STATISTICS WORKSHEET-1 (Answer Sheet)

Q1. (a)

Q2. (a)

Q3. (b)

Q4. (d)

Q5. (c)

Q6. (a)

Q7. (b)

Q8. (a)

Q9. (c)

Q10. What do you understand by the term Normal Distribution?

Ans.:- The Normal Distribution is also known as Gaussian distribution or bell curve. It refers to a statistical distribution that is characterized by a specific shape and set of properties. It's a continuous probability distribution that is symmetric bell shaped .The Probability Density Function (PDF) of the normal distribution produces a bell shaped curve. It's most use for analysis and other type of statistical analysis. The normal distribution has two parameters one is mean and second is standard deviation. The mean, median and mode of a normal distribution are all equal and located at the center of the distribution. In an Empirical rule a large proportion of the data (about 68%) falls within one standard deviation of the mean, an even larger proportion (about 95%) falls within two standard deviations, and an even larger proportion (about 99.7%) falls within three standard deviations. In Z-score values from a normal distribution can be transformed into standardized scores called z-scores, which represent the number of standard deviations a data point is away from the mean. It's plays a crucial role in the Central Limit Theorem, which states that the sum or average of a large number of independent and identically distributed random variables tends to follow a normal distribution, regardless of the original distribution of the variables.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans.:- Handling missing data is an important step in data preprocessing and analysis. Missing data can arise due to various reasons such as data collection errors, non-responses in surveys, or technical issues. Imputation is the process of filling in missing values with estimated or predicted values to create a complete dataset for analysis. The choice of imputation technique depends on the nature of the data and the underlying assumptions. Here are some common imputation techniques:

1. Mean and median imputation:-Replace missing values with the mean (or median) of the available data for that variable. This method is simple but assumes that the missing values are missing completely at random (MCAR) and do not bias the distribution.
2. Mode imputation :- For categorical variables, replace missing values with the mode (most frequent value) of the available data for that category.
3. Regression imputation:- Use regression models to predict missing values based on other variables. Fit a regression model with the variable containing missing values as the dependent variable and other relevant variables as predictors.
4. KNN imputation:- Replace missing values with the average of the k-nearest neighbors in terms of other variables. This method considers relationships between variables and is useful when dealing with multiple missing values.
5. Multiple imputation:- This involves creating multiple datasets where missing values are imputed multiple times using a suitable technique. Analysis is performed on each imputed dataset, and results are combined to produce more accurate estimates and account for imputation uncertainty.
6. Hot-deck imputation:- Select a similar individual from the dataset with complete data and impute the missing value using that individual's value. This method mimics the "donor's" value for imputation.
7. Interpolation and Extrapolation:- For time series data, missing values can be imputed using interpolation (within the known range) or extrapolation (outside the known range) techniques.
8. Expectation-Maximization (EM) algorithm:- This iterative algorithm estimates missing values by maximizing the likelihood of the observed data. It's particularly useful for dealing with missing values in multivariate data.

Q12. What is A/B testing?

Ans.:- A/B testing, also known as split testing, is a method used in marketing, product development, and other fields to compare two versions of a webpage, app, marketing campaign, or any other variable to determine which one performs better. The objective of A/B testing is to identify which version yields better outcomes in terms of user engagement, conversions, sales, or other relevant metrics.

The process of A/B testing involves the following step:

Hypothesis Formulation:- Define a clear hypothesis about the change you want to test. This could be a design element, copy, feature, or any other variable.

Randomization:- Divide your audience or sample into two groups: Group A (the control group) and Group B (the experimental group). Randomization helps ensure that any differences observed between the groups are due to the changes being tested and not other factors.

Implementation:- Implement the change you want to test in the Group B version. Group A remains unchanged and serves as a baseline for comparison.

Data collection:- Gather data on relevant metrics such as click-through rates, conversion rates, sales, or any other performance indicators.

Comparision:- Analyze the data to compare the performance of the two versions. This could involve statistical analysis to determine whether the observed differences are statistically significant.

Conclusion:- Based on the analysis, decide whether the changes in Group B have a significant impact on the desired metrics compared to Group A. If the experimental version performs significantly better, you might consider adopting the changes.

Note that the success of A/B testing depends on various factors, including the quality of the hypothesis, the size of the sample, the duration of the test, and the accuracy of data collection and analysis. Careful planning and proper execution are crucial to draw valid conclusions from A/B testing experiments.

Q13. Is mean imputation of missing data acceptable practice?

Ans.:- Mean imputation of missing data is a widely used method due to its simplicity, but it comes with certain limitations and potential pitfalls that you should be aware of before applying it. Here are some considerations:

Simple implementation:- Mean imputation is straightforward and easy to implement, making it a quick solution for handling missing data.

Preserves:- Imputing missing values with the mean allows you to retain the original sample size, which can be important for maintaining statistical power.

Works well for MCAR:- Mean imputation can work reasonably well when the missing data is Missing Completely at Random (MCAR), meaning that the missingness is unrelated to the values of the variable or any other variables in the dataset.

Q14. What is linear regression in statistics?

Ans.:- Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It's commonly used for prediction and understanding the relationship between variables.

Formula:- Y= $\beta_0+\beta_1 X+\epsilon$

Where:

Y is the dependent variable (also known as the response or target variable).

X is the independent variable (also known as the predictor variable or feature).

- 0β0 is the intercept, which represents the value of Y when X is 0.

1β1 is the coefficient (slope), indicating the change in Y for a unit change in X.

ϵ represents the error term, which accounts for the variability in Y that is not explained by the linear relationship with X.

Q15. What are the various branches of statistics?

Ans.:- Statistics is a broad field that encompasses various branches, each focusing on different aspects of data analysis, interpretation, and application. Some of the main branches of statistics include.

Descriptive Statistics:- Descriptive statistics involve methods for summarizing and presenting data in a meaningful and informative way. This includes measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and graphical representations (histograms, box plots, scatter plots) to describe and visualize data.

Inferential Statistics:- Inferential statistics involve drawing conclusions and making predictions about populations based on sample data. Techniques in this branch include hypothesis testing, confidence intervals, and regression analysis.

Probability:- Probability theory deals with the mathematical concepts and rules governing uncertainty and randomness.

Statistical Computing:- This branch focuses on developing and using computational tools and software to handle large datasets, perform complex calculations, and implement statistical analyses.

Time series analysis etc..

These are some branches of statistics often overlap and interact, as different areas of study require tailored approaches to data analysis.