STATISTICS WORKSHEET-1

Internship: Mitchelle Khanna | Date:25th July | MCQ & Subjective| Deadline of Submission: 28th July 2024

Note: Answers of each question in Green Highlight

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

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 variablesare 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

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

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

Ans. A normal distribution, also known as a Gaussian distribution, is a continuous probability distribution that is symmetric about its mean, with its shape determined by the mean and standard deviation. It is characterized by its bell-shaped curve where most observations cluster around the central peak and taper off symmetrically towards the tails.

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

Ans. Handling missing data can involve several techniques, including:

- **Mean/Median Imputation:** Replacing missing values with the mean or median of the observed values.
- **Regression Imputation:** Using a regression model to predict and fill in missing values based on other variables.
- **K-Nearest Neighbors (KNN):** Imputing missing values based on similar cases in the dataset.
- **Multiple Imputation:** Creating multiple datasets with different imputed values and combining results to account for uncertainty.
- The choice of method depends on the data structure and the missing data mechanism.

12. What is A/B testing?

Ans. A/B testing, also known as split testing, involves comparing two versions of a variable (A and B) to determine which performs better. It is commonly used in marketing, web design, and product development to optimize outcomes by comparing two different versions of a webpage, email, or advertisement.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation is a simple and common technique, but it can be problematic because it can reduce variability in the data and may lead to biased estimates. It is generally acceptable for small amounts of missing data but might not be ideal for larger datasets or when the data are not missing completely at random.

14. What is linear regression in statistics?

Ans. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It estimates the coefficients of a linear equation that predicts the dependent variable based on the independent variables, aiming to minimize the sum of squared differences between the observed and predicted values.

15. What are the various branches of statistics?

Ans. The main branches of statistics include:

- **Descriptive Statistics:** Methods for summarizing and presenting data.
- **Inferential Statistics:** Techniques for making predictions or inferences about a population based on a sample.
- **Probability Theory:** The mathematical foundation for analyzing random events and uncertainty.
- **Biostatistics:** Application of statistics to biological and medical research.
- **Econometrics:** Use of statistical methods to analyze economic data.
- **Data Science:** An interdisciplinary field that uses statistical techniques along with computational methods to analyze and interpret complex data.