The Importance of Statistics in Data Science Practices

The global rise in the usage of internet has witnessed a positive increase, so is the amount of data generated. Researches from [IBM](https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/), a top business firm, shows 2.5 quintillion bytes of data is created daily. Nowadays, data science takes a fusion of several fields that seek to simplify core development processes. Such fields include computer science, statistics, econometrics and other domain-related fields.

Also, the usage of algorithms, procedures, and systems help make better deductions from collected data. The quality of statistical data create a better informed conclusion and support decision-making. Although, the management of large volumes of data may still be considered challenging, data science now offers several benefits to solving these problems.

Before advancements in computing technology, statisticians rely solely on statistical modeling techniques for predictions and data modeling. These traditional statistical modeling techniques use regression models like Bayesian regression and penalized regression while fast processing speed on computers allows for the use of data science libraries like random forests, support vector machines, neural network and deep learning. In the process of statistical modeling, statisticians are often concerned with selecting the appropriate model family, specifying the model, check how the model fits, analyzing the accuracy of the predictions, and interpreting the effects. Recently, high computational power equips the modern data scientist with the ability to focus less on the mathematics and more on choosing algorithms, writing code, specifying tuning parameters and analyzing the result of the predictions. The process of practical statistical modeling is often blurred from the data scientist in data science packages like Python, R etc. as all they need to do is to specify which algorithm to use and feed it with parameters. This makes it friendlier for people without deep statistics background.

In statistical analysis, assumptions are made about the data characteristics. Violating these assumptions will change the outcome of the project and analysis of results. Assumptions are made depending on the parametric analysis of the data. Such assumptions require certain tests be carried out before forming an hypothesis.

To check for randomness in data, a certain test known as the “run test” is used. Sometimes, data are not random but can be said to be multicollinear. This means that some varibales of interest in the data are correlated. It has been established that if the Variance Inflation Factor (VIF) is higher than 10, the data is multicollinear. Some other factors to consider before forming assumptions about data include Skewness and Kurtosis whose standard values are between ±2 and ±7 respectively; Heteroscedasticity, a circumstance where the variability of a value is unequal across the range of values of a second variable that predicts it; and autocorrelation, which measures the degree of similarity between a variable’s current value and its past value. The resulting output of this test ranges between +1 (perfect positive correlation) and -1 (perfect negative correlation). It can be used to show how impactful the past event has on the future event.

Executing a data science project involves different steps that uses one or more statistical approaches in solving the problem. To prepapre data for processing, newly acquired data need to be cleaned and missing values imputed. Using the mean or median value of that variable in previous time steps solves the problem of missing values. After cleaning the data, the next step is to explore the dataset. This step is vital because it gives the data scientist an apriori knowledge of paramers, the theory behind Bayesian statistics. It is from this data exploration stage; a data scientist selects a suitable analytic method or model. To model the data, a statistical method called fitting is employed. Fitting is achieved by looking for statistical relationships in data by examining the importance of variables in a given dataset. A statistical model takes a mathematical representation of analyzed data. A predicted model can fit with an existing mathematical model, Gaussian or Sigmoid function for example.

As the de-facto standard, only 70% of data is used for modeling while 30% will be reserved for model testing. After testing the model, the accuracy of such model is evaluated using purely statistical techniques. Error evaluation techniques are backed by classical formulae. The most prominent ones include: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and more.

Several models can be developed to solve a particular predictive problem. The models will be evaluated using one or a combination of the above-listed techniques with the best being the one with the lowest error or highest accuracy. A feat which will be impossible, without statistics.

After selecting the best performing model that suits the accuracy standard of the project, deductions will be made from the dataset and newly predicted value. It can be in the form of bar charts, line graphs, scatter plots or other data visualization methods.

The project execution stages explained shows that statistics is the bedrock of data science projects. Classical statistical techniques are being written as computer programming packages to aid easy integration of statistical techniques with data science projects. A data science project is not complete without applying one or more statistical techniques.