Answers for 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 called Gaussian Distribution is a probability distribution of a continuous random variables.
- These distributions are represented by probability density functions with bell shaped curves.
- Most of the data on normal distribution are concentrated towards the centre of the bell-shaped curve.
- For any normally distributed variables the mean, median and modes are almost same.

- For finding the probability of any random variable, we use interval in which that random variable may lie in terms of probability rather than using point estimates.

11. Missing values can be imputed by following ways:
    - **Mean imputation:** For a normally distributed continuous variables that do not have outliers, simple mean imputation can be used where missing values will be replaced by mean of the distribution.
    - **Median imputation:** For a highly skewed data set containing possible outliers, median imputation can be much more robust than the simple mean imputation. This method may be also be favourable for imputation of missing discrete continuous variables.
    - **Mode imputation:** This method is generally used for imputing missing values in case the variable is categorical in nature. In this method the missing category is replaced by most frequent (mode) of that categorical variable.
    - **Regression imputation:** Provided any two numerical variables are correlated to each other, then a simple linear regression can be fitted as a feature having missing values as target variable and the other feature as the predictor variable. This is a model-based imputation technique where we establish a linear relationship between these two variables to predict the missing values.
    - **KNN imputation:** This imputation technique also uses model-based approach to predict the missing values.

The idea behind this method is to find all the K nearest neighbours/data points to consider similar or relative enough to group them and each missing value in these groups/samples are imputed by simply taking the means of these respective groups. This technique can be used for imputing missing values in both numerical and categorical variables.

The recommended way to impute the missing values is to first analyse the distribution of the given variable that contains missing values, which further gives us the understanding of what imputation method to use, for instance, if the distribution is not normal and contains multiple peaks, then the missing values closer to those peaks should be imputed with mean with respect to those peaks.
Also, by analysing a single continuous distribution with respect to categorical variable like gender (for example), may give different distribution with respect to male and female category, that can be used to impute missing values with either mean or median independently for these categories by using groupby methods.
And last but not least, it is recommended to use KNN imputation, as this method is robust to data distribution and generally effective for most of the data.

12. A/B testing is application of the statistical analysis and method of hypothesis testing between two variants of the same variable. A/B testing involves in conducting experiments simultaneously where we compare the performance of one or more variants on the test group,

and by doing so we can statistically find out which variant performs the best for decision making process.

13. Mean imputation of the missing data is only acceptable when the distribution of the variable/feature is normally distributed about its mean. On the other hand, if data is not normally distributed, then it is better to first understand the given data distribution and replace the missing data accordingly or we can simply use KNN imputer for that matter.

14. Linear regression is one of the basic statistical models to establish a linear relationship between the target/output variable and the predictor/input variables. If there are only one predictor variable then it is called simple linear regression whereas, for more than one it is called multiple linear regression. The main idea behind linear regression models is to find the best fit line for the given data by minimizing the sum of squared errors between the actual and predicted output.

15. Statistics is mainly divided into two branches and they are:

Descriptive Statistics: This branch of statistics helps to describe and summarize the data in meaningful way. The approach for descriptive statistics is to find the central tendencies of the data like mean, median or mode and even the spread of the data with the help of standard deviations, variance, range or quantiles of the data distribution.

Inferential Statistics: This branch of statistics uses the various statistical analysis and methods like hypothesis testing to infer about the population estimates using the sample and draw useful conclusions about the population.