**STATISTICS WORKSHEET-1**

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

**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?

**Answer** a) Central Limit Theorem

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

**Answer** b) Modeling bounded count data

**4.** Point out the correct statement.

**Answer** d) All of the mentioned

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

**Answer** d) All of the mentioned

6. 10 Usually replacing the standard error by its estimated value does change the CLT.

**Answer** b) False

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

**Answer** b) Hypothesis

8. 4. Normalized data are centered at\_\_\_\_\_\_and have units equal to standard deviations of the original data.

**Answer** a) 0

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

**Answer** c) Outliers cannot conform to the regression relationship

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

**Answer** Normal Distribution also called Gaussian distribution is symmetric about the mean. It shows that values near the mean will occur more frequent than the values that are farther away from mean. The Middle of the range is also knows as the mean of the distribution and it is always the peak in the bell curve.

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

**Answer** We have different techniques to handle missing data. Some common imputation techniques are

a). We can delete rows having missing values

b). Mean/Median/Mode imputation – Missing values here are replaced with either Mean/Median or most frequent value of the variable.

c). Regression Imputation – Using linear regression techniques to estimate missing value.

d). K-Nearest Neighbours – KNN imputes missing values by averaging the values of the K nearest neighbours in the feature space.

e). Multiple imputation – It is the most advanced methodology for performing missing data imputation. It generates multiple datasets with imputed values and combine the results. Multiple imputation accounts for uncertainty and provides more accurate estimates.

12. What is A/B testing?

**Answer** Also known as Split testing. It involves comparing two de virtually identical except for one variation to determine which performs better. This testing method evaluates the efficiency of proposed changes again an existing design.

13. Is mean imputation of missing data acceptable practice?

**Answer** Mean imputation is not acceptable practice as it ignores feature correlation and also decrease the variance of our data while increasing bias.

14. What is linear regression in statistics?

**Answer** It is a statistical model that estimates the linear relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting straight line that represents the general trend of a dataset.

15. What are the various branches of statistics?

**Answer** The two major areas of statistics are known as [descriptive statistics](https://www.investopedia.com/terms/d/descriptive_statistics.asp), which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions. Descriptive statistics include mean (average), variance, [skewness](https://www.investopedia.com/terms/s/skewness.asp), and [kurtosis](https://www.investopedia.com/terms/k/kurtosis.asp). Inferential statistics include linear regression analysis, analysis of variance (ANOVA), and null hypothesis testing.