STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False

   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
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
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
5. _____ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
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

Answer of question 2 : A) Central Limit Theorem

Answer of question 3: b) Modelling bounded count data

Answer of question 4: c) The square of a standard normal random variable follows what is called chi-squared distribution

Answer of question 5: c)Poisson

Answer of question 6: b) false

Answer of question 7: d) none of the mentioned

Answer of question 8: a) 0

Answer of question 9: c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer: Normal distribution, also known as the 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. In graphical form, the normal distribution appears as a "bell curve".

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

Answer: Removing entire rows with missing values. This is suitable when missing data is random and does not significantly impact the analysis. However, it may lead to a loss of valuable information.
Imputation Techniques:

Mean/Median/Mode Imputation: Replace missing values with the mean, median, or mode of the observed Values in that variable. This is a simple method but may not be suitable for variables with skewed distributions.
Forward Fill or Backward Fill (for time series data): Fill missing values with the previous (forward fill) or next (backward fill) observed value. This is appropriate when missingness has a temporal pattern.
Linear Regression Imputation: Predict missing values based on the relationship with other variables using linear regression. This method is suitable when the variable with missing values has a linear relationship with other variables.
K-Nearest Neighbors (KNN) Imputation: Estimate missing values based on the values of their k-nearest neighbors. This method is effective when observations with similar feature values tend to have similar target variable values.
Multiple Imputation: Generate multiple imputed datasets, perform analysis on each, and combine results. This accounts for uncertainty in imputed values.
Domain-Specific Imputation:
Custom Imputation: Impute missing values based on domain-specific knowledge. For example, imputing missing temperature values based on the season or geographic location.
Imputation by Group Statistics:
Group-wise Imputation: Compute group-specific statistics and use them to impute missing values within each group. This is useful when data can be naturally grouped.

Machine Learning-Based Imputation:

Use machine learning models: Train a machine learning model to predict missing values based on other features in the dataset. This method is effective when relationships between variables are complex.

12. What is A/B testing?

Answer: A/B testing, also known as split testing, is a statistical method used in marketing, product development, and other fields to compare two versions of a product or service. The objective is to determine which version performs better and yields better results, often measured in terms of user engagement, conversion rates, or other key performance indicators (KPIs). The process involves dividing a sample population into two groups: Group A and Group B. Each group is exposed to a different version of a variable, such as a webpage, email, advertisement, or product feature. The variations between the two groups are typically minor and represent changes in design, content, layout, or functionality.

13. Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data's mean is known as mean imputation.Mean imputation is typically considered terrible practice since it ignores feature correlation Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Linear regression is the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables(predictors) using a straight line. There are two types of linear regression - **Simple** and **Multiple**. Linear regression is only dealing with **continuous variables** instead of **Bernoulli variables**. The problem of Linear Regression is that these predictions are not sensible for classification since the true probability must fall between 0 and 1, but it can be larger than 1 or smaller than 0. Noted that classification is not normally distributed which is violated assumption 4: **Normality**. Moreover, both mean and variance depend on the underlying probability. Any factor that affects the probability will change not just the mean but also the variance of the observations, which means the variance is no longer constantly violating the assumption

15. What are the various branches of statistics?

Answer: Statistics is a broad field that encompasses various branches, each focusing on specific aspects of data analysis and interpretation. Some of the key branches of statistics include:

Descriptive Statistics: Descriptive statistics involves methods for summarizing and describing the main features of a dataset. Measures such as mean, median, mode, range, and standard deviation fall under descriptive statistics.

Inferential Statistics: Inferential statistics involves making inferences and predictions about a population based on a sample of data. It includes hypothesis testing, confidence intervals, and regression analysis.

Probability Theory: Probability theory is the foundation of statistics and deals with the likelihood of events occurring. It provides the theoretical basis for statistical methods and models.

Biostatistics: Biostatistics applies statistical methods to biological and medical data. It is used in clinical trials, epidemiology, and other health-related research.

Econometrics: Econometrics applies statistical methods to economic data. It is used to analyze economic relationships, test hypotheses, and make predictions in the field of economics.

Actuarial Science: Actuarial science applies statistical and mathematical methods to assess risk and uncertainty in the fields of insurance, finance, and pension planning.

Social Statistics: Social statistics involves the application of statistical methods to social science research, including sociology, psychology, and political science.

Environmental Statistics: Environmental statistics deals with the analysis of environmental data, including pollution levels, climate data, and ecological studies.

Statistical Computing: Statistical computing involves the development and application of computational methods for statistical analysis. This includes the use of software tools like R, Python, and SAS.
Quality Control and Reliability: Statistical methods are used in quality control to monitor and improve processes. Reliability analysis assesses the reliability of systems and products.
Spatial Statistics: Spatial statistics focuses on the analysis of data with spatial components, such as geographical data. It is used in fields like geography, ecology, and urban planning.
Bayesian Statistics: Bayesian statistics is a branch that applies Bayesian methods to statistical analysis. It involves updating probabilities based on new evidence and is particularly used in decision-making and machine learning.