**STATISTICS WORKSHEET-1**

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 centred 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**

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

**Answer:** The Normal Distribution, also known as the Gaussian distribution, is a bell-shaped probability distribution characterized by its symmetrical, bell-curve shape. In this distribution:

1. The mean, median, and mode are all equal and located at the center of the curve.

2. The curve is symmetric around the mean.

3. The spread or width of the distribution is determined by its standard deviation.

4. About 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and approximately 99.7% within three standard deviations.

It is a fundamental concept in statistics and is often used in various fields to model naturally occurring phenomena due to its simplicity and numerous desirable properties.

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

**Answer:** Handling missing data is crucial to ensure the validity of analyses and models. Here are some common techniques for imputing missing data:

1. Mean/Median/Mode Imputation: Replace missing values with the mean, median, or mode of the available data for that feature.
2. Forward Fill (or Backward Fill): Use the last known value to fill in missing data (or the next known value).
3. Linear Interpolation: Estimate missing values based on the values before and after the missing point using a linear function.
4. Predictive Imputation: Use regression models, k-NN, or other machine learning algorithms to predict and fill in missing values based on other variables.
5. Multiple Imputation: Create multiple imputations and analyze each dataset separately. Combine results to get a single result. This accounts for the uncertainty of the imputation process.
6. Hot Deck Imputation: Replace missing values with values from a similar record based on other variables.
7. Dummy Variable Imputation: Create a dummy variable indicating the presence of a missing value and impute the missing value using mean, median, or mode.

The choice of technique depends on the nature and distribution of the missing data, the underlying process that generated the missingness, and the specific requirements of the analysis or model. Always evaluate the implications of imputation on your data and consider sensitivity analyses to assess the robustness of your results.

12. What is A/B testing?

**Answer:** A/B testing, also known as split testing, is a method of comparing two versions of a webpage, app, or other assets to determine which one performs better in terms of a specific metric or goal.

13. Is mean imputation of missing data acceptable practice?

**Answer:** Mean imputation is a commonly used method to handle missing data, but its acceptability depends on the context and nature of the data:

1. Simplicity: Mean imputation is straightforward and easy to implement.
2. Bias: Mean imputation can introduce bias, especially if the data are not missing completely at random. It assumes that the missing values have the same distribution as the observed values.
3. Loss of Variability: By replacing missing values with the mean, the variance of the dataset can be underestimated.
4. Assumption Violation: If the data are not normally distributed or if outliers are present, mean imputation may not be appropriate.

Alternative methods like multiple imputation or using machine learning algorithms that can handle missing values (e.g., tree-based methods) might be more suitable depending on the specific situation. Always consider the nature of your data and the potential impact of imputation methods on your analyses.

14. What is linear regression in statistics?

**Answer:** Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

15. What are the various branches of statistics?

**Answer:** There are four type of Statistics which are mentioned here:

1. Descriptive Statistics
2. Inferential Statistics
3. Predictive Statistics
4. Prescriptive Statistics