STATISTICS WORKSHEET -1

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

Ans.- 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?

Ans.- central limit theorem

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

Ans.- modelling bounded count data

4. Point out the correct statement.

Ans.- all of the mentioned

5. _____ random variables are used to model rates.

Ans.- poisson

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

ANS.- false

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

Ans.- hypothesis

8. Normalized data are centered at _____ and have units equal to standard deviations of the original data.

Ans.- 0

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

Ans.- outliers cannot conform to the regression relationship

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

Ans.- Normal distribution is also known as a 'Gaussian distribution' or 'probability distribution'. It is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. The normal distribution appears as a "bell curve" because of its flared shape. Normal distribution describes the distribution of values for many natural phenomena in a wide range of areas, including biology, physical science, mathematics, finance and economics. It can also represent these random variables accurately.

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

Ans.- Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you. Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea. Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed value.

STATISTICS WORKSHEET -1

The following are some of the most prevalent methods:

Mean imputation, Hot deck imputation, Cold deck imputation, Regression imputation, stochastic imputation & Single or Multiple imputation.

12. What is A/B testing?

Ans.- A/B testing is also known as split testing or bucket testing. It is a methodology for comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal. Running an A/B test that directly compares a variation against a current experience lets you ask focused questions about changes to your website or app and then collect data about the impact of that change. For example, retail website would run more tests to optimize for purchases, where a B2B website might run more experiments to optimize for lead. A test results dashboard shows 2 (or more) variants, their respective audience and it's goal completions. Say you optimize for clicks on a call-to-action (CTA) on a website, a typical view would contain visitors and clicks, as well as a conversion rate — the percentage of visitors that resulted in a conversion.

13. Is mean imputation of missing data acceptable practice?

Ans.- The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Ans.- In statistics, linear regression is a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. If the explanatory variables are measured with error then errors-in-variables models are required, also known as measurement error models.

15. What are the various branches of statistics?

Ans.- there are two various branches of statistics are Descriptive and Inferential Statistics. Descriptive statistics- 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, and kurtosis. Inferential statistics include linear regression analysis, analysis of variance, logit/Probit models, and null hypothesis testing.

Inferential Statistics- Inferential statistics is a tool that statisticians use to draw conclusions about the characteristics of a population, drawn from the characteristics of a sample, and to

STATISTICS WORKSHEET -1

determine how certain they can be of the reliability of those conclusions. Based on the sample size and distribution, statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.