STATISTICS WORKSHEET 1

1. Answer is option a)True.
2. Answer is option d)All of the mentioned.
3. Answer is option b)Modeling bounded count data.
4. Answer is option d)All of the mentioned.
5. Answer is option c)Poisson.
6. Answer is Option b)False.
7. Answer is Option b)Hypothesis.
8. Answer is Option a)0.
9. Answer is option c)Outliers cannot conform to the regression relationship.
10. Normal distribution: The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems. Many things actually are normally distributed, or very close to it. For example: height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution.
11. Missing data can be dealt with in a variety of ways. The most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you. Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea. Another common strategy among those who pay attention is imputation.

    Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values. Imputation techniques recommended are as follows:

- Imputation with constant value : As the title hints — it replaces the missing values with either zero or any constant value.
- Imputation using Statistics : The syntax is the same as imputation with constant only the SimpleImputer strategy will change. It can be “Mean” or “Median” or “Most_Frequent”.

  Mean” will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed.

“Median” will replace missing values using the median in each column. It is preferred if data is numeric and skewed.
“Most_frequent” will replace missing values using the most_frequent in each column. It is preferred if data is a string(object) or numeric.
Before using any strategy, the foremost step is to check the type of data and distribution of features (if numeric).

12. A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B. For instance, 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. In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.
13. The process of replacing null values in a data collection with the data’s mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.
14. Linear regression in statistics: If we want to use a variable x to draw conclusions concerning a variable y: y is called dependent or responsive variable. x is called independent, predictor, or explanatory variable. If the relationship between two variables is linear can be summarized by a

straight line. A straight line can be described by an equation: y=a+bx where a is called the intercept and b the slope of the equation. The slope is the amount by which y increases when x increases by 1 unit.

15. Various branches of statistics.
    - Mathematical or theoretical statistics: It helps in forming the experimental and statistical distribution.
    - Statistical methods or functions: It helps in the collection, tabulation and interpretation of the data. It helps in analysing the data and returns insight from the data.
    - Descriptive statistics : It helps in summarizing and organizing any data set characteristics. It also helps in the representation of data in both classification and diagrammatic way.
    - Inferential statistics : It helps in finding the conclusion regarding the population after analysis on the sample drawn from it.