**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**

2. Which of the following theorem states that the distribution of averages of id variables, properly normalized, becomes that of a standard normal as the sample size increases?

**a) Central Limit Theorem**

3. Which of the following is incorrect with respect to use of Poisson distribution?

**c) Modeling contingency tables**

4. Point out the correct statement.

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

5. \_\_\_\_\_\_ random variables are used to model rates.

**c) Poisson**

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

**a) True**

7. Which of the following testing is concerned with making decisions using data?

**b) Hypothesis**

8. Normalized data are centered at\_\_\_\_\_\_and have units equal to standard deviations of the original data. **a) 0**

9. Which of the following statements is incorrect with respect to outliers?

**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?**

A Normal Distribution is one of the types of probability distribution that is symmetric around its mean. It is also called the Gaussian distribution. It depicts that the data near the mean are more frequent in occurrence than data far from the mean. In graphical form the normal distribution appears as a bell curve. Mean, Median and Mode are all equal and located at the center of the distribution. About 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.

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

Handling missing data is crucial in data analysis to avoid biases and to maintain integrity of the dataset. Some recommended imputation techniques are:

1. **Mean/Median/Mode Imputation:** Replace missing values with the mean, median, or mode of the observed data.
2. **Random Sample Imputation:** Fill missing values with randomly selected samples from the dataset, maintaining the original distribution.
3. **Forward/Backward Fill:** Use the value from the previous or next non-missing data point to impute missing values, useful for time series data.
4. **K-Nearest Neighbors (KNN):** Predict missing values based on similar data points using this algorithm, effective for datasets with complex patterns.
5. **Multiple Imputation:** Generate multiple plausible values for each missing data point, incorporating variability due to uncertainty.

**12. What is A/B testing?**

A/B testing is a controlled experiment in statistics where two variants (A & B) are compared to determine which one performs well based on specific metrics. It involves splitting an audience into two groups and showing each group a different version (A and B) of the item being tested. By measuring the responses of each group, typically through metrics like click-through rates, conversion rates, or sales figures, businesses can assess which version produces better results.

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

Mean imputation may be acceptable in cases where missing data is minimal and missingness is completely at random. However, in scenarios where data are missing not at random or for predictive modeling, more advanced imputation techniques such as multiple imputation or machine learning-based methods are recommended.

**14. What is linear regression in statistics?**

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It considers that there is a linear relationship between the variables, meaning that changes in dependent variable are directly proportional to the change in independent variables.

There are different types of Linear Regressions, such as:

1) Simple Linear Regression where only one independent variable is present

2) Multiple Linear regression where more than one independent variable is present.

**15. What are the various branches of statistics?**

There are two branches in statistics:

**Descriptive Statistics**: This branch involves methods of organizing, summarizing, and presenting data in a form that is interpretable and informative. It includes measures of central tendency (mean, median, mode), measures of dispersion (variance, standard deviation), and graphical representation of data.

**Inferential Statistics**: This branch deals with drawing conclusions and making inferences about a population based on sample data. It includes techniques such as hypothesis testing, regression analysis, analysis of variance (ANOVA), and probability distributions.