WORKSHEET 1 – STATISTICS ANSWERS

1. (a) True

2. (a)Central limit theorem

3. (c) Modeling bounded count data

4. (d)All 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 continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

11. There are several ways to handle missing data:

a. Delete rows with missing data

b. Mean/Median/Mode imputation

c. Assigning a unique value

d. Predicting the missing values

e. Using an algorithm which supports missing values, like random forests.

The best method is to delete rows with missing data as it ensures that no bias or variance is added or removed, and ultimately results in a robust and accurate model. However, this is only recommended if there's a lot of data to start with and the percentage of missing values is low.

12. A/B testing is a form of hypothesis testing and two-sample hypothesis testing to compare two versions. The control and variant, of a single variable. It is commonly used to improve and optimize user experience and marketing.

13. Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year- old will appear to have a much higher fitness score that he actually should.

14. Linear regression is one of the statistical techniques used in predictive analysis, in this technique will identify the strength of the impact that the independent variables show on deepened variables.

15. Statistics have two main branches, namely:

a. Descriptive Statistics: This usually summarizes the data from the sample by making use of an index like mean or standard deviation. The methods which are used in the descriptive statistics are displaying, organizing, and describing the data.

b. Inferential Statistics: These conclude from data which are subject to random variations like observation mistakes and other sample variation.