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 centered 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 term Normal Distribution is a probability distribution which is symmetric about the mean and also known as Gaussian distribution. It shows the data close to the mean are more frequent to occur as comparison to the data which is far from the mean. It appear as a bell curve in graphical form and with most of the values cluster around the central region & also tapering off as they go far from the centre.

Some of the properties of Normal Distribution are:

- Mean, mode, & median are same.
- About the mean, distribution is symmetric.
- Distribution is referred by 2 values i.e. standard deviation & mean.

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

Answer. Missing data is a big problem for analysis of data as it distorts findings. According to data scientist, there are 3 types of missing data i.e MCAR, MAR, & NMAR.

MCAR- when data is completely missing at random across the dataset with no discernable pattern.

MAR- when data is not missing randomly, but only within sub-samples of data

NMAR- when there is a noticeable trend in the way data is missing.

The best way to handle missing data is to use deletion technique & use regression technique to remove missing data.

- The deletion technique can be used for some datasets where we have missing fields.
- By using regression technique, we can predict the null value with the help of other info from the dataset.

12.)What is A/B testing?

Answer.) A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

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

Answer. 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 and also 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?

Answer.) To predict a dependent variable value (y) based on a given independent variable (x) we perform Linear regression. This regression technique finds out a linear relationship between x (input) and y(output). The relationship between two variables by fitting a linear equation to observed data on which only a variable is considered to be an explanatory variable and the rest is the dependent one.

15.)What are the various branches of statistics?

Answer.) The various branches of static are :-

a) Descriptive Analytics -> The interpretation of historical data to better understand changes that have occurred in a business is called Descriptive Analytics.

b)Predictive Analytics ->The advanced analytics that makes predictions about the future outcomes by using historical data combined with statistical modeling, machine learning, and data mining.

c) Prescriptive Analytics-> The data analytics that uses algorithms and analysis of raw data so that the output will be better and effective decisions for a long and also for short span of time.