STATISTICS WORKSHEET-1

Q1
A
Q2
A
Q3
C
Q4
D
Q5
C
Q6
B
Q7
B
Q8
A
Q9
C

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

The normal distribution, also known as the Gaussian distribution or the bell curve, is a continuous probability distribution that is defined by its mean (μ) and standard deviation (σ). It is a symmetrical distribution where the mean, median, and mode are all equal, and where the majority of values fall within one to three standard deviations from the mean.

Normal distributions are symmetrical about the mean and show that data near the mean are more likely to be observed than data off the mean. In graph form, the normal distribution will appear as a bell curve. The normal distribution is widely used in statistics because of its simplicity, tractability, and wide applicability.

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

Here are some common imputation techniques used to handle missing data: 1. Deletion methods: This involves deleting or removing the missing data points entirely. This can be done in three ways: pairwise deletion, listwise deletion, and case deletion. These methods are easy to implement but can lead to biased results. 2. Mean/median imputation: This involves replacing missing values with the mean or median of the available data. This method is simple and easy to implement but can produce biased results. 3. Regression imputation: This involves predicting missing values by using a regression model based on the other variables in the dataset. This method can produce accurate results but can be computationally expensive. 4. Multiple imputations: This involves creating multiple imputed datasets and then performing the analysis separately on each data set. This method produces accurate and unbiased results but can be computationally expensive and requires advanced statistical software. The choice of imputation technique depends on the nature of the missing data and the goals of the analysis. The key is to carefully consider the pros and cons of each method and choose the one that is most appropriate for the dataset and analysis.

12. **What is A/B testing?**

A/B testing, also known as split testing, is a statistical method used to compare two different versions of a particular product or service to determine which one performs better. In A/B testing, the two versions, version A and version B, are randomly assigned to a group of users or customers, and their behaviour is tracked and analyzed to determine which version is more effective in achieving the desired outcome, such as higher click-through rates, more conversions, or more sales.

Typically, A/B testing involves creating two different versions that differ in only one element, such as the colour of a button, the text of a message, or the layout of a webpage. The users are randomly assigned to either version A or version B, and their behaviour, such as clicking on a button or making a purchase, is recorded and analyzed. The data is then statistically analyzed to determine which version is more effective.

A/B testing is commonly used in web design, marketing, and product development to optimize user experience, improve conversion rates, and increase revenue. It allows businesses to make data-driven decisions and continually improve their products and services based on user feedback and behaviour.

13. **Is mean imputation of missing data acceptable practice?**

Mean imputation involves replacing missing values with the mean of the available data for that variable. While this method is simple and easy to implement, it has some drawbacks:

Loss of variability: Mean imputation reduces the variability in the dataset, as it replaces missing values with a constant value (the mean). This can lead to biased estimates of variances and covariances, which can affect the performance of some statistical models.

Distortion of relationships: If the missing data is not missing at random, mean imputation can distort the relationships between variables. This can lead to incorrect inferences and conclusions.

Ignoring the underlying reasons for missing data: Mean imputation does not consider the reasons why data is missing. If there is a systematic reason for the missing data, mean imputation may not be appropriate.

In some cases, mean imputation can be an acceptable practice, especially when the proportion of missing data is small, and the data is missing at random. However, it is essential to consider alternative methods for handling missing data, such as:

Median or mode imputation: Replacing missing values with the median or mode can be more robust to outliers than mean imputation.

Regression imputation: Using regression models to predict missing values based on the relationships between variables.

Multiple imputation: Creating multiple datasets with different imputed values and combining the results to account for the uncertainty introduced by imputation.

K-Nearest Neighbors (KNN) imputation: Replacing missing values with the average of the K-nearest neighbors in the dataset.

14. **What is linear regression in statistics?**

Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (also known as predictors, features, or explanatory variables). The goal of linear regression is to find the best-fitting straight line that describes the relationship between the variables, allowing us to make predictions or understand the associations between them.

15. **What are the various branches of statistics?**

Statistics is a broad field that is used in many areas of science, engineering, business, and social sciences. Here are some of the main branches of statistics:

1. Descriptive statistics: This branch of statistics is concerned with summarizing and describing data using measures such as mean, median, mode, variance, and standard deviation.

2. Inferential statistics: This branch of statistics involves drawing conclusions or making inferences about a population based on a sample of data.

3. Probability theory: This branch of statistics is concerned with the study of random events and their outcomes and is used to model and analyze complex systems.

4. Biostatistics: This is the application of statistical methods to problems in biology and medicine, such as clinical trials, epidemiology, and genetics.

5. Econometrics: This is the application of statistical methods to problems in economics, such as predicting economic trends and evaluating policy interventions.

6. Data mining: This involves using statistical techniques to extract patterns and knowledge from large datasets.

7. Statistical learning: This involves using statistical techniques to build predictive models from data, such as machine learning algorithms.

8. Bayesian statistics: This is a branch of statistics that uses Bayes' theorem to update prior beliefs in the face of new evidence and is used in a wide range of fields, including medicine, finance, and engineering.