

---

# UNIT 1 INTRODUCTION TO DATA SCIENCE

---

- 1.0 Introduction
- 1.1 Objective
- 1.2 Data Science - Definition
- 1.3 Types of Data
  - 1.3.1 Statistical Data Types
  - 1.3.2 Sampling
- 1.4 Basic Methods of Data Analysis
  - 1.4.1 Descriptive Analysis
  - 1.4.2 Exploratory Analysis
  - 1.4.3 Inferential Analysis
  - 1.4.4 Predictive Analysis
- 1.5 Common Misconceptions of Data Analysis
- 1.6 Applications of Data Science
- 1.7 Data Science Life cycle
- 1.8 Summary
- 1.9 Solutions/Answers

---

## 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

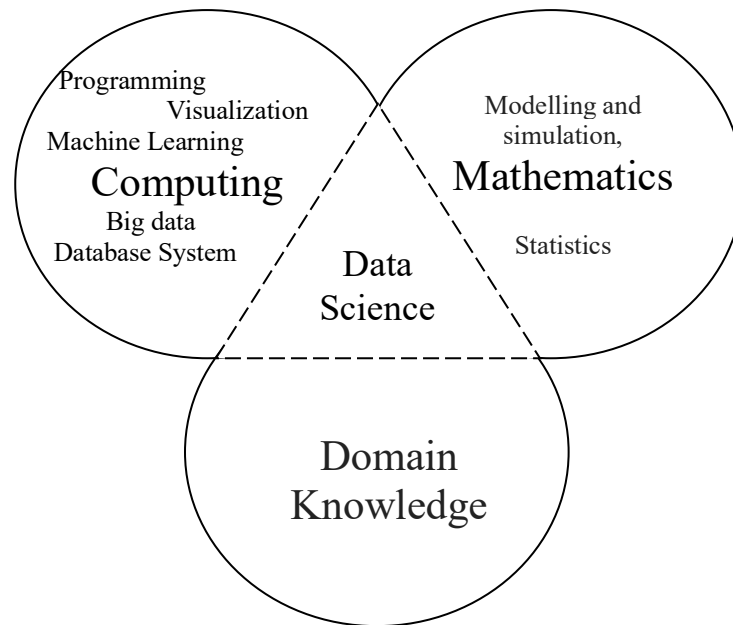
---

## 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 **Common Business-Oriented 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.

<pre>&lt;Book&gt;   &lt;title&gt;Data Science and Big Data&lt;/title&gt;   &lt;author&gt;R Raman&lt;/author&gt;   &lt;author&gt;C V Shekhar&lt;/author&gt;   &lt;yearofpublication&gt;2020&lt;/yearofpublicatio n&gt; &lt;/Book&gt;</pre>	<pre>"Book": {   "Title":  "Data Science",   "Price":   5000,    "Year":   2020 }</pre>
---	---

**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
	<i>IDV</i>	<i>M</i>	<i>EI</i>	<i>MZV</i>	
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.

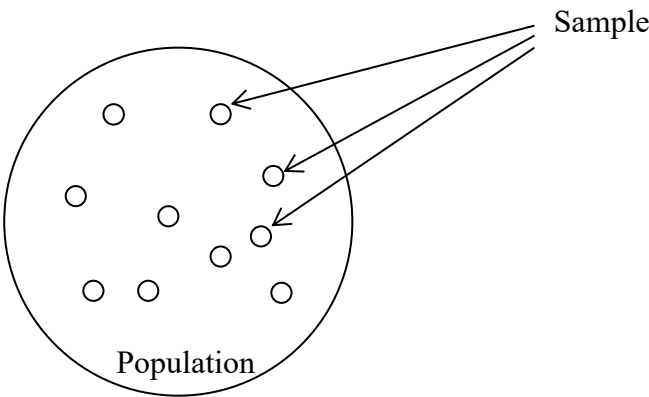


Figure 5: Population and Sample

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 ( $\mu$ ), Population Standard deviation ( $\sigma$ ), 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.

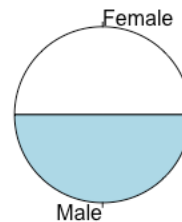


Figure 7: The Pie Chart

#### *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:

$$\bar{x} = \frac{\sum x}{n}$$

For the given data  $\bar{x} =$

$$(4 + 21 + 25 + 10 + 18 + 9 + 7 + 14 + 11 + 19 + 14) / 11$$

$$\text{Mean } \bar{x} = 13.82$$

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



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

If  $n$  is odd, then

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

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:

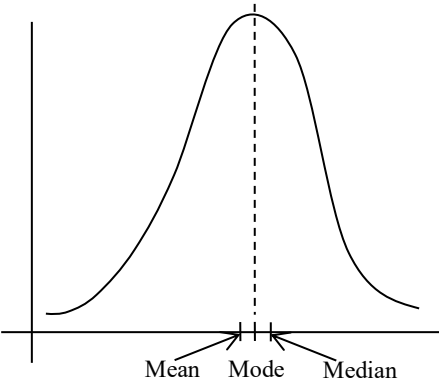
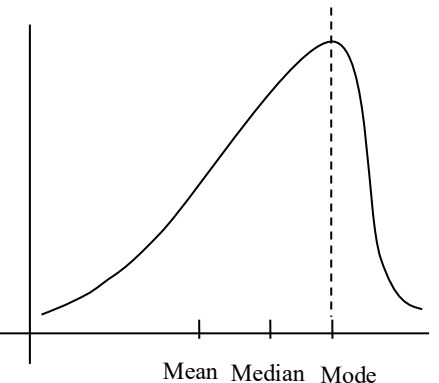
$$\text{median} = (Valueof(\frac{11+1}{2})^{th}) position = 14$$

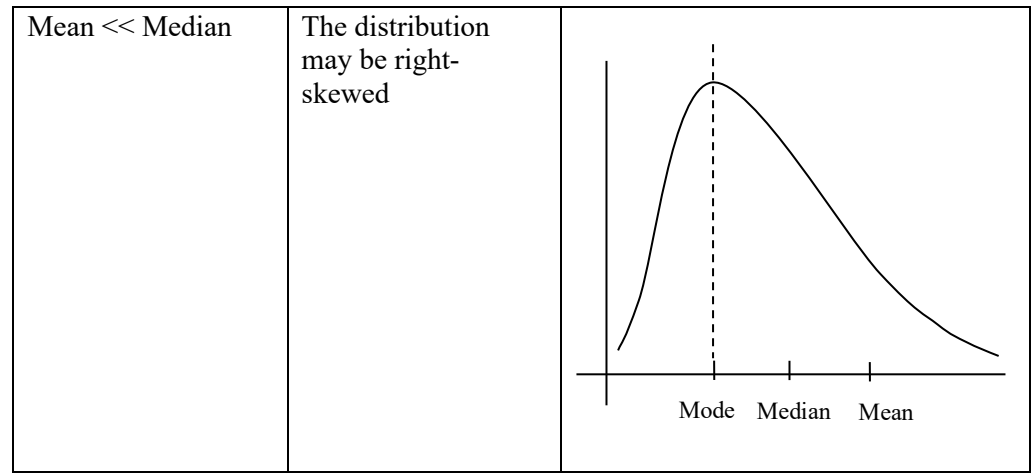
You may please note that outliers, which are defined as values highly different from most other values, can impact the mean but not the median. For example, if one observation in the data, as shown in example 2 is changed as:

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	



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

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

**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  $Q1 - 1.5 \times IQR$

Observation/values more than  $Q3 + 1.5 \times IQR$

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

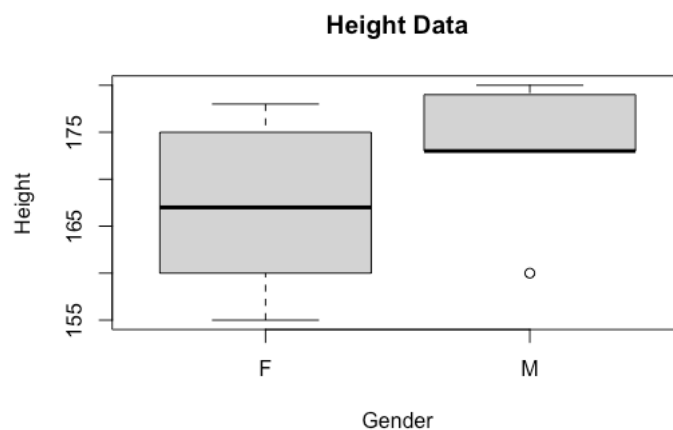
determining the suitability of data for the purpose of analysis. This may also help you in data cleaning, modification and transformation of data.

- 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 Learning 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 height 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

	<i>Female</i>	<i>Male</i>
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 -1.07 and 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.

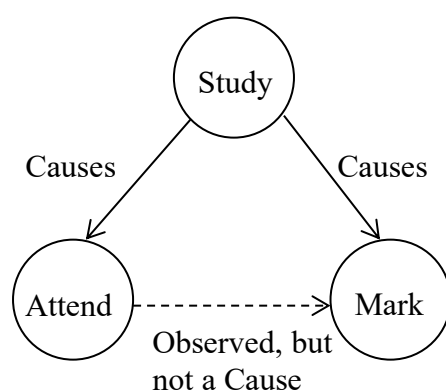
---

## 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 a non-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 validity 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.

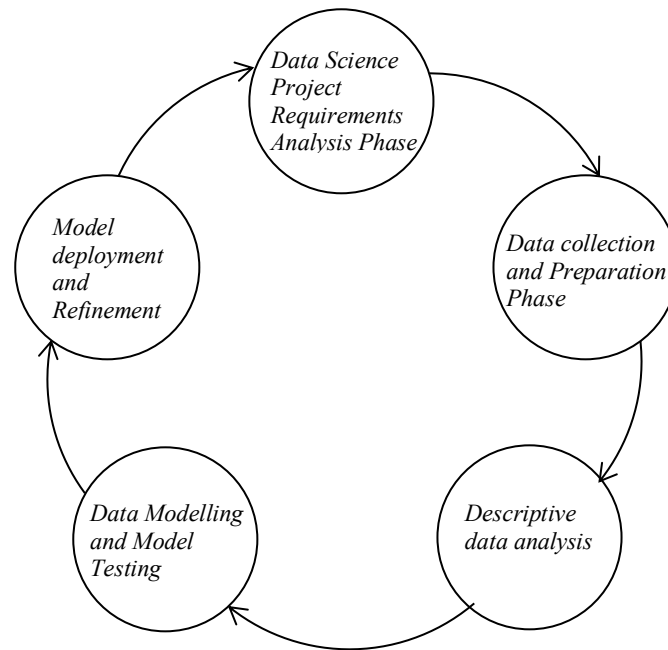
### *Data Modelling and Model Testing*

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 3<sup>rd</sup> quartile or 1.5 times IQR lower than 1<sup>st</sup> 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.

---

## UNIT 2    PORTABILITY AND STATISTICS FOR DATA SCIENCE

---

- 2.0 Introduction
- 2.1 Objectives
- 2.2 Probability
  - 2.2.1 Conditional Probability
  - 2.2.2 Bayes Theorem
- 2.3 Random Variables and Basic Distributions
  - 2.3.1 Binomial Distribution
  - 2.3.2 Probability Distribution of Continuous Random Variable
  - 2.3.3 The Normal Distribution
- 2.4 Sampling Distribution and the Central Limit Theorem
- 2.5 Statistical Hypothesis Testing
  - 2.5.1 Estimation of Parameters of the Population
  - 2.5.2 Significance Testing of Statistical Hypothesis
  - 2.5.3 Example using Correlation and Regression
  - 2.5.4 Types of Errors in Hypothesis Testing
- 2.6 Summary
- 2.7 Solution/Answers

---

### 2.0 INTRODUCTION

---

In the previous unit of this Block, you were introduced to the basic concepts of data science, which include the basic types of data, basic methods of data analysis and applications and life cycle of data science. This Unit introduces you to the basic concepts related to probability and statistics related to data science.

It introduces the concept of conditional probability and Bayes Theorem. It is followed by discussion on the basic probability distribution, highlighting their significance and use. These distributions includes Binomial and Normal distributions, the two most used distributions from discrete and continuous variables respectively. The Unit also introduces you to the concept of sampling distribution and central limit theorem. Finally, this unit covers the concepts of statistical hypothesis testing with the help of an example of correlation. You may refer to further readings for more details on these topics, if needed.

---

### 2.1 OBJECTIVES

---

After going through thus unit, you should be able to:

- Compute the conditional probability of events;
- Use Bayes theorem in problem solving
- Explain the concept of random variable;
- Explain the characteristics of binomial and normal distribution;
- Describes the sampling distribution and central limit theorem;
- State the statistical hypothesis; and
- Perform significance testing.

## 2.2 PROBABILITY

Probability is a possible measure of occurrence of a specific event amongst a group of events, if the occurrence of the events is observed for a large number of trials. For example, possibility of getting 1, while rolling a fair die is  $1/6$ . You may please note that you need to observe this event by repeatedly rolling a die for a large number of trials to arrive at this probability or you may determine the probability subjectively by finding the ratio of this outcome and the total number of possible outcomes, which may be equally likely. Thus, the probability of an event (E) can be computed using the following formula:

$$P(E) = \frac{\text{Number of outcomes in the set of all possible outcomes that result in event } E}{\text{Number outcomes in the set of all possible outcomes}} \quad (1)$$

In the equation above, the set of all possible outcomes is also called sample space. In addition, it is expected that all the outcomes are equally likely to occur.

Consider that you decided to roll two fair dice together at the same time. Will the outcome of first die will affect the outcome of the second die? It will not, as both the outcomes are independent of each other. In other words both the trials are independent, if the outcome of the first trial does not affect the outcome of second trial and vice-versa; else the trials are dependent trials.

How to compute the probability for more than one events in a sample space. Let us explain this with the help of example.

**Example 1:** A fair die having six equally likely outcomes is to be thrown, then:

- (i) What is the sample space:  $\{1, 2, 3, 4, 5, 6\}$
- (ii) An Event A is die shows 2, then outcome of event A is  $\{2\}$ ; and probability  $P(A)=1/6$
- (iii) An Event B is die shows odd face, then Event B is  $\{1, 3, 5\}$ ; and probability of Event B is  $P(B) = 3/6 = 1/2$
- (iv) An Event C is die shows even face, then Event C is  $\{2, 4, 6\}$ ; and probability of Event C is  $P(C) = 3/6 = 1/2$
- (v) Event A and B are disjoint events, as no outcomes between them is common. So are the Event B and C. But event A and C are not disjoint.
- (vi) Intersection of Events A and B is a null set  $\{\}$ , as they are disjoint events, therefore, probability that both events A and B both occur, viz.  $P(A \cap B) = 0$ . However, intersection of A and C is  $\{2\}$ , therefore,  $P(A \cap C) = 1/6$ .
- (vii) The union of the Events A and B would be  $\{1, 2, 3, 5\}$ , therefore, the probability that event A or event B occurs, viz.  $P(A \cup B) = 4/6 = 2/3$ . Whereas, the Union of events B and C is  $\{1, 2, 3, 4, 5, 6\}$ , therefore,  $P(B \cup C) = 6/6 = 1$ .

Please note that the following formula can be derived from the above example. Probability of occurrence of any of the two events A or B (also called union of events) is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (2)$$

For example 1, you may compute the probability of occurrence of event A or C as:

$$\begin{aligned} P(A \cup C) &= P(A) + P(C) - P(A \cap C) \\ &= 1/6 + 1/2 - 1/6 = 1/2. \end{aligned}$$

In the case of **disjoint events**, since  $P(A \cap B)$  is zero, therefore, the equation (2) will reduce to:

$$P(A \cup B) = P(A) + P(B) \quad (3)$$

*Probability of Events in independent trials:*

This is explained with the help of the following example.

**Example 2:** A fair die was thrown twice, what is the probability of getting a 2 in the first throw and 4 or 5 in the second throw.

The probability of getting a 2 in the first throw (say event X) is  $P(X) = 1/6$

The probability of getting {4, 5} in the second throw (say event Y) is  $P(Y) = 2/6$ .

Both these events are independent of each other, therefore, you need to use the formula for *intersection of independent events*, which is:

$$P(X \cap Y) = P(X) \times P(Y) \quad (4)$$

Therefore, the probability  $P(X \cap Y) = \frac{1}{6} \times \frac{2}{6} = \frac{1}{18}$

This rule is applicable even with more than two independent events. However, this rule will not apply if the events are not independent.

### 2.2.1 Conditional Probability

Conditional probability is defined for the probability of occurrence of an event, when another event has occurred. Conditional probability addresses the following question.

Given two events X and Y with the probability of occurrences  $P(X)$  and  $P(Y)$  respectively. What would be the probability of occurrence of X if the other event Y has actually occurred?

Let us analyse the problem further. Since the event Y has already occurred, therefore, sample space reduces to the sample space of event Y. In addition, the possible outcomes for occurrence of X could be the occurrences at the intersection of X and Y, as that is the only space of X, which is part of sample space of Y. Figure 1 shows this with the help of a Venn diagram.

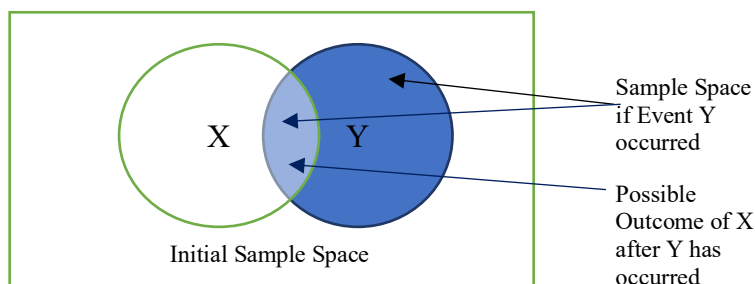


Figure 1: The conditional Probability of event A given that event B has occurred

You can compute the conditional probability using the following equation.

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)} \quad (5)$$

Where  $P(X/Y)$  is the conditional probability of occurrence of event X, if event Y has occurred.

For example, in example 1, what is the probability of occurrence of event A, if event C has occurred?

You may please note that  $P(A \cap C) = 1/6$  and  $P(C) = 1/2$ , therefore, the conditional probability  $P(A/C)$  would be:

$$P(A/C) = \frac{P(A \cap C)}{P(C)} = \frac{1/6}{1/2} = 1/3$$

What would be conditional probability of disjoint events? You may find the answer, by computing the  $P(A/B)$  for the Example 1.

What would be the conditional probability for Independent events?

The equation (5) of conditional probability can be used to derive a very interesting results, as follows:

You can rewrite equation (5) as,

$$P(X \cap Y) = P(X/Y) \times P(Y) \quad (5a)$$

Similarly, you can rewrite equation (5) for  $P(Y/X)$  as,

$$P(Y/X) = \frac{P(X \cap Y)}{P(X)} \quad \text{or} \quad P(X \cap Y) = P(Y/X) \times P(X) \quad (5b)$$

Using equation 5(a) and equation 5(b) you can conclude the following:

$$P(X \cap Y) = P(X/Y) \times P(Y) = P(Y/X) \times P(X) \quad (6)$$

Independent events are a special case for the conditional probability. As the two events are independent of each other, therefore, occurrence of the any one of the event does not change the probability or occurrence of the second event. Therefore, for independent events X and Y

$$P(X/Y) = P(X) \text{ and } P(Y/X) = P(Y) \quad (7)$$

In fact, the equation (7) can be used to determine the independent events

## 2.2.2 Bayes Theorem

Bayes theorem is one of the important theorem, which deals with the conditional probability. Mathematically, Bayes theorem can be written using equation (6) as,

$$\begin{aligned} P(X/Y) \times P(Y) &= P(Y/X) \times P(X) \\ \text{Or } P(X/Y) &= \frac{P(Y/X) \times P(X)}{P(Y)} \end{aligned} \quad (8)$$

**Example 3:** Assume that you have two bags namely Bag A and Bag B. Bag A contains 5 green and 5 red balls; whereas, Bag B contains 3 green and 7 red balls. Assume that you have drawn a red ball, what is the probability that this red ball is drawn from Bag B.

In this example,

Let the event X be “Drawing a Red Ball”. The probability of drawing a red ball can be computed as follows;

You may select a Bag and then draw a ball. Therefore, the probability will be computed as:

(Probability of selection of Bag A)  $\times$  (Probability of selection of red ball in Bag A) + (Probability of selection of Bag B)  $\times$  (Probability of selection of red ball in Bag B)

$$P(\text{Red}) = (1/2 \times 5/10 + 1/2 \times 7/10) = 3/5$$

Let the event Y be “Selection of Bag B from the two Bags, assuming equally likely selection of Bags. Therefore,  $P(\text{BagB}) = 1/2$ .

In addition, if Bag B is selected then the probability of drawing Red ball  $P(\text{Red}/\text{BagB}) = 7/10$ , as Bag B has already been selected and it has 3 Green and 7 Red balls.

As per the Bayes Theorem:

$$P(\text{BagB}/\text{Red}) = \frac{P(\text{Red}/\text{BagB}) \times P(\text{BagB})}{P(\text{Red})}$$

$$P(\text{BagB}/\text{Red}) = \frac{\frac{7}{10} \times \frac{1}{2}}{\frac{3}{5}} = \frac{7}{12}$$

Bayes theorem is a powerful tool to revise your estimate provided a given event has occurred. Thus, you may be able to change your predictions.

### Check Your Progress 1

1. Is  $P(Y/X) = P(Y/X)$ ?
2. How can you use probabilities to find, if two events are independent.
3. The MCA batches of University A and University B consists of 20 and 30 students respectively. University A has 10 students who have obtained more than 75% marks and University B has 20 such students. A recruitment agency selects one of these student who has more than 75% marks out of the two Universities. What is the probability that the selected student is from University A?

## 2.3 RANDOM VARIABLES AND BASIC DISTRIBUTIONS

In statistics, in general, you perform random experiments to study particular characteristics of a problem situation. These random experiments, which are performed in almost identical experimental setup and environment, determine the attributes or factors that may be related to the problem situation. The outcome of these experiments can take different values based on the probability and are termed as the random variables. This section discusses the concept of random variables.

Example 4: Consider you want to define the characteristics of random experiment toss of the coin, say 3 tosses, you selected an outcome “Number of Heads” as your variable, say  $X$ . You may define the possible outcomes of sample space for the tosses as:

Outcomes	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Number of Heads ( $X$ )	3	2	2	1	2	1	1	0

Figure 2: Mapping of outcomes of sample space to Random variable.

Using the data of Figure 2, you can create the following frequency table, which can also be converted to probability.



$X$	Frequency	Probability $P(X)$
0	1	$1/8$
1	3	$3/8$
2	3	$3/8$
3	1	$1/8$
Total	8	Sum of all $P(X) = 1$

Figure 3: The Frequency and Probability of Random Variable  $X$

The Random variables are of two kinds:

- Discrete Random Variable
- Continuous Random Variable

Discrete random variables, as the name suggests, can take discrete values only. Figure 3 shows a discrete random variable  $X$ . A discrete random variable, as a convention, may be represented using a capital alphabets. The individual values are represented using lowercase alphabet, e.g., for the discrete variable  $X$  of Figure 3, the discrete values are  $x_0, x_1, x_2$  and  $x_3$ . Please note that their values are also 0, 1, 2 and 3 respectively. Similarly, to represent individual probability, you may use the value names  $p_0, p_1, p_2$  and  $p_3$ . Please also note that the sum of all these probabilities is 1, e.g. in Figure 3,  $p_0 + p_1 + p_2 + p_3 = 1$ .

#### *Probability Distribution of Discrete Random Variable*

For the discrete random variable  $X$ , which is defined as the number of head in three tosses of coin, the pair  $(x_i, p_i)$ , for  $i=0$  to 3, defines the probability distribution of the random variable  $X$ . Similarly, you can define the probability distribution for any discrete random variable. The probability distribution of a discrete random variable has two basic properties:

- The  $p_i$  should be greater than or equal to zero, but always less than or equal to 1.
- The sum of all  $p_i$  should be 1.

Figure 4 shows the probability distribution of  $X$  (the number of heads in three tosses of coin) in graphical form.

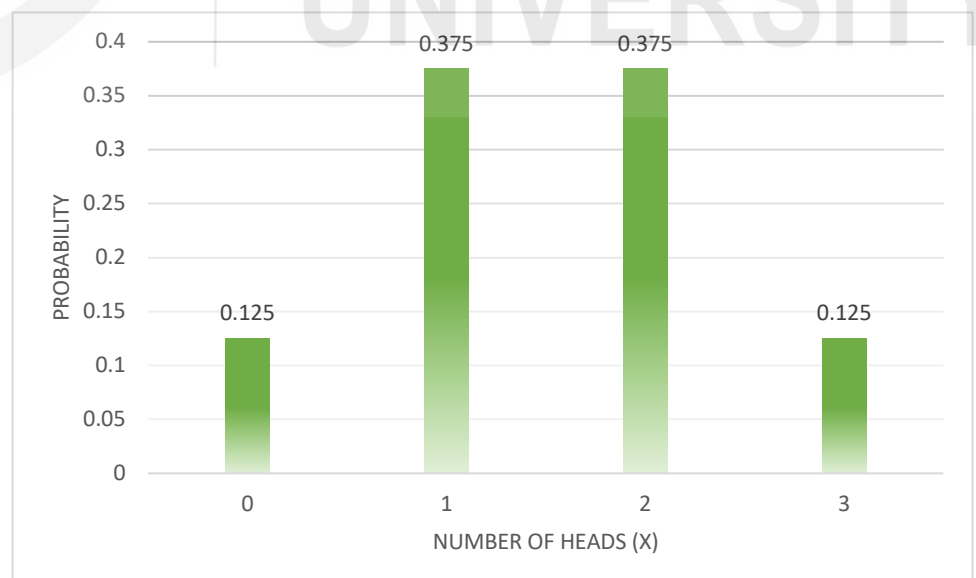


Figure 4: Probability Distribution of Discrete Random Variable  $X$  (Number of heads in 3 tosses of coin)

Another important value defined in probability distribution is the mean or expected value, which is computed using the following equation (9) for random variable X:

$$\mu = \sum_{i=0}^n x_i \times p_i \quad (9)$$

Thus, the mean or expected number of heads in three trials would be:

$$\begin{aligned} \mu &= x_0 \times p_0 + x_1 \times p_1 + x_2 \times p_2 + x_3 \times p_3 \\ \mu &= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8} = 1.5 \end{aligned}$$

Therefore, in a trail of 3 tosses of coins, the mean number of heads is 1.5.

### 2.3.1 Binomial Distribution

Binomial distribution is a discrete distribution. It shows the probability distribution of a discrete random variable. The Binomial distribution involves an experiment involving Bernoulli trials, which has the following characteristics:

- A number of trials are conducted, say  $n$ .
- There can be only two possible outcomes of a trial – Success(say  $s$ ) or Failure (say  $f$ ).
- Each trial is independent of all the other trials.
- The probability of the outcome Success ( $s$ ), as well as failure ( $f$ ), is same in each and every independent trial.

For example, in the experiment of tossing three coins, the outcome success is getting a head in a trial. One possible outcome for this experiment is THT, which is one of outcome of the sample space shown in Figure 2.

You may please note that in case of  $n=3$ , the for the random variable X, which represents the number of heads, the success is getting a Heads, while failure is getting a Tails. Thus, THT is actually Failure, Success, Failure. The probability for such cases, thus, can be computed as shown earlier. In general, in Binomial distribution, the probability of  $r$  successes is represented as:

$$P(X = r) \text{ or } p_r = {}^nC_r \times s^r \times f^{n-r} \quad (10)$$

Where  $s$  is the probability of success and  $f$  is the probability of failure in each trial. The value of  ${}^nC_r$  is computed using the combination formula:

$${}^nC_r = \frac{n!}{r!(n-r)!} \quad (11)$$

For the case of three tosses of the coins, where X is represented as number of heads in the three tosses of coins  $n = 3$  and both  $s$  and  $f$  are  $1/2$ , the probability as per Binomial Distribution would be:

$$P(X = 0) \text{ or } p_0 = {}^3C_0 \times s^0 \times f^{3-0} = \frac{3!}{0!(3-0)!} \times \left(\frac{1}{2}\right)^0 \times \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

$$P(X = 1) \text{ or } p_1 = {}^3C_1 \times s^1 \times f^{3-1} = \frac{3!}{1!(3-1)!} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^2 = \frac{3}{8}$$

$$P(X = 2) \text{ or } p_2 = {}^3C_2 \times s^2 \times f^{3-2} = \frac{3!}{2!(3-2)!} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^1 = \frac{3}{8}$$

$$P(X = 3) \text{ or } p_3 = {}^3C_3 \times s^3 \times f^{3-3} = \frac{3!}{3!(3-3)!} \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^0 = \frac{1}{8}$$

Which is same as Figure 2 and Figure 3.

Finally, the mean and standard deviation of Binomial distribution for  $n$  trials, each having a probability of success as  $s$ , can be defined using the following formulas:

$$\mu = n \times s \quad (12a)$$

$$\sigma = \sqrt{n \times s \times (1 - s)} \quad (12b)$$

Therefore, for the variable  $X$  which represents number of heads in three tosses of coin, the mean and standard deviation are:

$$\mu = n \times s = 3 \times \frac{1}{2} = 1.5$$

$$\sigma = \sqrt{n \times s \times (1 - s)} = \sqrt{3 \times \frac{1}{2} \times (1 - \frac{1}{2})} = \frac{\sqrt{3}}{2}$$

Distribution of a discrete random variable, thus, is able to compute the probability of occurrence of specific number successes, as well as the mean or expected value of a random probability experiment.

### 2.3.2 Probability Distribution of Continuous Random Variable

A continuous variable is measured using scale or interval measures. For example, height of the students of a class can be measured using an interval measure. You can study the probability distribution of a continuous random variable also, however, it is quite different from the discrete variable distribution. Figure 5 shows a sample histogram of the height of 100 students of a class. You may please notice it is typically a grouped frequency distribution.

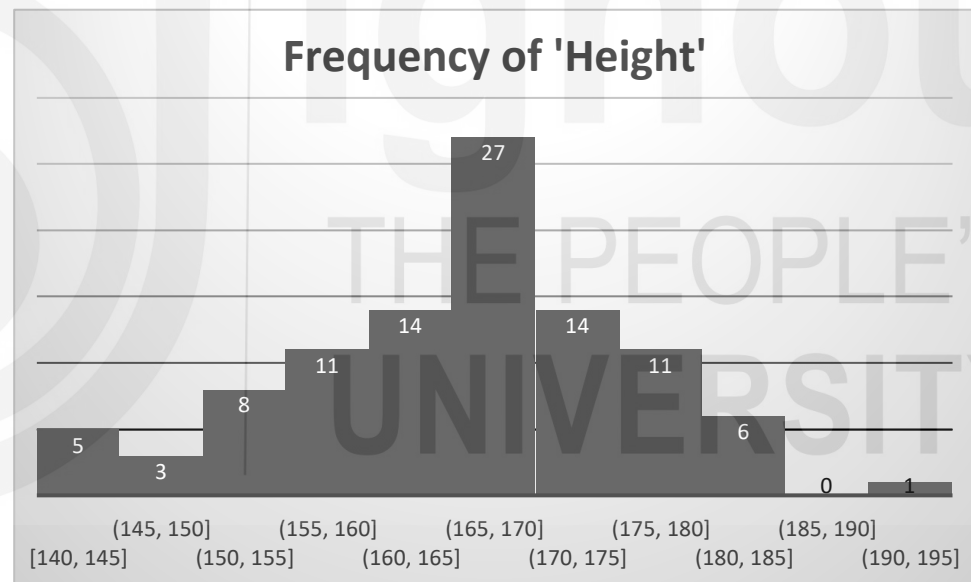


Figure 5: Histogram of Height of 100 students of a Class

The mean of the height was 166 and the standard distribution was about 10. The probability for a student height is in between 165 to 170 interval is 0.27.

In general, for large data set continuous random variable distribution is represented as a smooth curve, which has the following characteristics:

- The probability in each interval would be between 0 and 1. To compute the probability in an interval you need to compute the area of the curve between the starting and end points of that interval.
- The total area of the curve would be 1.

### 2.3.3 The Normal Distribution

An interesting frequency distribution of continuous random variable is the Normal Distribution, which was first demonstrated by a German Scientist C.F.

Gauss. Therefore, it is sometime also called the Gaussian distribution. The Normal distribution has the following properties:

- The normal distribution can occur in many real life situations, such as height distribution of people, marks of students, intelligence quotient of people etc.
- The curve looks like a bell shaped curve.
- The curve is symmetric about the mean value ( $\mu$ ). Therefore, about half of the probability distribution curve would lie towards the left of the mean and other half would lie towards the right of the mean.
- If the standard deviation of the curve is  $\sigma$ , then about 68% of the data values would be in the range  $(\mu-\sigma)$  to  $(\mu+\sigma)$  (Refer to Figure 6)
- About 95% of the data values would be in the range  $(\mu-2\sigma)$  to  $(\mu+2\sigma)$  (Refer to Figure 6)
- About 99.7% of the data values would be in the range  $(\mu-3\sigma)$  to  $(\mu+3\sigma)$  (Refer to Figure 6).
- Skewness and Kurtosis of normal distribution is closer to zero.
- The probability density of standard normal distribution is represented using a mathematical equation using parameters  $\mu$  and  $\sigma$ . You may refer to the equation in the further readings.

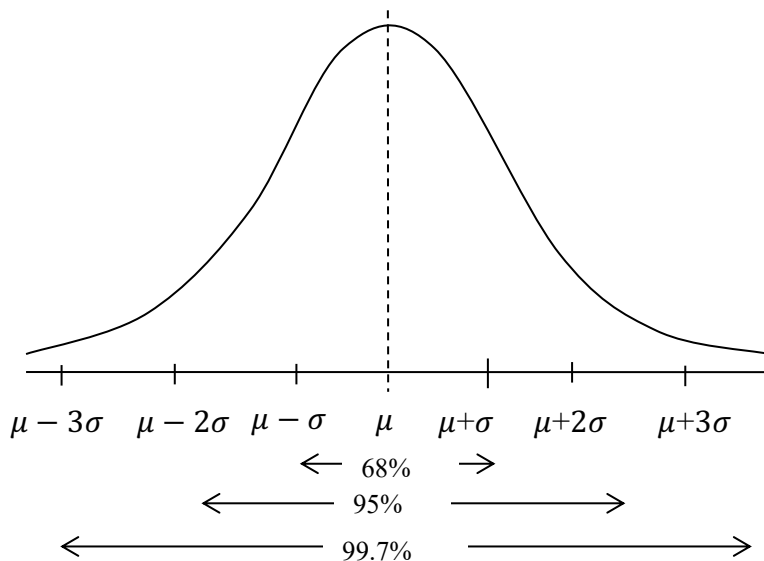


Figure 6: Normal Distribution of Data

#### Computing probability using Normal Distribution:

The Normal distribution can be used to compute the z-score, which computes the distance of a value  $x$  from its mean in terms of its standard deviation.

For a given continuous random variable  $X$  and its value  $x$ ; and normal probability distribution with parameters  $\mu$  and  $\sigma$ ; the z-score would be computed as:

$$z = \frac{(x-\mu)}{\sigma} \quad (13)$$

You can find the cumulative probabilities at a particular z-value using Normal distribution, for example, the shaded portion of the Figure 7 shows the cumulative probabilities at  $z = 1.3$ , the probability of the shaded portion at this point is 0.9032

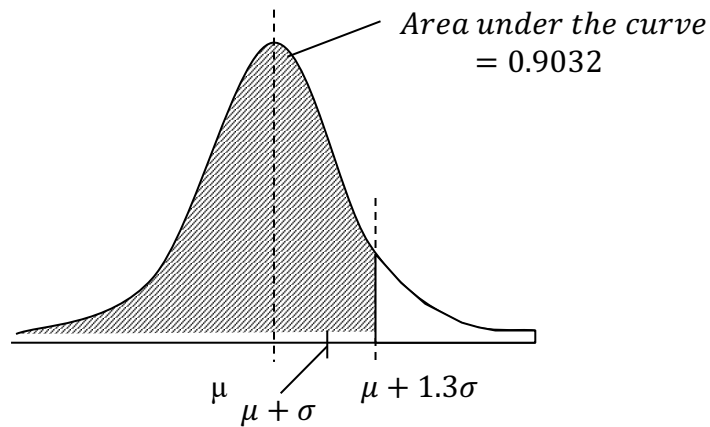


Figure 7: Computing Probability using Normal Distribution

*Standard Normal Distribution* is a standardized form of normal distribution, which allows comparison of various different normal curves. A standard normal curve would have the value of mean ( $\mu$ ) as zero and standard deviation ( $\sigma$ ) as 1. The z-score for standard normal distribution would be:

$$z = \frac{(x-0)}{1} = x$$

Therefore, for standard normal distribution the z-score is same as value of  $x$ . This means that  $z = \pm 2$  contains the 95% area under the standard normal curve.

In addition to Normal distribution a large number of probability distributions have been studied. Some of these distributions are – Poisson distribution, Uniform Distribution, Chi-square distribution etc. Each of these distribution is represented by a characteristics equation involving a set of parameters. A detailed discussion on these distributions is beyond the scope of this Unit. You may refer to Further Reading for more details on these distributions.

---

## 2.4 SAMPLING DISTRIBUTION AND THE CENTRAL LIMIT THEOREM

---

With the basic introduction, as above, next we discuss one of the important aspect of sample and population called sampling distribution. A typical statistical experiment may be based on a specific sample of data that may be collected by the researcher. Such data is termed as the primary data. The question is – Does the statistical results obtained by you using the primary data can be applied to the population? If yes, what may be the accuracy of such a collection? To answer this question, you must study the sampling distribution. Sampling distribution is also a probability distribution, however, this distribution shows the probability of choosing a specific sample from the population. In other words, a sampling distribution is the probability distribution of means of the random samples of the population. The probability in this distribution defines the likelihood of the occurrence of the specific mean of the sample collected by the researcher. Sampling distribution determines whether the statistics of the sample falls closer to population parameters or not. The following example explains the concept of sampling distribution in the context of a categorical variable.

Example 5: Consider a small population of just 5 person, who vote for a question “Data Science be made the Core Course in Computer Science? (Yes/No)”. The following table shows the population:

P1	P2	P3	P4	P5	Population Parameter (proportion) (p)
Yes	Yes	No	No	No	0.4

Figure 8: A hypothetical population

Suppose, you take a sample size ( $n$ ) = 3, and collects random sample. The following are the possible set of random samples:

Sample	Sample Proportion ( $\hat{p}$ )
P1, P2, P3	0.67
P1, P2, P4	0.67
P1, P2, P5	0.67
P1, P3, P4	0.33
P1, P3, P5	0.33
P1, P4, P5	0.33
P2, P3, P4	0.33
P2, P3, P5	0.33
P2, P4, P5	0.33
P3, P4, P5	0.00

Frequency of all the sample proportions is:

$\hat{p}$	Frequency
0	1
0.33	6
0.67	3

Figure 9: Sampling proportions

The mean of all these sample proportions =  $(0 \times 1 + 0.33 \times 6 + 0.67 \times 3) / 10$   
 $= 0.4$  (ignoring round off errors)

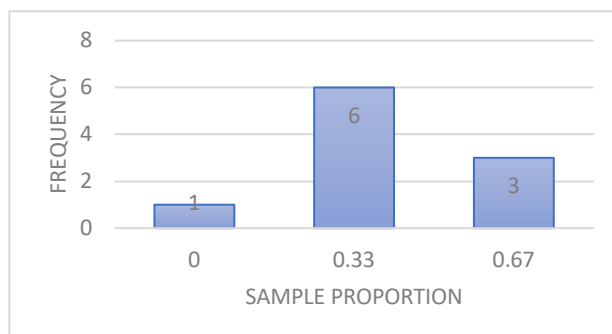


Figure 10: The Sampling Proportion Distribution

Please notice the nature of the sampling proportions distribution, it looks closer to Normal distribution curve. In fact, you can find that out by creating an example with 100 data points and sample size 30.

Given a sample size  $n$  and parameter proportion  $p$  of a particular category, then the sampling distribution for the given sample size would fulfil the following:

$$\text{mean proportion} = p \quad (14a)$$

$$\text{Standard Deviation} = \sqrt{\frac{p \times (1-p)}{n}} \quad (14b)$$

Let us extend the sampling distribution to interval variables. Following example explains different aspects sampling distribution:

Example 6: Consider a small population of age of just 5 person. The following table shows the population:

P1	P2	P3	P4	P5	Population mean ( $\mu$ )
20	25	30	35	40	30

Figure 8: A hypothetical population

Suppose, you take a sample size ( $n$ ) = 3, and collects random sample. The following are the possible set of random samples:

Sample	Sample Mean ( $\bar{x}$ )
P1, P2, P3	25
P1, P2, P4	26.67
P1, P2, P5	28.33
P1, P3, P4	28.33
P1, P3, P5	30
P1, P4, P5	31.67
P2, P3, P4	30
P2, P3, P5	31.67
P2, P4, P5	33.33
P3, P4, P5	35

Figure 11: Mean of Samples

The mean of all these sample means = 30, which is same as population mean  $\mu$ . The histogram of the data is shown in Figure 12.

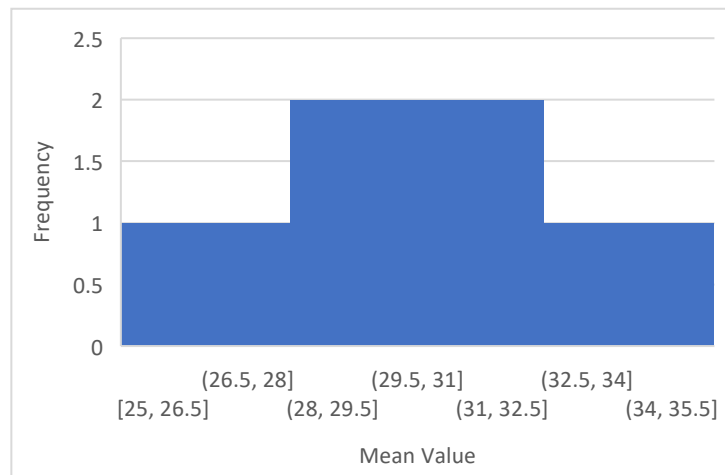


Figure 12: Frequency distribution of sample means

Given a sample size  $n$  and population mean  $\mu$ , then the sampling distribution for the given sample size would fulfil the following:

$$\text{Mean of sample means} = \mu \quad (15a)$$

$$\text{Standard Deviation of Sample Means} = \frac{\sigma}{\sqrt{n}} \quad (15b)$$

Therefore, the z-score computation for sampling distribution will be as per the following equation:

Note: You can obtain this equation from equation (13), as this is a distribution of means, therefore,  $x$  of equation (13) is  $\bar{x}$ , and standard deviation of sampling distribution is given by equation (15b).

$$z = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}} \quad (15c)$$

Please note that the histogram of the mean of samples is close to normal distribution.

Such exponentiations led to the Central limit Theorem, which proposes the following: *Central Limit Theorem*: Assume that a sample of size is drawn from a population that has the mean  $\mu$  and standard deviation  $\sigma$ . The central limit theorem states that with the increase in  $n$ , the sampling distribution, i.e. the distribution of mean of the samples, approaches closer to normal distribution.

However, it may be noted that the central limit theorem is applicable only if you have collected independent random samples, where the size of sample is sufficiently large, yet it is less than 10% of the population. Therefore, the Example 5 and Example 6 are not true representations for the theorem, rather are given to illustrate the concept. Further, it may be noted that the central limit theorem does not put any constraint on the distribution of population. Equation 15 is a result of central limit theorem.

Does the Size of sample have an impact on the accuracy of results?

Consider that a population size is 100,000 and you have collected a sample of size  $n=100$ , which is sufficiently large to fulfil the requirements of central limit theorem. Will there be any advantage of taking a higher sample size say  $n=400$ ? The next section addresses this issue in detail.

## Check Your Progress 2

1. A fair dice is thrown 3 times, compute the probability distribution of the outcome number of times an even number appears on the dice.
2. What would be the probability of getting different number of heads, if a fair coin is tossed 4 times.
3. What would be the mean and standard deviation for the random variable of Question 2.
4. What is the mean and standard deviation for standard normal distribution?
5. A country has the population of 1 billion, out of which 1% are the students of class 10<sup>th</sup>. A representation sample of 10000 students of class 10 were asked a question "Is Mathematics difficult or easy?". Assuming that the population proportion of this question was reported to be 0.36, what would be possible standard deviation of the sampling distribution?
6. Given a quantitative variable, what is the mean and standard deviation of sampling distribution?



## 2.5 STATISTICAL HYPOTHESIS TESTING

In general, statistical analysis is mainly used in the two situations:

- S1. To determine if students of class 12 plays some sport, a sample random survey collected the data from 1000 students. Of these 405 students, stated that they play some sport. Using this information, can you infer that students of class 12 give less importance to sports? Such a decision would require you to estimate the population parameters.
- S2. In order to study the effect of sports on the performance of class 12<sup>th</sup> marks, a study was performed. It performed random sampling and collected the data of 1000 students, which included information of Percentage of marks obtained by the student and hours spent by the student in sports per week during class 12<sup>th</sup>. This kind of decision can be made through hypothesis testing.

In this section, let us analyse both these situations.

### 2.5.1 Estimation of Parameters of the Population

One of the simplest ways to estimate the parameter value as a point estimation. Key characteristics of this estimate should be that it should be unbiased, such as mean or median that lies towards the centre of the data; and should have small standard deviation, as far as possible. For example, a point estimate for situation S1 above would be that 40.5% students play some sports. This point estimate, however, may not be precise and may have some margin of error. Therefore, a better estimation would be to define an interval that contains the value of the parameter of the population. This interval, called confidence interval, includes the point estimate along with possible margin of error. The probability that the chosen confidence interval contains the population parameter is normally chosen as 0.95. This probability is called the confidence level. Thus, you can state with 95% confidence that the confidence interval contains a parameter. Is the value of confidence level as 0.95 arbitrary? As you know that sampling distribution for computing proportion is normal if the sample size ( $n$ ) is large. Therefore, to answer the question asked above, you may study Figure 13 showing the probability distribution of sampling distribution.

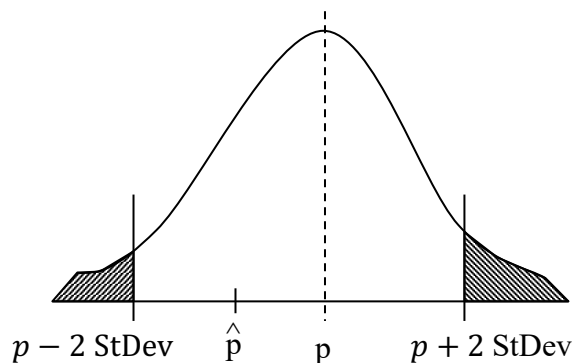


Figure 13: Confidence Level 95% for a confidence interval (non-shaded area).

Since you have selected a confidence level of 95%, you are expecting that proportion of the sample ( $\hat{p}$ ) can be in the interval—(population proportion ( $p$ )) -

$2 \times (\text{Standard Deviation})$ ) to (population proportion ( $p$ ) +  $2 \times (\text{Standard Deviation})$ ), as shown in Figure 13. The probability of occurrence of  $\hat{p}$  in this interval is 95% (Please refer to Figure 6). Therefore, the confidence level is 95%. In addition, note that you do not know the value of  $p$  that is what you are estimating, therefore, you would be computing  $\hat{p}$ . You may observe in Figure 13, that the value of  $p$  will be in the interval  $(\hat{p} - 2 \times (\text{Standard Deviation}))$  to  $(\hat{p} + 2 \times (\text{Standard Deviation}))$ . The standard deviation of the sampling distribution can be computed using equation (14b). However, as you are estimating the value of  $p$ , therefore, you cannot compute the exact value of standard deviation. Rather, you can compute standard error, which is computed by estimating the standard deviation using the sample proportion ( $\hat{p}$ ), by using the following formula:

$$\text{Standard Error}(StErr) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

Therefore, the confidence interval is estimated as  $(\hat{p} - 2 \times StErr)$  to  $(\hat{p} + 2 \times StErr)$ . In general, for a specific confidence level, you can specify a specific  $z$ -score instead of 2. Therefore, the confidence interval, for large  $n$ , is:  $(\hat{p} - z \times StErr)$  to  $(\hat{p} + z \times StErr)$

In practice, you may use confidence level of 90% or 95% and 99%. The  $z$ -score used for these confidence levels are 1.65, 1.96 (not 2) and 2.58 respectively.

Example 7: Consider the statement S1 of this section and estimate the confidence interval for the given data.

For the sample the probability that class 12<sup>th</sup> students play some sport is:

$$\hat{p} = 405/1000 = 0.405$$

The sample size ( $n$ ) = 1000

$$StErr = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} = \sqrt{\frac{0.405 \times (1 - 0.405)}{1000}} = 0.016$$

Therefore, the Confidence Interval for the confidence level 95% would be:

$$(0.405 - 1.96 \times 0.016) \text{ to } (0.405 + 1.96 \times 0.016)$$

$$0.374 \text{ to } 0.436$$

Therefore, with a confidence of 95%, you can state that the students of class 12<sup>th</sup>, who plays some sport is in the range 37.4% to 43.6%

How can you reduce the size of this interval? You may please observe that  $StErr$  is inversely dependent on the square root of the sample size. Therefore, you may have to increase the sample size to approximately 4 times to reduce the standard error to approximately half.

#### Confidence Interval to estimate mean

You can find the confidence interval for estimating mean in a similar manner, as you have done for the case of proportions. However, in this case you need estimate the standard error in the estimated mean using the variation of equation 15b, as follows:

$$\text{Standard Error in Sample Mean} = \frac{s}{\sqrt{n}}$$

; where  $s$  is the standard deviation of the sample

Example 8: The following table lists the height of a sample of 100 students of class 12 in centimetres. Estimate the average height of students of class 12.

170	164	168	149	157	148	156	164	168	160
149	171	172	159	152	143	171	163	180	158
167	168	156	170	167	148	169	179	149	171
164	159	169	175	172	173	158	160	176	173

159	160	162	169	168	164	165	146	156	170
163	166	150	165	152	166	151	157	163	189
176	185	153	181	163	167	155	151	182	165
189	168	169	180	158	149	164	171	189	192
171	156	163	170	186	187	165	177	175	165
167	185	164	156	143	172	162	161	185	174

Figure 14: Random sample of height of students of class 12 in centimetres

The sample mean and sample standard deviation is computed and is shown below:

Sample Mean ( $\bar{x}$ ) = 166; Standard Deviation of sample ( $s$ ) = 11

Therefore, the estimated height confidence interval of the mean height of the students of class 12<sup>th</sup> can be computed as:

Mean height ( $\bar{x}$ ) = 166

The sample size ( $n$ ) = 100

Standard Error in Sample Mean =  $\frac{11}{\sqrt{100}} = 1.1$

The Confidence Interval for the confidence level 95% would be:

(166 – 1.96 × 1.1) to (166 + 1.96 × 1.1)  
163.8 to 168.2

Thus, with a confidence of 95%, you can state that average height of class 12<sup>th</sup> students is in between 163.8 to 168.2 centimetres.

You may please note that in example 8, we have used t-distribution for means, as we have used sample's standard deviation rather than population standard deviation. The t-distribution of means is slightly more restrictive than z-distribution. The t-value is computed in the context of sampling distribution by the following equation:

$$t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \quad (16)$$

## 2.5.2 Significance Testing of Statistical Hypothesis

In this section, we will discuss about how to test the statement S2, given in section 2.5. A number of experimental studies are conducted in statistics, with the objective to infer, if the data support a hypothesis or not. The significance testing may involve the following phases:

1. Testing Pre-condition on Data:

Prior to performing the test of significance, you should check the pre-conditions on the test. Most of the statistical test require random sampling, large size of data for each possible category being tested and normal distribution of the population.

2. Making the statistical Hypothesis: You make statistical hypothesis after the parameters of the population. There are two basic hypothesis in statistical testing – the Null Hypothesis and the Alternative Hypothesis.

Null Hypothesis: Null hypothesis either defines a particular value for the parameter or specifies there is no difference or no change in the specified parameters. It is represented as  $H_0$ .

Alternative Hypothesis: Alternative hypothesis specifies the values or difference in parameter values. It is represented as either  $H_1$  or  $H_a$ . We use the convention  $H_a$ .

For example, for the statement S2 of Section 2.5, the two hypothesis would be:

$H_0$ : There is no effect of hours of study on the marks percentage of 12<sup>th</sup> class.

$H_a$ : The marks of class 12<sup>th</sup> improves with the hours of study of the student.

Please note that the hypothesis above is one sided, as your assumption is that the marks would increase with hours of study. The second one sided hypothesis may relate to decrease in marks with hours of study. However, most of the cases the hypothesis will be two sided, which just claims that a variable will cause difference in the second. For example, two sided hypothesis for statement S2 would be hours of study of students makes a difference (it may either increase or decrease ) the marks of students of class 12<sup>th</sup>. In general, one sided tests are called one tailed tests and two sided tests are called two tailed tests.

In general, alternative hypothesis relates to the research hypothesis. Please also note that the alternative hypothesis given above is one way hypothesis as it only states the effect in terms of increase of marks. In general, you may have alternative hypothesis which may be two way (increase or decrease; less or more etc.).

### 3. Perform the desired statistical analysis:

Next, you perform the exploratory analysis and produce a number of charts to explore the nature of the data. This is followed by performing a significance statistical test like chi-square, independent sample t-test, ANOVA, non-parametric tests etc., which is decided on the basis of size of the sample, type and characteristics of data. These tests generate assumes the null hypothesis to be True. A test may generate parameter values based on sample and the probability called p-value, which is an evidence against the null hypothesis. This is shown in Figure 15.

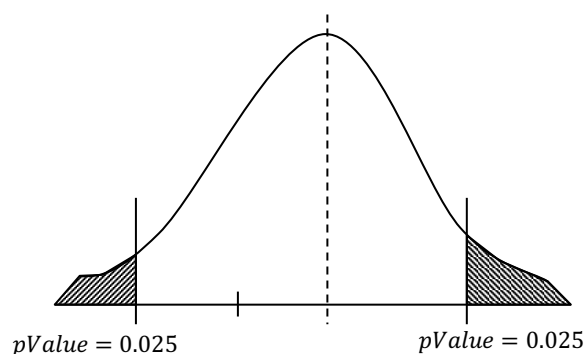


Figure 15: p-value of test statistics

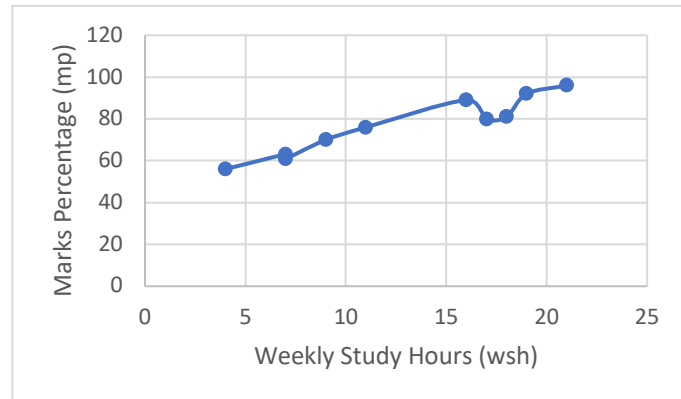
### 4. Analysing the results:

In this step, you should analyse your results. As stated in Unit 1, you must not just draw your conclusion based on statistics, but support it with analytical reasoning.

**Example 9:** We demonstrate the problem of finding a relationship between study hours and Marks percentage (S2 of section 2.5), however, by using only sample data of 10 students (it is hypothetical data and just used for the illustration purpose), which is given as follows:

Weekly Study Hours ( <i>wsh</i> )	96	92	63	76	89	80	56	70	61	81
Marks Percentage ( <i>mp</i> )	21	19	7	11	16	17	4	9	7	18

In order to find such a relationship, you may like to perform basic exploratory analysis. In this case, let us make a scatter plot between the two variables, taking *wsh* as an independent variable and *mp* as a dependent variable. This scatter plot is shown in Figure 16



**Figure 16: Scatter plot of Weekly Study Hours vs. Marks Percentage.**

The scatter plot of Figure 16 suggests that the two variables may be associated. But how to determine the strength of this association? In statistics, you use Correlation, which may be used to determine the strength of linear association between two quantitative variables. This is explained next.

### 2.5.3 Example using Correlation and Regression

As stated correlation is used to determine the strength of linear association. But how the correlation is measured?

Consider two quantitative variables  $x$  and  $y$ , and a set of  $n$  pairs of values of these variables (for example, the *wsh* and *mp* values as shown in example 9), you can compute a correlation coefficient, denoted by  $r$  using the following equation:

$$r_{xy} = \frac{\sum_{i=1}^n \left( \frac{(x-\bar{x})}{s_x} \right) \times \left( \frac{(y-\bar{y})}{s_y} \right)}{(n-1)} \quad (16)$$

The following are the characteristics of the correlation coefficient ( $r$ ):

- The value of  $r$  lies between +1 and -1.
- A positive value of  $r$  means that value of  $y$  increases with increase in value of  $x$  and the value of  $y$  decreases with decrease in value of  $x$ .
- A negative value of  $r$  means that value of  $y$  increases with decrease in value of  $x$  and the value of  $y$  decreases with increase in value of  $x$ .
- If the value of  $r$  is closer to +1 or -1, then it indicates that association is a strong linear association.
- Simple scaling one of the variable does not change the correlation.
- Correlation does not specify the dependent and independent variables.
- Please remember a correlation does not mean cause. You have to establish it with reasoning.

The data of Example 9 shows a positive correlation. It can be computed as follows:

Mean of *wsh* = 12.9; Standard Deviation of *wsh* (Sample) = 5.98980616

Mean of *mp* = 76.4; Standard Deviation of *mp* (Sample) = 13.7210301

$$r_{wsh,mp} = \frac{8.61944034}{(10-1)} = 0.95771559$$

Therefore, the data shows strong positive correlation.

You may also use any statistical tool to find the correlation, we used MS-Excel, which gave the following output of correlation:

	Weekly Study Hours (wsh)	Marks Percentage (mp)
Weekly Study Hours (wsh)	1	
Marks Percentage (mp)	0.957715593	1

**Figure 17: The Correlation coefficient**

As the linear correlation between *wsh* and *mp* variables is strong, therefore, you may like to find a line, called linear regression line, that may describe this association. The accuracy of regression line, in general, is better for higher correlation between the variables.

*Single Linear Regression:*

A single linear regression predicts a response variable or dependent variable (say *y*) using one explanatory variable or independent variable (say *x*). The equation of single linear regression can be defined by using the following equation:

$$y_{predicted} = a + bx \quad (17)$$

Here,  $y_{predicted}$  is the predicted value of response variable (*y*), *x* is the explanatory variable, *a* is the intercept with respect to *y* and *b* is called the slope of the regression line. In general, when you fit a linear regression line to a set of data, there will be certain difference between the  $y_{predicted}$  and the observed value of data (say  $y_{observed}$ ). This difference between the observed value and the predicted value, that is  $(y_{observed} - y_{predicted})$ , is called the residual. One of the most used method of finding the regression line is the method of least square, which minimises the sum of squares of these residuals. The following equations can be used for computing residual:

$$Residual = y_{observed} - y_{predicted} \quad (18)$$

The objective of least square method in regression is to minimise the sum of squares of the residual of all the *n* observed values. This sum is given in the following equation:

$$SumOfResidualSquares = \sum_{i=1}^n (y_{observed} - y_{predicted})^2 \quad (19)$$

Another important issue with regression model is to determine the predictive power of the model, which is computed using the square of the correlation ( $r^2$ ). The value of  $r^2$  can be computed as follows:

- In case, you are not using regression, then you can predict the value of *y* using the mean. In such a case, the difference in predicted value and observed value would be given by the following equation:

$$ErrorUsingMean = y_{observed} - \bar{y} \quad (20)$$

- The total of sum of square of this error can be computed using the following equation:

$$TotalSumOfSquare = \sum_{i=1}^n (y_{observed} - \bar{y})^2 \quad (21)$$

The use of regression line reduces the error in prediction of the value of *y*. Equation (19) represents this square error. Thus, use of regression results helps in reducing the error. The proportion  $r^2$  is actually the predictive power of the regression and is represented using the following equation:

$$r^2 = \frac{\sum_{i=1}^n (y_{observed} - \bar{y})^2 - \sum_{i=1}^n (y_{observed} - y_{predicted})^2}{\sum_{i=1}^n (y_{observed} - \bar{y})^2} \quad (22)$$

As stated earlier,  $r^2$  can also be computed by squaring the value of *r*.

On performing regressing analysis on the observed data of Example 9, the statistics as shown in Figure 18 is generated.

<i>Regression Statistics</i>				
Multiple R	0.9577			
R Square	0.9172			
Adjusted R Square	0.9069			
Standard Error	4.1872			
Observations	10.0000			

ANOVA				
	<i>df</i>	<i>SS</i>	<i>F</i>	<i>Significance F</i>
Regression	1.0000	1554.1361	88.6407	0.0000
Residual	8.0000	140.2639		
Total	9.0000	1694.4000		

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
<i>Intercept</i>	48.0991	3.2847	14.6435	0.0000
Weekly Study Hours ( <i>wsh</i> )	2.1939	0.2330	9.4149	0.0000

**Figure 18: A Selected Regression output**

The regression analysis results, as shown above are discussed below:

- Assumptions for the regression model:
  - Data sample is collected using random sampling.
  - For every value of  $x$ , the value of  $y$  in the population
    - is normally distributed
    - has same standard deviation
  - The mean value if  $y$  in the population follows regression equation (17)
- Various Null hypothesis related to regression are:
  - For the analysis of variance (ANOVA) output in the regression:
    - $H_{0A}$ : All the coefficients of model are zero, therefore, the model cannot predict the value of  $y$ .
  - For the *Intercept*:
    - $H_{0I}$ : *Intercept* = 0.
  - For the *wsh*:
    - $H_{0wsh}$ : *wsh* = 0.
- The *Significance F* in ANOVA is 0, therefore, you can reject the Null hypothesis  $H_{0A}$  and determine that the this model can predict the value of  $y$ . Please note high  $F$  value supports this observation.
- The p-value related to *intercept* and *wsh* are almost 0, therefore, you can reject the Null hypothesis  $H_{0I}$  and  $H_{0wsh}$ .
- The regression line has the equation:  

$$mp_{predicted} = 48.0991 + 2.1939 \times wsh$$
- You can compute the sum of squares ( $SS$ ) using Equation (19) and Equation (21).
- The degree of freedom in the context of statistics is the number of data items required to compute the desired statistics.

- The term “Multiple R” in *Regression Statistics* defines the correlation between the dependent variable (say  $y$ ) with the set of independent or explanatory variables in the regression model. Thus, multiple R is similar to correlation coefficient ( $r$ ), except that it is used when multiple regression is used. Most of the software express the results in terms of Multiple R, instead of  $r$ , to represent the regression output. Similarly, R Square is used in multiple regression, instead of  $r^2$ . The proposed model has a large  $r^2$ , therefore, can be considered for deployment.

You can go through further readings for more details on all the terms discussed above.

Figure 19 shows the regression line for the data of Example 9. You may please observe that residuals is the vertical difference between the Marks Percentage and Predicted marks percentage. These residuals are shown in Figure 20.

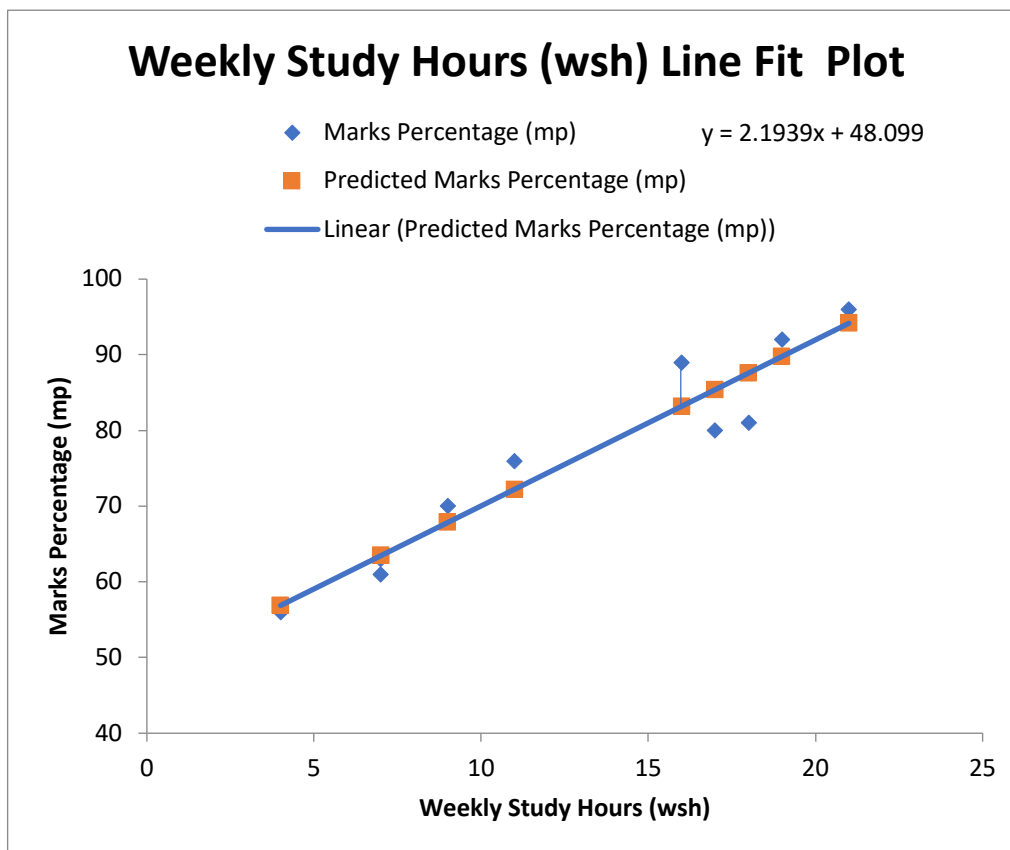


Figure 19: The Regression Line

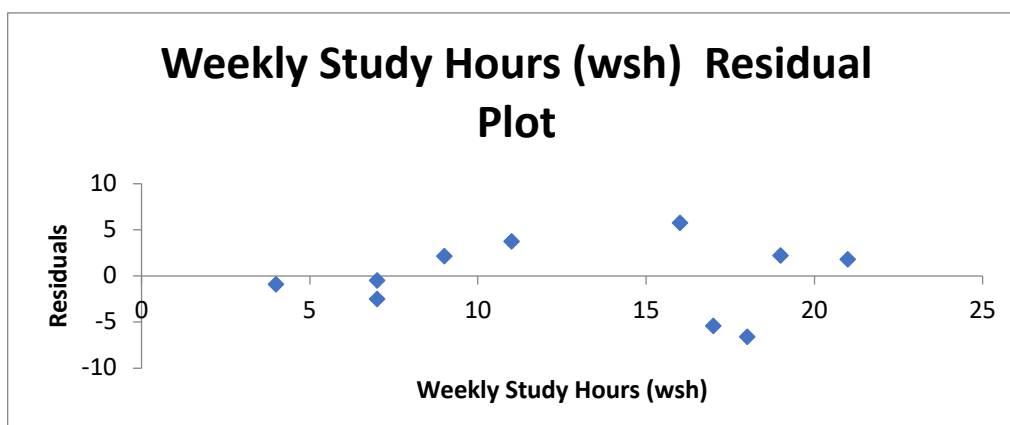


Figure 20: The Residual Plot



## 2.5.4 Types of Errors in Hypothesis Testing

In the section 2.5.1 and section 2.5.2, we have discussed about testing the Null hypothesis. You either Reject the Null hypothesis and accepts alternative hypothesis based on the computed probability or p-value; or you fail to Reject the Null hypothesis. The decisions in such hypothesis testing would be:

- You reject Null hypothesis for a confidence interval 95% based on the p-value, which lies in the shaded portion, that is  $p\text{-value} < 0.05$  for two tailed hypothesis (that is both the shaded portions in Figure 15, each area of probability 0.025). Please note that in case of one tailed test, you would consider only one shaded area of Figure 15, therefore, you would be considering  $p\text{-value} < 0.05$  in only one of the two shaded areas.
- You fail to reject the Null hypothesis for confidence interval 95%, when  $p\text{-value} > 0.05$ .

The two decisions as stated above could be incorrect, as you are considering a confidence interval of 95%. The following Figure shows this situation.

<i>The Actual Scenario</i>	<i>Final Decision</i>	
	<i><math>H_0</math> is Rejected, that is, you have accepted the Alternative hypothesis</i>	<i>You fail to reject <math>H_0</math>, as you do not have enough evidence to accept the Alternative hypothesis</i>
$H_0$ is True	This is called a TYPE-I error	You have arrived at a correct decision
$H_0$ is False	You have arrived at a correct decision	This is called a TYPE-II error

For example, assume that a medicine is tested for a disease and this medicine is NOT a cure of the disease. You would make the following hypotheses:

$H_0$ : The medicine has no effect for the disease

$H_a$ : The medicine improves the condition of patient.

However, if the data is such that for a confidence interval of 95% the p-value is computed to be less than 0.05, then you will reject the null hypothesis, which is Type-I error. The chances of Type-I errors for this confidence interval is 5%. This error would mean that the medicine will get approval, even though it has no effect on curing the disease.

However, now assume that a medicine is tested for a disease and this medicine is a cure of the disease. Hypotheses still remains the same, as above. However, if the data is such that for a confidence interval of 95% the p-value is computed to be more than 0.05, then you will not be able to reject the null hypothesis, which is Type-II error. This error would mean that a medicine which can cure the disease will not be accepted.

### Check Your Progress 3

1. A random sample of 100 students were collected to find their opinion about whether practical sessions in teaching be increased? About 53 students voted for increasing the practical sessions. What would be the confidence interval of the population proportions of the students who would favour increasing the population percentage. Use confidence levels 90%, 95% and 99%.

2. The Weight of 20 students, in Kilograms, is given in the following table

65 75 55 60 50 59 62 70 61 57  
62 71 63 69 55 51 56 67 68 60

Find the estimated weight of the student population.

3. A class of 10 students were given a validated test prior and after completing a training course. The marks of the students in those tests are given as under:

Marks before Training ( <i>mbt</i> )	56	78	87	76	56	60	59	70	61	71
Marks after training ( <i>mat</i> )	55	79	88	90	87	75	66	75	66	78

With a significance level of 95% can you say that the training course was useful?

## 2.6 SUMMARY

This Unit introduces you to the basic probability and statistics related to data science. The unit first introduces the concept of conditional probability, which defines the probability of an event given a specific event has occurred. This is followed by discussion on the Bayes theorem, which is very useful in finding conditional probabilities. Thereafter, the unit explains the concept of discrete and continuous random variables. In addition, the Binomial distribution and normal distribution were also explained. Further, the unit explained the concept of sampling distribution and central limit theorem, which forms the basis of the statistical analysis. The Unit also explain the use of confidence level and intervals for estimating the parameters of the population. Further, the unit explains the process of significance testing by taking an example related to correlation and regression. Finally, the Unit explains the concept of errors in hypothesis testing. You may refer to further readings for more details on these concepts.

## 2.7 SOLUTION/ANSWERS

### Check Your Progress – 1

1. Is  $P(Y/X) = P(Y/X)$ , No. Please check in Example 3, the probability  $P(\text{Red}/\text{BagB})$  is  $7/10$ , whereas,  $P(\text{BagB}/\text{Red})$  is  $7/12$ .
2. Consider two independent events A and B, first compute  $P(A)$  and  $P(B)$ . The probability of any one of these events to occur would be computed by equation (2), which is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The probability of occurrence of both the events will be computed using the equation (4), which is:

$$P(X \cap Y) = P(X) \times P(Y)$$

3. Let us assume Event X, as “A student is selected from University A”. Assuming, any of the University can be selected with equal probability,  $P(\text{UniA}) = 1/2$ .

Let the Event Y, as “A student who has obtained more that 75% marks is selected”. This probability  $P(\text{StDis}) = \frac{1}{2} \times \frac{10}{20} + \frac{1}{2} \times \frac{20}{30} = \frac{7}{12}$

In addition,  $P(\text{StDis}/\text{UniA}) = \frac{10}{20} = \frac{1}{2}$

$$P(\text{UniA}/\text{StDis}) = \frac{P(\text{StDis}/\text{UniA}) \times P(\text{UniA})}{P(\text{StDis})} = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{7}{12}} = \frac{3}{7}$$

### Check Your Progress 2

1. As the probability of getting the even number (E) or odd number (O) is equal in each two of dice, the following eight outcomes may be possible:

Outcomes	EEE	EEO	EOE	EOO	OOE	OEO	OOE	OOO
Number of times Even number appears (X)	3	2	2	1	2	1	1	0

Therefore, the probability distribution would be:

X	Frequency	Probability P(X)
0	1	1/8
1	3	3/8
2	3	3/8
3	1	1/8
Total	8	Sum of all P(X) = 1

2. This can be determined by using the Binomial distribution with  $X=0, 1, 2, 3$  and 4, as follows ( $s$  and  $f$  both are  $1/2$ ):

$$P(X = 0) \text{ or } p_0 = {}^4C_0 \times s^0 \times f^{4-0} = \frac{4!}{0!(4-0)!} \times \left(\frac{1}{2}\right)^0 \times \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

$$P(X = 1) \text{ or } p_1 = {}^4C_1 \times s^1 \times f^{4-1} = \frac{4!}{1!(4-1)!} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^3 = \frac{4}{16}$$

$$P(X = 2) \text{ or } p_2 = {}^4C_2 \times s^2 \times f^{4-2} = \frac{4!}{2!(4-2)!} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^2 = \frac{6}{16}$$

$$P(X = 3) \text{ or } p_3 = {}^4C_3 \times s^3 \times f^{4-3} = \frac{4!}{3!(4-3)!} \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^1 = \frac{4}{16}$$

$$P(X = 4) \text{ or } p_4 = {}^4C_4 \times s^4 \times f^{4-4} = \frac{4!}{4!(4-4)!} \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^0 = \frac{1}{16}$$

3. The number of tosses ( $n$ ) = 4 and  $s = 1/2$ , therefore,

$$\mu = n \times s = 4 \times \frac{1}{2} = 2$$

$$\sigma = \sqrt{n \times s \times (1-s)} = \sqrt{4 \times \frac{1}{2} \times \left(1 - \frac{1}{2}\right)} = 1$$

4. Mean = 0 and Standard deviation = 1.

5. Standard deviation of sampling distribution =

$$\sqrt{\frac{p \times (1-p)}{n}} = \sqrt{\frac{0.36 \times (1-0.36)}{10000}} = \frac{0.6 \times 0.8}{100} = 0.0048$$

The large size of sample results in high accuracy of results.

6. Mean of sample means =  $\mu$

$$\text{Standard Deviation of Sample Means} = \frac{\sigma}{\sqrt{n}}$$

### Check Your Progress 3

1. The value of sample proportion  $\hat{p} = 53/100 = 0.53$

$$\text{Therefore, } StErr = \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}} = \sqrt{\frac{0.53 \times (1-0.53)}{100}} = 0.05$$

The Confidence interval for 90%:

$$(0.53 \pm 1.65 \times 0.05), \text{ which is } 0.4475 \text{ to } 0.6125$$

The Confidence interval for 95%:

$$(0.53 \pm 1.96 \times 0.05), \text{ which is } 0.432 \text{ to } 0.628$$

The Confidence interval for 99%:

$$(0.53 \pm 2.58 \times 0.05), \text{ which is } 0.401 \text{ to } 0.659$$

2. Sample Mean ( $\bar{x}$ ) = 61.8; Standard Deviation of sample ( $s$ ) = 6.787

Sample size ( $n$ ) = 20

$$\text{Standard Error in Sample Mean} = \frac{6.787}{\sqrt{20}} = 1.52$$

The Confidence Interval for the confidence level 95% would be:

$$(61.8 \pm 1.96 \times 1.52) = 58.8 \text{ to } 64.8$$

3. Analysis: This kind of problem would require to find, if there is significant difference in the mean of the test results before and after the training course. In addition, the data size of the sample is 10 and the same group of person are tested twice, therefore, paired sample t-test may be used to find the difference of the mean. You can follow all the steps for this example of hypothesis testing.

1. Testing Pre-condition on Data:

- The students who were tested through this training course were randomly selected.
- The population test scores, in general, are normally distributed.
- The sample size is small, therefore, a robust test may be used.

2. The Hypothesis

$$H_0: \overline{mbt} = \overline{mat}$$

$$H_1: \overline{mbt} < \overline{mat}$$

3. The results of the analysis are given below

(Please note  $H_1$  is one sided hypothesis, as you are trying to find if training was useful for the students)

#### **t-Test: Paired Two Sample for Means**

	Marks before Training (mbt)	Marks after training (mat)
Mean	67.4	75.9
Variance	112.9333333	124.1
Observations	10	10
df	9	
t Stat	-2.832459252	
P(T<=t) one-tail	0.009821702	
t Critical one-tail	1.833112933	

4. Analysis of results: The one tail p-value suggests that you reject the null hypothesis. The difference in the means of the two results is significant enough to determine that the scores of the student have improved after the training.