STATISTICS WORKSHEET-1

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

Ans: a) True

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

Ans: a) Central Limit Theorem

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

Ans: b) Modelling 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

Ans: d) All of the mentioned

5. _____ random variables are used to model rates

Ans: c) Poisson

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

Ans: b) False

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

Ans: b) Hypothesis

8. Normalized data are centred at_____ and have units equal to standard deviations of the original data.

Ans: a) 0

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

Ans: c) Outliers cannot conform to the regression relationship

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

Ans: Normal distribution, often referred to as a Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. Here are some key characteristics:

1. Bell-shaped Curve: The graph of a normal distribution is bell-shaped and is centred around the mean.

2. Mean, Median, Mode: In a normal distribution, the mean, median, and mode are all equal and located at the centre of the distribution.

3. Standard Deviation: The spread of the distribution is determined by the standard deviation. A smaller standard deviation results in a steeper curve, while a larger standard deviation results in a flatter curve.

4. 68-95-99.7 Rule: Approximately 68% of the data falls within one standard deviation of the mean, about 95% within two standard deviations, and about 99.7% within three standard deviations.

5. Central Limit Theorem: The normal distribution is significant in statistics because of the central limit theorem, which states that the sum of a large number of independent, identically distributed variables tends toward a normal distribution, regardless of the original distribution of the variables.

Normal distributions are widely used in statistics, natural and social sciences for various analyses and inferential statistics.

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

Ans: Handling missing data is a crucial aspect of data analysis, as it can significantly affect the results of your models. Here are some common strategies and imputation techniques:

1. Understand the Nature of Missing Data

- Missing Completely at Random (MCAR): The missingness is unrelated to the data itself.

- Missing at Random (MAR): The missingness is related to other observed data but not to the missing data.

- Missing Not at Random (MNAR): The missingness is related to the missing data itself.

2. Techniques for Handling Missing Data

- Deletion Methods:

- Listwise Deletion: Remove entire records with missing values. This can lead to loss of data and potential bias.

- Pairwise delegation: Use available data for calculations, ignoring missing values only for specific analyses.

- Imputation Techniques:

-Mean/Median/Mode Imputation: Replace missing values with the mean, median, or mode of the observed data. Simple but can reduce variability.

- K-Nearest neighbours (KNN) Imputation: Replace missing values based on the values of their nearest neighbours in the dataset.

- Regression Imputation: Use regression models to predict and fill in missing values based on other variables.

- Multiple Imputation: Generate several imputed datasets and combine results to account for uncertainty around the missing values.

- Interpolation/Extrapolation: Suitable for time-series data; fill missing values based on trends in the data.

- Machine Learning Models: Use algorithms like random forests or neural networks to predict and impute missing values based on other features.

3. Assessing Impact

- Always assess how the imputation affects the overall results, using techniques like sensitivity analysis to evaluate how robust your findings are to the method of imputation.

4. Software Tools

- Many programming languages and libraries (e.g., Python's `pandas`, `scikit-learn`, R's `mice`, `miss Forest`) offer built-in functions for various imputation techniques.

Choosing the right method depends on the context of the data, the proportion of missingness, and the specific analysis goals. It's important to document your approach and justify your choices based on the data's characteristics.

12. What is A/B testing?

Ans: A/B testing, also known as split testing, is a method used to compare two versions of a webpage, app, or other content to determine which one performs better. In an A/B test, you create two variants (A and B) that are identical except for one key element—such as a headline, button colour, or layout.

Here's how it typically works:

1. Hypothesis: Identify what you want to test and formulate a hypothesis about how a change might improve performance.

2. Segmentation: Randomly split your audience into two groups. One group sees version A (the control), and the other sees version B (the variation).

3. Data Collection: Measure how each version performs based on predefined metrics, like click-through rates, conversion rates, or user engagement.

4. Analysis: After sufficient data has been collected, Analyse the results to determine which version performed better.

5. Implementation: If one version outperforms the other significantly, you can implement that change more broadly.

A/B testing helps businesses make data-driven decisions, optimizing user experience and increasing conversion rates.

13. Is mean imputation of missing data acceptable practice?

Ans: Mean imputation is a common method for handling missing data, but it's generally considered a suboptimal practice for several reasons:

1. Bia: It can introduce bias, particularly if the data are not missing completely at random. This can distort relationships between variables.

2. Underestimation of Variance: By replacing missing values with the mean, you reduce the variability in the dataset, which can lead to misleading results, particularly in statistical analyses.

3. Distorted Distributions: Mean imputation can affect the distribution of the data, leading to skewness or kurtosis that doesn’t reflect the true underlying data.

4. Loss of Information: It ignores the underlying data patterns and relationships, essentially discarding information about the missing values.

Alternative methods, such as multiple imputation, k-nearest neighbours, or model-based approaches, often provide better results by preserving the underlying data structure and variability. It’s usually advisable to explore and use these methods instead of mean imputation when dealing with missing data.

14. 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. The basic form of linear regression involves one independent variable and is represented by the equation:

$$ y = mx + b $$

Where:

- $ y $ is the dependent variable (the outcome you're trying to predict).

- $ x $ is the independent variable (the predictor).

- $ m $ is the slope of the line (indicating the change in $ y $ for a one-unit change in $ x $)).

- $ b $ is the y-intercept (the value of $ y $ when $ x $ is zero).

When there are multiple independent variables, the equation extends to:

$$ y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n $$

Where $ b_0 $ is the intercept and $ b_1, b_2, ..., b_n $ are the coefficients for each independent variable.

Key Concepts:

1. Assumptions: Linear regression assumes linearity, independence of errors, homoscedasticity (constant variance of errors), and normally distributed errors.

2. Least Squares Method: This is the most common method for estimating the parameters of the model. It minimizes the sum of the squares of the differences between the observed and predicted values.

3. R-squared: This statistic indicates the proportion of variance in the dependent variable that can be explained by the independent variables.

4. Applications: Linear regression is widely used in various fields such as economics, biology, engineering, and social sciences for prediction and forecasting.

Overall, linear regression provides a simple yet powerful tool for understanding relationships between variables and making predictions based on those relationships.

15. What are the various branches of statistics?

Ans: Statistics is a broad field with several branches, each focusing on different aspects of data analysis and interpretation. Here are the main branches of statistics:

1. Descriptive Statistics: This branch deals with summarizing and describing data through measures such as mean, median, mode, variance, and standard deviation. It often involves graphical representations like histograms and box plots.

2. Inferential Statistics: This branch focuses on making predictions or generalizations about a population based on a sample. It includes hypothesis testing, confidence intervals, and regression analysis.

3. Bayesian Statistics: This approach involves using Bayes' theorem to update the probability of a hypothesis as more evidence or information becomes available. It incorporates prior knowledge along with new data.

4. Non-parametric Statistics: This branch deals with statistical methods that do not assume a specific distribution for the data. It is useful when data do not meet the assumptions required for parametric tests.

5. Multivariate Statistics: This area involves the analysis of more than one variable at a time. Techniques include multivariate regression, factor analysis, and cluster analysis.

6. Experimental Design: This branch focuses on designing experiments to ensure that the data collected can lead to valid conclusions. It involves planning how to manipulate variables and control for confounding factors.

7. Quality Control and Six Sigma: This area applies statistical methods to monitor and improve the quality of processes and products. Techniques include control charts and process capability analysis.

8. Time Series Analysis: This branch deals with data that is collected over time, focusing on identifying trends, seasonal patterns, and forecasting future values.

9. Survival Analysis: This area focuses on time-to-event data, often used in medical research to analyse the time until an event such as death or failure occurs.

10. Spatial Statistics: This branch deals with data that has a geographical or spatial component, Analysetion patterns and relationships based on location.

11. Statistical Machine Learning: This emerging field combines statistics with machine learning techniques to develop models that can learn from data and make predictions.

Each of these branches plays a crucial role in various applications, from scientific research to business analytics, helping to extract meaningful insights from data.