STATISTICS WORKSHEET-1

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

3. Which of the following is incorrect with respect to use of Poisson distribution?
   b) Modeling bounded count data

4. Point out the correct statement. (Correct Ans: d)
   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

5. _____random variables are used to model rates.
   c) Poisson

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

7. Which of the following testing is concerned with making decisions using data?
   b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0

9. Which of the following statement is incorrect with respect to outliers?
   c) Outliers cannot conform to the regression relationship

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

Ans) A probability function that specifies how the values of a variable are distributed is called the normal distribution. It is symmetric since most of the observations assemble around the central peak of the curve. When data are normally distributed, plotting them on a graph results a bell-shaped and symmetrical image often called the bell curve. The mean, median and mode of a normal distribution are equal

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

Ans) I will handle the data based on the reason why data goes missing. Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

a) Missing at Random (MAR)
b) Missing Completely at Random (MCAR)
c) Missing not at Random (MNAR)

As I'm just as beginner I'm learning the techniques and the best imputation techniques known are:

a) Mean or Median Imputation :- When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations.
b) Regression Imputation
c) Pred. Mean Matching.
d) Hot Deck Imputation
e) Listwise Deletion

Note: Listwise deletion is technically not an imputation method. However, since the method is quite often used in practice, I included it to this comparison.

12. What is A/B testing?

Ans) A/B tests, also known as split tests, we can learn which version of something is more effective by comparing 2 of them.

In research, the concept is similar to the scientific method. If we want to see what happens when we change one thing, we must create an environment where just that one thing changes

A/B testing is a basic randomized control experiment

13. Is mean imputation of missing data acceptable practice?

Ans) Imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. Outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model and hence gets ruled out. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?

Ans) Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables: One variable, denoted x, is regarded as the predictor, explanatory, or independent variable. The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

Regression equation of Y on X
Y=a+bX
Where −
Y = Dependent variable
X = Independent variable
a = Constant showing Y-intercept
b = Constant showing slope of line

15. What are the various branches of statistics?

Ans) There are two main branches of statistics:-

1) Descriptive Statistics : It deals with the presentation and collection of data . This is generally the first segment of statistical analysis . It focuses on collecting , summarizing and presenting a set of data. It mainly works on : Central tendency (Mean , Median and Mode) and Dispersion of data (Range, Variance , Standard Deviance ,Percentile, Skewness)
2) Inferential Statistics: as the name proposes, includes making the right determinations from the statistical analysis that has been performed utilizing descriptive statistics. The branch of statistics that analyzes sample data to draw conclusions about a population. It work on Z-score , Hypothesis testing(t-test , Chi-square test, z-test , ANOVA test etc.,)