STATISTICS WORKSHEET-1

Answers

1. b) 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
10. Normal Distribution is a probability distribution of the continuous datasets which is symmetric around the mean of data. Normal Distributions are defined by two parameters mean and Standard Deviation, we can see the mean, median and mode of normal distributions will be the same. If the graph of Normal distribution is drawn, it will be appearing in a bell curve. The normal distributions are denser at the Centre part and less dense at the tail ends. 65% of data falls under first standard deviation and around 95% of data falls under 2^{nd} standard deviation of mean. Most real-life examples fall under Normal Distributions.

    Ex: If we want to calculate the height of class students, we observe that very few students with extreme heights(i.e., too height, too short) but a very large number of students height falls under an average height. Hence, we can visualize that the tail ends of the curve are less dense, but the central part is denser.

11. Missing data occurs when no data value is stored for the variable in observations. There are a few ways to handle the missing data. They are as follows:

    a. Checking with data collection source(people who collected data) if they can find the value.
    b. Dropping the missing values. If the missing values are less, then dropping the particular data entry can be implemented. If there are many missing data, we can go with dropping the variable.

Using *DataFrame.dropna(subset=None, axis=0, inplace=False)* we can drop the missing values in datasets.

c. Replacing the missing values by statistical functions (mean or median or mode) or any functions based on different scenarios. Replacing the missing data is better since no data will be wasted

Using *DataFrame.replace(missing values, new_value, inplace=False)* we can replace the missing value with calculated data.

d. Imputation with an additional column.

Imputation is method of fill the missing values with some numbers using specific strategy. There are few imputers

- Simple Imputer
- KNN Imputer
- Iterative Imputer

12. A/B Testing is a randomized control experiment. It is a way to compare 2 variables and find out which variable performs better in controlled environment. It is a popular method to test the products which is gaining popularity in Data Science. It can be illustrated by following example: Suppose a person want to increase the sales of product, so we can go with random experiment or Scientific and statistical methods.
    a) We can divide products into 2 groups i.e., Control group and Test group. The control group remains the same but changes to be brought can me applied on Test group(quality of product, packaging method etc.) and delivered to Random sample of customer.
    b) Now on basis of response from customer groups who used Control and Test products. We can decide better performing product(By applying Statistical Significance Tests).

13. Mean imputation method has been most popular solution being used for handling missing data for it being easy to use, but it may not be the best solution possible as it does not provide relationship between strong parameter estimates. The mean imputation will bias the standard error invalidating the hypothesis testing and calculations of confidence intervals.

14. Linear Regression is a process of predicting dependent variable(target) referred as Y based on independent variables(features) referred as X. When there is only one independent variable it is called simple linear regression. Here the predictions of Y when plotted as function of X will form a straight line. In Linear regression we find the best fit straight line through the given data points. The formula for a regression line is:

$$Y = mx + c$$

m = intercept

c = slope

Y = predicted score

The target variable can be predicted if we know the values of coefficients. These are estimated using 'least square criterion' i.e., best fit line is to be calculated that minimize "sum of square residuals(distance between actual and predicted value)".

15. Statistics is all about interpretation of the data. It is the science concerned with developing, studying method of collecting, analyzing, interpreting and presenting the empirical data. There are 2 branches of statistics.
i. Descriptive statistics ii. Inferential Statistic

Descriptive Statistics:
Descriptive statistics are the numbers that are used to summarize / describe the data in ways more meaningful and useful ways. The data collected from experiments, surveys, records all these will have numbers; all these data will help us analyze and interpret the data and provide decisions to be made. It helps in describing visible characteristics of the large datasets.
Ex: If we want to analyze the birth certificates in India. A descriptive statistic would be average Age of mothers, or number of boy or girl baby percentage etc.

Inferential Statistics:
In inferential statistics the sample of data from whole dataset is collected at pure random chance and generalize about the entire population dataset. It focuses on making predictions/generalizations about the larger datasets based on random sample of data. Choosing the sample datasets from population is most crucial stage in making inferences. If the sample data chosen to represent only particular group of population then it may bias the Data and predictions may go wrong. Here are few things we need to keep in mind while selecting random samples.

a. The random sampling of data should be such that it is pure random and represent all the population data set without any bias.
b. The size of sample dataset chosen should be large enough to represent the population dataset so that every member of the population has an equal chance of being selected.
c. The sampling data should follow random assignments.