statistics worksheet 01 assignment by Mohsin Hilal Qadri

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

a) True

b) False

My 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

My 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

My 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

My 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

My answer= c) Poisson

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

a) True

b) False

My 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

My 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

My 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

My answer= c) Outliers cannot conform to the regression relationship

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

My answer=

The normal distribution, is a bell-shaped curve that is symmetric around its mean. It's defined by its mean (μ) and standard deviation (σ), and about 68%, 95%, and 99.7% of the data falls within one, two, and three standard deviations from the mean, respectively. It's widely used in statistics to model many natural phenomena and measurement processes due to its simplicity and prevalence.

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

My answer=

In statistics, handling missing data is critical to ensure accurate analysis and interpretation of results. Here are some commonly recommended imputation techniques:

1. **Mean/Median/Mode Imputation**:
    - **Mean Imputation**: Replace missing values in numerical data with the mean of the non-missing values of that variable.
    - **Median Imputation**: Replace missing values with the median, which is less sensitive to outliers compared to the mean.
    - **Mode Imputation**: Replace missing categorical data with the most frequent category (mode) of that variable.
2. **Forward Fill or Backward Fill (for time series data)**:
    - **Forward Fill**: Use the last observed value to fill missing values.
    - **Backward Fill**: Use the next observed value to fill missing values.
3. **Hot Deck Imputation**:
    - Replace missing values with values from similar cases based on a similarity measure such as Euclidean distance or Mahalanobis distance.
4. **Regression Imputation**:
    - Predict missing values using a regression model fitted on other variables that are not missing.
5. **Multiple Imputation**:
    - Generate multiple plausible values for each missing data point, based on the distribution of the observed data and the uncertainty about the missing values. Analyze each dataset and combine results using specialized methods.
6. **K-Nearest Neighbors (KNN) Imputation**:
    - Predict missing values based on the values of its nearest neighbors in the feature space.
7. **Expectation-Maximization (EM) Algorithm**:
    - Iterative method to estimate parameters of a statistical model when there are missing data.

The choice of imputation technique depends on various factors including the type of data (numerical or categorical), the distribution of missing values, the analysis context, and assumptions about the missing data mechanism (e.g., Missing Completely at Random, Missing at Random, or Missing Not at Random).

If I am asked, I will recommend **Multiple Imputation technique**

12. What is A/B testing?

My answer=

In statistical terms, A/B testing involves experimenting with two variations (A and B) to understand which one delivers better results. It uses random assignment of participants to each variation and statistical methods to analyze which version achieves superior outcomes based on measurable metrics. This method guides decision-making by providing empirical evidence of effectiveness between the compared variations.

13. Is mean imputation of missing data acceptable practice?

My answer=

Mean imputation is commonly used but may not always be the best approach. It fills missing values with the average of existing data, which is easy, but it can skew results if the missing data isn't random or if there are extreme values. Depending on the situation, other methods might be more accurate.

14. What is linear regression in statistics?

My answer=

Linear regression in statistics is a method used to explore relationships between variables. It assumes a linear (straight-line) relationship between an independent variable (the predictor) and a dependent variable (the outcome). The goal is to find the best-fitting line that summarizes the relationship between these variables.

In simple terms, linear regression helps us understand how changes in one variable are associated with changes in another. For example, it can help predict sales based on advertising spending or understand how temperature affects crop yield. By fitting a line to the data points, linear regression allows us to make predictions and draw insights about these relationships.

15. What are the various branches of statistics?

My answer=

Statistics branches into several key areas, include:

1. **Descriptive Statistics**: Involves summarizing and describing data using measures like mean, median, and standard deviation.
2. **Inferential Statistics**: Uses data from a sample to make conclusions or predictions about a larger population.
3. **Business Statistics**: Applies statistical methods to business and management data for decision-making and strategy.
4. **Social Statistics**: Analyzes social phenomena using statistical methods, such as surveys and census data analysis.
5. **Machine Learning and Data Mining**: Uses statistical techniques to develop algorithms and models for learning patterns and making predictions from large datasets.