Stats Assignment Fliprobo internship assignment 3

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

a) True

b) False

Answer : a

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

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

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

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer : C

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

a) True

b) False

Answer : B

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

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

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

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

Answer : Normal distribution refers to a continuous probability distribution that is symmetrical around its mean.

Shape: The distribution has a bell-shaped curve, with the highest point at the mean, and symmetrically tapering off on either side.
Parameters: It is characterized by two parameters: the mean which determines the center of the distribution, and the standard deviation, which measures the spread or dispersion of the distribution.
Probability Density Function (PDF): The probability of observing a value within a certain range is given by the area under the curve of the normal distribution's PDF.
Properties: Normal distributions are characterized by their smooth, symmetric, and unimodal shape. Many natural phenomena and measurements in fields such as physics, biology, economics, and social sciences approximate a normal distribution due to the central limit theorem, which states that the sum or average of many independent and identically distributed random variables tends to follow a normal distribution.

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

Answer : Handling missing data is a crucial step in data analysis to ensure that the results are reliable and unbiased. Here are some common approaches to handle missing data:

1. Identify and Understand Missing Data: First, identify the patterns and reasons for missing data. It could be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Understanding this can guide appropriate handling strategies.
2. Delete Missing Data: One straightforward approach is to delete rows or columns with missing data (listwise deletion or pairwise deletion). This approach is simple but may lead to loss of valuable information and potentially biased results if the missing data are not random.
3. Imputation Techniques: Imputation involves replacing missing values with estimated values. Some common imputation techniques include:

   Mean/Median/Mode Imputation: Replace missing values with the mean, median, or mode of the observed data for that variable. This is simple but may not preserve relationships between variables.

    - Regression Imputation: Predict missing values using a regression model based on other variables that are not missing.
    - Multiple Imputation: Generate multiple plausible values for each missing data point, based on uncertainty about the true value. This accounts for variability and provides more accurate estimates.
    - K-nearest Neighbors (KNN) Imputation: Replace missing values based on values of nearest neighbors in the feature space.
    - Expectation Maximization (EM) Algorithm: Estimate missing values using iterative algorithms that maximize the likelihood of observed data.

4. Domain-Specific Knowledge: Use domain-specific knowledge to inform imputation. For example, for time series data, missing values might be imputed based on historical trends or seasonal patterns.
5. Evaluate Impact: After imputation, evaluate the impact on your analysis. Assess how sensitive your results are to the imputation method chosen.

The choice of imputation technique depends on the nature of the missing data, the amount of missingness, the distribution of the data, and the specific goals of the analysis. It's often recommended to compare results from multiple imputation methods to assess robustness.

12. What is A/B testing?

Answer: A/B testing is a method used to compare two different versions of something (like a webpage, advertisement, or product feature) to determine which one leads to better outcomes. Here's how it typically works:

- Experimental Setup: Two variants, often labeled A and B, are created.
    - Control Group (A): Receives the current standard or existing version.
    - Treatment Group (B): Receives a modified version with changes (such as design, content, or layout).
- Randomization: Participants or subjects are randomly assigned to either the control or treatment group to ensure that any differences in outcomes are due to the changes made and not to differences in the characteristics of the groups.

Outcome Measurement: Statistical metrics, such as conversion rates, average order value, or click-through rates, are measured for each group to assess how each variant performs.
Statistical Analysis: Statistical methods are applied to determine if the differences observed between the groups are statistically significant. This helps in understanding whether the changes made in the treatment group have a meaningful impact compared to the control group.

13. Is mean imputation of missing data acceptable practice?

Answer : Mean imputation can be a useful initial step in handling missing data, especially under the assumption of MCAR and in situations where simplicity and speed are priorities. However, researchers should be aware of its limitations and consider more advanced imputation methods depending on the nature of the data and the goals of the analysis. It's often beneficial to compare results from different imputation techniques to assess robustness and ensure the reliability of conclusions drawn from the data.

14. What is linear regression in statistics?

Answer : In statistics, linear regression is a fundamental approach used to model the relationship between a dependent variable (often denoted as Y) and one or more independent variables (often denoted as X). The goal of linear regression is to find the best-fitting linear relationship that describes how changes in the independent variables are associated with changes in the dependent variable.

Types of Linear Regression:

Simple Linear Regression: Involves a single independent variable
Multiple Linear Regression: Involves multiple independent variables
Polynomial Regression: Extends linear regression to include polynomial terms of the independent variables
Generalized Linear Models (GLMs): A broader class that includes linear regression and extends to non-normal error distributions or non-linear relationships.

15. What are the various branches of statistics?

Answer : Main branches of statistics include:

1. Descriptive Statistics: Descriptive statistics involve methods for summarizing and describing data sets. This branch includes measures such as mean, median, mode, variance, standard deviation, and graphical representations like histograms, box plots, and scatter plots.
2. Inferential Statistics: Inferential statistics involves making inferences or generalizations about a population based on sample data. This branch includes hypothesis testing, confidence intervals, regression analysis, and analysis of variance (ANOVA).
3. Bayesian statistics: is a framework for statistical inference where probabilities are assigned to hypotheses. It involves updating beliefs or probabilities based on new evidence and prior knowledge
4. ML and Data science : While not traditional branches of statistics, machine learning and data science heavily rely on statistical methods for modeling, prediction, and pattern recognition. Techniques such as supervised learning (e.g., regression, classification) and unsupervised learning (e.g., clustering, dimensionality reduction) are used for data-driven decision-making and knowledge discovery.