STATISTICS – WORKSHEET-1

Q1 to Q9

1. a) True

2. a) Central Limit Theorem

3. b) Modeling bounded count data

4. d) All of the mentioned

5. c) Poisson

6. b) False

7. b) Hypothesis

8. a) 0

9. c) Outliers cannot conform to the regression relationship

Q10 to Q15

10. The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions.

11. Best techniques to handle missing data:

a. Using deletion methods to eliminate missing data

It only works for certain datasets where participants have missing fields. Two common deleting methods include List-wise Deletion and Pair-wise Deletion. List-wise means deleting any participants or data entries with missing values. Pair-wise deletion is the process of eliminating information when a particular data point, vital for testing, is missing.

b. Using regression analysis to systematically eliminate data

Regression is useful for handling missing data as it can be used to predict the null value using other information from the dataset

c. Using data imputation techniques

Two data imputation techniques to handle missing data are average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. Common-point imputation is when we utilise the middle point or the most commonly chosen value.

Techniques used In Imputation…

1. Complete Case Analysis (CCA): It directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing. This method is also known as "Listwise deletion".

2. Arbitrary Value Imputation: This imputation can handle both the Numerical and Categorical variables. It states that we group the missing values in a column and assign them to a new value that is far away from the range of that column. Mostly we use values like 99999999 or -9999999 or "Missing" or "Not defined" for numerical & categorical variables.

3. Frequent Category Imputation: This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as Mode Imputation.

12. It is a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

13. Yes, imputing the mean preserves the mean of the observed data. If the data are missing completely at random, the estimate of the mean remains unbiased. And also by imputing the mean, you are able to keep your sample size up to the full sample size. This is the logic involved in mean imputation.

14. In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

15. Statistics may be divided into two main branches: - **(1)** Descriptive Statistics **(2)** Inferential Statistics

(1) Descriptive Statistics
Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

(2) Inferential Statistics
Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.