STATISTICS WORKSHEET 1

Answers of Question 1-9

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.

Answers of Question 10 to Question 15

Q.10 What do you understand by the term Normal Distribution?

Ans - Normal Distribution is also called Gaussian Distribution. Normal distribution is the most commonly seen continuous distribution in nature. Just as the binomial distribution, every event is independent from one another. In the normal distribution the mean, median and mode all line up such that the center of the distribution is the mean. Because of this, exactly half of the results fall to either side of the mean. The normal distribution is also identifiable by its Bell shape and may sometimes be referred to as a Bell Curve.

Q. 11 How do you handle missing data? What imputations techniques do you recommend?

Ans- Missing data can occur due to many reasons. The data is collected from various sources and, while mining the data, there is a chance to lose the data. However, most of the time cause for missing data is **item non response,** which means people are not willing(Due to a lack of knowledge about the question) to answer the questions in a survey, and some people unwillingness to react to sensitive questions like age, salary, gender.

Types of Missing data

Before dealing with the missing values, it is necessary to understand the category of missing values. There are 3 major categories of missing values.

Missing Completely at Random (MCAR):A variable is missing completely at random (MCAR)if the missing values on a given variable (Y) don't have a relationship with other variables in a given data set or with the variable (Y) itself. In other words, When data is MCAR, there is no relationship between the data missing and any values, and there is no particular reason for the missing values.

Missing at Random (MAR):MAR occurs when the missingness is not random, but there is a systematic relationship between missing values and other observed data but not the missing data.

Missing Not at Random (MNAR): The final and most difficult situation of missingness. MNAR occurs when the missingness is not random, and there is a systematic relationship

between missing value, observed value, and missing itself. To make sure, if the missingness is in 2 or more variables holding the same pattern, you can sort the data with one variable and visualize it.

Detecting missing data

Detecting missing values numerically: First, detect the percentage of missing values in every column of the dataset will give an idea about the distribution of missing values.

Detecting missing values visually using Missingno library : Missingno is a simple Python library that presents a series of visualizations to recognize the behavior and distribution of missing data inside a pandas data frame. It can be in the form of a barplot, matrix plot, heatmap, or a dendrogram.

To use this library, we require to install and import it.

Treating missing data

After classifying the patterns in missing values, it needs to treated.

Deletion: The Deletion technique deletes the missing values from a dataset. followings are the types of missing data.

List wise deletion: List wise deletion is preferred when there is a Missing Completely at Random case. In List wise deletion entire rows(which hold the missing values) are deleted. It is also known as complete-case analysis as it removes all data that have one or more missing values.

List wise deletion is not preferred if the size of the dataset is small as it removes entire rows if we eliminate rows with missing data then the dataset becomes very short and the machine learning model will not give good outcomes on a small dataset.

Pair wise Deletion: Pair wise Deletion is used if missingness is missing completely at random i.e MCAR. Pair wise deletion is preferred to reduce the loss that happens in List wise deletion. It is also called an available-case analysis as it removes only null observation, not the entire row. All methods in pandas like mean, sum, etc. intrinsically skip missing values.

Dropping complete columns: If a column holds a lot of missing values, say more than 80%, and the feature is not meaningful, that time we can drop the entire column.

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:

Imputation with constant value:

As the title hints — it replaces the missing values with either zero or any constant value.

We will use the SimpleImputer class from sklearn.

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).

Advanced Imputation Technique:

Unlike the previous techniques, Advanced imputation techniques adopt machine learning algorithms to impute the missing values in a dataset. Following are the machine learning algorithms that help to impute missing values.

K Nearest Neighbour Imputation:

The KNN algorithm helps to impute missing data by finding the closest neighbors using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbours.

The fundamental weakness of KNN doesn’t work on categorical features. We need to convert them into numeric using any encoding method. It requires normalizing data as KNN Imputer is a distance-based imputation method and different scales of data generate biased replacements for the missing values.

Conclusion There is no single method to handle missing values. Before applying any methods, it is necessary to understand the type of missing values, then check the data type and skewness of the missing column, and then decide which method is best for a particular problem.

Q.12 What is A/B Testing?

Ans- An AB test is an example of **statistical hypothesis testing**, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

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.

A/B testing is a form of statistical and two-sample hypothesis testing. **Statistical hypothesis testing** is a method in which a sample dataset is compared against the population data. **Two-sample hypothesis testing** is a method in determining whether the differences between the two samples are statistically significant or not.

Q.13 Is mean imputation of missing data acceptable practice?

Ans- There are three problems with using mean-imputed variables in statistical analyses:

- Mean imputation reduces the variance of the imputed variables.
- Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- Mean imputation does not preserve relationships between variables such as correlations.

Conclusion:

Although imputing missing values by using the mean is a popular imputation technique, there are serious problems with mean imputation. The variance of a mean-imputed variable is always biased downward from the variance of the un-imputed variable. This bias affects standard errors, confidence intervals, and other inferential statistics. Experts agree that mean imputation should be avoided when possible.

Q.14 What is Linear Regression in Statistics?

Ans Linear regression strives to show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased.

It is not necessary that here one variable is dependent on others, or one causes the other, but there is some critical relationship between the two variables. In such cases, we use a scatter plot to imply the strength of the relationship between the variables. If there is no relation or linking between the variables, the scatter plot does not indicate any increasing or decreasing pattern. For such cases, the linear regression design is not beneficial to the given data.

Linear Regression Equation:- The measure of the extent of the relationship between two variables is shown by the **correlation coefficient**. The range of this coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data for two variables. A linear regression line equation is written in the form of:
Y = a + bX where X is the independent variable and plotted along the x-axis
Y is the dependent variable and plotted along the y-axis
The slope of the line is b, and a is the intercept (the value of y when x = 0).
Properties of Linear Regression:-For the regression line where the regression parameters b_0 and b_1 are defined, the properties are given as:

- The line reduces the sum of squared differences between observed values and predicted values.
- The regression line passes through the mean of X and Y variable values.
- The regression constant (b_0) is equal to y-intercept the linear regression.
- The regression coefficient (b_1) is the slope of the regression line which is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X).

Q.15 What are the various branches of Statistics?

Ans 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.

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.
While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.