ASSIGNMENT - 1

STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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: - c) Poisson

6. **Usually replacing the standard error by its estimated value does change the CLT**
   Answer: - b) False

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

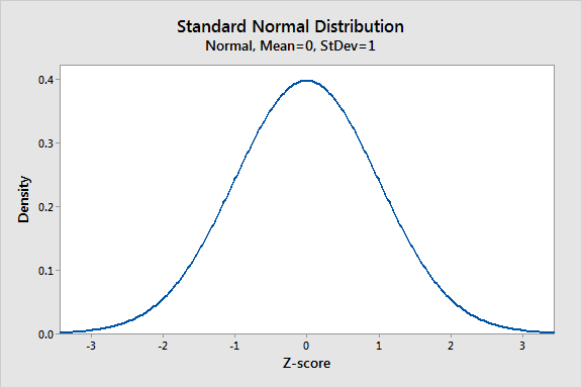
8. **Normalized data are centered at and have units equal to standard deviations of the original data.**
   Answer: - a) 0

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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer: - Normal Distribution in statistics can be defined as the probability distribution of the data which is symmetric about the mean that is showing the data near mean as compared to data far from the mean. This distribution is also known as Gaussian distribution and appear as bell curve.



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

Answer: - Missing data can be defined as not available values like missing sequence, incomplete features, files missing, information incomplete, and data entry error.

The following are imputation techniques are recommended while handling missing data

1. Dropping rows with null values
2. Mean or median
3. Deletion(Listwise, Pairwise, Dropping Variables)
4. Multiple Imputation(Imputation, Analysis, Pooling)
5. KNN (K Nearest Neighbors)

12. What is A/B testing?

Answer: - A/B testing can be defined as the controlled experiment to compare the two version of a variable to find out which performs better in a controlled environment. By taking an example let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods.

A/B testing is one of the most prominent and widely used statistical tools.

13. Is mean imputation of missing data acceptable practice?

Answer: - Mean Imputation of missing data can't be acceptable in practice as it has major drawbacks which are as follows:-

1. Mean imputation does not preserve the relationships among variables.
2. Mean Imputation Leads to an Underestimate of Standard Errors.

Although it has drawbacks but it is widely used method mainly because it's easy, also imputing the mean preserves the mean of the observed data and if the data are missing completely at random, the estimate of the mean remains unbiased.

14. What is linear regression in statistics?

Answer: - Linear Regression in statistics can be defined as a method to quantifies the relationship between one or more predictor variable(s) and one outcome variable. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

For example: - a modeler might want to relate the weights of individuals to their heights using a linear regression model.

$$\mathbf{Y = a + bX}$$

A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0)

15. What are the various branches of statistics?
Answer: - The two main branches of statistics are:-

1. Descriptive statistics
2. Inferential statistics

1. **<u>Descriptive Statistics</u>: -** Descriptive statistics is the first part of statistics that deals with the collection of data. Descriptive statistics have two parts:

   - Central tendency measures(Mean, Median, Mode)
   - Variability measures (The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set)

2. **Inferential Statistics**: - The inference statistics are techniques which is used to enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

   The following are the different inferential statistics:-

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis