**Statistics– WORKSHEET 4 Answer sheet**

**Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.**

**Answer 1.** a) True

**Answer 2.** a) Central Limit Theorem

**Answer 3.** b) Modeling bounded count data

**Answer 4.** d) All of the mentioned

**Answer 5.** c) Poisson

**Answer 6.** b) False

**Answer 7.** b) Hypothesis

**Answer 8.** a) 0

**Answer 9.** c) Outliers cannot conform to the regression relationship

**Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.**

**Answer 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.

**Properties of a normal distribution**

* The mean, mode and median are all equal.
* The curve is symmetric at the center (i.e. around the mean, μ).
* Exactly half of the values are to the left of center and exactly half the values are to the right.
* The total area under the curve is 1.

**Answer 11.** The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the pre-processing of the dataset as many machine learning algorithms do not support missing values.

To handle missing values in the dataset we use following imputation techniques:

1. **Deleting Rows with missing values-**Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.

**Pros:**

* A model trained with the removal of all missing values creates a robust model.

**Cons:**

* Loss of a lot of information.
* Works poorly if the percentage of missing values is excessive in comparison to the complete dataset.

1. **Impute missing values for continuous variable-** Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. This method can prevent the loss of data compared to the earlier method. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.

**Pros:**

* Prevent data loss which results in deletion of rows or columns
* Works well with a small dataset and easy to implement.

**Cons:**

* Works only with numerical continuous variables.
* Can cause data leakage
* Does not factor the covariance between features.

1. **Impute missing values for categorical variable-** When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large then it can be replaced with a new category.

**Pros:**

* Prevent data loss which results in deletion of rows or columns
* Works well with a small dataset and easy to implement.
* Negates the loss of data by adding a unique category

**Cons:**

* Works only with categorical variables.
* Addition of new features to the model while encoding, which may result in poor performance

1. **Other Imputation Methods-** Depending on the nature of the data or data type, some other imputation methods may be more appropriate to impute missing values.

For example, for the data variable having longitudinal behavior, it might make sense to use the last valid observation to fill the missing value. This is known as the Last observation carried forward (LOCF) method.

For the time-series dataset variable, it makes sense to use the interpolation of the variable before and after a timestamp for a missing value.

1. **Using Algorithms that support missing values-** All the machine learning algorithms don’t support missing values but some ML algorithms are robust to missing values in the dataset. The k-NN algorithm can ignore a column from a distance measure when a value is missing. Naive Bayes can also support missing values when making a prediction. These algorithms can be used when the dataset contains null or missing values.

The sklearn implementations of naive Bayes and k-Nearest Neighbors in Python does not support the presence of the missing values.

Another algorithm that can be used here is RandomForest that works well on non-linear and the categorical data. It adapts to the data structure taking into consideration the high variance or the bias, producing better results on large datasets.

**Pros:**

* No need to handle missing values in each column as ML algorithms will handle it efficiently

**Cons:**

* No implementation of these ML algorithms in the scikit-learn library.

1. **Prediction of missing values-** In the earlier methods to handle missing values, we do not use correlation advantage of the variable containing the missing value and other variables. Using the other features which don’t have nulls can be used to predict missing values.

The regression or classification model can be used for the prediction of missing values depending on nature (categorical or continuous) of the feature having missing value.

**Pros:**

* Gives a better result than earlier methods
* Takes into account the covariance between missing value column and other columns.

**Cons:**

* Considered only as a proxy for the true values

1. **Imputation using Deep Learning Library — Datawig -**This method works very well with categorical, continuous, and non-numerical features. Datawig is a library that learns ML models using Deep Neural Networks to impute missing values in the datagram.

**Install datawig library,**

**pip3 install datawig**

Datawig can take a data frame and fit an imputation model for each column with missing values, with all other columns as inputs.

**Pros**:

* Quite accurate compared to other methods.
* It supports CPUs and GPUs.

**Cons:**

* Can be quite slow with large datasets.

**There is no thump rule to handle missing values in a particular manner**, the method which gets a robust model with the best performance. One can use various methods on different features depending on how and what the data is about. Having a domain knowledge about the dataset is important, which can give an insight into how to pre-process the data and handle missing values.

Choosing best method to impute the missing values of data is based on applying trial and error.

1.First we need to create a subset of data from the population.

2.Then delete some of the values manually.

3.Impute those deleted values with Imputation methods which are mentioned above.

4.Compare those impute data and actual data (the ones we have deleted manually)

5.Atlast we will be able to see imputed data from which method is close to the actual value and choose that method for the model.

**Answer 12. A/B testing** (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

**Advantages of A/B Testing**

• They are fast

• Advanced analytics can be installed and evaluated for each variation (e.g. click tracking, heatmaps, etc.)

• Can achieve more dramatic conversion rate lift results

• Requires less traffic

**Disadvantages of A/B Testing**

• More dramatic failures

• Less specific understanding of element effects

**Answer 13.**

* Bad practice in general
* If just estimating means: mean imputation preserves the mean of the observed data
* Leads to an underestimate of the standard deviation
* Distorts relationships between variables by “pulling” estimates of the correlation toward zero

**Answer 14.** In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

* **Simple linear regression:** Y = a + bX + u
* **Multiple linear regression:** Y = a + b1X1+ b2X2 + b3X3 + ... + btXt + u

Where:

* Y = the variable that you are trying to predict (dependent variable).
* X = the variable that you are using to predict Y (independent variable).
* a = the intercept.
* b = the slope.
* u = the regression residual.

**Answer 15.** The two main branches of statistics are **descriptive statistics** and **inferential 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.

**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.