# STATISTICS WORKSHEET-1

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

1. Bernoulli random variables take (only) the values 1 and 0

Ans: True.

1. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases

Ans: Central Limit Theorem

1. Which of the following is incorrect with respect to use of Poisson distribution ?

Ans: Modeling bounded count data

1. Point out the correct statement.

Ans : All of the mentioned

1. \_\_\_\_\_\_ random variables are used to model rates.

Ans: Poisson

1. Usually replacing the standard error by its estimated value does change the CLT.

Ans: False

1. Which of the following testing is concerned with making decisions using data?

Ans: Hypothesis

1. Normalized data are centered at\_\_\_\_\_\_and have units equal to standard deviations of the original data.

Ans: 0

1. Which of the following statement is incorrect with respect to outliers?

Ans: Outliers cannot conform to the regression relationship

1. What do you understand by the term Normal Distribution?

Ans: Normal Distribution is also known as “ Gaussian Distribution “. It is the probability distribution that is symmetric about the mean , showing the data near the mean are more frequent in occurrence of the data far from the mean . This normal distribution will appear in Bell form in Graph when plotted. Norma; distribution are symmetric , but not all symmetric distributions are Normal. These values are equally distributed on Left and right side of the central tendency.

1. How do you handle missing data? What imputation techniques do you recommend?

Ans: Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.

Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

Common Methods

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

There may not be enough observations with non-missing data to produce a reliable analysis

In predictive analytics, missing data can prevent the predictions for those observations which have missing data

External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2. Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

To set up the data for MICE, it is important to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.

3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data. For example, if the distribution of race/ethnicity for non-missing data is similar to the distribution of race/ethnicity for missing data, overfitting is not likely to throw off results. However, if the two distributions differ, the accuracy of imputations will suffer.

The MICE library in R also allows imputations by random forest by setting the method to “rf”. The authors of the MICE library have provided an example on how to implement the random forest method here.

To sum up data imputations is tricky and should be done with care. It is important to understand the nature of the data that is missing when deciding which algorithm to use for imputations. While using the above algorithms, predictor variables should be set up carefully to avoid confusion in the methods implemented during imputation. Finally, you can test the quality of your imputations by normalized root mean square error (NRMSE) for continuous variables and proportion of falsely classified (PFC) for categorical variables.

1. What is A/B testing?

Ans: A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

1. Is mean imputation of missing data acceptable practice?

Ans: The process of replacing null values in a data collection with the data’s mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

1. What is linear regression in statistics?

Ans: Linear regression is a basic and commonly used type of predictive analysis.These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.There are many names for a regression’s dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

1. What are the various branches of statistics?

Ans: Statistics have majorly categorised into two types:

* Descriptive Statistics :

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.Descriptive statistics are also categorised into four different categories:

* Measure of frequency
* Measure of dispersion
* Measure of central tendency
* Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

* Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc. Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.