UNIT 1 INTRODUCTION TO DATA SCIENCE

1.0 INTRODUCTION

The Internet and communication technology has grown tremendously in the past decade leading to generation of large amount of unstructured data. This unstructured data includes data such as, unformatted textual, graphical, video, audio data etc., which is being generated as a result of people use of social media and mobile technologies. In addition, as there is a tremendous growth in the digital eco system of organisation, large amount of semi-structured data, like XML data, is also being generated at a large rate. All such data is in addition to the large amount of data that results from organisational databases and data warehouses. This data may be processed in real time to support decision making process of various organisations. The discipline of data science focuses on the processes of collection, integration and processing of large amount of data to produce useful decision making information, which may be useful for informed decision making.

This unit introduces you to the basic concept of data sciences. This unit provides an introduction to different types of data used in data science. It also points to different types of analysis that can be performed using data science. Further, the Unit also introduces some of the common mistakes of data science.

1.1 OBJECTIVES

At the end of this unit you should be able to:

- Define the term data science in the context of an organization
- explain different types of data
- list and explain different types of analysis that can be performed on data
- explain the common mistakes about data size
- define the concept of data dredging
- List some of the applications of data sites
- Define the life cycle of data science

1.2 DATA SCIENCE-DEFINITION

Data Science is a multi-disciplinary science with an objective to perform data analysis to generate knowledge that can be used for decision making. This knowledge can be in the form of similar patterns or predictive planning models, forecasting models etc. A data science application collects data and information from multiple heterogenous sources, cleans, integrates, processes and analyses this data using various tools and presents information and knowledge in various visual forms.

As stated earlier data science is a multi-disciplinary science, as shown in Figure 1.

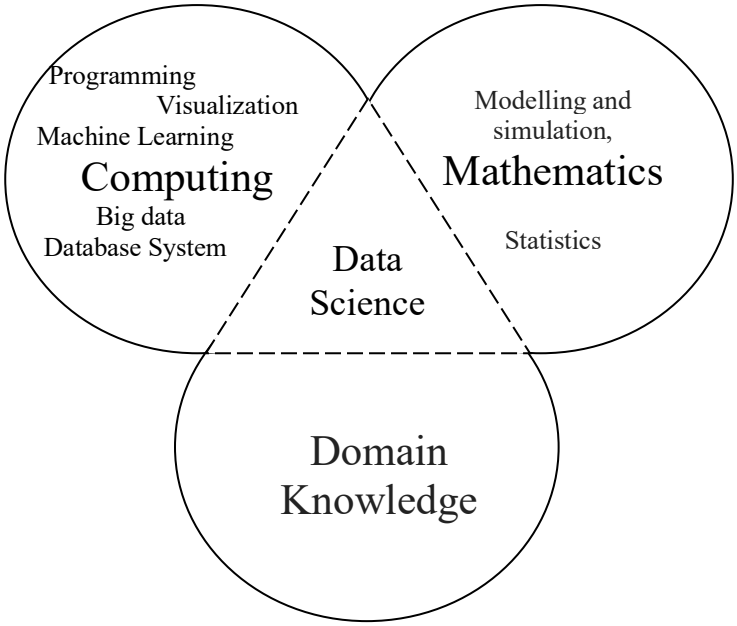


Figure 1: Data Science

What are the advantages of Data science in an organisation? The following are some of the areas in which data science can be useful.

- It helps in making business decisions such as deciding the health of companies with whom they plan to collaborate,
- It may help in making better predictions for the future such as making strategic plans of the company based on present trends etc.
- It may identify similarities among various data patterns leading to applications like fraud detection, targeted marketing etc.

In general, data science is a way forward for business decision making, especially in the present day world, where data is being generate at the rate of Zetta bytes.

Data Science can be used in many organisations, some of the possible usage of data science are as given below:

- It has great potential for finding the best dynamic route from a source to destination. Such application may constantly monitor the traffic flow and predict the best route based on collected data.
- It may bring down the logistic costs of an organization by suggesting the best time and route for transporting foods
- It can minimize marketing expenses by identifying the similar group buying patterns and performing selective advertising based on the data obtained.
- It can help in making public health policies, especially in the cases of disasters.

- It can be useful in studying the environmental impact of various developmental activities
- It can be very useful in savings of resources in smart cities

1.3 TYPES OF DATA

Type of data is one of the important aspect, which determines the type of analysis that has to be performed on data. In data science, the following are the different types of data, that are required to be processed:

1. Structured Data
2. Semi-Structured Data
3. Unstructured data
4. Data Streams

Structured Data

Since the start of the era of computing, computer has been used as a data processing device. However, it was not before 1960s, when businesses started using computer for processing their data. One of the most popular language of that era was **Co**mmon **B**usiness-**O**riented Language (COBOL). COBOL had a data division, which used to represent the structure of the data being processed. This was followed by a disruptive seminal design of technology by a E.F. Codd. This lead to creation of relational database management systems (RDBMS). RDBMS allows structured storage, retrieval and processing of integrated data of an organisation that can be securely shared among several applications. The RDBMS technology also supported secure transaction, thus, became a major source of data generation. Figure 2 shows the sample structure of data that may be stored in a relational database system. One of the key characteristics of structured data is that it can be associated with a schema. In addition, each schema element may be related to a specific data type.

Customer (*custID, custName, custPhone, custAddress, custCategory, custPAN, custAadhar*)
Account (*AccountNumber,custIDoffirstaccountholder,AccountType, AccountBalance)*
JointHolders (AccountNumber, custID)
Transaction(transDate, transType, AccountNumber, Amountoftransaction)

Figure 2: A sample schema of structured data

The relational data is structured data and large amount of this structured data is being collected by various organisations, as backend to most applications. In 90s, the concept of a data warehouse was introduced. A data warehouse is a time-invariant, subject-oriented aggregation of data of an organisation that can be used for decision making. A data in a data warehouse is represented using dimension tables and fact tables. The dimensional tables classifies the data of fact tables. You have already studied various schemas in the context of data warehouse in MCS221. The data of data warehouse is also structured in nature and can be used for analytical data processing and data mining. In addition, many different types of database management systems have been developed, which mostly store structured data.

However, with the growth of communication and mobile technologies many different applications became very popular leading to generation of very large amount of semi-structured and unstructured data. These are discussed next.

Semi-structured Data

As the name suggest Semi-structured has some structure in it. The structure of semi-structured data is due to the use of tags or key/value pairs The common form of semi-structured data is produced through XML, JSON objects, Server logs, EDI data, etc. The example of semi-structured data is shown in the Figure 3.

```
<Book>
        <title>Data Science and Big Data</title>
        <author>R Raman</author>
        <author>C V Shekhar</author>
        <yearofpublication>2020</yearofpublicatio
n>
</Book>
```

```
"Book": {
  "Title":    "Data
Science",
  "Price":    5000,

  "Year":   2020
}
```

Figure 3: Sample semi-structured data

Unstructured Data

The unstructured data does not follow any schema definition. For example, a written text like content of this Unit is unstructured. You may add certain headings or meta data for unstructured data. In fact, the growth of internet has resulted in generation of Zetta bytes of unstructured data. Some of the unstructured data can be as listed below:

- Large written textual data such as email data, social media data etc..
- Unprocessed audio and video data
- Image data and mobile data
- Unprocessed natural speech data
- Unprocessed geographical data

In general, this data requires huge storage space, newer processing methods and faster processing capabilities.

Data Streams

A data stream is characterised by a sequence of data over a period of time. Such data may be structured, semi-structured or unstructured, but it gets generated repeatedly. For example, IoT devices like weather sensors will generate data stream of pressure, temperature, wind direction, wind speed, humidity etc for a particular place where it is installed. Such data is huge for many applications are required to be processed in real time. In general, not all the data of streams is required to be stored and such data is required to be processed for a specific duration of time.

1.3.1 Statistical Data Types

There are two distinct types of data that can be used in statistical analysis. These are – Categorical data and Quantitative data

Categorical or qualitative Data:

Categorical data is used to define the category of data, for example, occupation of a person may take values of the categories "Business", "Salaried". "Others" etc. The categorical data can be of two distinct measurement scales called Nominal and Ordinal, which are given in Figure 4. If the categories are not related, then categorical data is of Nominal data type, for example, the Business category and Salaried categories have no relationship, therefore it is of Nominal type. However, a categorical variable like age category, defining age in categories "0 or more but less than 26", "26 or more but less than 46", "46 or more but less than 61", "More than 61", has a specific relationship. For example, the person in age category "More than 61" are elder to person in any other age category.

Quantitative Data:

Quantitative data is the numeric data, which can be used to define different scale of data. The qualitative data is also of two basic types –discrete, which represents distinct numbers like 2, 3, 5,… or continuous, which represent a continuous values of a given variable, for example, your height can be measured using continuous scale.

Measurement scale of data

Data are raw facts, for example, student data may include name, Gender, Age, Height of student, etc. The name typically is a distinguishing data that tries to distinctly identify two data items, just like primary key in a database. However, the name data or any other identifying data may not be useful for performing data analysis in data science. The data such as Gender, Age, Height may be used to answer queries of the kind: Is there a difference in the height of boys and girls in the age range 10-15 years? One of the important question is how do you measure the data so that it is recorded consistently? Stanley Stevens, a psychologist, defined the following four characteristics that any scale that can be measured:

- Every representation of the measure should be unique, this is referred to as identify of a value (*IDV*).
- The second characteristics is the magnitude (*M*), which clearly can be used to compare the values, for example, a weight of 70.5 kg is more than 70.2 kg.
- Third characteristics is about equality in intervals (*EI*) used to represent the data, for example, the difference between 25 and 30 is 5 intervals, which is same as the difference between 41 to 46, which are also 5 intervals.
- The final characteristics is about a defined minimum or zero value(*MZV*), for example, in Kelvin scale, temperature have an

absolute zero value, whereas, the Intelligent quotient cannot be defined as zero.

Based on these characteristics four basic measurement scales are defined. Figure 4 defines these measurements, their characteristics and examples.

Measurement Scale	Characteristics				Example
	IDV	*M*	*EI*	*MZV*	
Nominal	Yes	No	No	No	Gender F - Female M- Male
Ordinal	Yes	For rank ordering	No	No	A hypothetical Income Category: 1 - "0 or more but less than 26" 2 -"26 or more but less than 46" 3 - "46 or more but less than 61" 4 - "More than 61"
Interval	Yes	Yes	Yes	No	IQ, Temperature in Celsius
Ratio	Yes	Yes	Yes	Yes	Temperature in K, Age

Figure 4: Measurement Scales of Data

1.3.2 Sampling

In general the size of data that is to be processed today is quite large. This leads you to the question, whether you would use the entire data or some representative sample of this data. In several data science techniques sample data is used to develop an exploratory model also. Thus, even in the data science sample is one of the ways, which can enhance the speed of exploratory data analysis. The population in this case may be the entire set of data that you may be interested. Figure 5 shows the relationships between population and sample. One of the question, which is asked in this context is what should be the size of a good sample. You may have to find the answer in the literature. However, you may please note that a good sample is representative of its population.

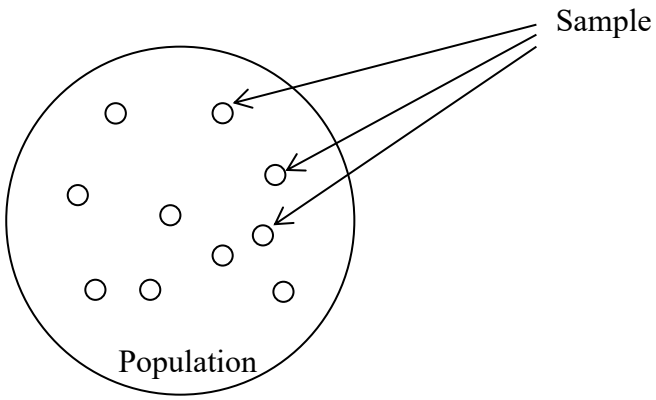


Figure 5: Population and Sample

One of the key objectives of statistics, which uses sample data, is to determine the statistic of the sample and find the probability that the statistic developed for the sample would determine the parameters of population with a specific percentage of accuracy. Please note the terms stated above are very important and explain in the following table:

Term	Used for	Example
Statistic	Statistic is computed for the Sample	Sample mean ($\bar{x}$), Sample Standard deviation (s), Sample size (n)
Parameter	Parameters are predicted from sample and are about the Population	Population mean (μ), Population Standard deviation (σ), Population size (N)

Next, we discuss different kind of analysis that can be performed on data.

Check Your Progress 1:

1. Define the term data science.

2. Differentiate between structured, semi-structured, unstructured and stream data.

3. What would be the measurement scale for the following? Give reason in support of your answer.
   Age, AgeCategory, Colour of eye, Weight of students of a class, Grade of students, 5-point Likert scale

1.4 BASIC METHODS OF DATA ANALYSIS

The data for data science is obtained from several data sources. This data is first cleaned of errors, duplication, aggregated and then presented in a form that can be analysed by various methods. In this section, we define some of the basic methods used for analysing data. These are: Descriptive analysis, Exploratory data analysis and Inferential data analysis.

1.4.1 Descriptive Analysis

Descriptive analysis is used to present basic summaries about data; however, it makes no attempt to interpret the data. These summaries may include different statistical values and certain graphs. Different types of data are described using different ways. The following example illustrates this concept:

Example 1: Consider the data given in the following Figure 6. Show the summary of categorical data in this Figure.

Enrolment Number	Gender	Height
S20200001	F	155
S20200002	F	160
S20200003	M	179
S20200004	F	175

S20200005	M	173
S20200006	M	160
S20200007	M	180
S20200008	F	178
S20200009	F	167
S20200010	M	173

Figure 6: Sample Height Data

Please note that enrolment number variable need not be used in analysis, so no summary data for enrolment number is to be found.

Descriptive of Categorical Data:

The Gender is a categorical variable in Figure 6. The summary in this case would be in terms of frequency table of various categories. For example, for the given data the frequency distribution would be:

Gender	Frequency	Proportion	Percentage
Female (F)	5	0.5	50%
Male (M)	5	0.5	50%

In addition, you can draw bar chart or pie chart for describing the data of Gender variable. The pie chart for such data is shown in Figure 7. Details of different charts are explained in Unit 4. In general, you draw a bar graph, in case the number of categories is more.

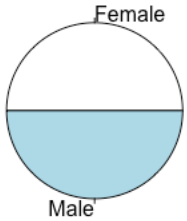


Figure 7: The Pie Chart

Descriptive of Quantitative Data:

The height is a quantitative variable. The descriptive of quantitative data is given by the following two ways:

1. Describing the central tendencies of the data
2. Describing the spread of the data.

Central tendencies of Quantitative data: Mean and Median are two basic measures that define the centre of data though using different ways. They are defined below with the help of an example.

Example 2: Find the mean and median of the following data:

Data Set (n observations)	1	2	3	4	5	6	7	8	9	10	11
x	4	21	25	10	18	9	7	14	11	19	14

The mean can be computed as:

$$\bar{x} = {\Sigma x}/{n}$$

For the given data $\bar{x} = {(4 + 21 + 25 + 10 + 18 + 9 + 7 + 14 + 11 + 19 + 14)}/{11}$

Mean $\bar{x} = 13.82$

The median of the data would be the mid value of the sorted data. First data is sorted and the median is computed using the following formula:

If n is even, then

median = $[(Valueof(\frac{n}{2})^{th})\ position + Valueof((\frac{n}{2}+1)^{th} position]/2$

If *n* is odd, then

median =$(Valueof(\frac{n+1}{2})^{th})\ position$

For this example, the sorted data is as follows:

Data Set (*n* observations)	1	2	3	4	5	6	7	8	9	10	11
x	4	7	9	10	11	14	14	18	19	21	25

So, the median is:

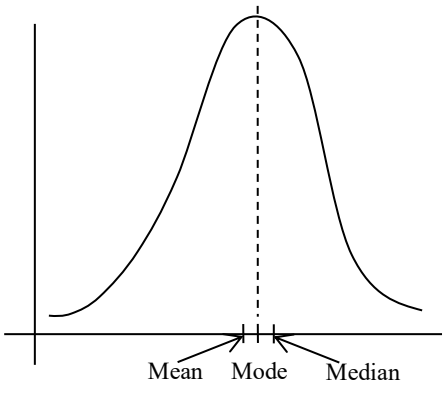
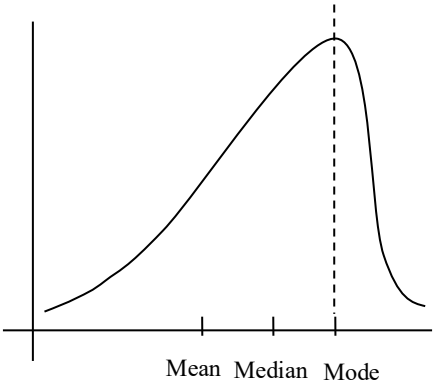
median = $(Valueof(\frac{11+1}{2})^{th})\ position$ = 14

You may please note that outliers, which are defined as values highly different from most other values, can impact the mean but not the median. For example, if one observation in the data, as shown in example 2 is changed as:

Data Set (*n* observations)	1	2	3	4	5	6	7	8	9	10	11
x	4	7	9	10	11	14	14	18	19	21	100

Then the median will still remain 14, however, mean will change to 20.64, which is quite different from the earlier mean. Thus, you should be careful about the presence of outliers while data analysis.

Interestingly, mean and mode may be useful in determining the nature of data. The following table describes these conditions:

Relationship between mean and mode	**Comments about observations**	**A possible Graph of Data Distribution**
Almost Equal values of mean and median	The distribution of data may be symmetric in nature	
Mean >> Median	The distribution may be left-skewed	

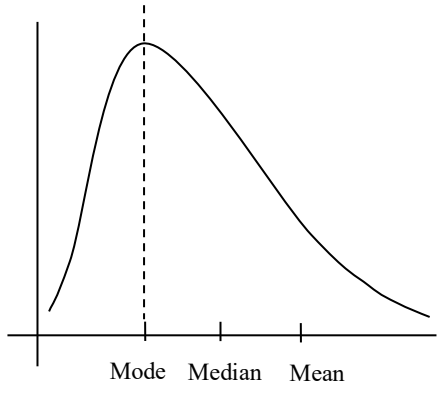
Mean << Median	The distribution may be right-skewed	

Figure 8: Mean and Median for possible data

The concept of data distribution is explained in the next Unit.

Mode: Mode is defined as the most frequent value of a set of observation. For example, in the data of example 2, the value 14, which occurs twice, is the mode. The mode value need not be a mid-value rather it can be any value of the observations. It just communicates the most frequently occurring value only. In a frequency graph, mode is represented by peak of data. For example, in the graphs shown in Figure 8, the value corresponding to the peaks is the mode.

Spread of Quantitative data: Another important aspect of defining the quantitative data is its spread or variability of observed data. Some of the measures for spread of data are given in the Figure 9.

Measure	**Description**	**Example (Please refer to Data of Example 2)**
Range	Minimum to Maximum Value	4 to 25
Variance	Sum of the squares of difference between the observations and its sample mean, which is divided by (n-1), as the difference of n^{th} value can be determined from (n-1) computed difference, as overall sum of differences has to be zero. Formula of Variance for sample is: $$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x - \bar{x})^2$$ However, in case you are determining the Population variance, then you can use to following formula: $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x - \mu)^2$$	Try both the formula and then match the answer: 40.96
Standard Deviation	Standard deviation is one of the most used measure for finding the spread or variability of data. It can be computed as: For Sample:	Try both the formula and then match the answer: 6.4

	$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x - \bar{x})^2}$ For Population: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x - \mu)^2}$	
5-Point Summary and Interquartile Range (IQR)	For creating 5-point summary first, you need to sort the data. The five point summary is defined as follows: Minimum Value (Min) 1st Quartile<=25% values (Q1) 2nd Quartile is median (M) 3rd Quartile is <=75% values (Q3) Maximum Value (Max) IQR is the difference between 3rd and 1st quartiles values.	Use Sorted data of Example 2 Min = 4 Q1=(9+10)/2=9.5 M = 14 Q3=(18+19)/2=18.5 Max=25 IQR= 18.5-9.5=9

Figure 9: The Measure of Spread or Variability

The IQR can also be used to identify suspected outliers. In general, a suspected outlier can exist in the following two ranges:

Observation/values less than Q1 – 1.5×IQR

Observation/values more than Q3+1.5 × IQR

For the example 2,

IQR is 9, therefore the outliers may be: Values < (9.5 – 9) or Values < 0.5.

or Values > (18.5 – 9) or Values > 27.5.

Thus, there is no outlier in the initial data of Example 2.

For the qualitative data, you may draw various plots, such as histogram, box plot etc. These plots are explained in Unit 4 of this block.

Check Your Progress 2

1. Age category of student is a categorical data. What information would you like to show for its descriptive analysis.

2. Age is a quantitative data; how will you describe its data?

3. How can you find that given data is left skewed?

4. What is IQR? Can it be used to find outliers?

1.4.2 Exploratory Analysis

Exploratory data analysis was suggested by John Turkey of Princeton University in 1960, as a group of methods that can be used to learn possibilities of relationships amongst data. After you have obtained relevant data for analysis, instead of performing the final analysis, you may like to explore the data for possible relationships using exploratory data analysis. In general, graphs are some of the best ways to perform exploratory analysis. Some of the common methods that you can perform during exploratory analysis are as follows:

1. As a first step, you may perform the descriptive of various categorical and qualitative variables of your data. Such information is very useful in

determining the suitability of data for the purpose of analysis. This may also help you in data cleaning, modification and transformation of data.

   a. For the qualitative data, you may create frequency tables and bar charts to know the distribution of data among different categories. A balanced distribution of data among categories is most desirable. However, such distribution may not be possible in actual situations. Several methods has been suggested to deal with such situations. Some of those will be discussed in the later units.
   b. For the quantitative data, you may compute the mean, median, standard deviation, skewness and kurtosis. The kurtosis value relates to peaks (determined by mode) in the data. In addition, you may also draw the charts like histogram to look into frequency distribution.

2. Next, after performing the univariate analysis, you may try to perform some bi-variate analysis. Some of the basic statistics that you can perform for bi-variate analysis includes the following:
   a. Make two way table between categorical variables and make related stacked bar charts. You may also use chi-square testing find any significant relationships.
   b. You may draw side-by-side box plots to check if the data of various categories have differences.
   c. You may draw scatterplot and check the correlation coefficient, if that exists between two variables.
3. Finally, you may like to look into the possibilities of multi-variate relationships amongst data. You may use dimensionality reduction by using techniques feature extraction or principle component analysis, you may perform clustering to identify possible set of classes in the solution space or you may use graphical tools, like bubble charts, to visualize the data.

It may be noted that exploratory data analysis helps in identifying the possibilities of relationships amongst data, but does not promises that a causal relationship may exist amongst variables. The causal relationship has to be ascertained through qualitative analysis. Let us explain the exploratory data analysis with the help of an example.

Example 3: Consider the sample data of students given in Figure 6 about Gender and Height. Let us explore this data for an analytical question: Does Height depends on Gender?

You can perform the exploratory analysis on this data by drawing a side-by-side box plot for Male and Female students height. This box plot is shown in Figure 10.

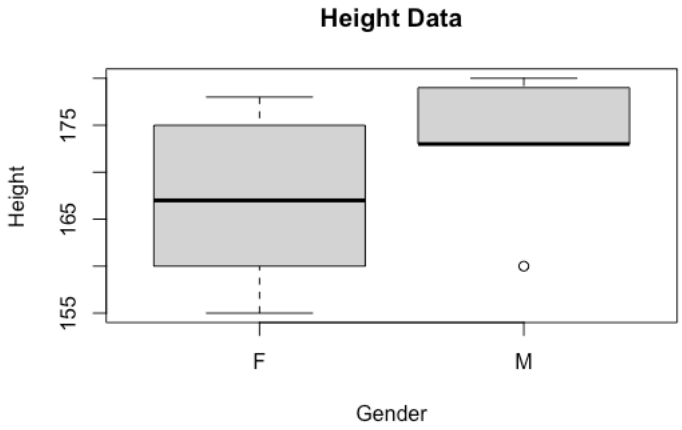


Figure 10: Exploratory Data Analysis

Please note that box plot of Figure 10 shows that on an average height of male students is more than the female student. Does this result applies, in general for the population? For answering this question, you need to find the probability of

occurrence of such sample data. need to determine the probability , therefore, Inferential analysis may need to be performed.

1.4.3 Inferential Analysis

Inferential analysis is performed to answer the question that what is the probability of that the results obtained from an analysis can be applied to the entire population. A detailed discussion on various terms used in inferential analysis in the context of statistical analysis had been done in Unit 2. You can perform many different types of statistical tests on data. Some of these well-known tools for data analysis are listed in the Figure 11.

Test	Why Performed?
Univariate Analysis: Z-Test or T-test	To determine, if the computed value of mean of a sample can be applicable for the population and related confidence interval.
Bivariate Chi-square test	To test the relationship between two categorical variables or groups of data
Two sample T-Test	To test the difference between the means of two variables or groups of data
One way ANOVA	To test the difference in mean of more than two variables or groups of data
F-Test	It can be used to determine the equality of variance of two or more groups of data
Correlation analysis	Determines the strength of relationship between two variables
Regression analysis	Examines the dependence of one variable over a set of independent variables
Decision Trees	Supervised learning used for classification
Clustering	Non-supervised Learning

Figure 11: Some tools for data analysis

You may have read about many of these tests in Data Warehousing and Data Mining and Artificial Intelligence and Machine Leaning course. In addition, you may refer to further readings for these tools. The following example explains the importance of Inferential analysis.

Example 4: Figure 10 in Example 3 shows the box plot of height of male and female students. Can you infer from the boxplot and the sample data (Figure 6), if there is difference in the height of male and female students.

In order to infer, if there is a difference between the hight of two groups (Male and Female Students), a two-sample t-test was run on the data. The output of this t-test is shown in Figure 12.

t-Test (two tail): Assuming Unequal Variances

	Female	*Male*
Mean	167	173
Variance	94.5	63.5
Observations	5	5
Computed t-value	-1.07	
p-value	0.32	
Critical t-value	2.30	

Figure 12: The Output of two sample t-test (two tail)

Figure 12 shows that the mean height of the female students is 167 cm, whereas for the male students it is 173 cm. The variance of female candidates is 94.5, whereas for male candidate it is 63.5. Each group is interpreted on the basis of 5 observations. The computed t-value is -1.07and p-value is 0.32. As the p-value is greater than 0.05, therefore you can conclude that you cannot conclude that the average male student height is different from average female student height.

1.4.4 Predictive Analysis

With the availability of large amount of data and advanced algorithms for mining and analysis of large data have led the way to advanced predictive analysis. The predictive analysis of today uses tools from Artificial Intelligence, Machine Learning, Data Mining, Data Stream Processing, data modelling etc. to make prediction for strategic planning and policies of organisations. Predictive analysis uses large amount of data to identify potential risks and aid the decision-making process. It can be used in several data intensive industries like electronic marketing, financial analysis, healthcare applications, etc. For example, in the healthcare industry, predictive analysis may be used to determine the support for public health infrastructure requirements for the future based on the present health data.
Advancements in Artificial intelligence, data modeling, machine learning has also led to Prescriptive analysis. The prescriptive analysis aims to take predictions one step forward and suggest solutions to present and future issues.
A detailed discussion on these topics is beyond the scope of this Unit. You may refer to further readings for more information on these.

1.5 COMMON MISCONCEPTIONS OF DATA ANALYSIS

In this section, we discuss three misconception that can affect the result of a data science. These misconceptions are explained with the help of an example, only.

Correlation is not Causation: Correlation analysis establishes relationship between two variables. For example, consider three variables, namely attendance of student (attend), marks obtained by student (marks) and weekly hours spent by a student for the study (study). While analysing data, you found that there is a strong correlation between the variables attend and marks. However, does it really mean that higher attendance causes students to obtain better marks? There is another possibility that both study and marks, as well as study and attend are correlated. A motivated student may be spending higher number of hours at home, which may lead to better marks. Similarly, a motivated student who is putting a greater number of hours in his/her study may be attending to school regularly. Thus, the correlation between study to marks and study to attend results in anon-existing correlation attend to marks. This situation is shown in Figure 13.

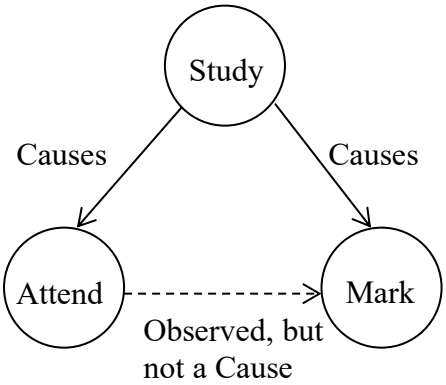


Figure 13: Correlation does not mean causation

Simpsons Paradox: Simson paradox is an interesting situation, which sometimes leads to wrong interpretations. Consider two Universities, say University 1 and University 2 and the pass out data of these Universities:

University	Student Passed	Passed %	Student Failed	Failed %	Total
U1	4850	97%	150	3%	5000
U2	1960	98%	40	2%	2000

Figure 14: The Results of the Universities

As you may observe from the data as above, the University U2 is performing better as far as passing percentage is concerned. However, during a detailed data inspection, it was noted that both the Universities were running Basic Adult Literacy Programme, which in general has a slightly poor result. In addition, the data of the literacy Programme is to be compiled separately. Therefore, the be data for the University would be:

General Programmes:

University	Student Passed	Passed %	Student Failed	Failed %	Total
U1	1480	98.7%	20	3%	1500
U2	1480	98.7%	20	2%	1500

Adult Literacy Programme:

University	Student Passed	Passed %	Student Failed	Failed %	Total
U1	3370	96.3%	130	3.7%	3500
U2	480	96%	20	2%	500

Figure 15: The result after a new grouping is added

You may observe that due to the additional grouping due to adult literacy programme, the corrected data shows that U1 is performing better than U2, as the pass out rate for General programme is same and pass out rate for Adult literacy programme is better from U1. You must note the changes in the percentages. This is the Simpson's paradox.

Data Dredging: Data Dredging, as the name suggest, is extensive analysis of very large data sets. Such analysis results in generation of large number of data associations. Many of those associations may not be casual, thus, requires further exploration through other techniques. Therefore, it is essential that every data association with large data set should be investigated further before reporting them as conclusion of the study.

1.6 APPLICATIONS OF DATA SCIENCE

Data Science is useful in analysing large data sets to produce useful information that can be used for business development and can help in decision making process. This section highlights some of the applications of data science.

Applications using Similarity analysis
These applications use similarity analysis of data using various algorithms, resulting into classification or clustering of data into several categories. Some of these applications can be:

- Spam detection system: This system classifies the emails into spam and non-spam categories. It analyses the IP addresses of mail origin, word patterns used in mails, word frequency etc. to classify a mail as spam or not.
- Financial Fraud detection system: This is one of the major applications for online financial services. Basic principle is once again to classify the transactions as safe or unsafe transactions based on various parameters of the transactions.
- Recommendation of Products: Several e-commerce companies have the data of your buying patterns, information about your searches to their portal and other information about your account with them. This information can be clustered into classes of buyers, which can be used to recommend various products for you.

Applications related to Web Searching
These applications primarily help you in finding content on the web more effectively. Some of the applications in this section would be the search algorithms used by the various search engines. These algorithms attempt to find the good websites based on the search terms. They may use tools related to semantic of your term, indexing of important website and terms, link analysis etc. In addition, the predictive text use of browser is also an example of use of

Applications related to Healthcare System
The data science can be extremely useful for healthcare applications. Some of the applications may involve processing and analysing images for neonatal care or to detect possibilities of tumors, deformities, problems in organs etc. In addition, there can be applications to establishing relationships of diseases to certain factors, creating recommendations for public health based on public health data. Genomic analysis, creation and testing of new drugs etc. The possibilities of use of streaming data for monitoring the patients is also a potential area for use of data science in healthcare.

Applications related to Transport sector
These applications may investigate the possibilities of finding best routes – air, road etc., for example, many e-commerce companies need to plan the most economical ways of logistic support from their warehouses to the customer. Finding the best dynamic route from a source to destination with dynamic load on road networks etc. This application will be required to process the streams of data.

In general, data science can be used for the benefit of society. It should be used creatively to improve the effective resource utilization, which may lead to sustainable development. The ultimate goal of data science applications should be to help us protect our environment and human welfare.

1.7 DATA SCIENCE LIFE CYCLE

So far, we have discussed about various aspects of data science in the previous sections. In this section, we discuss about the life cycle of a data science based application. In general, a data science application development may involve the following stages:

Data Science Project Requirements Analysis Phase
The first and foremost step for data science project would be to identify the objectives of a data science project. This identification of objectives is also coupled with the study of benefits of the project, resource requirements and cost of the project. In addition, you need to make a project plan, which includes project deliverables and associated time frame. In addition, the data that is required to be used for the project is also decided. This phase is similar as that of requirement study and project planning and scheduling.

Data collection and Preparation Phase
In this phase, first all the data sources are identified, followed by designing the process of data collection. It may be noted that data collection may be a continuous process. Once the data sources are identified then data is checked for duplication of data, consistency of data, missing data, and availability timeline of data. In addition, data may be integrated, aggregated or transformed to produce data for a defined set of attributes, which are identified in the requirements phase.

Descriptive data analysis
Next, the data is analysed using univariate and bivariate analysis techniques. This will generate descriptive information about the data. This phase can also be used to establish the suitability and validly of data as per the requirements of data analysis. This is a good time to review your project requirements vis-à-vis collected data characteristics.

Data Modelling and Model Testing
Next, a number of data models based on the data are developed. All these data models are then tested for their validity with test data. The accuracy of various models are compared contrasted and a final model is proposed for data analysis.

Model deployment and Refinement
The tested best model is used to address the data science problem, however, this model must be constantly refined, as the decision making environment keeps changing and new data sets and attributes may change with time. The refinement process goes through all the previous steps again.

Thus, in general, data science project follows a spiral of development. This is shown in Figure 16.

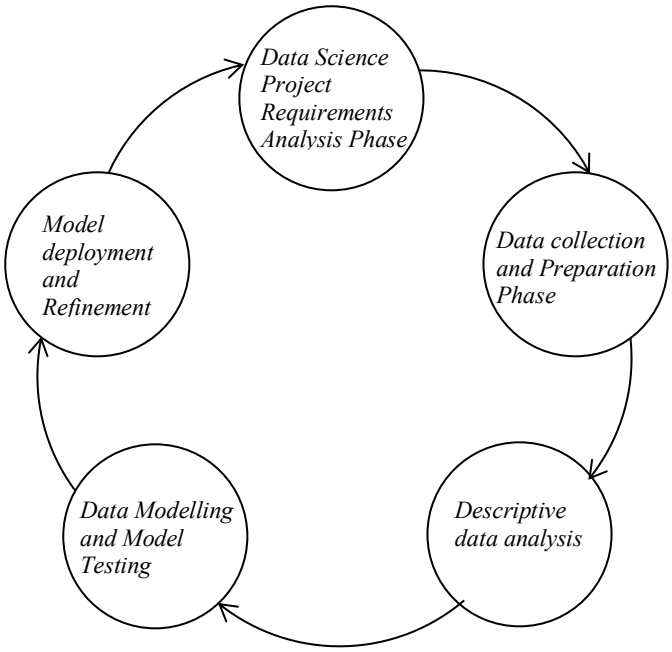


Figure 16: A sample Life cycle of Data Science Project

Check Your Progress 3

1. What are the advantage of using boxplot?

2. How is inferential analysis different to exploratory analysis?

3. What is Simpson's paradox?

1.8 SUMMARY

This Unit introduces basic statistical and analytical concepts of data science. This Unit first introduces you to the definition of the data science. Data science as a discipline uses concepts from computing, mathematics and domain knowledge. The types of data for data science is defined in two different ways. First, it is defined on the basis of structure and generation rate of data, next it is defined as the measures that can be used to capture the data. In addition, the concept of sampling has been defined in this Unit.

This Unit also explains some of the basic methods used for analysis, which includes descriptive, exploratory, inferential and predictive. Few interesting misconceptions related to data science has also been explained with the help of example. This unit also introduces you to some of the applications of data science and data science life cycle. In the ever-advancing technology, it is suggested to keep reading about newer data science applications

1.9 SOLUTIONS/ANSWERS

Check Your Progress 1:

1. Data science integrates the principles of computer science and mathematics and domain knowledge to create mathematical models

that shows relationships amongst data attributes. In addition, data science uses data to perform predictive analysis.

2. Structured data has a defined dimensional structure clearly identified by attributes, for example, tables, data cubes, etc. Semi-structure data has some structure due to use of tags, however, the structure may be flexible, for example, XML data. Unstructured data has no structure at all, like long texts. Data streams on the other hand may be structured, semi-structured or unstructured data that are being produced continuously.
3. Age category would be a categorical data, it will be of ordinal scale, as there are differences among categories, but that difference cannot be defined quantitatively. Weight of the students of a class is ration scale. Grade is also a measure of ordinal scale. 5-point Likert scale is also ordinal data.

Check Your Progress 2:

1. Descriptive of categorical data may include the total number of observations, frequency table and bar or pie chart.
2. The descriptive of age may include mean, median, mode, skewness, kurtosis, standard deviation and histogram or box plot.
3. For left skewed data mean is substantially higher than median and mode.
4. The difference between the Quartile 3 and Quartile 1 is interquartile range (IQR). In general, suspected outliers are at a distance of 1.5 times IQR higher than 3rd quartile or 1.5 times IQR lower than 1st quartile.

Check Your Progress 3:

1. Box plots shows 5-point summary of data. A well spread box plot is an indicator of normally distributed data. Side-by-side box blots can be used to do a comparison of scale data values of two or more categories.
2. Inferential analysis also computes p-value, which determines if the result obtained by exploratory analysis are significant enough, such that results may be applicable for the population.
3. Simpson's paradox signifies that grouped data sometimes statistics may produce results that are contrary to when same statistics is applied to ungrouped data.