STATISTICS WORKSHEET-1

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

Ans: 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?

Ans a) Central Limit Theorem

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

Ans: 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

Ans: d) All of the mentioned

5. ______ random variables are used to model rates.

Ans: c) Poisson

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

Ans: b) False

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

Ans: b) Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.

Ans: a) 0

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

Ans: c) Outliers cannot conform to the regression relationship

Ans-10: Normal Distribution:-Also known as Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that 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". In a normal distribution, the mean is 0 and the standard deviation is 1. Normal distribution are symmetrical, but not all symmetrical distribution are normal.

Ans 11: Handling missing values: - Missing data appear when no value is available in one or more variables of an individual. Missing data can occur due to many reasons. The data is collected from various sources and, while mining the data, there is a chance to lose the data. However, most of the time cause for missing data is item nonresponse, which means people are not willing to answer the questions. There are many ways to treat the missing values following are mention below:

-Deleting the missing data.

-Imputation Technique (by using Simple imputer, multiple imputer, stochastic regression imputation, Hot and cold deck imputation)

-k nearest neighbor Technique

Ans 12: A/B Testing- This is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. In hypothesis testing, we have to make two hypotheses i.e. Null hypothesis and the alternative hypothesis.

Null hypothesis (H0): There's no effect in the population.

Alternative hypothesis (Ha or H1): There's an effect in the population. In other words, it's the claim that you expect or hope will be true.

There are two types of errors that may occur in our hypothesis testing-

Type 1 error: We reject the null hypothesis when it is true. That is we accept the variant B when it is not performing better than A

Type 2 error: We failed to reject the null hypothesis when it is false. It means we conclude variant B is not good when it performs better than A.

There are a few key mistakes we Avoid While Conducting A/B Testing-

-Invalid hypothesis.

-Testing too many elements together.

-Not considering the external factor.

Ans 13: Mean imputation of missing data is not acceptable because Mean imputation does not preserve the relationships among variables. It leads to an underestimate of Standard Errors you're making Type 1 errors without realizing it.

Biased estimates of variances and covariance.

In high dimensions, mean substitution cannot account for dependence structure among features.

Ans 14: Linear Regression- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + m*x, where y = estimated dependent variable score, c = constant, m = regression coefficient, and x = score on the independent variable. There are two kinds of Linear Regression Models:-

Simple Linear Regression: A linear regression model with one independent and one dependent variable.

Multiple Linear Regression: A linear regression model with more than one independent variable and one dependent variable.

Ans 15: Branches of Statistics: Descriptive statistics and inferential statistics are the two main branches of statistics.

*Descriptive Statistics- The first aspect of statistics is descriptive statistics, which deals with the presentation and collection of data. Generally, descriptive statistics can be categorized into

-Measures of Central Tendency: Are used by statisticians to examine the value distribution center. These are the measures of tendency: Mean, Median, Mode

-Measures of Variability: The measure of variability helps the statisticians analyze the distribution from a particular data set. Quartiles, ranges, variances, and standard deviations are the variability variables.

*Inference statistics- Are statistical techniques that allow statisticians to utilize data from a sample to conclude, predict the behavior of a given population, and make judgments or decisions. There are some different types of inferential statistics, which include the following, which are: Regression analysis, Analysis of variance (ANOVA), Analysis of covariance (ANCOVA), Statistical significance (t-test), and Correlation analysis.