STATISTICS WORKSHEET-1 Answers

Ans.1 :- True

Ans.2 :- Central Limit Theorem

Ans.3 :- Modeling bounded count data

Ans.4 :- All of the mentioned

Ans.5 :- Poisson

Ans.6 :- False

Ans.7 :- Hypothesis

Ans.8 :- 0

Ans.9 :- Outliers can conform to the regression relationship

Ans.10 :-

Normal distribution is also called as Gaussian distribution.

It is the most important probability distribution in statistics for independent, random variables.

It 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.

The normal distribution has two parameters, the mean and standard deviation.

The ND is the most widely known and used of all distribution because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.

For example, heights, blood pressure, measurement error and IQ scores follow the normal distribution.

Ans.11 :-

Analyze each column with missing values carefully to understand the reasons behind the missing values as it is crucial to find out the strategy for handling the missing values.

There are 2 primary ways or techniques of handling missing values:

- Deleting the Missing values
- Imputing the Missing values

1) **Deleting the Missing values** :-

   Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.

   If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.

   If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.

   The disadvantage of this method is one might end up deleting some useful data from the dataset.

   There are 2 ways one can delete the missing values:

   Deleting the entire row :-

   If a row has many missing values then you can choose to drop the entire row.

   If every row has some (column) value missing then you might end up deleting the whole data.

   Deleting the entire column :-

   If a certain column has many missing values then you can choose to drop the entire column.

2) **Imputing the Missing values** :-

There are different ways of replacing the missing values. You can use the python libraries Pandas and Sci-kit learn as follows:

Replacing With Arbitrary Value :-

we are replacing the missing values of the 'Dependents' column with '0'.

Replacing With Mean :-

This is the most common method of imputing missing values of numeric columns.

If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first.

Replacing With Mode :-

Mode is the most frequently occurring value. It is used in the case of categorical features.

Replacing With Median :-

Median is the middlemost value. It's better to use the median value for imputation in the case of outliers.

so, there are some recommend technique to handle missing data.

Ans.12 :-

A/B testing is also known as bucket testing or split-run testing.

A/B testing is basically statistical hypothesis testing or statistical inference.

It is an analytical method for making decision that estimate population parameters based on sample statistics.

A/B testing is a basic randomized control experiment with two variants, A and B.

It is a way to compare the two versions of a variable to find out which performance better in a controlled environment.

It's a method of comparing two versions of a webpage or app against each other to determine which one performs better.

A/B testing are useful for understanding user engagement and satisfaction of online features like a new feature or product.

Large social media sites like LinkedIn, Facebook and Instagram use A/B testing to make user experience more successful and as a way to streamline their services.

Ans.13 :-

True, imputing the mean preserves the mean of the observed data.

So if the data are missing completely at random, the estimate of the mean remains unbiased.

By imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.

If all you are doing is estimating and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

Ans.14 :-

Linear regression is one of the simplest supervised learning algorithms. It is used to predict the value of variable based on the value of another variable.

In linear regression, the objective is to predict numerical features like real estate or stock price, temperature, marks in an examination, sales revenue, etc.

The underlying predictor variable and the target variable are continuous in nature.

In case of linear regression, a straight line relationship is 'fitted' between the predictor variables and the target variables, using the statistical concept of least squares method.

In case of simple linear regression, there is only one predictor variable whereas in case of multiple linear regression, multiple predictor variables can be included in the model.

A typical linear regression model can be represented in the form :- $y = \alpha + \beta x$

where 'x' is the predictor variable and 'y' is the target variable.

Typical applications of regression can be seen in

- Typical applications of regression can be seen in :-
- Demand forecasting in retails.
- Sales prediction for managers.
- Price prediction in real estate.
- Weather forecast.
- Skill demand forecast in job market.

Ans.15 :-

Statistics is mainly divided into two branches.

- Descriptive Statistics
- Inferential Statistics

1. **Descriptive Statistics** :-

   Descriptive statistics deals with the collection of data, its presentation in various forms,

   such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

   For example: Industrial statistics, population statistics, trade statistics, etc. Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

   Descriptive statistics involves the organization, summarization and display of data.

2. **Inferential Statistics** :-

   Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.

   For example: Suppose we want to have an idea about the percentage of the illiterate population of our country. We take a sample from the population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion. This study belongs to inferential statistics.

   Inferential Statistics involves using a sample to draw conclusions about a population.