. Where is inferential statistics used?

Ans: Inferential statistics is used in a wide variety of fields, including:

- Science: Inferential statistics is used by scientists to test hypotheses about the natural world. For example, a scientist might use inferential statistics to test the hypothesis that a new drug is effective in treating a certain disease.
- Business: Inferential statistics is used by businesses to make decisions about marketing, product development, and other areas. For example, a business might use inferential statistics to test the hypothesis that a new advertising campaign is effective in increasing sales.
- Government: Inferential statistics is used by governments to make decisions about public policy. For example, a government might use inferential statistics to test the hypothesis that a new education program is effective in improving student test scores.

- Medicine: Inferential statistics is used by medical researchers to develop new treatments and to test the effectiveness of existing treatments. For example, a medical researcher might use inferential statistics to test the hypothesis that a new drug is effective in reducing the risk of heart disease.
- Social sciences: Inferential statistics is used by social scientists to study human behaviour and society. For example, a social scientist might use inferential statistics to test the hypothesis that there is a relationship between income and education level.

69. What are population and sample in Inferential Statistics, and how are they different?
70. What is the relationship between the confidence level and the significance level in statistics?

Ans:
Confidence level and significance level are two important concepts in statistical hypothesis testing.

Confidence level is the probability that a confidence interval will contain the true population parameter. For example, a 95% confidence interval means that we have a 95% chance of capturing the true population parameter in our interval.

Significance level is the probability of rejecting the null hypothesis when it is true, also known as a Type I error. For example, a significance level of 0.05 means that we have a 5% chance of rejecting the null hypothesis when it is true.

The relationship between confidence level and significance level is inverse. This means that as the confidence level increases, the significance level decreases, and vice versa.

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

Ans: A point estimate is a single value that is used to estimate a population parameter. For example, the sample mean is a point estimate of the population mean.

A confidence interval estimate is a range of values that is likely to contain the true population parameter. For example, a 95% confidence interval for the population mean is a range of values that has a 95% probability of containing the true population mean.

72. What do you understand about biased and unbiased terms?

Ans:
A biased term is a word or phrase that expresses a particular point of view or prejudice and can influence people's thinking. An unbiased term is a word or phrase that is neutral and does not express any particular point of view or prejudice.

Here are some examples of biased and unbiased terms:

Biased term	Unbiased term
Homeless person	Person experiencing homelessness
Crackhead	Person who uses crack cocaine
Welfare queen	Person who receives welfare benefits
Illegal alien	Undocumented immigrant
Mentally retarded	Person with intellectual disabilities

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

Ans: The width of a confidence interval decreases as the length of the sample increases. This is because a larger sample provides more information about the population, which leads to a more precise estimate of the population parameter.

74. What is the meaning of standard error?

Ans: The standard error (SE) is a measure of the precision of an estimate. It is calculated by dividing the standard deviation of the sampling distribution of an estimate by the square root of the sample size.

The standard error is important because it tells us how much we can expect our estimate to vary from the true population parameter. A smaller standard error indicates a more precise estimate, while a larger standard error indicates a less precise estimate.

The standard error can be used to calculate confidence intervals for population parameters. A confidence interval is a range of values that is likely to contain the true population parameter. The width of the confidence interval depends on the standard error and the desired level of confidence.

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

Ans: Sampling errors can be reduced by:

- Increasing the sample size: A larger sample size is more likely to be representative of the population, and therefore the sampling error will be smaller.
- Using a random sampling method: A random sampling method ensures that all members of the population have an equal chance of being selected in the sample, which reduces the risk of bias.
- Stratifying the population: Stratifying the population involves dividing the population into groups, or strata, based on a shared characteristic, such as age, gender, or income level. A random sample is then taken from each stratum. This ensures that all groups in the population are represented in the sample.

- Using cluster sampling: Cluster sampling involves dividing the population into groups, or clusters, and then randomly selecting a sample of clusters. All members of the selected clusters are then included in the sample. This method is often used when it is difficult or expensive to sample the entire population.

76. How do the standard error and the margin of error relate?

Ans: The standard error and the margin of error are two related concepts in statistics.

The standard error is a measure of the precision of an estimate. It is calculated by dividing the standard deviation of the sampling distribution of an estimate by the square root of the sample size.

The margin of error is a measure of the accuracy of an estimate. It is calculated by multiplying the standard error by a critical value, which is determined by the desired level of confidence.

The standard error and the margin of error are related by the following formula:

Margin of error = z * standard error

where:

z is the critical value, which is determined by the desired level of confidence

standard error is the standard error of the estimate

77. What is hypothesis testing?

Ans: Hypothesis testing is a statistical procedure that is used to determine whether there is enough evidence to reject a null hypothesis. The null hypothesis is a statement about the population parameter that is being tested.

Hypothesis testing is based on the following steps:

- Formulate the null hypothesis (H0) and the alternative hypothesis (H1). The null hypothesis is the statement that is being tested. The alternative hypothesis is the opposite of the null hypothesis.
- Choose a test statistic. The test statistic is a measure of how different the sample data is from the null hypothesis.
- Calculate the p-value. The p-value is the probability of obtaining a test statistic as extreme or more extreme than the one observed, assuming that the null hypothesis is true.
- Make a decision. If the p-value is less than the significance level, the null hypothesis is rejected. Otherwise, the null hypothesis is not rejected.

78. What is an alternative hypothesis?

Ans: The alternative hypothesis (Ha) in hypothesis testing is the statement that is being tested. It is the opposite of the null hypothesis (H0).

The alternative hypothesis can be one-tailed or two-tailed. A one-tailed alternative hypothesis specifies the direction of the expected relationship between the variables. For example, a one-tailed alternative hypothesis might be that there is a positive relationship between two variables.

A two-tailed alternative hypothesis does not specify the direction of the expected relationship between the variables. For example, a two-tailed alternative hypothesis might be that there is a relationship between two variables, but it does not specify whether the relationship is positive or negative.

79.What is the difference between one-tailed and two-tail hypothesis testing?
Ans: In one-tailed hypothesis testing, the researcher predicts that the sample statistic will be either greater than or less than the hypothesized population parameter. In two-tailed hypothesis testing, the researcher predicts that the sample statistic will be different from the hypothesized population parameter, but does not specify the direction of the difference.

One-tailed hypothesis tests are more powerful than two-tailed hypothesis tests, meaning that they are more likely to detect a significant difference between the sample statistic and the hypothesized population parameter, if one exists. However, one-tailed hypothesis tests are also more likely to produce a type I error, which is the error of rejecting the null hypothesis when it is true.

Two-tailed hypothesis tests are less powerful than one-tailed hypothesis tests, but they are also less likely to produce a type I error.

The choice of whether to use a one-tailed or two-tailed hypothesis test depends on the research question being asked. If the researcher has a strong prior belief about the direction of the expected difference, then a one-tailed hypothesis test is appropriate. However, if the researcher does not have a strong prior belief about the direction of the expected difference, then a two-tailed hypothesis test is more appropriate.

80. What is one sample t-test?

Ans:
A one-sample t-test is a statistical test that is used to determine whether the mean of a sample is different from a specific value. This value is known as the hypothesized population mean.

The one-sample t-test is a parametric test, which means that it assumes that the sample data is normally distributed. If the sample data is not normally distributed, the one-sample t-test may not be reliable.

To perform a one-sample t-test, the following steps are taken:

- Calculate the sample mean and standard deviation.
- Calculate the t-statistic.
- Calculate the p-value.
- Make a decision.

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

Ans: Degrees of freedom (DF) in statistics is the number of independent values in a statistical analysis. It is an essential concept in many statistical tests, including hypothesis testing, confidence intervals, and linear regression.

The degrees of freedom can be calculated by subtracting one from the sample size. This is because one of the values in the sample is used to estimate the population mean, which reduces the number of independent values.

82. What is the p-value in hypothesis testing?

Ans: A p-value is a number that describes the probability of finding the observed or more extreme results when the null hypothesis (H0) is True.

P-values are used in hypothesis testing to help decide whether to reject the null hypothesis or not. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

83. How can you calculate the p-value?

Ans: The p-value can be calculated using a variety of methods, including:

- Using a statistical software package: Most statistical software packages, such as R, SPSS, and SAS, have built-in functions for calculating p-values.
- Using a table of p-values: P-value tables can be found in many statistics textbooks and online.
- Using a calculator: Some calculators have built-in functions for calculating p-values.

84. 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)?

Ans: Hypothesis testing is a type of statistical inference that uses data from a sample to conclude about the population data.

Before performing the testing, an assumption is made about the population parameter. This assumption is called the null hypothesis and is denoted by H0. An alternative hypothesis (denoted Ha), which is the logical opposite of the null hypothesis, is then defined.

The hypothesis testing procedure involves using sample data to determine whether or not H0 should be rejected. The acceptance of the alternative hypothesis (Ha) follows the rejection of the null hypothesis (H0).

85. How would you describe a ‘p-value’?

Ans: A p-value is a probability that is used to measure the strength of evidence against the null hypothesis in a statistical test. It is calculated by assuming that the null hypothesis is true and then determining the probability of obtaining a test statistic as extreme or more extreme than the one observed.

The p-value is a number between 0 and 1. A lower p-value indicates stronger evidence against the null hypothesis. For example, a p-value of 0.05 means that there is a 5% chance of obtaining a test statistic as extreme or more extreme than the one observed if the null hypothesis is true.

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

Ans: Type I and type II errors are two types of errors that can occur in statistical hypothesis testing.

A type I error is the error of rejecting the null hypothesis when it is true. This is also known as a "false positive".

A type II error is the error of failing to reject the null hypothesis when it is false. This is also known as a "false negative".

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

Ans:

A t-test and a z-test are both statistical tests used to compare two groups. However, they differ in some important ways.

A t-test is used when the population standard deviation is unknown or when the sample size is small (less than 30). A z-test is used when the population standard deviation is known and the sample size is large (greater than 30).

Another difference between the two tests is that the t-test uses a t-distribution to calculate the p-value, while the z-test uses a normal distribution. The t-distribution is more robust to violations of the normality assumption than the normal distribution.

88. What is the difference between the f test and anova test?

Ans:

The F-test and ANOVA (Analysis of Variance) are both statistical tests that are used to compare multiple groups. However, they differ in some important ways.

The F-test is a statistical test that is used to compare two variances. It is often used to test the null hypothesis that the variances of two populations are equal. The F-test is also used in ANOVA to test the significance of the difference between group means.

ANOVA is a statistical procedure that is used to compare multiple group means. It is based on the principle of partitioning the total variance in a data set into two components: between-group variance and within-group variance. The F-test is then used to test the significance of the between-group variance.

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

Ans : 1. K-fold cross-validation
2. Bootstrapping

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

Ans:The proportion of confidence intervals that will not contain the population parameter is equal to the significance level of the test. The significance level is the probability of rejecting the null hypothesis when it is true.

For example, if we construct a 95% confidence interval, the significance level is 0.05. This means that there is a 5% chance that the confidence interval will not contain the population parameter.

In other words, if we construct 100 confidence intervals, we can expect that approximately 95 of them will contain the population parameter and 5 of them will not.

91. What is a confounding variable?

Ans:

A confounding variable in statistics is an 'extra' or 'third' variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provides useless results.

For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather can be another confounding variable that may later the experiment design.

92.What are the steps we should take in hypothesis testing?
Ans:

1. State the null hypothesis

2. State the alternate hypothesis

3. Which test and test statistic to be performed

4. Collect Data

5. Calculate the test statistic

6. Construct Acceptance / Rejection regions

7. Based on steps 5 and 6, draw a conclusion about H0

83. What is the relationship between standard error and the margin of error?

Ans:

The standard error and the margin of error are two related concepts in statistics.

The standard error is a measure of the precision of an estimate. It is calculated by dividing the standard deviation of the sampling distribution of an estimate by the square root of the sample size.

The margin of error is a measure of the accuracy of an estimate. It is calculated by multiplying the standard error by a critical value, which is determined by the desired level of confidence.

The following formula shows the relationship between the standard error and the margin of error:

Margin of error = z * standard error

where:

- z is the critical value, which is determined by the desired level of confidence
- standard error is the standard error of the estimate

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

Ans: The best way to describe the p-value in simple terms is with an example. In practice, if the p-value is less than the alpha, say of 0.05, then we're saying that there's a probability of less than 5% that the result could have happened by chance. Similarly, a p-value of 0.05 is the same as saying "5% of the time, we would see this by chance."

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

Ans: Interpolation is a prediction made using inputs that lie within the set of observed values. Extrapolation is when a prediction is made using an input that's outside the set of observed values.

Generally, interpolations are more accurate.

86. What is an inlier?

Ans: An inlier is a data observation that lies within the rest of the dataset and is unusual or an error. Since it lies in the dataset, it is typically harder to identify than an outlier

87. You roll a biassed coin (p(head)=0.8) five times. What’s the probability of getting three or more heads?

and requires external data to identify them. Should you identify any inliers, you can simply remove them from the dataset to address them.

Ans:

The probability of getting three or more heads in five flips of a biased coin with a probability of heads of 0.8 is 0.43008000000000001.

This can be calculated using the following formula:

P(getting three or more heads) = 1 - P(getting two or fewer heads)

The probability of getting two or fewer heads can be calculated using the following formula:

P(getting two or fewer heads) = (1 - 0.8)^5 + 5(1 - 0.8)^4(0.8) + 10(1 - 0.8)^3(0.8)^2

This gives us a probability of 0.5699200000000001.

Therefore, the probability of getting three or more heads is 1 - 0.5699200000000001 = 0.4300800000000001.

88. 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.

Ans: To calculate the p-value of a one-sided test, we need to know the null hypothesis and the alternative hypothesis. The null hypothesis is the hypothesis that we are trying to reject. In this case, the null hypothesis is that the infection rate at the hospital is equal to or greater than 1 infection per 100 person-days at risk. The alternative hypothesis is the hypothesis that we are trying to support. In this case, the alternative hypothesis is that the infection rate at the hospital is less than 1 infection per 100 person-days at risk.

To test the null hypothesis, we can use a one-sided binomial test. The one-sided binomial test is a statistical test that is used to test the null hypothesis that the probability of success in a binomial experiment is equal to or greater than a certain value. In this case, the binomial experiment is the occurrence of an infection, and the success is the occurrence of a head.

```
pbinom(10, 1787, 1/100, lower.tail = FALSE)
```

This code outputs a p-value of 0.0127.

89. 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?
Ans:
To calculate a 95% Student's T confidence interval for the mean brain volume in this new population, we can use the following formula:

Confidence interval = mean ± t * standard error

where:

- mean is the sample mean brain volume (1,100cc)
- standard error is the standard deviation of the sample mean, which can be calculated using the following formula:

standard error = standard deviation / sqrt(sample size)

In this case, the standard error is 30cc / sqrt(9) = 10cc.

The t-value for a 95% confidence interval with 9 - 1 = 8 degrees of freedom is 2.262.

Therefore, the 95% confidence interval for the mean brain volume in this new population is 1,100cc ± 2.262 * 10cc = 1,077cc to 1,123cc.

This means that we are 95% confident that the true mean brain volume in this new population is between 1,077cc and 1,123cc.

90. What Chi-square test?

Ans: A statistical method is used to find the difference or correlation between the observed and expected categorical variables in the dataset.

Example: A food delivery company wants to find the relationship between gender, location and food choices of people in India.

It is used to determine whether the difference between 2 categorical variables is:

- Due to chance or
- Due to relationship

91. What is the ANOVA test?
Ans:

Alpha is the portion of confidence interval that will not contain the population parameter

$\alpha = 1 - CL$

92. How to calculate p-value using a manual method?
Ans:
To calculate a p-value using a manual method, you need to:

- Identify the correct test statistic for your test.
- Calculate the test statistic using the data from your sample.
- Look up the p-value in a table of p-values for your test statistic and degrees of freedom.
- Compare the p-value to your significance level.

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

What is the goal of A/B testing?
Ans:
Making a decision based on comparing p-value with significance level

In hypothesis testing, the p-value is the probability of obtaining a test statistic as extreme or more extreme than the observed test statistic if the null hypothesis is true. The significance level is the probability of rejecting the null hypothesis when it is true, and is typically set to 0.05.

To make a decision based on comparing the p-value with the significance level, we follow these steps:

1. Calculate the p-value for the test statistic.
2. Compare the p-value to the significance level.
3. If the p-value is less than the significance level, we reject the null hypothesis.
4. If the p-value is greater than or equal to the significance level, we fail to reject the null hypothesis.

Goal of A/B testing:

A/B testing is a method of comparing two versions of a website, app, or other digital product to see which one performs better. It is a type of experiment that uses statistical methods to compare the results of two different versions of a variable.

The goal of A/B testing is to identify the version of the product that is more effective in achieving the desired goal, such as increasing conversions or improving user engagement. A/B testing can be used to test a variety of elements of a product, such as the headline, call to action, or layout.

94. What is the difference between a box plot and a histogram.
Ans: Box plot:

A box plot is a graphical representation of a five-number summary of a data set. The five-number summary is a set of five statistics that describe the distribution of the data: minimum, first quartile, median, third quartile, and maximum.

Histogram:

A histogram is a graphical representation of the distribution of a data set. It is constructed by dividing the data into bins and plotting the number of data points in each bin.

Difference between box plots and histograms:

The main difference between box plots and histograms is that box plots are used to summarize a data set using five statistics, while histograms are used to visualize the shape of a distribution.

Another difference is that box plots can be used to compare the distributions of two or more data sets, while histograms are typically used to visualize the distribution of a single data set.

95. 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?
Ans:
The probability of getting 10 heads in 10 tosses of a fair coin is ½ power 10 =1/1024. However, given that we see 10 heads, we know that the coin is not fair. In fact, the only way to get 10 heads in 10 tosses is if the coin is double-headed. Therefore, the probability that the next toss of the coin is also a head is 1.

96. What is a confidence interval and how do you interpret it?
Ans: A confidence interval is a statistical method for estimating the range of values that a population parameter is likely to fall within. Confidence intervals are typically calculated using sample data and are expressed as a range of values with a certain level of confidence.

For example, a 95% confidence interval for the mean height of adults in the United States might be 69 inches to 71 inches. This means that we are 95% confident that the true mean height of adults in the United States is between 69 inches and 71 inches.

Confidence intervals can be used to estimate a variety of population parameters, such as the mean, median, proportion, and variance. They can also be used to test hypotheses about population parameters.

97. What is correlation?

Ans:

Correlation is a statistical measure of the relationship between two variables. It is used to measure how strongly two variables are related and in what direction.

There are two main types of correlation: positive and negative. A positive correlation means that the two variables tend to move in the same direction. For example, there is a positive correlation between height and weight, meaning that taller people tend to be heavier.

A negative correlation means that the two variables tend to move in opposite directions. For example, there is a negative correlation between temperature and ice cream sales, meaning that when the temperature is higher, ice cream sales tend to be lower.

98. What types of variables are used for Pearson’s correlation coefficient?

Ans:
Pearson's correlation coefficient is used to measure the linear correlation between two continuous variables. This means that the two variables must be measured on a numerical scale, such as height, weight, or temperature. The variables must also be normally distributed, which means that they should follow a bell-shaped curve.

Here are some examples of variables that can be used for Pearson's correlation coefficient:

- Height and weight
- Age and test scores
- Temperature and humidity
- Income and education level
- Sales and advertising spending

99.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?

Ans: A high correlation between the time a person sleeps and the amount of productive work he does can be inferred to mean that sleep is important for productivity. However, it is important to note that correlation does not imply causation. Just because two variables are correlated does not mean that one variable causes the other.

There are a few possible explanations for the correlation between sleep and productivity:

- Sleep may improve cognitive function, such as attention, memory, and decision-making.
- Sleep may help to reduce stress and improve mood, which can lead to increased productivity.
- Sleep may help to regulate hormones that affect energy levels and motivation.
- Sleep deprivation may lead to fatigue, which can make it difficult to focus and concentrate on work.

100. What is the meaning of covariance?

Ans: Covariance is a statistical measure of the relationship between two random variables. It measures how much the two variables tend to move together. For example, if the two variables are positively correlated, then the covariance will be positive. If the two variables are negatively correlated, then the covariance will be negative.

The covariance is calculated by taking the average of the product of the deviations of the two variables from their means. In other words, it is the average of the cross products of the two variables.

The covariance is a useful measure of the relationship between two variables, but it is important to note that it is not a normalized measure. This means that it is not possible to interpret the covariance directly.

101. What does autocorrelation mean?

Ans: Autocorrelation is a statistical measure of the correlation between a variable and its lagged values. In other words, it measures how much a variable is correlated with itself over time.

Autocorrelation is often used in time series analysis to identify patterns in data. For example, if a time series has a high autocorrelation, then it is likely that the future values of the series will be similar to the current and past values.

102. What types of variables are used for Pearson’s correlation coefficient?

Ans:
Pearson's correlation coefficient is used to measure the linear correlation between two continuous variables. This means that the two variables must be measured on a numerical scale, such as height, weight, or temperature. The variables must also be normally distributed, which means that they should follow a bell-shaped curve.

Here are some examples of variables that can be used for Pearson's correlation coefficient:

- Height and weight
- Age and test scores

- Temperature and humidity
- Income and education level
- Sales and advertising spending

103. How will you determine the test for the continuous data?
Ans:
To determine the test for continuous data, you need to consider the following factors:

- The type of data: Is the data normally distributed? Is it paired or unpaired?
- The research question: Are you comparing two groups or testing a hypothesis about a population mean?
- The significance level: What is the probability of rejecting a true null hypothesis?

Once you have considered these factors, you can choose the appropriate test. Here are some common tests for continuous data:

- t-test: This test is used to compare the means of two groups or to test a hypothesis about a population mean. The t-test can be used for paired or unpaired data, and it assumes that the data is normally distributed.
- ANOVA: This test is used to compare the means of three or more groups. ANOVA can be used for paired or unpaired data, and it assumes that the data is normally distributed.
- Chi-squared test: This test is used to test for independence between two categorical variables. The chi-squared test does not assume that the data is normally distributed.

104. What can be the reason for non normality of the data?
Ans:
There are a number of reasons why data may not be normally distributed. Some of the most common reasons include:

- The underlying distribution is non-normal. Some phenomena, such as bacterial growth, naturally follow a non-normal distribution.

- Outliers or mixed distributions are present. Outliers are extreme values that are far from the rest of the data. Mixed distributions are distributions that are made up of two or more different normal distributions.
- A low discrimination gauge is used. A low discrimination gauge is a measurement tool that does not have a high enough resolution to accurately measure the data.
- Skewness is present in the data. Skewness is a measure of the asymmetry of a distribution. A skewed distribution has more values on one side than the other.
- You have a small sample size. If you have a small sample size, the data is more likely to be non-normal, even if the underlying distribution is normal.

105. why is there no such thing like 3 samples t- test?? why t-test failed with 3 samples
Ans:
There is no such thing as a 3-sample t-test because the t-test is designed to compare the means of two groups. When you have three or more groups, you need to use a different statistical test, such as ANOVA (analysis of variance).

The t-test works by assuming that the data is normally distributed and that the two groups have equal variances. When you have three or more groups, it is difficult to meet these assumptions. Additionally, the t-test is not very powerful for detecting differences between more than two groups.

ANOVA is a more powerful test for detecting differences between three or more groups. ANOVA works by comparing the variance between groups to the variance within groups. If the variance between groups is significantly greater than the variance within groups, then we can conclude that there is a significant difference between the means of the groups.

Here are some examples of when to use the t-test and ANOVA:

- t-test: You want to compare the average height of men and women. You want to test the hypothesis that the average test score in your class is greater than 70.

- ANOVA: You want to compare the average height of people from different countries. You want to test the hypothesis that the average test score in your class is different for students who received different types of tutoring.