Submitted By: Arjun B
Batch Number: DS0082

WORKSHEET 1

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

10. **Normal distribution** is the continuous probability distribution with a probability density function that gives you a symmetrical bell curve. Simply put, it is a plot of the probability function of a variable that has maximum data concentrated around one point and a few points taper off symmetrically towards two opposite ends. The Normal distribution is also known as Gaussian or Gauss distribution.

11. Missing data can be dealt with in several ways. Listwise deletion is the default when a programme encounters an item missing from a questionnaire or data set. This approach deletes all items which are missing for any given subject, even if they are only missing for one question.

    - Interpolation and extrapolation - An estimated value from other observations from the same individual. It usually only works in longitudinal data.
    - Regression imputation - It is an estimation method that tries to predict the missing values based on other factors known about the data. As a result, instead of utilising the mean value as a substitute for missing values, you're relying on an estimated value. This predicted value can be influenced by other factors known about your data and helps prevents bias in your analysis.
    - Stochastic regression imputation - It is a combination of regression imputation and randomness. This has all the advantages of regression imputation but adds in the advantages of the random component, so you get accurate predictions with less bias.
    - Substitution - Impute the value from a new individual who was not selected to be in the sample.

12. **A/B testing** is a simple experiment in which two variants are compared to determine which performs better. Typically, both variations of the same thing are run simultaneously on different sets of customers to see if there is a significant difference in the metrics like sessions, click-through rate, and/or conversions. Using our visual example above as an example, we could randomly split our customer base into two groups, a control group and a variant group. Then, we can expose our variant group with a red banner website and see if we get a significant increase in conversions.

13. Imputation is not a recommended practice in general. It works by just estimating the mean values. The mean imputation preserves the mean of the observed data, but does not preserve any other characteristics like variance and relationship between variables. This leads to an underestimation of standard deviation by pulling estimates toward zero.

14. **Linear regression** is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Since linear regression shows the linear relationship between a dependent (y) variable and one or more independent (x) variables, hence called as linear regression.

15. **Descriptive statistics** - The distribution, variability, and central tendency of sample data are primarily the focus of descriptive statistics. A sample's or population's central tendency is an estimation of its features, which includes descriptive statistics like mean, median, and mode. The term "variability" refers to a collection of statistics, including measures like range, variance, and standard deviation, that illustrate how much variation there is among the components of a sample or population along the attributes examined.

    Inferential statistics - Using inferential statistics, statisticians may determine how confident they can be in the accuracy of their results and draw inferences about the features of a population from the characteristics of a sample. Statistics experts can determine the likelihood that statistics, which quantify the central tendency, variability, distribution, and relationships between characteristics within a data sample, give an accurate representation of the corresponding parameters of the entire population from which the sample is drawn based on sample size and distribution.