Statistics Worksheet_1

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

a) True

b) False

Ans: A

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

Ans: A

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

Ans: B

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

Ans: C

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans: C

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

a) True

b) False
Ans: B

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Ans: B

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

Ans: 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

Ans: C

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

Ans: The term "Normal Distribution" refers to a probability distribution that is symmetric, bell-shaped, and characterized by its mean and standard deviation. It is also known as the Gaussian distribution. In a normal distribution, the data is centered around the mean, and the majority of the observations fall close to the mean with decreasing frequency as they move away from it.

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

Ans: Handling missing data is an important aspect of data analysis. There are several approaches to handle missing data, and the choice of imputation technique depends on the nature and characteristics of the data. Here are some commonly used imputation techniques:

1. Mean/Median/Mode Imputation: In this approach, missing values are replaced with the mean, median, or mode of the available data for that variable. This technique is simple and commonly used when the missing values are assumed to be missing at random.

2. Hot Deck Imputation: This technique involves randomly selecting a value from a similar donor observation (based on certain matching criteria) and using it to impute the missing value. It preserves the relationship between variables and can be useful when there are patterns in the missing data.

3. Multiple Imputation: Multiple imputation involves creating multiple plausible imputations for missing values based on the observed data and the estimated relationships between variables. This technique takes into account the uncertainty associated with the imputed values and is often recommended when the missingness is non-random.

4. Regression Imputation: In this approach, missing values are imputed using regression models. A regression model is created using variables with complete data, and the missing values are predicted based on the relationship between the variables.

5. K-Nearest Neighbors Imputation: This technique involves identifying the K-nearest neighbors based on variables with complete data and using their values to impute the missing values. It is useful when the missingness is related to the values of other variables.

The choice of imputation technique depends on factors such as the amount of missing data, the pattern of missingness, the distribution of the variables, and the specific requirements of the analysis. It is important to consider the limitations and assumptions of each technique and to assess the potential impact of imputation on the results of the analysis.

12. What is A/B testing?

Ans: A/B testing, also known as split testing, is a statistical method used to compare two or more variations of a webpage, advertisement, or any other element to determine which one performs better in terms of a desired outcome or goal. It is commonly used in marketing, user experience (UX) research, and website optimization.

In A/B testing, two or more versions of a webpage or design element, referred to as the control (A) and the variant (B), are presented to different segments of the target audience simultaneously. The

performance of each variation is measured and compared based on specific metrics, such as conversion rate, click-through rate, or engagement.

The goal of A/B testing is to identify the variation that yields the highest conversion or engagement rate. By randomly assigning participants to different variations, A/B testing helps eliminate biases and provides a reliable comparison between different options. The statistical significance of the results is usually assessed to ensure the observed differences are not due to chance.

A/B testing allows businesses and researchers to make data-driven decisions by systematically testing and refining their designs, content, or marketing strategies. It helps uncover insights about user preferences, behavior, and the effectiveness of different elements, ultimately leading to improved performance, higher conversion rates, and better user experiences.

13. Is mean imputation of missing data acceptable practice?

Ans: Mean imputation of missing data is a commonly used method for handling missing values. However, its acceptability as a practice depends on the specific context and the nature of the data being analyzed.

Mean imputation involves replacing missing values with the mean value of the available data for that variable. While it is a simple and straightforward approach, it has certain limitations. One major concern is that mean imputation assumes that the missing values are missing completely at random (MCAR) and that the mean represents a reasonable estimate for the missing values. However, if the missingness is related to the underlying data or if the variable has substantial variability, mean imputation may introduce bias and distort the distribution of the variable.

In some cases, mean imputation may be acceptable if the missingness is minimal, the variable has low variability, and the assumptions of MCAR are reasonable. Additionally, mean imputation may be more appropriate for variables that are not the primary focus of the analysis or for exploratory analyses where the goal is to get a general understanding of the data.

However, in more complex analyses or situations where missingness is significant, other imputation techniques such as multiple imputation, regression imputation, or sophisticated machine learning-based imputation methods may be more suitable. These methods take into account the relationships between variables and can provide more accurate estimates.

Ultimately, the acceptability of mean imputation depends on the specific context and the goals of the analysis. It is important to carefully consider the potential limitations and implications of mean imputation before deciding to use it as a missing data handling technique.

14. What is linear regression in statistics?

Ans: Linear regression is a statistical modeling technique used to understand and analyze the relationship between a dependent variable and one or more independent variables. It aims to find a linear equation that best fits the data points and predicts the value of the dependent variable based on the independent variables.

In linear regression, the dependent variable is assumed to be a continuous numerical variable, while the independent variables can be either numerical or categorical. The goal is to estimate the coefficients of the linear equation that minimize the difference between the observed values and the predicted values.

15. What are the various branches of statistics?

Ans: Statistics, as a field of study, encompasses various branches that focus on different aspects of data analysis, interpretation, and inference. Some of the major branches of statistics include:

1. Descriptive Statistics: This branch involves summarizing and describing data through measures such as mean, median, mode, variance, and standard deviation. Descriptive statistics provide insights into the central tendency, variability, and distribution of data.

2. Inferential Statistics: Inferential statistics involves making inferences and drawing conclusions about a population based on a sample of data. It includes techniques such as hypothesis testing, confidence intervals, and estimation.

3. Probability Theory: Probability theory deals with the study of uncertainty and the likelihood of events occurring. It provides a mathematical framework to quantify and analyze randomness, and forms the foundation of statistical inference.

4. Biostatistics: Biostatistics is the application of statistical methods to biological and health-related data. It involves designing clinical trials, analyzing epidemiological data, and conducting studies in areas such as genetics, public health, and biomedical research.

5. Econometrics: Econometrics applies statistical methods to economic data to analyze economic relationships, test hypotheses, and forecast economic variables. It combines elements of economics, mathematics, and statistics to study economic phenomena.

6. Data Mining and Machine Learning: Data mining and machine learning involve using statistical techniques to discover patterns, extract knowledge, and make predictions from large datasets. These branches focus on developing algorithms and models for data analysis and automated decision-making.

7. Bayesian Statistics: Bayesian statistics is an approach that incorporates prior knowledge and beliefs into statistical analysis. It uses Bayes' theorem to update beliefs based on new evidence, providing a framework for decision-making under uncertainty.

8. Multivariate Analysis: Multivariate analysis deals with the analysis of datasets that involve multiple variables. It includes techniques such as principal component analysis, factor analysis, cluster analysis, and multivariate regression.

9. Time Series Analysis: Time series analysis focuses on analyzing and forecasting data that is collected sequentially over time. It includes methods to understand patterns, trends, and seasonality in time-dependent data.

These branches of statistics are interconnected and often overlap in their applications and methodologies. They provide a comprehensive toolkit for understanding data, drawing meaningful conclusions, and making informed decisions across various disciplines.