Application of Computational Statistics in Data Science

1st Dr. Kamlesh M. Jani,
[1]Principal, 1st Author,
Shree P.D.M. College of Commerece Rajkot 1st Author,

Abstract - This article has described critical task in the education of future data scientists is to instill data acumen. This requires exposure to key concepts in data science, real-world data and problems that can reinforce the limitations of tools, and ethical considerations that permeate many applications. Key concepts involved developing data acumen include the following:

Mathematical foundations,
Computational foundations,
Statistical foundations,
Data management and curation,
Data description and visualization,
Data modeling and assessment.
Workflow and reproducibility

Index Terms - Introduction To Statistics,Terminologies In Statistics,Categories In Statistics,Understanding Descriptive Analysis,Descriptive Statistics In R,Understanding Inferential Analysis,Inferential Statistics In R

I. INTRODUCTION

Data analysis is computing Statisticians have always been heavy on whatever is available ior computing facilities. As compute-in features have become more powerful over the years, those features have clearly decreased the amount of the statistician will have to afford the door out analysis. As computing facilities have become more powerful, however, an opposite has resulted; The computational aspect of Statistician's work has increased. The reason for this is that paradigm gifts in computer-enabled statistical science.

"Statistical Computation" to refer to statistic methods of computational methods. Satirical computing this includes numerical analysis, database methodology; computer graphics, software engineering, and computer , human Interface. We use the term "computational statistics" Statistical computing, but also statistical methods that are computationally intensive. Thus, to some extent, "Computational Statistics" refers to a large class of modern statistical methods. Computational statistics is Based on mathematical statistics, statistical computing and applied statistics. While we distinguish from "statistical computing" to "computational statistics", the emergence of all of computational statistics was Coincidence with statistical computation, and is not possible without development in statistical computing.

Computational statistics is a subset of data science. arguably the most important one. It can be loosely described as "traditional statistics using computers".

Data science includes a number of other elements that are not in the scope of computational statistics, most notably: . machine learning.

Big data analytics.

The business aspect

While computational statistics is a subarea of scientific computing that follows scientific tiger', data scientist are usually content with accepting whatever method that provides the best business value.

Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems. At the core is data. Troves of raw information, streaming in and stored in enterprise data warehouses. Much to learn by mining it. Advanced capabilities we can build with it. Data science is ultimately about using this data in creative ways to generate business value:

"Data Scientist is a person who is better at statistics than any programmer and better at programming than any statistician."

Math and Statistics for Data Science are essential because these disciples form the basic foundation of all the Machine Learning Algorithms. In fact, Mathematics is behind everything around us, from shapes, patterns and colors, to the count of petals in a flower. Mathematics is embedded in each and every aspect of our lives.

Although having a good understanding of programming languages, Machine Learning algorithms and following a data-driven approach is necessary to become a Data Scientist, Data Science isn't all about these fields.

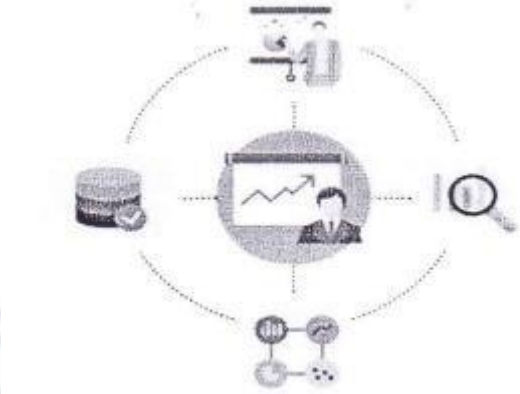
you will understand the importance of Math and Statistics for Data Science and how they can be used to build Machine Learning models.

Here's a list of topics I'll be covering in this Math and Statistics for Data Science:

- ✓ Introduction To Statistics
- ✓ Terminologies In Statistics
- ✓ Categories In Statistics
- ✓ Understanding Descriptive Analysis
- ✓ Descriptive Statistics In R
- ✓ Understanding Inferential Analysis
- ✓ Inferential Statistics In R

II. INTRODUCTION TO STATISTICS

To become a successful Data Scientist you must know your basics. Math and Stats are the building blocks of Machine Learning algorithms. It is important to know the techniques behind various Machine Learning algorithms in order to know how and when to use them. Now the question arises, what exactly is Statistics? Statistics is a Mathematical Science pertaining to data collection, analysis, interpretation and presentation.



Statistics is used to process complex problems in the real world so that Data Scientists and Analysts can look for meaningful trends and changes in Data. In simple words, Statistics can be used to derive meaningful insights from data by performing mathematical computations on it. Several Statistical functions, principles and algorithms are implemented to analyses raw data, build a Statistical Model and infer or predict the result.

Statistics application Math's and statistics for Data Science

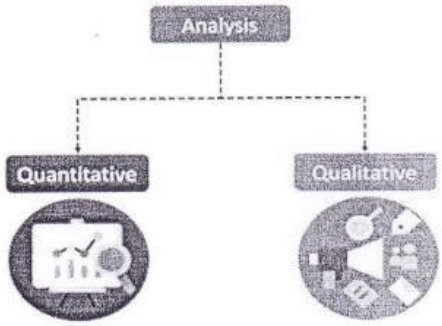
The field of Statistics has an influence over all domains of life, the Stock market, life sciences, weather, retail, insurance and education are but to name a few. Moving ahead. let's discuss the basic terminologies in Statistics. Terminologies In Statistics - Statistics For Data Science One should be aware of a few key statistical terminologies while dealing with Statistics for Data Science. I've discussed these terminologies below:

- ✓ Population is the set of sources from which data has to be collected.
- ✓ A Sample is a subset of the Population .
- ✓ A Variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item.
- ✓ Also known as a statistical model, A statistical Parameter or population parameter is a quantity that indexes a family of probability distributions. For example, the mean, median, etc of a population.

Before we move any further and discuss the categories of Statistics, let's look at the types of analysis.

III. Types 0f Analysis

Analysis of any event can be done in one of two ways

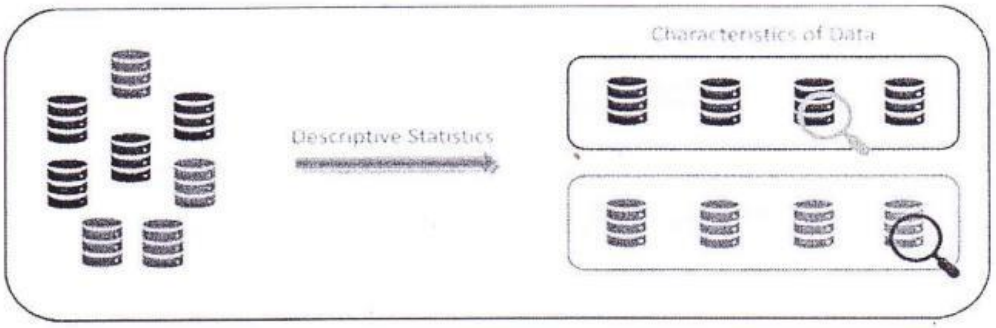


Types of Analysis Math and Statistical for data Science

Quantitative Analysis: Quantitative Analysis ol the Statistical Analysis is the science of collecting and interpreting data with numbers and graphs to identified patterns and trends
Qualitative Analysis: Qualitative or Non-Statistical Analysis gives generic information and uses text sound and other forms of media to do so.
For example. If I want purchase a coffee from starbucks , it is available in Short tall and Grande. This is an example of Qualitative Analysis. But if a store sells 70 regular coffee a week. it is Quantitative Analysis because we have a number of representing the coffee sold per week

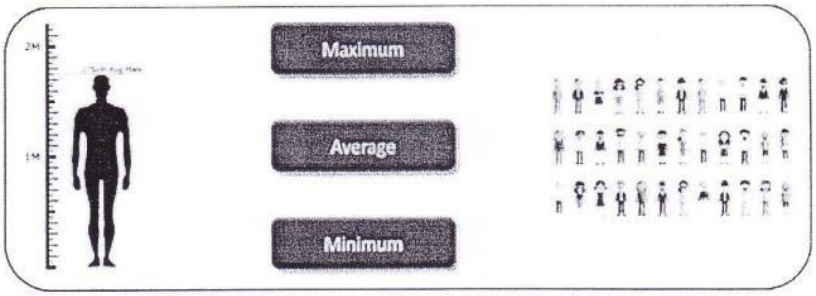
Categories in Statistics
There are two main categories in statistics namely:
1 Descriptive statistics
2 inferential statistics

Descriptive Statistics
Descriptive Statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables. Descriptive Statistics helps organize data and focuses on the characteristics of data providing parameters.



Descriptive Statistics – Math And Statistics For Data Science

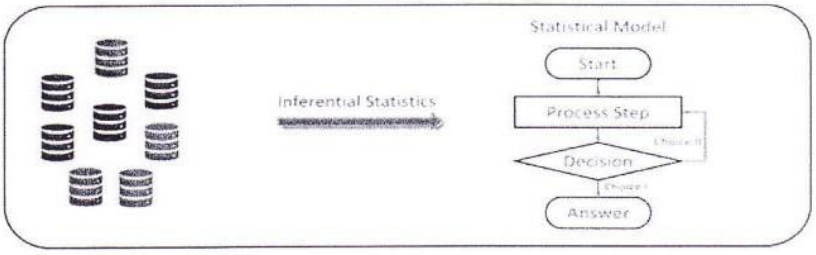
Suppose you want to study the average height of students in a classroom, in descriptive statistics you would record the heights of all students in the class and then you would find out the maximum, minimum and average height of the class.



Descriptive Statistics Example – Math And Statistics For Data Science

Inferential Statistics

Inferential Statistics makes inferences and predictions about a population based on a sample of data taken from the population in question. Inferential statistics generalizes a large data set and applies probability to arrive at a conclusion. It allows you to infer parameters of the population based on sample stats and build models on it.



Understanding Descriptive Analysis

When we try to represent data in the form of graphs, like histograms, line plots, ete, the data is represented based on some kind of central tendency. Central tendency measures like, mean, median, or measures of the spread, etc are used for statistical analysis. To better understand Statistics lets discuss the different measures in Statistics with the help of an example.

Cars	mpg	cyl	disp	hp	drat
A	21	6	160	110	3.9
B	21	6	160	110	3.9
C	22.8	4	108	93	3.85
D	21.3	6	108	96	3
E	23	4	150	90	4
F	23	6	108	110	3.9
G	23	4	160	110	3.9
H	23	6	160	110	3.9

Cars Data Set – Math And Statistics For Data Science

Here is a sample data set of cars containing the variables:

1. Cars
2. Mileage per Gallon (mpg)
3. Cylinder Type (cyl)
4. Displacement (disp)
5. Horse Power (hp) 6. Real Axle Ratio (drat).

Before we move any further, let's define the main Measures of the Center or Measures of Central tendency.

Measures Of The Center

1. Mean: Measure of average of all the values in a sample is called Mean.
2. Median: Measure of the central value of the sample set is called Median.
3. Mode: The value most recurrent in the sample set is known as Mode.

Using descriptive Analysis, you can analyze each of the variables in the sample data set for mean, standard deviation, minimum and maximum.

- ✓ If we want to find out the mean or average horsepower of the cars among the population of cars, we will check and calculate the average of all values, In this case, we'll take the sum of the Horse Power of each car, divided by the total number of cars:

Mean (110+110+93+96+90+110+110+110)/8=103.625

- ✓ . If we want to find out the center value of mpg among the population of cars, we will arrange the mpg values in ascending or descending order and choose the middle value. In this case, we have 8 values which is an even entry. Hence we must take the average of the two middle values.

The mpg for 8 cars: 21.21.21.3.22.8,23,23,23,23
Median (22.8+23)/2-22.9

- ✓ If we want to find out the most common type of cylinder among the population of cars, we will check the value which is repeated most number of times. Here we can see that the cylinders come in two values, 4 and 6. Take a look at the data set, you can see that the most recurring value is 6. Hence 6 is our Mode.

Measures Of The Spread

Just like the measure of center, we also have measures of the spread, which comprises of the following measures:

Range: It is the given measure of how spread apart the values in a data set are.
Inter Quartile Range (IQR): It is the measure of variability, based on dividing a data set into quartiles.
Variance: It describes how much a random variable differs from its expected value. It entails computing squares of deviations.
Deviation is the difference between each element from the mean Population Variance is the average of squared deviations
Sample Variance is the average of squared differences from the mean
Standard Deviation: It is the measure of the dispersion of a set of data from its mean.

Descriptive statistics In IR
It's always best to perform practical implementation to better understand a concept. In this section, we'll be executing a small demo that will show you how to calculate the Mean, Median, Mode, Variance, Standard Deviation and how to study the variables by plotting a histogram. This is quite a simple demo but it also form the foundation that every Machine Learning algorithm is built upon

Step 1: Import data for computation

```
1 >set.seed(1)
2 #Generate random numbers and store it in a variable called data
 3 >data runif(20,1,10)
```

Step 2: Calculate Mean for the data

```
1 #Calculate Mean
2 >mean mean(data)
3 >print(mean)
4 [1] 5.996504
```

Step 3: Calculate the Median for the data

```
1 #Calculate Median
2 >median median(data)
3 >print(median)
4
5 [1] 6.408853
```

Step 4: Calculate Mode for the data

```
1 #Create a function for calculating Mode
2 >mode<- function(x) (>ux < unique(x) >ux/which.max(tabulate(match(x, ux)))]
3 }
4>result mode(data) >print(data)
5 [1] 3.389578 4.349115 6.155680 9.173870 2.815137 9.085507 9.502077 6.947180 6.662026
6 [10] 1.556076 2.853771 2.589011 7.183206 4.456933 7.928573 5.479293 7.458567 9.927155
7 [19] 4.420317 7.997007
8>cat("mode ()", result) mode=(3.389578
```

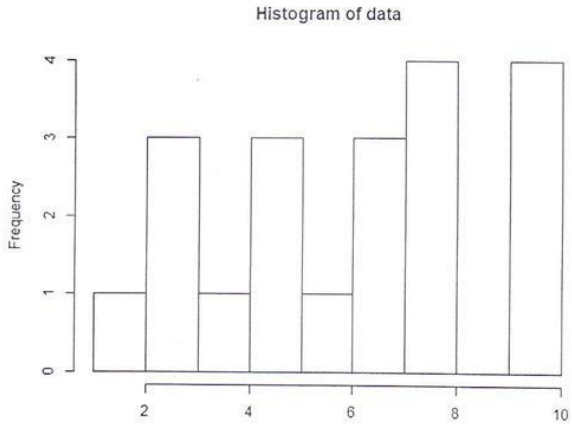
Step 5: Calculate Variance & Std Deviation for the data

```
1 #Calculate Variance and std Deviation
2>variance var(data)
3 >standardDeviation = sqrt(var(data))
4>print(standardDeviation)
5 12.575061
```

Step 6: Plot a Histogram

```
1 #Plot Histogram
2 >hist(data, bins-10, range= c(0,10), edgecolor="black")
```

The Histogram is used to display the frequency of data points:



Math and Statistics For Data Science-Histogram

Data science

This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviors, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions. For example:

- ✓ Netflix data mines movie viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce.
- ✓ Target identifies what are major customer segments within it's base and the unique shopping behaviors within those segments, which helps to guide messaging to different market audiences.
- ✓ Procter & Gamble utilizes time series models to more clearly understand fisture demand, which help plan for production levels more optimally

How do data scientists mine out insights? It starts with data exploration. When given a challenging question,data scientists become detectives. They investigate leads and try to understand pattern or characteristics within the data. This requires a big dose of analytical creativity.

Then as needed, data scientists may apply quantitative technique in order to get a level deeper-eg inferential models, segmentation analysis, time series forecasting, synthetic control experiments, etc. The intent is to scientifically piece together a forensic view of what the data is really saying

This data-driven insight is central to providing strategic guidance. In this sense, data scientists act as consultants, guiding business stakeholders on how to act on findings.

Data science-development of data product A "data product" is a technical asset that:

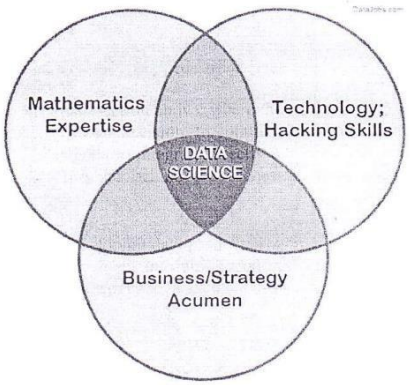
(1) utilizes data as input, (2) processes that data to return algorithmically-generated results. The classic example of a data product is a recommendation engine, which ingests user data, and makes personalized recommendations based on that data. Here are some examples of data products:

- ✓ Amazon's recommendation engines suggest items for you to buy, determined by their algorithms. Netflix recommends movies to you. Specify recommends music to you.
- ✓ Gmail's spam filter is data product - an algorithm behind the scenes processes incoming mail and determines if a message is junk or not.
- ✓ Computer vision used for self-driving cars is also data product-machine learning algorithms are able to recognize traffic lights, other cars on the road, pedestrians, etc.

This is different from the "data insights" section above, where the outcome to that is to perhaps provide advice to an executive to make a smarter business decision. In contrast, a data product is technical functionality that encapsulates an algorithm, and is designed to integrate directly into core applications. Respective examples of upplications that incorporate data product behind the scenes: Amazon's homepage, Gmail's inbox, and autonomous driving software.

Data scientists play a central role in developing data product. This involves building out algorithms, as well as testing, refinement, and technical deployment into production systems. In this sense, data scientists serve as technical developers, building assets that can be leveraged at wide scale.

What is data science-the requisite skill set
Data science is a blend of skills in three major areas:



Mathematics Expertise

At the heart of mining data insight and building data product is the ability to view the data through a quantitative lens. There are textures, dimensions, and corrier in data that can be expressed mathematically. Finding solutions utilizing data becomes a ben tesser of tics and quantitative technique. Solutions to many business problems involve building analytic models grounded in the hard math, where being able to understand the underlying mechanics of those models is key to success in building them.

Also, a misconception is that data science all about statistics. While statistics is important, it is not the only type of math utilized. First, there are two branches of statistics-classical statistics and Bayesian statistics. When most people refer to stars they are generally referring to classical stats, but knowledge of both types is helpful. Furthermore, many inferential techniques and machine learning algorithms lean on knowledge of linear algebra. For example, a popular method to discover hidden characteristics in a data set is SVD, which is grounded in matrix math and has much less to do with classical stats. Overall, it is helpful for data scientists to have breadth and depth in their knowledge of mathematics.

Technology and Hacking

First, let's clarify on that we are not talking about hacking as in breaking into computers. We're referring to the tech programmer subculture meaning of hacking-i.e, creativity and ingenuity in using technical skills to build things and find clever solutions to problems;

Why is hacking ability important? Because data scientists utilize technology in order to wrangle enormous data sets and work with complex algorithms, and it requires tools far more sophisticated than Excel. Data scientists need to be able to code-prototype quick solutions, as well as integrate with complex data systems. Core languages associated with data science include SQL, Python, R, and SAS. On the periphery are Java, Scala, Julia, and others. But it is not just knowing language fundamentals. A hacker is a technical ninja, able to creatively navigate their way through technical challenges in order to make their code work.

Along these lines, a data science hacker is a solid algorithmic thinker, having the ability to break down messy problems and recompose them in ways that are solvable. This is critical because data scientists operate within a lot of algorithmic complexity. They need to have a strong mental comprehension of high-dimensional data and tricky data control flows. Full clarity on how all the pieces come together to form a cohesive solution.

Strong Business Acumen

It is important for a data scientist to be a tactical business consultant. Working so closely with data, data scientists are positioned to learn from data in ways no one else can. That creates the responsibility to translate observations to shared knowledge, and contribute to strategy on how to solve core business problems. This means a core competency of data science is using data to cogently tell a story. No data-puking- rather, present a cohesive narrative of problem and solution, using data insights as supporting pillars, that lead to guidance.

Having this business acumen is just as important as having acumen for tech and algorithms. There needs to be clear alignment between data science projects and business goals. Ultimately, the value doesn't come from data, math, and tech itself. It comes from leveraging all of the above to build valuable capabilities and have strong business influence.

What is a data scientist-curiosity and training

The Mindset

A common personality trait of data scientists is they are deep thinkers with intense intellectual curiosity. Data science is all about being inquisitive - asking new questions, making new discoveries, and learning new things. Ask data soientists most obsessed with their work what drives them in their job, and they will not say "money". The real motivator is being able to use their creativity and ingenuity to solve hard problems and constantly indulge in their curiosity. Deriving complex reads from data is beyond just making an observation, it is about uncovering "truth" that lies hidden beneath the surface. Problem solving is not a task, but an intellectually-stimulating journey to a solution. Data scientists are passionate about what they do, and reap great satisfaction in taking on challenge.

Training

There is a glaring misconception out there that you need a sciences or math to become a legitimate data scientist. That view misses the point that data science is multidisciplinary. Highly-focused study in academia is certainly helpful, but doesn't guarantee that graduates have the full set of experiences and abilities to succeed.

Statistician may still need to pick up a lot of programming skills and gain business experience, to complete the trifecta.

In fact, data science is such a relatively new and rising discipline that universities have not caught up in developing comprehensive data science degree programs-meaning that no one can really claim to have "done all the schooling to be become a data scientist. Where does much of the training come from? The unyielding intellectual curiosity of data scientists push them to be motivated autodidacts, driven to self-learn the right skills, guided by their own determination.

What is Analytics?

Analytics has risen quickly in popular business lingo over the past several years; the term is used loosely, but generally meant to describe critical thinking that is quantitative in nature, Technically, analytics in the "science of analysis"-put another way, the practice of analyzing information to make decisions.

Is "analytics" the same thing as data science? Depends on context. Sometimes it is synonymous with the definition of data science that we have described, and sometimes it represents something else. A data scientist using raw data to build a predictive algorithm falls into the scope of analytics.

At the same time, a non-technical business user interpreting pre-built dashboard reports is also in the realm of analytics, but does not cross into the skill set needed in data science. Analytics has come to have fairly broad meaning. At the end of the day, as long as you understand beyond the buzzword level, the exact semantics don't matter much.

Why We Need Data Science ?

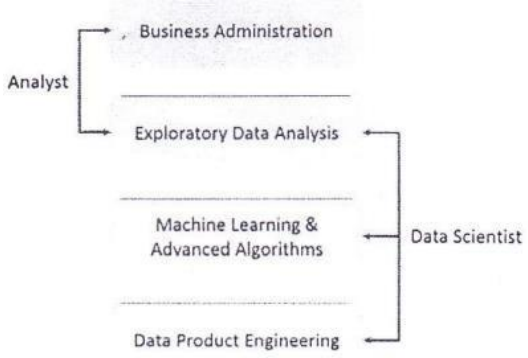
Traditionally, the data that we had was mostly structured and small in size, which could be analyzed by using the simple BI tools. Unlike data in the traditional systems which was mostly structured, today most of the data is unstructured or semi-structured. Let's have a look at the data trends in the image given below which shows that by 2020, more than 80% of the data will be unstructured. This data is generated from different sources like financial logs, text files, multimedia forms, sensors, and Instruments. Simple BI tools are not capable of processing this huge volume and variety of data. This is why we need more complex and advanced analytical tools and algorithms for processing, analyzing and drawing meaningful insights out of it This is not the only reason why Data Science has become so popular. Let's dig deeper and see how Data Science is being used in various domains.

How about if you could understand the precise requirements of your customers from the existing data like the customer's past browsing history, purchase history, age and income. No doubt you had all this data earlier too, but now with the vast amount and variety of data, you can train models more effectively and recommend the product to your customers with more precision. Wouldn't it be amazing as it will bring more business to your organization?

Let's take a different scenario to understand the role of Data Science in decision making. How about if your car had the intelligence to drive you home? The self-driving cars collect live data from sensors, including radars, cameras and lasers to create a map of its surroundings. Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where to take a turn-making use of advanced machine learning algorithms.

Let's see how Data Science can be used in predictive analytics. Let's take weather forecasting as an example. Data from ships, aircrafts, radars, satellites can be collected and analyzed to build models. These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities. It will help you to take appropriate measures beforehand and save many precious lives.

Use of the term Data Science is increasingly common, but what does it exactly mean? What skills do you need to become Data Scientist? What is the difference between BI and Data Science? How are decisions and predictions made in Data Science? These are some of the questions that will be answered further.



As you can see from the above image, a Data Analyst usually explains what is going on by processing history of the data. On the other hand, Data Scientist not only does the exploratory analysis to discover insights from it. but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future. A Data Scientist will look at the data from many angles, sometimes angles not known earlier.

So, Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning

Predictive causal analyties - If you want a model which can predict the possibilities of a particular event in the future, you need to apply predictive causal analytics. Say, if you are providing money on credit, then the probability of customers making finture credit payments on time is a matter of concern for you. Here, you can build a model which can perform predictive analytics on the payment history of the customer to predict if the future payments will be on time or not.

Prescriptive analytics: If you want a model which has the intelligence of taking its own decisions and the ability to modify it with dynamic parameters, you certainly need prescriptive analytics for it. This relatively new field is all about providing advice. In other terms, it not only predicts but suggests a range of prescribed actives and associated outcomes The best example for this is Google's self-driving car which I had discussed earlier too. The data gathered by vehicles can be used to train self-driving cars. You can run algorithms on this data to bring intelligence to it. This will enable your car to take decisions like when to turn, which path to take, when to slow down or speed up.

Machine learning for making predictions-If you have transactional data of a finance company and need to build a model to determine the future trend, then machine learning algorithms are the best bes. This falls under the paradigm of supervised learning. It is called supervised because you already have the data based on which you can train your machines. For example, a fraud detection model can be trained using a historical record of fraudulent purchases.

Machine learning for pattern discovery-If you don't have the parameters based on which you can make predictions, then you need to find out the hidden patterns within the dataset to be able to make meaningful predictions. This is nothing but the unsupervised model as you don't have any predefined labels for grouping. The most common algorithm used for pattern discovery is Clustering Let's say you are working in a telephone company and you need to establish a network by putting towers in a region. Then, you can use the clustering technique to find those tower locations which will ensure that all the users receive optimum signal strength.

Let's see how the proportion of above-described approaches differ for Data Analysis as well as Data Science. As you can see in the image below, Data Analysis includes descriptive analytics and prediction to a certain extent. On the other hand, Data Science is more about Predictive Causal Analytics and Machine Learning

1. R has a complete set of modeling capabilities and provides a good environment for building interpretive models.
2. SQL Analysis services can perform in-database analytics using common data mining functions and basic predictive models.
3. SAS/ACCESS can be used to access data from Hadoop and is used for creating repeatable and reusable model flow diagrams.

Although, many tools are present in the market but R is the most commonly used tool.

Computational statistics, or statistical computing, is the interface between statistics and computer science. It is the area of computational science (or scientific computing) specific to the mathematical science of statistics. This area is also developing rapidly, leading to calls that a broader concept of computing should be taught us part of general statistical education.

As in traditional statistics the goal is to transform raw data into knowledge, but the focus lies on computer intensive statistical methods, such as cases with very large sample size and non-homogeneous data sets.

The terms 'computational statistics' and 'statistical computing' are often used interchangeably, although Carlo Lauro (a former president of the International Association for Statistical Computing) proposed making a distinction, defining statistical computing as "the application of computer science to statistics", and computational statistics as "aiming at the design of algorithm for implementing statistical methods on computers, including the ones unthinkable before the computer age as well as to cope with analytically intractable problems.

IV. CONCLUSIONS

This article has described critical task in the education of future data scientists is to instill data acumen. This requires exposure to key concepts in data science, real-world data and problems that can reinforce the limitations of tools, and ethical considerations that permeate many applications. Key concepts involved developing data acumen include the following:

Mathematical foundations,
Computational foundations,
Statistical foundations,
Data management and curation,
Data description and visualization,
Data modeling and assessment.
Workflow and reproducibility

This report aims to increase the level of awareness of the intellectual and technical issues surrounding the analysis of massive data. This is not the first report written on massive data, and it will not be the last, but given the major attention currently being paid to massive data in science, technology, and govemment, the committee believes that it is a particularly appropriate time to be considering these issues. The aim of this report is to help Application of Computational statistics in Data Science and improve their assessment practices for the benefit of students.

V. REFERENCES

[1] Introduction to Statistical Process Control by Peihua Qiu (Author)
[2] Principles of Data Science by Sinan Ozdemir (Author)
[3] Bayesian Data Analysis (Chapman & Hall/CRC Texts in Statistical Science) by Andrew Gelman (Author)
[4] Matrix Algebra: Theory, Computations, and Applications by James E. Gentle (Author)
[5] Computational Statistics: An Introduction to R by Günther Sawitzki (Author) 6. Statistics for Data Science by James D. Miller ((Author)
[6] What is DATA Science? https://datajobs.com/what-is-data-science
[7]Computational Statistics https://www.researchgate.com
[8]CS Became the Backbone of Modern Data Science Computing in the Statistics https://en.wikipedia.org/wiki/Computational statistics