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 refers to a continuous and symmetric probability distribution that is characterised by a bell-shaped curve. It is a fundamental concept in statistics and is most widely used to model natural phenomena and random variables. In a normal distribution, the data is symmetrically distributed around the mean (μ) of the distribution. The mean, median and mode are all equal and is the central value of the distribution. The standard deviation (σ) of the distribution measures the spread of the distribution. The probability distribution function of a normal distribution is a bell-shaped curve, symmetric around the mean, while gradually tailing off on both sides. The curve depends on the mean and standard deviation. Higher standard deviation produces wider and flatter curves. Normal distribution is widely popular because of its prevalence in nature. Real-world phenomena such as height, intelligence, errors in measurement etc. all tend to follow normal distribution. Normal distribution can be used in statistical inference, hypothesis testing, probability calculation and data modeling.

11. Missing data can be handled in various ways: The most common approaches include:
    - Deletion: In this approach, rows or columns with missing values are removed from the dataset. This approach is appropriate only when the missing data is minimal and does not affect the analysis significantly. It can, however, lead to loss of information
    - Imputation: In this approach, missing values are replaced with another value, most likely the mean, median or mode of the available values for that attribute. This is known as mean imputation, median imputation and mode imputation based on the value used for imputation. This method is quick and simple. There are other imputation techniques such as Hot deck and Cold deck imputation, Multiple imputation etc.

The most widely used imputation techniques are as follows:

- Mean Imputation: In this method, the mean of the observed values for the attribute is computed and is used to replace the missing values. The advantage is that it maintains the same mean and sample size. Mean imputation should be considered for symmetric data distribution.
- Median Imputation: It is similar to mean imputation, but instead of the mean, the median of the observed values for that attribute is computed and is used instead of the missing values. It should be considered when the data is skewed and numeric.
- Mode Imputation: Mode imputation is similar to mean and median imputations, but in this case the missing values are replaced with the mode of all the observed values for the attribute. It is a good technique to consider when the distribution is skewed and can be done with both numeric and categorical data.
- Hot deck Imputation: In this approach, missing values are replaced using values from similar individuals in the dataset. It involves finding cases comparable in other attributes to the missing value individual and using their observed values to impute the missing value. Hot deck imputation randomly selects the value from the list of comparable cases and uses the randomly selected value for imputation. This method preserves the correlation between features.
- Cold deck Imputation: It is similar to hot deck imputation, except it does not randomly select the imputed value from a list of comparable cases. It selects the most comparable individual and uses that value to replace the missing values.
- Multiple Imputation: It is a more advanced technique and involves creating multiple imputed datasets based on statistical models. Each dataset includes imputations for the missing values, and are used to perform analyses. The results of the analyses are then combined to provide more robust estimation. This method accounts for the errors associated with imputation of missing values.
- Regression Imputation: This technique involves using regression on the observed values to predict the missing value using the other attributes as inputs. This technique learns from the observed values and can capture relationships in the data.

12. A/B testing is a fundamental randomised control experiment. It is a method for comparing two variants of a process to determine which one performs better in a controlled environment. The basic idea behind A/B testing, is to randomly divide the population into two or more groups – a control group and the experimental group(s). The control group represents the existing process, while the experimental group(s) are provided with a modified version of the process. The responses of the different groups are compared to determine if there are statistically significant differences between them. A/B testing helps in decision-making by providing empirical evidence. It enables data-driven decision making, optimisation and improvement by comparing different versions and measuring their effects.

13. Mean imputation is a method where missing values are replaced with the mean of the observed values. It is simple and the most commonly used method for handling missing data. However, whether it is acceptable or not depends on the context and nature of the data.
    For instance, mean imputation is suitable for data with a symmetric distribution. It preserves the sample size, is easy to implement and is also computationally efficient.

On the other hand, for an asymmetric distribution, it is unsuitable as the imputed value might be very high or very low depending on the direction of the skew. It ignores feature correlation and decreases the variance of the data while increasing the bias. Hence mean imputation results in a less accurate model.

14. Linear regression is a statistical modeling technique used to establish a relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variable and the dependent variables, which means that changes in the independent variables would change the dependent variable proportionally. The goal of linear regression is to fit a straight line that best represents the relationship between the variables. The line is defined by coefficients assigned to each independent variable and an intercept. The process of determining best-fitting line involves estimating the coefficients that minimise the sum of the squared differences between the predicted and the actual values of the dependent variable. This is done using a method called ordinary least squares (OLS). Linear regression is usually used for prediction, where the values of independent variables are provided, and a linear regression model is used to predict the values of the dependent variable. Linear regression has some assumptions: linearity, independence of errors and normality of errors. Violations of these assumptions can affect the accuracy of the model. Linear regression is a widely used statistical technique for understanding relationships between variables and predicting values for dependent variables.

15. The major branches of statistics are: Descriptive Statistics and Inferential Statistics.
    - Descriptive Statistics: It involves describing and summarising data using various measures like mean, median, variance, standard deviation. It focuses on collecting, organising, presenting and summarising data to provide insights and patterns.
    - Inferential Statistics: It is concerned with drawing conclusions about a population based on a sample. It involves techniques such as hypothesis testing, regression analysis, and confidence intervals to generalise findings from the sample to the larger population.