STATISTICS WORKSHEET

MCQs.

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: d) All of the mentioned

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

Answer: c) Poisson

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

Answer: b) False

7. 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. 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

Q10 to Q15: Subjective Answer Type Questions

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

A normal distribution is a type of continuous probability distribution for a real-valued random variable. It is characterized by a symmetric, bell-shaped curve where most of the observations cluster around the central peak. The mean, median, and mode of a normal distribution are all equal and located at the center of the distribution. This distribution is important in statistics because many real-world phenomena tend to follow this pattern, making it useful for various analytical methods.

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

Handling missing data requires careful consideration to avoid biasing the results. Some common strategies include:

Deletion: Removing rows or columns with missing values, which is effective if the amount of missing data is small.
Mean/Median/Mode Imputation: Filling missing values with the mean, median, or mode of the respective column. This is simple but can distort the data's variance.
K-Nearest Neighbors (KNN) Imputation: Estimating missing values based on the nearest neighbors' values in the dataset.
Multiple Imputation: Creating multiple datasets with different imputed values and then combining the results for more robust analysis.
Predictive Modeling: Using machine learning algorithms to predict and fill in missing values based on other available data.

12. What is A/B testing?

A/B testing is a method of comparing two versions of something to determine which one performs better. It involves dividing a sample into two groups: the control group (A) and the treatment group (B). Each group is exposed to a different version of a variable (e.g., a webpage, product feature). By measuring the performance of both groups, one can assess the impact of changes and decide which version yields better results. This method is commonly used in marketing, product development, and UX design.

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

Mean imputation, which involves replacing missing values with the mean of the observed data, is a simple and commonly used technique. However, it has limitations. While it preserves the mean of the data, it can underestimate the variance and distort the relationships between variables. Therefore, it is generally not recommended for datasets with a significant amount of missing data or when the missing values are not randomly distributed. More sophisticated methods like multiple imputation or model-based approaches are often preferred.

14. What is linear regression in statistics?

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. The model assumes that this relationship can be expressed as a linear equation. The goal of linear regression is to find the best-fit line through the data points that minimizes the sum of the squared differences between the observed values and the values predicted by the model. It is widely used for predictive analysis and to understand the strength and nature of relationships between variables.

15. What are the various branches of statistics?

Statistics can be broadly divided into several branches, each focusing on different aspects of data analysis:

- Descriptive Statistics: Summarizes and describes the features of a dataset using measures like mean, median, mode, and standard deviation.

- Inferential Statistics: Makes inferences and predictions about a population based on a sample of data through hypothesis testing, confidence intervals, and regression analysis.

- Probability Theory: Studies the likelihood of different outcomes and forms the mathematical foundation of statistics.

- Bayesian Statistics: Uses Bayes' theorem to update the probability of a hypothesis based on new evidence.

- Biostatistics: Applies statistical methods to biological and health-related research.

- Econometrics: Uses statistical methods to analyze economic data and test economic theories.

- Quality Control: Applies statistical techniques to monitor and improve manufacturing and production processes.

- Experimental Design: Involves planning experiments to ensure that data collection is structured and unbiased, allowing for valid and reliable conclusions.