STATISTICS - 1

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

Correct answer (a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Correct answer (a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

Correct answer (b) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log-normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

Correct answer (c) The square of a standard normal random variable follows what is called chi-squared distribution

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Correct answer (c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

Correct answer (b) False

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Correct answer (b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

Correct answer (a) 0

9. Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Correct answer (c) Outliers cannot conform to the regression relationship

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

Answer

Normal Distribution, also known as Gaussian Distribution, is a type of probability distribution that describes the pattern of a set of data that is distributed symmetrically around the mean. In simple language, it means that a normal distribution looks like a bell-shaped curve, where the majority of the data points are clustered around the middle (the mean) and decrease in frequency as they move away from the mean in either direction.

The normal distribution is characterized by two parameters: the mean (average) and the standard deviation (spread or variability) of the data. The mean is the central point of the distribution, while the standard deviation measures how spread out the data is from the mean.

Many real-world phenomena, such as the heights of people, weights of objects, and scores on standardized tests, follow a normal distribution. The normal distribution is an important concept in statistics and data analysis because it allows us to make predictions about the likelihood of events occurring based on the characteristics of the data.

11. How do you handle missing data? What imputation techniques do you recommend?

Answer

Handling missing data is an important task in data analysis, as missing values can affect the quality and accuracy of the results. In simple language, missing data refers to the absence of data for certain variables or observations in a dataset.

There are several ways to handle missing data, including:

Complete case analysis: This involves simply removing any observations that have missing data for any of the variables. This method is simple but can result in a loss of information.

Mean/median imputation: This involves replacing missing values with the mean or median of the non-missing values for that variable. This method is simple but can result in biased estimates and incorrect standard errors.

Multiple imputation: This involves creating multiple plausible values for the missing data based on a statistical model, then combining the results using a specific algorithm. This method is more complex but can provide more accurate and unbiased estimates.

Maximum likelihood estimation: This involves estimating the missing data values based on

the likelihood of the observed data, using a statistical model. This method is more complex but can provide more accurate estimates.

The choice of imputation technique depends on the nature of the data and the research question. However, multiple imputation is often recommended as it can provide more accurate and reliable results. It is important to keep in mind that imputing missing data introduces uncertainty and should be used with caution. It is also important to report any missing data and the method used for imputation in any analyses or publications.

12. What is A/B testing?

Answer

A/B testing is a statistical technique used to compare two or more versions of a product, service, or marketing campaign to determine which one performs better. In A/B testing, two versions (A and B) of a product or service are randomly presented to different groups of users, and their performance is compared based on a specific metric or outcome, such as click-through rates, conversion rates, or sales.

The goal of A/B testing is to determine which version of the product or service is more effective in achieving the desired outcome, and to identify the factors that contribute to the difference in performance. A/B testing can be used for a variety of purposes, including optimizing website design, improving email marketing campaigns, and testing new product features.

To conduct an A/B test, a sample size is first determined, and users are randomly assigned to either version A or B. The performance of each version is then measured using a specific metric, and the results are compared using statistical analysis. If there is a statistically significant difference in performance between the two versions, the better performing version is chosen as the winner, and the changes are implemented.

A/B testing can be a powerful tool for optimizing and improving products and services, as it allows for data-driven decision-making and can help identify the factors that contribute to success. However, it is important to ensure that the test is designed and executed properly, and that the results are interpreted correctly.

13. Is mean imputation of missing data acceptable practice?

Answer

Mean imputation is a common method for handling missing data in datasets. It involves replacing missing values with the mean value of the non-missing data for that variable. While mean imputation is a simple and straightforward method, it does have some drawbacks.

One of the main issues with mean imputation is that it assumes that the missing values are missing at random (MAR) or missing completely at random (MCAR), which may not be the

case in practice. If the missing values are related to other variables in the dataset, such as age, gender, or income, mean imputation can introduce bias into the analysis.

Another issue with mean imputation is that it reduces the variance of the variable, which can affect the results of subsequent analyses. For example, if the variable is used to predict an outcome variable, the standard errors of the coefficients will be underestimated, leading to incorrect statistical inferences.

Therefore, mean imputation is generally not recommended as the sole method for handling missing data. Other methods, such as multiple imputation or maximum likelihood estimation, may be more appropriate, depending on the nature of the missing data and the research question.

14. What is linear regression in statistics?

Answer

linear regression is a type of supervised learning algorithm used for predicting a continuous outcome variable based on one or more predictor variables.

The basic idea behind linear regression is to find a linear relationship between the predictor variables and the outcome variable by fitting a straight line to the data. The algorithm uses a cost function to calculate the difference between the predicted and actual values of the outcome variable, and then adjusts the parameters of the line to minimize this difference.

Once the line is fitted to the data, it can be used to make predictions for new data points by plugging in values of the predictor variables. The algorithm can also be used to identify the predictor variables that are most strongly associated with the outcome variable, and to test hypotheses about the relationship between the predictor variables and the outcome variable.

Linear regression is a popular and widely used machine learning algorithm due to its simplicity and interpretability. It is often used as a baseline method for more complex models, and can be extended to handle more complex relationships between the predictor variables and the outcome variable.

15. What are the various branches of statistics?

Answer

Statistics is a broad field that encompasses various sub-disciplines. Some of the major branches of statistics include:

- Descriptive Statistics: This branch deals with the methods used to summarize and describe the main features of a dataset, such as measures of central tendency, variability, and correlation.
- Inferential Statistics: This branch deals with the methods used to make inferences or

predictions about a larger population based on a sample of data. It involves techniques such as hypothesis testing and estimation.

- Probability Theory: This branch deals with the study of random events and their probabilities. It provides the foundation for statistical inference and modeling.
- Exploratory data analysis: It involves visualizing and analyzing data to identify patterns, trends, and anomalies. This branch includes techniques such as scatter plots, histograms, and box plots.
- Time series analysis: It involves the analysis of data collected over time, such as stock prices or weather data. Time series analysis is used in fields such as economics, finance, and meteorology.