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
Answer- a) Central Limit Theorem

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
Answer- b) Modeling 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
Answer- c) The square of a standard normal random variable follows what is called chi-squared distribution

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned
Answer- c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False
Answer-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
Answer- b) Hypothesis

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
Answer- a) 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

Answer- 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- The normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about its mean, forming a bell-shaped curve when plotted. In a normal distribution, the mean, median, and mode are all equal, and the curve is characterized by its mean and standard deviation. Many natural phenomena and measurements in fields such as statistics, physics, and social sciences tend to follow a normal distribution.

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

Answer- Handling missing data depends on the nature of the data and the specific analysis being conducted. Here are some common techniques for handling missing data:

1. *Deletion*: Deleting observations with missing data. This can be done list-wise (removing entire observations with missing data) or pair-wise (using available data for each analysis).

2. *Mean/Median/Mode Imputation*: Filling missing values with the mean, median, or mode of the observed data for that variable.

3. *Hot-Deck Imputation*: Replacing missing values with values from similar cases, based on other variables.

4. *Cold-Deck Imputation*: Replacing missing values with values from a different dataset, usually obtained from external sources.

5. *Regression Imputation*: Using regression models to predict missing values based on other variables.

6. *Multiple Imputation*: Generating multiple plausible values for missing data based on the observed data distribution. This involves creating multiple complete datasets with imputed values and combining results.

7. *Machine Learning Techniques*: Utilizing machine learning algorithms, such as k-nearest neighbors (KNN) or decision trees, to predict missing values based on other variables.

The choice of imputation technique depends on factors like the amount and pattern of missing data, the nature of the dataset, and the assumptions underlying the analysis. Multiple imputation is often preferred because it preserves variability and uncertainty in the imputed values. However, simpler methods like mean imputation may be sufficient for exploratory analysis or when data are missing completely at random.

12. What is A/B testing?

Answer- A/B testing, also known as split testing, is a method used to compare two versions of a product or service to determine which one performs better. In an A/B test, two versions, A and B, are compared by randomly assigning users to either version A or B and then measuring their responses to determine which version is more effective.

13. Is mean imputation of missing data acceptable practice?

Answer- Mean imputation of missing data is a common practice and can be acceptable in certain situations, particularly when the missing data are missing completely at random (MCAR) or missing at random (MAR). However, it has some limitations and potential drawbacks:

1. *Loss of Variability*: Mean imputation can underestimate the variability in the data, as it replaces missing values with the same value (the mean). This can lead to underestimation of standard errors and confidence intervals, potentially affecting the validity of statistical inference.

2. *Bias*: Mean imputation can introduce bias if the data are not missing completely at random. For example, if the missing values are related to certain characteristics or outcomes, imputing them with the mean of the observed data may distort the relationships and lead to biased estimates.

3. *Imprecise Estimates*: Mean imputation assumes that missing values are missing randomly, which may not always be the case. In situations where the missingness mechanism is not random, mean imputation can result in imprecise estimates and incorrect conclusions**.**

4. *Assumption Violation*: Mean imputation assumes that the missing data have a normal distribution. If this assumption is violated, imputing missing values with the mean may not accurately represent the underlying distribution of the data.

Despite these limitations, mean imputation can be a quick and simple method for handling missing data, especially when the missingness is minimal and the data are MCAR or MAR. However, it's essential to consider the potential implications and limitations of mean imputation and to explore alternative methods, such as multiple imputation, especially in more complex or sensitive analyses.

14. What is linear regression in statistics?
Answer- Linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as Y) and one or more independent variables (often denoted as $X_1, X_2, \ldots, X_p$). The relationship is modeled as a linear equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

where:
- Y is the dependent variable (the variable we want to predict).
- $X_1, X_2, \ldots, X_p$ are the independent variables (predictor variables).
- $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients (parameters) that represent the intercept and slopes of the linear equation.
- ϵ is the error term, representing the variability in Y that is not explained by the linear relationship with the independent variables.

The goal of linear regression is to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ that best fit the observed data, such that the linear equation provides the best prediction of the dependent variable Y. This is typically done by minimizing the sum of squared differences between the observed values of Y and the values predicted by the linear equation.

Linear regression can be used for various purposes, such as predicting future outcomes, understanding the relationship between variables, and testing hypotheses about the relationship between variables. It is a fundamental and widely used statistical technique in both research and practical applications.

15. What are the various branches of statistics?
Answer- Statistics is a broad field with various branches, each focusing on different aspects of data analysis and interpretation. Some of the main branches of statistics include:

1. *Descriptive Statistics*: Descriptive statistics involve methods for summarizing and describing the features of a dataset, such as measures of central tendency (mean, median, mode), measures of variability (standard deviation, variance), and graphical representations (histograms, box plots, etc.).

2. *Inferential Statistics*: Inferential statistics involves making inferences or predictions about populations based on sample data. This includes hypothesis testing, confidence intervals, and regression analysis.

3. *Probability Theory*: Probability theory is the foundation of statistics and deals with the study of random phenomena and uncertainty. It includes concepts such as probability distributions, random variables, and stochastic processes.

4. *Biostatistics*: Biostatistics applies statistical methods to biological and health-related data. It is used in fields such as epidemiology, public health, genetics, and clinical trials to analyze and interpret data related to diseases, treatments, and health outcomes.

5. *Econometrics*: Econometrics applies statistical methods to economic data to analyze and model economic relationships. It is used in areas such as finance, macroeconomics, microeconomics, and economic forecasting.

6. *Multivariate Statistics*: Multivariate statistics deals with the analysis of datasets with multiple variables. It includes techniques such as multivariate regression, principal component analysis, factor analysis, and cluster analysis.

7. *Bayesian Statistics*: Bayesian statistics is a framework for statistical inference that incorporates prior knowledge or beliefs about a phenomenon. It involves updating beliefs based on new evidence and is used in various fields, including machine learning, genetics, and decision making.

8. *Spatial Statistics*: Spatial statistics focuses on the analysis of data with spatial or geographic components. It includes techniques for spatial autocorrelation, spatial interpolation, and spatial regression, and is used in fields such as geography, environmental science, and urban planning.

These are just a few examples of the branches of statistics, and the field continues to evolve with advances in data science, machine learning, and interdisciplinary research.