STATISTICS - 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, also known as the Gaussian distribution, is **a 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) Real-world data is messy and usually holds a lot of missing values. Missing data can skew anything for data scientists and, A data scientist doesn't want to design biased estimates that point to invalid results. Behind, any analysis is only as great as the data. **Missing data appear when no value is available in one or more variables of an individual.** Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.

Imputation techniques:

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem. Imputation techniques can be broadly they can be classified as follows:

Basic Imputation Techniques:

1. Imputation with a constant value
2. Imputation using the statistics (mean, median, mode)
3. K-Nearest Neighbor Imputation.

12) A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. 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.

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 is **an attempt to model the relationship between two variables by fitting a linear equation to observed data, where one variable is considered to be an explanatory variable and the other as a dependent variable**. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15) The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics:

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment. Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence

in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.
Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.