BAOU
Education for All

DR.BABASAHEB AMBEDKAR OPEN UNIVERSITY

BBA

BACHELOR OF BUSINESS ADMINISTRATION

BBAR-304

Business Analytics

Business Analytics

ISBN 978-93-91071-43-1

Editorial Panel
Authors
Mr. Ankur Sharma
Senior Manager, TCS Limited,
Gandhinagar

Dr. Ashvin R. Dave
Professor and HOD,
Departmentof Business Administration & Commerce,
School of Liveral Studies, Pandit Deendayal Energy University. Gandhinagar

Editor
Dr Ashish Joshi
Associate Professor,
Pandit Deendayal Petroleum University,
Gandhinagar

Language Editor
Dr. Jagdish Anerao
Associate Professor,
Smt AP Patel Arts & NP Patel Commerce College, Naroda, Ahmedabad

Edition : 2021

Acknowledgment
Every attempt has been made to trace the copyright holders of material re–produced in this book. Should an infringement have occurred, we apologize for the same and will be pleased to make necessary correction/amendment in future edition of this book.
The content is developed by taking reference of online and print publications that are mentioned in Bibliography. The content developed represents the breadth of research excellence in this multidisciplinary academic field. Some of the information, illustrations and examples are taken "as is" and as available in the references mentioned in Bibliography for academic purpose and better understanding by learner.

ROLE OF SELF INSTRUCTIONAL MATERIAL IN DISTANCE LEARNING

The need to plan effective instruction is imperative for a successful distance teaching repertoire. This is due to the fact that the instructional designer, the tutor, the author (s) and the student are often separated by distance and may never meet in person. This is an increasingly common scenario in distance education instruction. As much as possible, teaching by distance should stimulate the student's intellectual involvement and contain all the necessary learning instructional activities that are capable of guiding the student through the course objectives. Therefore, the course / self–instructional material are completely equipped with everything that the syllabus prescribes.

To ensure effective instruction, a number of instructional design ideas are used and these help students to acquire knowledge, intellectual skills, motor skills and necessary attitudinal changes. In this respect, students' assessment and course evaluation are incorporated in the text.

The nature of instructional activities used in distance education self–instructional materials depends on the domain of learning that they reinforce in the text, that is, the cognitive, psychomotor and affective. These are further interpreted in the acquisition of knowledge, intellectual skills and motor skills. Students may be encouraged to gain, apply and communicate (orally or in writing) the knowledge acquired. Intellectual– skills objectives may be met by designing instructions that make use of students'prior knowledge and experiences in the discourse as the foundation on which newly acquired knowledge is built.

The provision of exercises in the form of assignments, projects and tutorial feedback is necessary. Instructional activities that teach motor skills need to be graphically demonstrated and the correct practices provided during tutorials. Instructional activities for inculcating change in attitude and behavior should create interest and demonstrate need and benefits gained by adopting the required change. Information on the adoption and procedures for practice of new attitudes may then be introduced.

Teaching and learning at a distance eliminates interactive communication cues, such as pauses, intonation and gestures, associated with the face–to–face method of teaching. This is particularly so with the exclusive use of print media. Instructional activities built into the instructional repertoire provide this missing interaction between the student and the teacher. Therefore, the use of instructional activities to affect better distance teaching is not optional, but mandatory.

Our team of successful writers and authors has tried to reduce this. Divide and to bring this Self Instructional Material as the best teaching and communication tool. Instructional activities are varied in order to assess the different facets of the domains of learning.

Distance education teaching repertoire involves extensive use of self–instructional materials, be they print or otherwise. These materials are designed to achieve certain pre–determined learning outcomes, namely goals and objectives that are contained in an instructional plan. Since the teaching process is affected over a distance, there is need to ensure that students actively participate in their learning by performing specific tasks that help them to understand the relevant concepts. Therefore, a set of exercises is built into the teaching repertoire in order to link what students and tutors do in the framework of the course outline. These could be in the form of students' assignments, a research project or a science practical exercise. Examples of instructional activities in distance education are too numerous to list. Instructional activities, when used in this context, help to motivate students, guide and measure student's performance (continuous assessment).

PREFACE

We have put in lots of hard work to make this book as user–friendly as possible, but we have not sacrificed quality. Experts were involved in preparing the materials. However, concepts are explained in easy language for you. We have included many tables and examples for easy understanding.

We sincerely hope this book will help you in every way you expect.

All the best for your studies from our team!

Business Analytics

Contents

Unit 2 : Continuous Probability Distributions

Introduction, Probability Density Function, The Normal Distribution, Binomial Distributions, Poisson Distribution, Student's t–Distribution, PDF and CDF for t–Distribution, Properties of t–Distribution

Unit 3 : Sampling and Confidence Intervals

Introduction, Introduction to Sampling Process, Important Steps in Designing a Sampling Strategy, Sampling Methods, Probabilistic Sampling Methods, Non-Probabilistic Sampling Methods, Central Limit Theorem, Confidence Interval

Unit 4 : Introduction to Hypothesis Testing

Introduction, Life Cycle of Hypothesis Testing, Hypothesis Testing Process Steps, Hypothesis Test Statistics, Two–Tailed and One–Tailed Hypothesis Test, Concept of p–Value, Type I, Type II Error and Power of the Hypothesis Test, Hypothesis Testing for a Population Mean with Known Population Variance : Z–Test, Hypothesis Testing for a Population Mean with Known Population Variance : t-Test

BLOCK 3 : CORRELATION AND REGRESSION

Unit 1 : Covariance and Correlation Analysis

Introduction, Covariance : Statistical Relationship between Variables, Mathematical Interpretation of the Covariance, Relationship between Covariance and Variance, Covariance Matrix, Relationship between Covariance and Correlation, Spearman Rank Correlation

Unit 2 : Simple Linear Regression

Introduction, Essence of Simple Linear Regression, Introduction to Simple Linear Regression, Determining the Equation of the Linear Regression Line, Baseline Prediction Model, Simple Linear Regression Model Building, Ordinary Least Square Method to Estimate Parameters, Calculation of Regression Parameters, Interpretation of Regression Equation, Measures of Variation, Comparison of Two Models, Coefficient of Determination, Mean Square Error and Root Mean Square Error (Standard Error), Simple Linear Regression in MS Excel, Residual Analysis to Test The Regression Assumptions

Unit 3 : Multiple Linear Regression

Introduction, Essence of Multiple Linear Regression, Introduction to Multiple Linear Regression, Understanding the Concept of Multiple Linear Regression with a Worked Example, The Correlation Coefficient for Multiple Linear Regression, Coefficient of Coefficient (R2), Adjusted R2, and Standard Error, Multiple Linear Regression in MS Excel, The Modified Regression Model in Excel, Residual Analysis to Test the Regression Assumptions

BLOCK 4 : TIME SERIES ANALYSIS

Unit 1 : Introduction to Forecasting Techniques

Introduction, Forecasting : Magical Crystal Ball of Statisticians, Time–Series Data and Components of Time–Series Data, Time–Series Data Modelling Techniques, Additive Model of Time–Series Modelling, Multiplicative Model of Time–Series Modelling, Measuring Forecasting Accuracy Techniques, Mean Absolute Error (MAE) / Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Root Mean Square Error (RMSE), Factors Affecting Forecasting Accuracy

Unit 2 : Moving Average and Single Exponential Smoothing Techniques

Introduction, Parts of Forecasting Techniques, Naïve Forecasting Models, Averaging Models, Simple Averages, Moving Averages, Weighted Moving Averages, Single Exponential Smoothing Forecasting Technique, Single Exponential Smoothing Forecasting Technique in MS Excel

Unit 3 : Regression Methods for Forecasting

Introduction, Forecasting Techniques with a Trend, How to Draw Trendline and Regression Equation in Time–Series Graph, Double Exponential Smoothing Constant Technique for Forecasting

Unit 4 : Auto–Regression (AR) and Moving Average (MA) Forecasting Models

Introduction, Introduction to Autocorrelation, Reasons for Autocorrelation, Impact of Autocorrelation on a Regression Model, Ways to Detect Autocorrelation : Durbin Watson Test, Autoregression : Remedy to Resolve Autocorrelation, Moving Average Model MA(q)

Dr. Babasaheb Ambedkar OpenUniversity

BBAR-304

Business Analytics

BLOCK-1 BUSINESS ANALYTICS FUNDAMENTALS

UNIT 1

INTRODUCTION TO BUSINESS ANALYTICS

UNIT 2

DESCRIPTIVE ANALYTICS

UNIT 3

VISUALIZATION TECHNIQUES FOR BUSINESS ANALYTICS

BLOCK 1 : BUSINESS ANALYTICS FUNDAMENTALS

Block Introduction

Every organization despite its size needs to measure important business metrics like profit, sales, cost, market share, return on investment, customer satisfaction, employee satisfaction, etc. It is utterly important to identify relationships among these metrics and important factors which impact these metrics directly or indirectly. Analytics is a discipline that is made up of statistics, mathematics, and computer science. It helps us to analyze these metrics periodically and convert all data into business information which leads to robust decision making.

Business analytics has been used in different industries for several decades but in the last 20 years company's dependencies on analytics have increased exponentially. There are four main reasons for this sharp rise :

- **Advanced software techniques** are available e.g. advanced data structures, advanced database systems, cloud computing, etc.
- **Clean data** is available, now most organizations have robust software infrastructure which helps in capturing clean customer, vendors, and sales data
- **Advanced hardwares** are available which can store huge data in such a way that it can be easily available for analysis without any time lag. Also, the cost is quite reasonable e.g. GPU/TPU processors, distributed networks, etc
- **Advanced business problem-solving techniques** are providing new alternatives to tackle business problems e.g. Agile and lean six sigma frameworks for business excellence

Every hour organizations take several decisions which decide their fate whether they will be profitable or suffer losses. Studies reveal that organizations that take more data-backed decisions are more probable to incur high profits. With the help of advanced analytics software and clean input data, most of these decisions are taken by systems that make processes more efficient and responsive. Forbes magazine stated in its famous report that the main difference between successful and not so successful organizations is that how much they use analytics in their decision making.

Block Objectives

After learning this block, you will be able to understand :

- Understanding how Business Analytics is changing the decision–making process in the business world
- Different components of business analytics
- Classifications and challenges of business analytics
- Basic concepts of descriptive analytics
- Data types and scale of measurements
- Learning various types of visualization techniques
- Interpretation of visualization techniques
- Understanding the most appropriate visualization technique as per the scenario

Block Structure

Unit 1

INTRODUCTION TO BUSINESS ANALYTICS

: UNIT STRUCTURE :

1.0 Learning Objectives :

After learning this unit, you will be able to understand :

- Understanding how Business Analytics is changing the decision–making process in the business world
- Different components of business analytics
- Evaluation of business analytics and how it has become the new language in the business world
- Classifications of business analytics
- Challenges of business analytics

1.1 Introduction :

In this unit we will study basic institution about business analytics and how it is changing the overall canvas of business world. At the end of the unit, you will understand how different components of business analytics are influencing decision–making capabilities. We will also see different type of business analytics and important tools and techniques used in each type of business analytics. We will also touch upon various challenges organizations are facing in current era and what are different approaches to overcome these challenges.

1.2 Importance of Data and its Sources

"Data is what you need to do analysis. Information is what you need to do business." – ***John Owen***

Today Businesses speak the language of analytics. We see a flood of analytics jargon in all business presentations. Analytics has covered a long journey from simple number crunching to solving complex business problems to create a competitive business strategy. In the last two decades, mostly all big and mid–size organizations have started business analytics as one of the primary functions, and that's why business analysts hold critical and well–paid roles in these organizations. World–renowned magazine, HBR mentioned it in one of the articles as "The Sexiest Job of the 21st Century".

When we apply structured and scientific tools/ approaches to convert raw data into meaningful business information which leads to better Business Decisions, we call it Business Analytics.

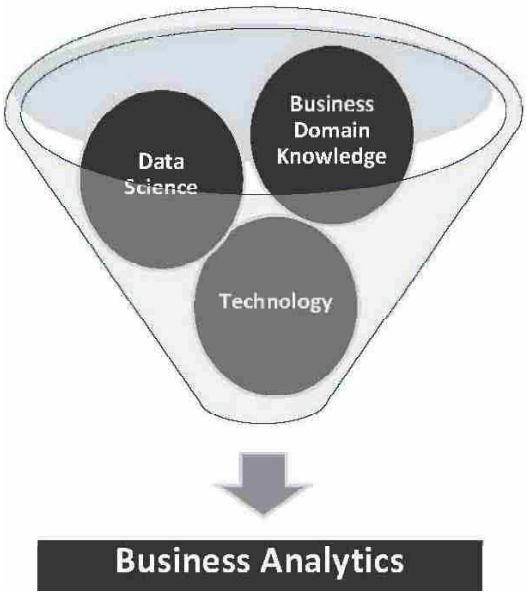


Fig 1.1 Components of Business Analytics

❖ **Business Analytics is made up of Three Crucial Components :**

1. **Technology :** In Business Analytics technology plays a significant role in capturing a large amount of complex data, sharing it simultaneously with different geographies, streaming it from various

sources, e.g. social media, sales systems, customer relationship management systems. Technology also helps in real–time data backup.

2. **Business Domain Knowledge :** Analytics projects always revolve around domain knowledge. An analyst with sound domain knowledge will have a great knack for asking the right questions; it helps them in selecting relevant data and the right tools. This would finally result in a good analytics storyboard.

3. **Data Science :** This is the heart of Business Analytics; it generally consists of statistical and machine learning concepts. It starts with the right tools/ approaches to framing the right business problem post that helps in analyzing data and drawing conclusions out of that.

1.2.1 Why has Data Suddenly become so Important :

Data is becoming the most critical commodity in today's world. It is not limited to the business world but also equally important in our personal life. In today's boundaryless business world, there is a famous saying, *"The last two decades were dedicated to **software** while the next two decades will be dedicated to **Statistics**"*. It means that in the last two decades, all types of organizations irrespective of their size and nature of business got the software installed in their vital departments. Even shops in our locality like grocery stores, medical shops, hardware stores, etc. also get the software installed to manage their inventory and customer information. This software is collecting important data about sales, inventory, customer demographics, vendor details, etc. But it is also raising a huge question in front of the world about how to convert this raw data into some meaningful information that can be further used in strategy formation and better decision making. This is where Business Analytics comes as a saviour. For example, if we download the feedback report from the software then it is data but if we see that feedback of a particular region is relatively low then it becomes information. Management will initiate root cause analysis for finding reasons for low feedback scores.

On the other side, the next generation of softwares are available, which can store a very complicated and massive amount of data. Below are a few examples of complicated or complex data :

- Emoticons, likes, and pictures, videos from social media sites
- Real–time data gathering from various points of sales at different locations (with the help of Big Data technologies)
- Data is getting updated in real–time across geographies (with the help of cloud technologies)
- Customer feedback from Facebook or Twitter in the form of raw texts with wrong spellings ?
- Customer demographics and sales related data from Customer Relationship Management (CRM) software

As currently softwares are installed in all important departments of the organization, these softwares capture data at a very granular level. This detailed and clean data is available for processing and analysis with the help of recent developments in the computational power of computers and advanced machine learning algorithms. Hence raw data is getting converted into meaningful information on a real–time basis, which is further consumed by decision–makers to get a competitive edge. In this way, data has become one of the most important currencies in the business world. Today most successful business leaders are trying to infuse analytical capabilities throughout their organizations to start the culture of taking data–based decisions, it also helps in optimizing business outcomes such as maximum revenue with minimum resources, passing benefits to their customers.

1.2.2 Different Sources of Data Accumulation in the Personal and Business World :

❖ **Data Sources in Personal Life :**

If you are not paying for it, you're not the customer; you are the product being sold. – ***Commentator Blue Beetle***

We use various software in our daily life, which captures our personal information, e.g. Facebook, Google, YouTube, Instagram, LinkedIn, etc. These softwares get access to our personal information like places we visit, brands we endorse, TV/radio channels we choose as entertainment/ information sources, type of blogs/ books we read, people we meet, etc.

Today softwares are the integrated part of most of the devices we use in our personal life. This also raises the risk that our personal data can be used for unethical purposes. Below are a few examples :

- **Mobile Phones :** When we install various apps, unknowingly we accept their terms and conditions to share our contact lists, to read our messages, to trace our locations, etc. If this data gets leaked, then it may be dangerous for our personal and monetary security, e.g. we have bank details, OTPs, password in our messages which can be misused
- **Health Fitness–Related Devices :** These devices get access to our locations, our diet, our sleeping and exercise patterns, etc
- **TV Setup Boxes :** Setup boxes get data about our entertainment quotient, e.g. which channels we watch and for how long. They may use this data to make customized offers for us or organizations may share this data with their vendors for marketing purpose
- **Smart Watches :** Smart watches also have most of the features of mobile phone hence above data problems may occur in the case of smartwatches also
- **Social Media Sites :** Websites/ apps like YouTube/ Facebook/ Twitter/ Instagram/ LinkedIn get access to our personal information like family photos, our views on various matters/ issues, our likes

or dislikes about multiple brands/ products. They also share this data with their customers for commercial purposes

These examples enlighten us about how data is being monetized and prove that "Data is one of the most important currencies of today's world".

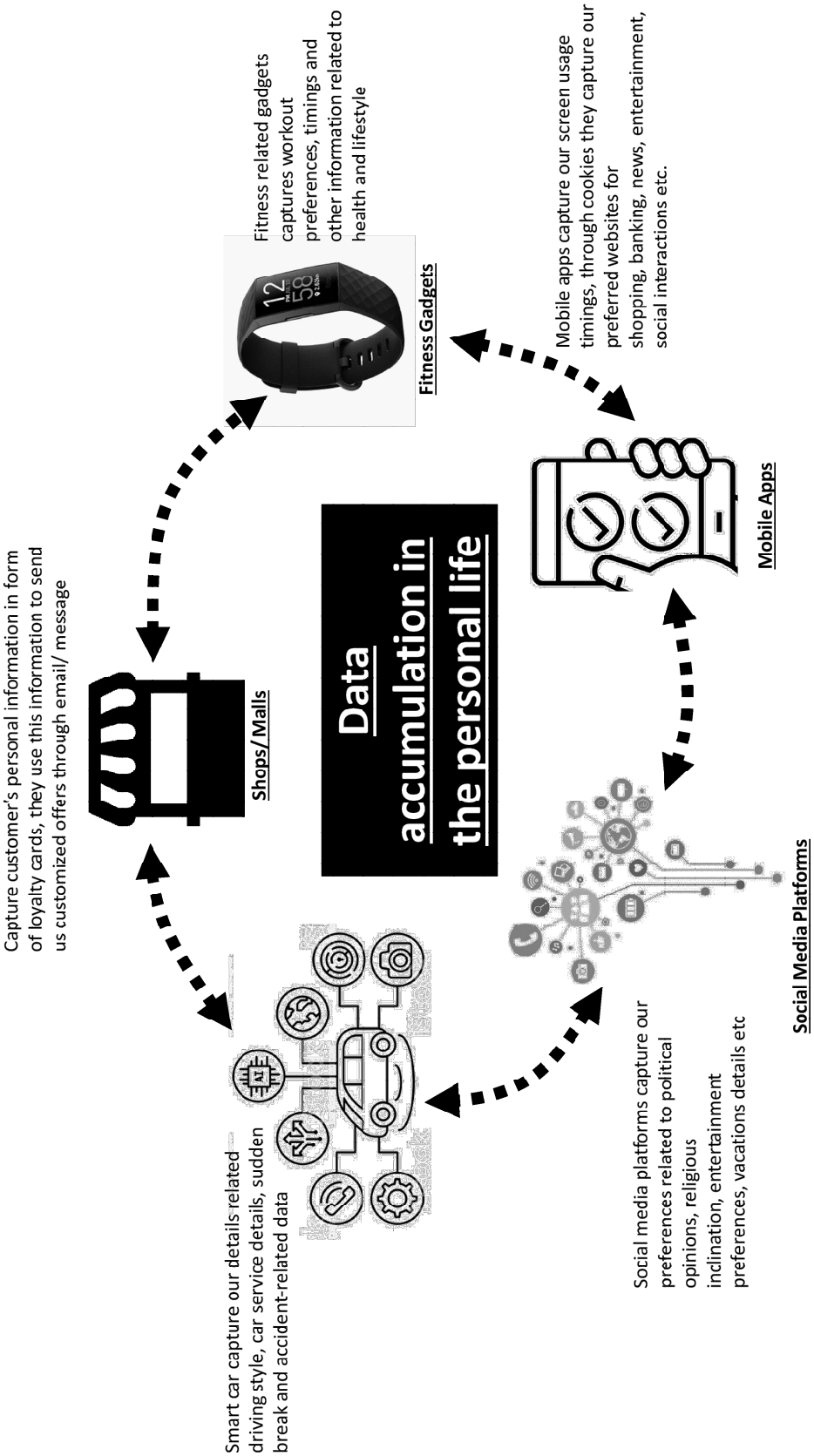


Fig 1.2 Sources of Data Accumulation in Personal Life

❖ Data Sources in Business World :

Data is the primary fuel in the business world as all planning and execution activities at different phases of the business life cycle generate and consume a large amount of data. Information about customers, vendors, competitors, employees, feedback plays an essential role in decision making in the business world.

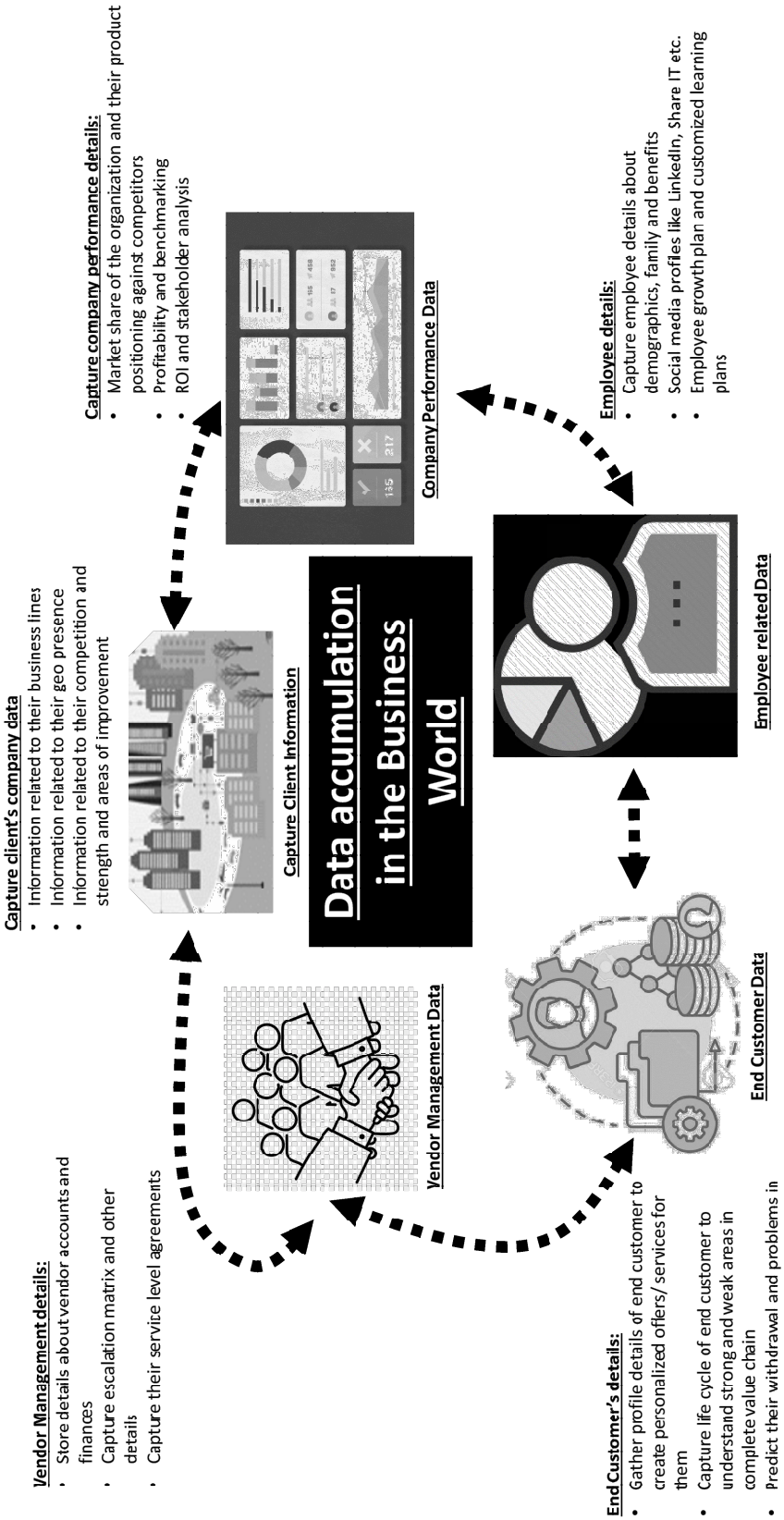


Fig 1.3 Sources of Data Accumulation in the Business World

1.3 Life Cycle of Business Analytics Process :

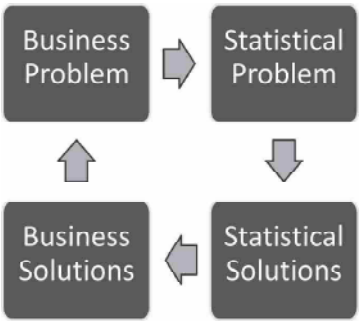


Fig 1.4 Business Problem-Solving Process

Business Analytics projects start with a correctly framed business problem; analysts convert this business problem into an analytical problem so that problem becomes measurable and its impact can be understood by management in terms of money, time, and resources. As per the nature of the problem, analysts figure out relevant data and tools to solve the statistical problem. In the end, they again summarise their statistical findings in terms of Business Solutions which can be easily understood by business executives and converted into a sustainable and replicable solution.

The high–level data–driven business analytics process is mapped in fig (1.5). These are very high–level guidelines; organizations use this process as per their requirements and customize it accordingly. It is an iterative process where we complete one analytics initiative and raise the bar to pick up a challenge of the next level to bring more value to business and stakeholders.

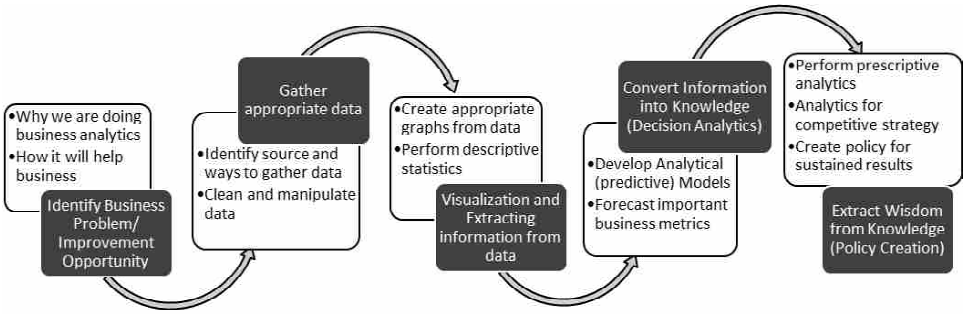


Fig 1.5 Life Cycle of Solving a Business Problem

In today's era of globalization, organizations are getting diversified in terms of lines of business and spreading their business across geographies. At that scale of business without a properly structured approach to execute analytics, it would be impossible to bring efficiency and effectiveness in their business operations. It leads to value generation to essential stakeholders. Below is the high–level description of the business analytics process steps :

1. Identify the opportunity for improvement to create value for business
2. Select significant sources to gather relevant data for analysis with the help of the right set of tools, here we also clean the data and put it in the right format
3. Post validation of data, we use suitable visualization which is easy to understand and also convey vital information about current business scenarios

4. Based on prior steps we take appropriate decisions by keeping budget, time, and resources in mind also predict the output in the short and long run
5. To get sustainable solutions, we put suitable controls in the system which results in optimizing outputs and helps the organization in creating a competitive strategy

1.3.1 Scope of Business Analytics – Where Does it Fit on Business Canvas :

Business Analytics works like a magical crystal ball that can solve tiny day–to–day problems like reducing packaging errors to enormous and complex business problems such as designing a space shuttle or setting up a nuclear power plant.

Let's try to understand the applications of Business Analytics at a different level with the help of fig 1.6

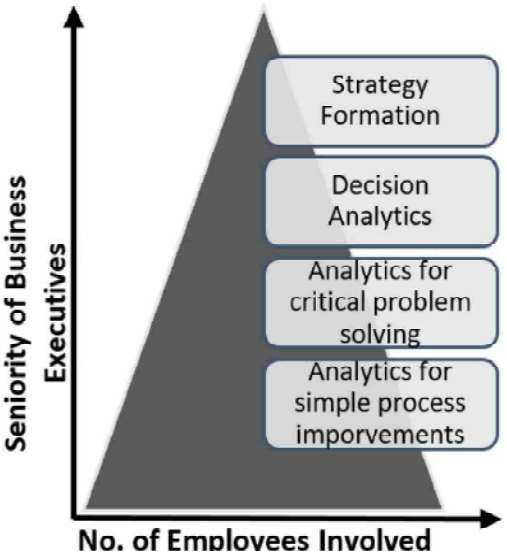


Fig 1.6 - Scope of Business Analytics at Different Levels

At the bottom, more and more workforce get involved in Business Analytics initiatives, and they identify smaller problems in their day–to–day operational activities with the help of quick applications of analytics, here analysts help them to find solutions and incorporate these solutions into their way of working. In business, we call it Kaizen, which is a Japanese word that means "change for good". In the next level generally, specialized analysts get involved in solving critical business problems; typically, they tackle these in terms of business projects. The next level is Decision Analytics, where departmental heads get engaged in executing more significant business initiatives, here the analytics team uses complex tools and massive data for impactful and detailed analysis. At the top–level leadership gets involved in using comprehensive analytics for strategy formation.

Check Your Progress – 1 :

1. Which of the below tasks can be handled by business analytics ?
   a. Predicting business results
   b. Finding patterns in the data
   c. Validating business assumptions
   d. All of the above

2. The business analysis process starts with :
   a. Analysis of data
   b. Collecting data
   c. Determine the need of the process
   d. Predict the business outcomes

1.4 Classification of Business Analytics :

"Errors using inadequate data are much less than those using no data at all". – ***Charles Babbage***

Business analytics is a structured approach that brings value to the business in a very systematic way. It is always great to start with the basics and build analytical capabilities step by step. Analytics can be classified into four levels which help the organizations to become mature in terms of analytical proficiency. Below are the brief descriptions :

1. **Descriptive Analytics :** This is the simplest form of analytics; it summarises current business status in the way of narrative and innovative visualization. It emphasizes *"what is going on in the business."*
2. **Diagnostic Analytics :** It provides the reasons for descriptive analytics; generally, analysts provide visual reasoning in terms of interactive dashboards. It emphasizes *"why did it happen."*

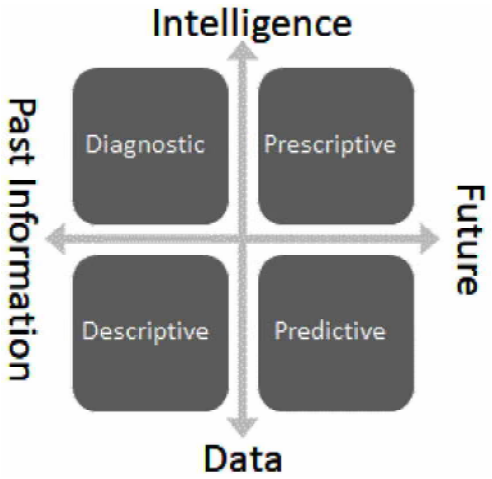


Fig 1.7 Classification of Business Analytics

3. **Predictive Analytics :** It predicts the business metrics in the short and long run, here we use advanced machine learning algorithms to analyze massive data from different sources to predict the future values of essential business metrics. It emphasizes *"what will happen in the future".*
4. **Prescriptive Analytics :** It suggests all favourable business outputs for any specified course of action, it also offers the pros and cons for each course of action. It optimizes the business results suggested by descriptive and predictive analytics. It emphasizes *"how can we make it happen."*

Among these different types of analytics, there is no superior or inferior to each other, but all play a significant role during different phases of solving a business problem. These levels of analytics are sequential and linear in nature which means an organization cannot implement diagnostic or predictive analytics without having proper descriptive analytics

implemented throughout the organization. There is always a strong relationship between levels of analytics used Vs business value derived. But we need to have useful descriptive analytics to start and later emphasize diagnostic analytics which provides a base for Predictive Analytics and finally, prescriptive analytics tells us the way to optimize solutions recommended by all three primary types of analytics. Let's discuss them in detail with a few examples and tools involved :

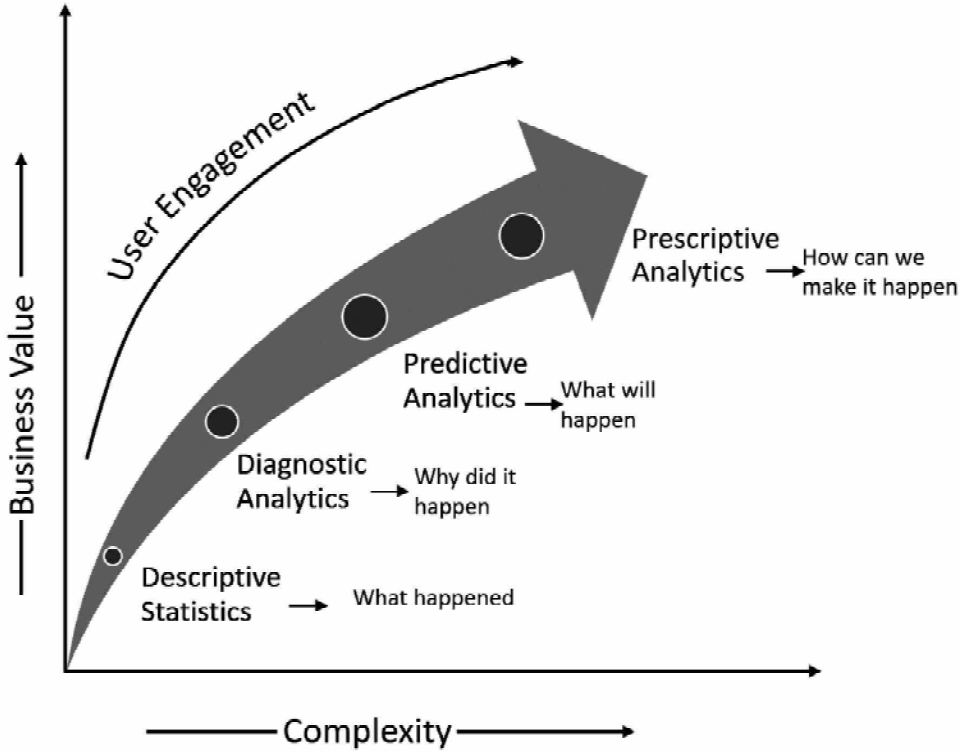


Fig 1.8 Types of Business Analytics

1.4.1 Descriptive Analytics :

If the statistics are boring, then you've got the wrong numbers.

– Edward R. Tufte

It is the most simplistic form of analytics; it digs deeper into past data and tells us "what has happened and when did it happen". Here, the main objective is to summarize data into useful insights with the help of easily interpretable visualization techniques. It highlights past trends that lead to valuable insights for business, but we do not emphasize here "why these trends happened". Here we try to connect the dots to make an exciting analytics storyboard, which helps us understand the problems and their impacts on business and stakeholders.

Descriptive statistics generally used to show the performance of an organization against its leading Key Performance Indicators (KPIs). It also indicates trend analysis for these KPIs against competitors and among different geographies. Information obtained during descriptive analytics can be used to make better business decisions.

We use Descriptive Analytics when we want to summarize the story of an organization's performance (mostly in the form of

Dashboards). It provides us with a comprehensive view by joining different things together to highlight hidden trends and insights.

Information extracted from descriptive analytics helps leadership to take actions to make things better, and now with the help of Big Data technologies, management sees the real–time progress of various vital business metrics. Management sees a complete picture by benchmarking company performance against the past few years and key competitors. Below are a few examples of knowledge extracted from descriptive analytics :

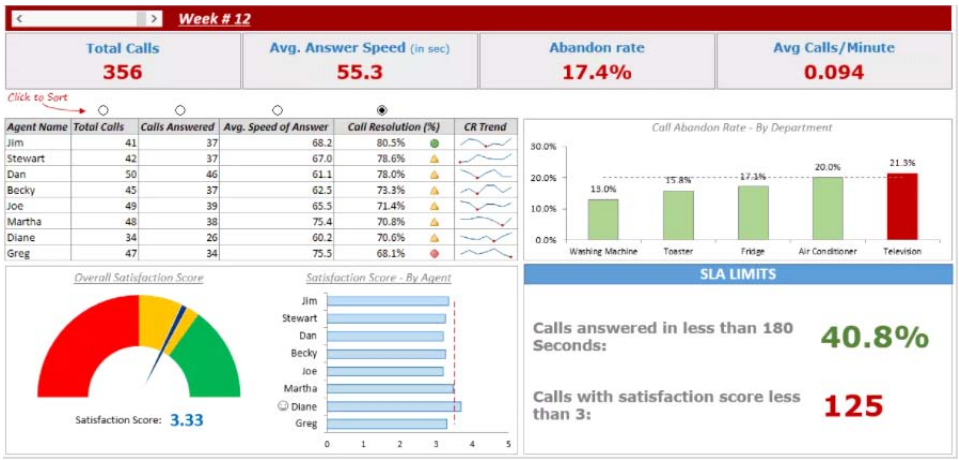


Fig 1.9 Sample Business Dashboard

1. Men convert credit card transactions into EMI more than women; banks should target men for EMI promotion as they are more likely to opt for the promotional campaign.
2. Internet routers show lots of information packets drop during 4–6 PM due to high congestion, support team to provide extra bandwidth during this time slot for seamless customer experience.
3. More cars come for servicing during monsoon due to water problems so garage should think about hiring part–time mechanics during monsoon to cater to the temporary demand.
4. Analysts find a strong correlation that customers quit websites on the checkout screen and customers who use mobiles to browse company website; it indicates customers do not see website mobile friendly hence the company should launch a mobile version of company website soon.
5. The health department observes a recurring hike in malaria disease in a particular locality every year during the rainy season; they find water bodies are open in that area which is causing mosquito breeding.

❖ **Essential Tools used in Descriptive Analytics :**

1. **Statistical Summary :** It provides statistical descriptions for a given business metric, e.g. Mean, Median, Standard Deviation, Percentile, Interquartile range, etc. (will study these techniques in detail in upcoming chapters).

2. **Z–Score :** Z Score tells us how far (in terms of standard deviation) is a particular value of x from its mean.
3. **Coefficient of Variance :** It is a ratio where we divide standard deviation with mean. Alone mean or standard deviation are not appropriate methods to measure to benchmark different company performance metrics. It is important to consider both centrality and spread of data to make it comprehensive.
4. **Interquartile Range :** It is an important measure to gauge the variation in the dataset. The height of the interquartile box is the difference between the third and first quartile of data. It is quite powerful as it removes very small and very big data points.

1.4.2 Diagnostic Analytics :

"There is wisdom in always exploring the counterpoint – sometimes a silver cloud has a dark lining too." **– *Gyan Nagpal***

Diagnostic analytics provides "Why did it happen in my business". It is a bit advanced where analysts examine data in order to find reasons for business problems or opportunities. One of the ways is drawing correlation among various business metrics as sometimes changes happening in one metric lead to change in other metrics, e.g. reduction in production because of higher absenteeism in that week or drop in quality because of new training batch going live in the previous week as new guys make more errors than tenured one. There are numerous ways to perform causal analysis. Below are a few examples :

Fig 1.10 5-Why Diagnostic Analysis

1. A company found that employees are not completing learning certifications, analyst diagnosed that most of the employees are stuck at programming assignments, where programming interface was not supportive/ flexible, and there was no way to get hints/ help to proceed further.
2. There was a low hotel check–in feedback score; analysts diagnosed that front office executive enters customer details which are not required fields during check–in itself. Typing speed and system navigation is also very slow which is resulting in a longer check–in time.

3. The product return rate was very high during last month, and it found that out of total return items more than 60% of products were supplied by two vendors only, where the vendor provided the wrong specification about products.

❖ **Essential tools used in Descriptive Analytics :**

1. **Correlation Analysis :** It is a statistical measure that indicates the strength of the relationship between two variables. It is a critical causal analysis technique that helps in identifying reasons in terms of relationship with other metrics.
2. **5 Why Analysis :** It is a very structured approach where we try to dig into a problem and peel it layer by layer to reach the root cause of the problem. Solutions to root cause provide us with sustainable solutions.
3. **Cause and Effect Analysis :** Here, we identify all possible reasons for one problem then we pick up all the reasons as a problem one by one and try to find other causes for that problem. In this way, we create a diagram that looks like the skeleton of a fish because of its looks. It is also known as the fishbone diagram.

1.4.3 Predictive Analytics :

"The future belongs to those who see possibilities before they become obvious". – ***John Scully***

Predictive analytics is the heart of business analytics, it aims to help the organization by predicting probabilities of occurrence of a future event or future values of any essential business metrics, e.g. sales in the next month/ quarter, employee attrition, and product return rate, etc.

Once organizations have a stable setup for descriptive analytics, means data sources have been identified, and those data sources are supplying data about important metrics continuously into leadership dashboards. Predictive analytics combines this historical data with advanced business protocols (policy and rules) to forecast future values of business events. Predictive analytics allows organizations to become forward–looking, providing an appetite to consume calculated risk by anticipating customer behaviour and business outcomes.

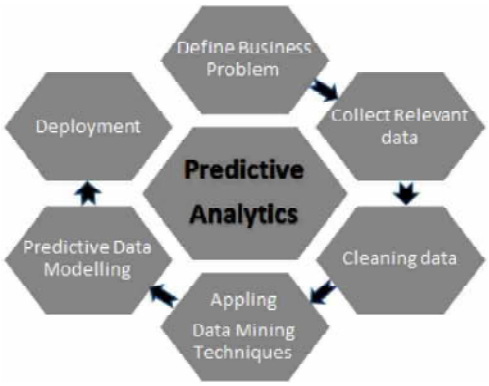


Fig 1.11 Predictive Analytics Cycle

❖ **Below is a List of Predictive Analytics Examples :**

- Financial analysts predict the share value/ gold prices/crude oil prices in the next few days or weeks with the help of predictive modelling.

- Airline companies predict competitive airfares to extraordinary and ordinary days also they indicate how much airfare should be increased as per the increased customer's traffic on their websites.
- Netflix predicts the next movie customers want to watch, more than 80% of customers select their next movie from their recommendation list. In this way, Netflix earns more rental income from regular customers by suggesting them the next film or programs.
- IRCTC predict the probability to confirm the seat which provides assurance to the customer about their seat confirmation, it helps to attract more customers to their portal.
- Taxi services predict the demand during different time slots and change their tariff accordingly.

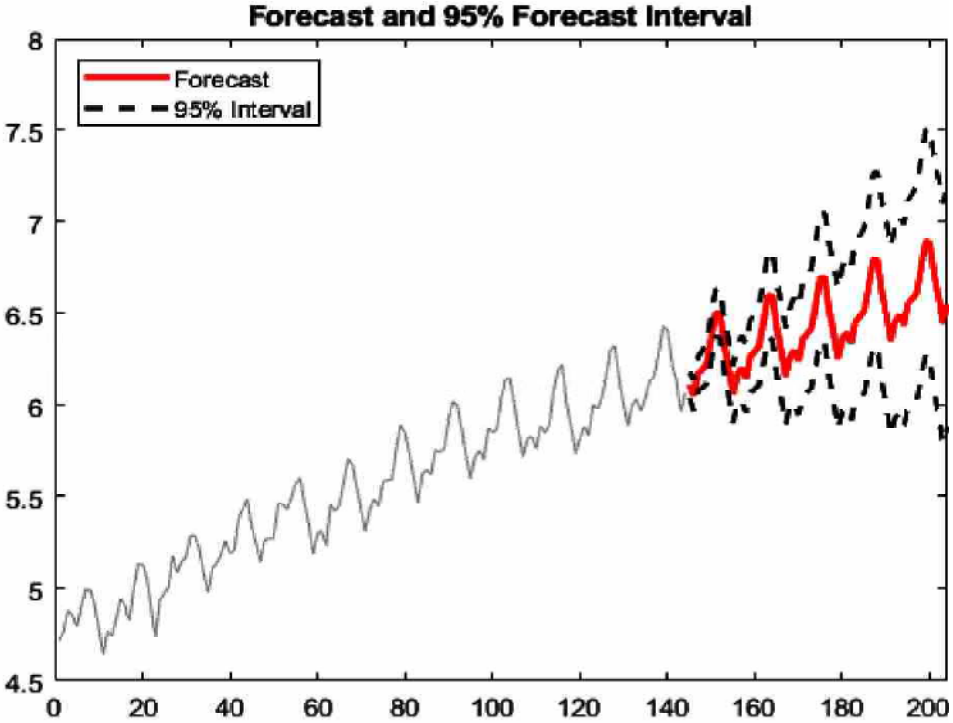


Fig 1.12 Example of Forecasted Demand

❖ **Important Tools used in Predictive Analytics :**

1. **Regression Analysis :** It establishes the mathematical relationship between input variables and output variables, which means if we can calculate the future value of output for any given input, e.g. sales forecast for next month.
2. **Logistic Regression :** It is a classification predictive analytics technique that can predict the output class for any given set of inputs. E.g. by providing customer demographics logistic regression can indicate whether the customer will default bank loan in the future or not.
3. **Decision Tree :** Most of the time, we use a decision tree as a classification technique; it tells us the output probability of the output variable for various permutations of our input variables. Although it can be used for continuous output variables also.

4. **Clustering Techniques :** These techniques segregate our customers into a few logical segments so that we can create tailored offers for a different type of customers as per their needs and interests.
5. **Random Forest :** It is another very famous business analytics technique that uses a collaborative approach to solve the problem by generating a large number of predictive models. Their accuracy is generally better.

1.4.4 Prescriptive Analytics :

"Without prescriptive analytics, you are navigating with rear–view mirror". ***– Martin Zync***

Prescriptive analytics solves the complex business problem as it is the most advanced form of analytics, where we have to choose the most optimal way to increase important business metrics. As we discussed earlier, all type of Business Analytics is sequential; therefore, perspective analytics can be applied once we have sound business knowledge from descriptive and predictive analytics.

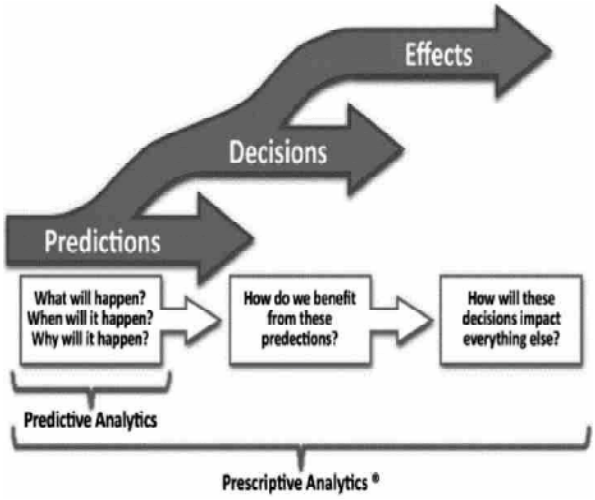


Fig 1.13 Prescriptive Analytics

Descriptive and predictive analytics suggest to us various ways to improve business performance while prescriptive analytics tells us the pros and cons of all alternatives and try to provide the optimal outputs by keeping minimum risk in execution.

Prescriptive analytics is not limited to predict "what will happen" and "when will it happen" but it also tries to reveal "why it will happen" and "what would be the impact on the business".

It relies on slightly complex machine learning algorithms, and its success depends on the organization's maturity to ask the right questions; otherwise, it may completely reverse the business outcomes, e.g. if because of complex prescriptive analytical model banks make the loan approval process too tricky then they may lose several potential customers.

❖ **Below are Examples of Prescriptive Analytics :**

- In 2019, there was a prediction of the cyclone on coastal areas of Gujarat (by predicting changing airspeed, varying wind direction, and mathematical relationship between low pressure in the ocean with changes in cyclone intensity) therefore Government and disaster management team had taken proactive actions in shifting citizens from coastal areas to save places, and they stopped fishermen from

going to sea and arrange comfortable camps. While in a similar situation in 1999 we lost approx. 10,000 lives due to cyclone.

- At the time of launching a new service or a product into the market, organizations have to keep various factors into the mind like the cost of the product, features of the product, geographies in which they will launch first, customer segments whom they want to attract, marketing channels for product promotion, etc. By getting analytical results from descriptive and predictive analytics, analysts apply prescriptive analytics to decide the right mix of all these factors to make a product launch successful.
- In agriculture crop yield depends on various factors like rainfall, soil type, demand in the market, etc. Analysts apply prescriptive analytics and suggest the best kind of crop in different regions as per the rainfall and demand forecast in that season.
- Banks use prescriptive analytics to identify investment options for their customers to maximize their returns and minimize risk. They balance customer's portfolio by having an optimized ratio of equity, debt, and other types of funds.

❖ **Important Tools used in Prescriptive Analytics :**

1. **Linear Programming :** In linear programming, we optimize the objective functions like revenue, market share, customer feedback ratings by also keeping constraints in the model like budget, no. of people deployed, etc. as linear functions.
2. **Analytical Hierarchy Process :** We apply these techniques in scenarios where we have to identify the best solution among various available options, and there is the list of criteria's to select the solution, e.g. select best cloud service providers among top 5 organizations by keeping multiple factors into consideration like budget, customer service, flexibility to upgrade, backup services, maintenance cost, etc.
3. **Combinational Optimization :** It involves identifying optimal solutions from a considerable number of finite solutions, e.g. the travelling salesman problem, vehicle routing problem, etc.

Check Your Progress – 2 :

1. Which type of business analytics generally uses statistical and machine learning algorithms.

   a. Descriptive b. Predictive c. Prescriptive d. Diagnostic

2. Which type of business analytics provides recommendation about optimization of business outcomes, includes simulation, etc.

   a. Descriptive b. Predictive c. Prescriptive d. Diagnostic

3. Which type of business analytics gain information from historic data and provide us various reports, visualization, and scorecards, etc.

   a. Descriptive b. Predictive c. Prescriptive d. Diagnostic

1.5 Challenges in Business Analytics :

Establishing profitable business analytical capabilities in any organization requires complete cultural changes; it is one of the most challenging tasks for leadership. Below is the list of few crucial challenges which organizations are facing today :

- Management is comfortable in taking decisions based on their experience and learning, and due to limited statistical knowledge they hesitate in adapting scientific calculations to predict the outcome for their respective business metrics.
- Investment in technological aspects for example cost of cloud infrastructure, licenses for data science software, cost of seasoned business analytics, etc.
- Availability of talent pool is another big challenge, as business analysts should have very good technical, business, and statistical skills hence it is difficult to find good professionals with the right combination of these skills.
- Developing in house talent so that essential functions can solve their analytical assignment themselves without taking the help of the centralized expert team.
- Reporting is the main time–consuming activity for middle management, to minimize its organizations have to build a culture of clean data feeding discipline into the systems so that reports can be published by software without human interventions.
- Most of the softwares are available in open–source form, one side these minimize huge cost on licenses, but the flip side is that there is minimum or no technical support hence learning curve for these technologies is slower than other.

Soon analytics will emerge even more vital driving force for enhancing business performance, and leadership will move to data–centric decisions than experiential ones alone. Finding more innovative ways to solve traditional and new business problems will be the critical success factor for deploying business analytics among various vital functions of the organization.

Check Your Progress – 3 :

1. Which type of business analytics emphasize "What did happen."
   a. Predictive analytics b. Diagnostic analytics
   c. Descriptive analytics d. Prescriptive analytics
2. The main essence of Diagnostic analytics is
   a. Choosing the best action among alternatives
   b. Thinking forward to avoid possible consequences
   c. Splitting big problem into smaller ones
   d. Finding the root cause of why something happened good or bad

3. Which of the following is NOT an important component of business analytics ?

   a. Technology b. Data Science

   c. Business Domain Knowledge d. People management

4. Out of the below analytical tool which one doesn't fall under prescriptive analytics

   a. Regression analysis b. Linear programming

   c. Analytic hierarchy process d. Combinational optimization

5. Which type of analytics provides visualization of all critical business metrics

   a. Predictive analytics b. Descriptive analytics

   d. Descriptive analytics d. Prescriptive analytics

6. Out of the following which level of business analytics, the maximum workforce is involved

   a. Analytics for critical problem solving

   b. Decision Analytics

   c. Analytics for simple process improvements

   d. Strategy Formation

7. Which type of analytics helps in choosing the best possible solution among alternatives

   a. Prescriptive analytics b. Descriptive analytics

   c. Descriptive analytics d. All of above

8. Out of the below analytical tool which one doesn't fall under predictive analytics

   a. Decision tree b. Clustering techniques

   c. Linear programming d. Regression analysis

9. Which technique establishes a mathematical relationship between input and output variable

   a. 5–Why analysis b. Z–Score

   c. Regression analysis d. Linear programming

10. Which of the following sequence is correct ?

    a. Data $\rightarrow$ Information $\rightarrow$ Decision $\rightarrow$ Optimization

    b. Information $\rightarrow$ Data $\rightarrow$ Decision $\rightarrow$ Optimization

    c. Optimization $\rightarrow$ Information $\rightarrow$ Decision $\rightarrow$ Data

    d. Decision $\rightarrow$ Information $\rightarrow$ Data $\rightarrow$ Optimization

1.6 Let Us Sum Up :

1. In the current era of business 4.0, clean and comprehensive data is available in all type of organization therefore the scope of business analytics has increased many folds.
2. Business analytics is a structured approach that brings value to the business in a very systematic way. It is always great to start with the basics and build analytical capabilities step by step. Analytics can be classified into four levels which help the organizations to become mature in terms of analytical proficiency.
3. Business Analytics projects start with a correctly framed business problem; analysts convert this business problem into an analytical problem (writing in terms of business metrics and quantitatively impact on business), analysts figure out relevant data and tools to solve the statistical problem. In the end, they again summarise their statistical findings in terms of Business Solutions which can be easily interpreted and converted into a sustainable and replicable solution.
4. Descriptive analytics is the most simplistic form of analytics, it digs deeper into past data and tells us "what has happened and when did it happen".
5. Diagnostic analytics provides "Why did it happen in my business". It is a bit advanced where analysts examine data in order to find reasons for business problems or opportunities.
6. Predictive analytics aims to help the organization by predicting probabilities of occurrence of a future event or future values of any essential business metrics, e.g. sales in the next month/ quarter, employee attrition, and product return rate, etc.
7. Prescriptive analytics solves the complex business problem as it is the most advanced form of analytics, where we have to choose the most optimal way to increase important business metrics.

1.7 Answers to Check Your Progress :

Check Your Progress – 1 :

1. d **2.** c

Check Your Progress – 2 :

1. b **2.** c **3.** a)

Check Your Progress – 3 :

1. b	**2.** d	**3.** d	**4.** a	**5.** b
6. c	**7.** a	**8.** c	**9.** c	**10.** a

1.8 Glossary :

Descriptive Analytics : This is the simplest form of analytics; it summarises current business status in the way of narrative and innovative visualization. It emphasizes on *"what is going on in the business."*

Diagnostic Analytics : It provides the reasons for descriptive analytics; generally, analysts provide visual reasoning in terms of interactive dashboards. It emphasizes on *"why did it happen."*

Predictive Analytics : It predicts the business metrics in the short and long run, here we use advanced machine learning algorithms to analyze massive data from different sources to predict the future values of essential business metrics. It emphasizes on *"what will happen in the future"*.

Prescriptive Analytics : It suggests all favourable business outputs for any specified course of action, it also offers the pros and cons for each course of action. It optimizes the business results suggested by descriptive and predictive analytics. It emphasizes on *"how can we make it happen."*

Z–Score : Z Score tells us how far (in terms of standard deviation) is a particular value of x from its mean.

Coefficient of Variance : It is a ratio where we divide standard deviation with mean. Alone mean or standard deviation are not appropriate methods to measure to benchmark different company performance metrics. It is important to consider both centrality and spread of data to make it comprehensive.

Regression Analysis : It establishes the mathematical relationship between input variables and output variables, which means if we can calculate the future value of output for any given input, e.g. sales forecast for next month.

Logistic Regression : It is a classification predictive analytics technique that can predict the output class for any given set of inputs. E.g. by providing customer demographics logistic regression can indicate whether the customer will default bank loan in the future or not.

Decision Tree : Most of the time, we use a decision tree as a classification technique; it tells us the output probability of the output variable for various permutations of our input variables. Although it can be used for continuous output variables also.

Linear Programming : In linear programming, we optimize the objective functions like revenue, market share, customer feedback ratings by also keeping constraints in the model like budget, no. of people deployed, etc. as linear functions.

1.9 Assignments :

1. Write down important tools used in descriptive analytics.
2. Write down important components of business analytics.
3. Write down important data sources in the personal and business world.
4. Mention the scope of business analytics at different levels in an organization.

1.10 Activities :

Suppose you have been hired as an analyst at SDFG Bank, your manager has provided you home loan data for the last 5 years. You have to build up an application with the help of the Data Science team to predict potential loan defaulters.

Write down various analyses you can do on this data from a Descriptive, diagnostic, predictive, and prescriptive business analytics perspective.

1.11 Case Study :

Nirav Soft drinks is the largest beverages bottler and distributor of carbonized non–alcoholic, bottled beverages in North Gujarat, and one of the largest bottlers in western India

How ABC Image Recognition for Retail is being used :

Before using ABC's scanning and imaging technology, Nirav Soft drinks was dependent on limited and completely manual measurements of products/ cold drink cans in–store, as well as delayed/ lost data sourced from either fax or phone conversations.

Nirav Soft drinks sales representatives/ manager after using ABC Retail Execution scanning image–based technology to take pictures of stores shopping shelves with their hand gadgets or mobile devices; these pictures were sent to the Trax Cloud and analysed, providing actionable reports within few minutes to sales representatives/ manager and providing more detailed online assessments to company management.

❖ **Value Proposition :**

Real–time images of inventory allowed the sales representative to identify performance gaps and apply corrective actions in store. Reports on shelf share and competitive insights also allowed reps to strategize on opportunities in–store and over the phone with store managers/ promoters.

Nirav Soft drinks gained a 7.3% market share in the western India region within five months.

Questions :

1. Write down the various type of Descriptive, diagnostic, predictive, and prescriptive business analytic you can imagine for Nirav Soft drinks ?
2. Which important business metrics will improve through this technology intervention ?
3. How this change will bring value for store owners, Nirav soft drinks, and end customers ?

1.12 Further Readings :

- "Mathematical Methods of Statistics," Princeton University Press, Carmer H. (1946)
- "Super Freakonomics," Penguin Press, Levitt S. D. and Dubner S. J. (2009)
- "An Explanation of the Persistent Doctor Mortality Association" Journal of Epidemiology and Community Hearlth, Yough F. W. (2001)
- "Data Strategy : How to Profit from A World of Big Data, Analytics and The Internet of Things", O'Reilly Media, Bernard Marr
- "Predictive Analytics : The Power to Predict Who Will Click, Buy, Lie, or Die", Wiley, Eric Siegel

Unit 2

DESCRIPTIVE ANALYTICS

: UNIT STRUCTURE :

2.0 Learning Objectives :

After learning this unit, you will be able to understand :

- Basic concepts of descriptive analytics and how it is influencing the decision process of organizations
- Data types and scale of measurements
- Understanding measures of data centrality and variability
- Important tools and techniques for data visualization

2.1 Introduction :

In this unit we will study about descriptive statistics and different type of measurement scales used in descriptive statistics. Basic concepts about sampling theory and how it helps business in saving huge revenue and time to conduct the study/ analysis. We will also understand various techniques to measures centrality and variability of data and how these help us to take better business decisions. At the end we will touch upon few data visualization techniques which helped us to understand the data and analysis better.

2.2 Introduction to Descriptive Statistics :

"Data are just summaries of thousands of stories – tell a few of those stories to help make the data meaningful" ***– Chip & Dan Heath***

Descriptive analytics is the foundation of any analytical project. It focuses on **"What has happened"** by visualizing the current business performance.

We develop dashboards to showcase important business metrics. By putting together historical performances, data helps us to see hidden inferences which lead to better business decisions. Innovative visualization plays an important role in displaying a comprehensive picture of organizational progress. We can also include competitor's information which will help management take appropriate strategic actions.

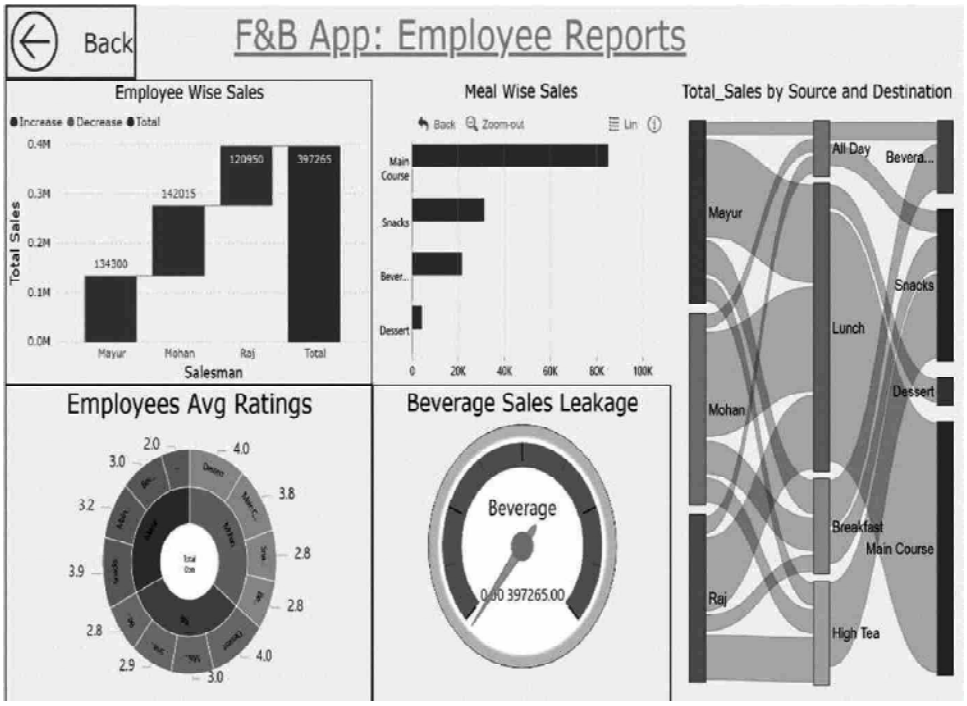


Fig 2.1 Sample Dashboard

2.3 Different Type of Data Measurement Scales :

Irrespective of the size and nature of an organization, data is getting generated at a high pace, e.g. sales data, employee data, inventory, customer–related data, etc. This data comes in various forms like numeric, text, alphanumeric and it may be captured in different scales, e.g. sales and year of experience both are numeric data, but they can get measured in rupees and years respectively. Different types of analytical tools and techniques are available for different types of data. We divide data into two important categories :

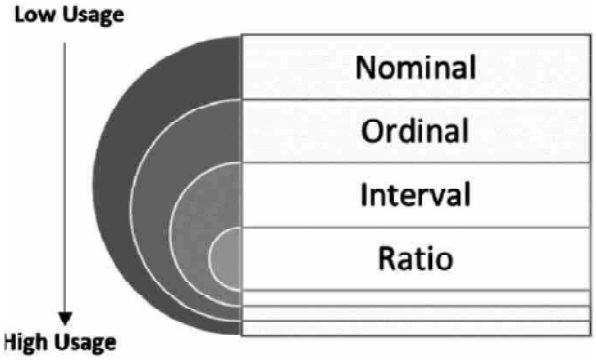


Fig 2.2 Usage Level of Data

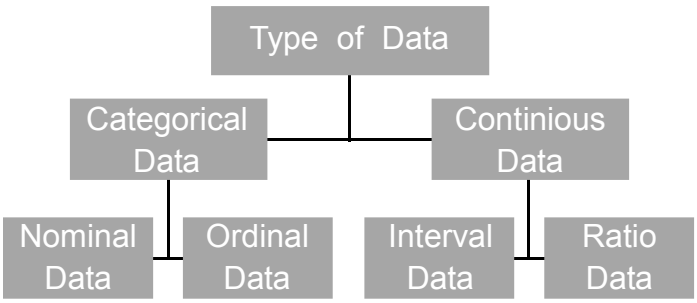


Fig 2.3 Data Types

2.3.1 Categorical Data :

This type of data is also known as discrete data. Categorical variables have a finite number of groups, e.g. payment method, gender, income group, property type, etc. Categorical data can be further divided into the following two types :

a. **Nominal Data :** In this type of categorical variable, there is no logical sequence among categories, which means one category is not superior or inferior to others. Categorization is just a type of segregation of data into different groups. For example, people wearing a dark colour shirt or light colour shirt, north Indian city and south Indian city, software or textile industry, etc. One of the limitations is that we cannot perform any mathematical operation on nominal data.

b. **Ordinal Data :** It is better than nominal data in terms of usage potential for any analysis. There is a logical sequence in terms of the superiority among categories. Ranked or ordered data generally come into this category. E.g. pass or fail status, feedback on Likert scale – Good, fair, bad, the risk level of investment bonds – high and low–risk bonds, etc. Mathematical operations are not possible on ordinal data.

2.3.2 Continuous Data :

Data that can be measured on a continuous scale–like height, weight, money, time, etc. It can be divided into halves any number of times. Continuous data can be further divided into the following two types.

a. **Interval Data :** Interval data is always numeric, and there is an equal distance between consecutive interval data points. Data comes from an interval range like e.g. temperature, percentage return of a share, percentage change in the gold price, intelligence quotient scores for an exam, etc. Another important point is that zero is like another number (it does not mean that value is none or missing). For example, a zero degree is a valid temperature. On such data, we can do summation and subtraction as mathematical operations, but division and multiplication wouldn't make sense. The difference in heat between 90^{o}C and 70^{o}C is the same as the difference between 50^{o}C and 30^{o}C, but we cannot say that 80^{o}C

is not twice as hot as 40°C or somebody with IQ scores 70 is not twice as smart as another student with IQ score 35.

b. Ratio Data : In terms of statistical relevance, this is the highest level of data which is desirable for the application of statistical tools and techniques. On such data, we can do all types of mathematical operations. For example, if product A's sale is ₹ 50,000 and product B's sale is ₹ 25,000, then we can interpret that sale of product A is twice the sale of product B.

2.4 Population and Sample Size :

For a statistical problem if we have access to all possible data sets and the entire dataset is used for analysis, then it is an analysis of population data. Below are a few examples

- The government analyses the base of census data to calculate important indices like per capita income, employment rate, GDP, etc.
- The election commission uses entire voters' data to provide information on the proportion of graduate voters, how many voted in the last election etc. Generally, the size of the population is huge in terms of data hence it requires large space to store and lots of computer memory to run the analytical queries.
- Big organizations like Indian Railway, Indian Army, etc. use entire data of their employees (which is in millions) to analyze their salary data.

It is very expensive, time–consuming, and requires advanced computation power to analyze population data. Therefore, it is not advisable to analyze population data until there is no other way to get the information.

The sample is a logical subset of the population, which mimics the population. Selecting a relevant sample out of the population is challenging, but it makes analysis faster, precise, and economical. There are standard guidelines from statisticians to calculate the relevant sample size, appropriate sampling methodology, and tool to analyze sampled data. Below are a few examples of sample–based analysis :

- During exit polls, companies want to predict the election winner much before the result is declared by the election commission, they ask the opinion of few sampled voters and declare their predictions.
- Companies ask preferences of a few sampled candidates to decide features for their upcoming mobile phones.
- Companies call few sampled customers in order to understand the area of improvements about their product or services.

Check Your Progress – 1 :

1. Final grades (A, A+, B, etc.) in university exams is an example of :

   a. Nominal data b. Ordinal data

   c. Ratio data d. None of above

2. Generally, we prefer to analyze the _____ data as it is not advisable to analyze the entire data even if we have access to it as it leads to consuming more time, resources and efforts.

2.5 Components of Descriptive Statistics :

The focus of Descriptive analytics is on two important dimensions of data : Central tendency of data and dispersion of data.

Either central tendency or dispersion alone cannot provide a complete picture of the dataset. E.g. if one river is 4.5 feet deep towards its banks and 5 feet deep for the next two meters from the bank and 6 feet deep in the middle. Let's say the average depth of the river is 5.3 feet. Can anybody with 5.6 feet of height cross the river (if he/she doesn't know swimming) ? Not ! Therefore, having information only about central tendency is not enough to make a better business decision we need to know the data dispersion also.

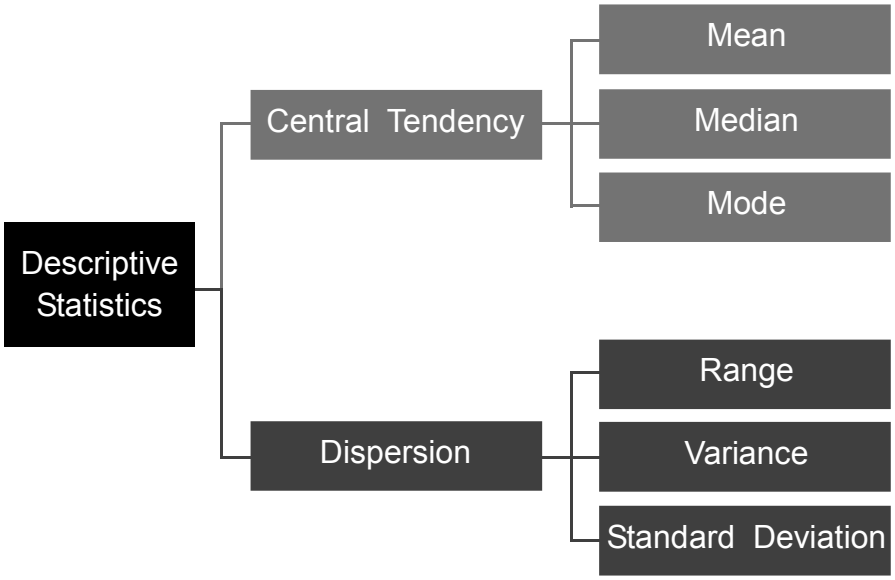


Fig 2.3 Components of Descriptive Analytics

2.5.1 The Measure of Central Tendency :

One of the most important measures used to describe a dataset statistically is the measure of central tendency. It provides us with information about the centre point of the dataset in terms of a single value. Mean, median, and mode are the most important measures, but frequently percentile and quartile are used as a measure of central tendency.

1. Mean (Average) Value :

Arithmetic mean is the most frequently used measure of central tendency. For simplicity, we refer to arithmetic mean as "mean". It is the summation of all numbers divided by the total numbers. The population means is represented by a Greek letter μ while the sample means is represented by $\overline{X}$ whereas N is the number of total data in population while n is the total number of data points in the sample.

$$\mu = \frac{\Sigma x}{N} = \frac{x_1 + x_2 + x_3 + x_4 + ... + x_n}{N}$$

MS Excel formula for the mean is **AVERAGE (array of numbers).** The array represents all numbers in sequence.

Example 2.1 : Calculate the mean for a student who has scored 90, 76, 62, 91 and 56.

Solution : Step1 : Add all numbers, which is 375

Step2 : Divide by the total number of data points, which is 5 here

$$\mu = \frac{90 + 76 + 62 + 91 + 56}{5} = 75$$

Example 2.2 : Data is available in the form of frequency

Age	16	18	20	25	28
Total Students	2	1	5	2	10

Weighted Average : If a few data points amongst the data set are to be given more importance than others, the weighted average method can be applied. In the below example, few subjects are more important. In that scenario, we calculate the weighted average; it's calculated the same as we have calculated the mean (average) from the frequency table in the above example.

Example 2.3 : Data is available in the form of frequency

Sr. No.	Subject Name	Credit (Weight)	Score (out of 10)
1	Statistical Analysis	5	7
2	Data Mining	4	8
3	Logical Reasoning	4	9
4	English	3	8

$$\text{Weighted Average} = \frac{5 \times 7 + 4 \times 8 + 4 \times 9 + 3 \times 8}{(5 + 4 + 4 + 3)} = \frac{127}{16} = 7.94$$

2. Median Value :

Median value is the midpoint of the data sets when data is sorted in ascending/descending order. For the odd number of data points, the

median value will be $\frac{(n+1)}{2}$ observation while for the even number of data points, an average of the middle two observations after sorting the data in ascending order.

MS Excel formula for mean is **median (array of numbers).**

Example 2.4 : Calculate median for following income data of 11 executives in company ABC Limited.

Annual Income
62000
64000
49000
324000
1264000
54330
64000
51000
55000
48000
53000

Solution :

Step1 : Arrange all data in ascending or descending order

Step2 : Check if total observations are odd or even

Step3 : As there is an odd number of data points hence median will be $\left[\frac{n+1}{2}\right]^{th}$ observation.

In this case, n = 11 hence $\left[\frac{n+1}{2}\right]^{th}$ observation will be the median.

Therefore 6th observation, which is 55000 is the median.

Example 2.5 : Calculate median for following income data of 12 executives in company ABC Limited.

Annual Income
62000
64000
49000
324000

1264000
54330
64000
51000
55000
48000
49000
53000

Solution : Step 1 : Arrange all data in ascending or descending order

Step 2 : Check if total observations are odd or even

Step 3 : As there are an even number of data points hence the median will be the average of the middle two observations

$$\frac{(54,330 + 55,000)}{2} = \frac{1,09,330}{2} = 54,665$$

3. Mode :

It is the value that most often occurs in the data; it can be applicable for both numeric and categorical data.

MS Excel formula for Mode is **MODE (array of numbers).**

Example 2.5 : Burger prices of 10 different food outlets in Vadodara and Ahmedabad

Food Outlet Number	Vadodara	Ahmedabad
1	35	23
2	55	25
3	65	36
4	45	39
5	35	42
6	62	49
7	48	55
8	35	67
9	40	72
10	42	80

Solution : As we can see 35 is the most repetitive price of Burger in Vadodara; hence it is the mode. In the case of Ahmedabad, all values are unique; hence we can say there is no mode. Although we can claim that there are 10 modes in the case of Ahmedabad, that will not make

any business sense. Therefore, there can be 0 or any number of modes in the data. For example, there are two modes, 25 and 30 in the below data as both numbers occur thrice in the data set :

25
30
36
25
25
30
30
45

Now the question arises which measure of central tendency is best, and how can we determine which one is most appropriate in any given situation ?

Mean is the best measure of central tendency as it includes all data points of the dataset, but it is highly sensitive towards outliers (data points that are significantly different from others).

In the case of outliers, the **median** will be more appropriate as it does not matter what minimum or maximum observations are. The Median is always calculated by middle observation(s).

Mode is generally used when we want to estimate the central tendency, and we do not have the opportunity to calculate the mean or median of data.

To understand the distribution of our data, whether it is homogenous (all data points are close to each other and do not have outliers), we plot the data distribution. One of the easy ways to visualize data distribution is seeing a Histogram of data. We will cover it in detail in the next unit.

2.5.2 Measure of Dispersion of Data :

There are three most important measures for dispersion of data which are Range, Variance, and Standard Deviation.

1. **Range :** Range is the difference between the maximum and minimum observation in a dataset. For example, in example 2.4, the range of annual salary is :

   12,64,000 (maximum data point) – 48,000 (minimum data point) = 12,16,000

2. **Variance :** Variance is a measure of the distance of data points from its mean. Population variance is generally represented by σ^2 and the formula is as below :

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{N}$$

While the variance of a sample is given by :

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{N - 1}$$

MS Excel formula for variance for population, sample is **var.p(array)** and **var.s(array)** respectively.

In statistics, we have a different formula for the population (when we can analyze entire data) and sample (when we can access limited data to save time and effort). In the case of the population, we are 100% sure about the measures like mean, variance, etc. While the sample is an approximation of the population parameters as if we get 10 different samples from the same population, we will get 10 different measures. Hence statisticians tried to solve this by adjusting the formula. This will be clearer in the next block when we introduce the concept of ***Degree of Freedom.***

Variance tells us how far data points on an average are from the mean. In a way, it tells us the spread or dispersion of the dataset. Let's try to understand why we do the square of all distances in the variance formula without going much in–depth into its mathematical interpretation. There are two reasons; first, it magnifies the distance and second is that it helps to escape from positive and negative terms cancellations. For example, suppose the mean is 16, and there are two points 12 and 20. Now distance of 12 from 16 is –4 while the distance of 20 from 16 is +4. Now if we see the total distance of these two points from mean 16 then it will become 0, which is incorrect as both data points are 4 units away from the mean 16. Hence, we do square as it helps in getting rid out of negative signs.

Example 2.6 : Calculate the variance of the below data points, consider it as a sample (not population)

No of Toys
1
1
1
2
3
4
5
5
5
5

Solution :

No. of Toys (x)	$X_i - \mu$	$(X_i - \mu)^2$
1	1 – 3.2 = –2.20	4.84
1	1 – 3.2 = –1.96	4.84
1	1 – 3.2 = 1.00	4.84
2	2 – 3.2 = 2.00	1.44
3	3 – 3.2 = 3.00	0.04
4	4 – 3.2 = 4.00	0.64
5	5 – 3.2 = 5.00	3.24
5	5 – 3.2 = 5.00	3.24
5	5 – 3.2 = 5.00	3.24
5	5 – 3.2 = 5.00	3.24
$\Sigma x = 32$		**$\Sigma(X_i - \mu)^2 = 29.6$**

Mean = 3.2

The formula for sample variance

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{N - 1}$$

$$\text{Variance} = \frac{29.6}{9} = 3.29$$

While if it had been population, then variance would be $\frac{29.6}{10} = 2.96$

3. Standard Deviation :

Standard deviation is the square root of variance. The formula for the standard deviation of the population (σ) and sample standard deviation (s) is as below :

The standard deviation for a population :

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{N}}$$

The standard deviation for a sample :

$$s = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{N - 1}}$$

The standard deviation of no of toys given in example 2.6 is as below :

$$\sigma = 1.72 \qquad s = 1.81$$

Variance is squared; hence it is quite large compared to the data points. Another challenge is it is square the unit of measurement. Because of that, most of the time, we use standard deviation instead of variance.

One challenge with variance and mean is that we can't compare these measures of different datasets as these depend on absolute values of datasets. For example, we can't compare the share price of one company with another company as one company may have a share price of less than 100 while others may have more than 20,000. To counter this, we get another measure known as the coefficient of variance.

4. Coefficient of Variance :

It is a ratio of standard deviation and means. It is unitless; hence we can compare the coefficient of variance of two completely different datasets. Similar to other measures, we have a different formula for the coefficient of variance for population and sample.

$$CV = \frac{\sigma}{\mu} \times 100\%; \qquad CV = \frac{S}{\overline{x}} \times 100$$

Check Your Progress – 2 :

1. If we have only 4 data points, then which one would be the best way to calculate the variation in the data

   a. Variance b. Standard deviation

   c. Range d. All of above

2. Coefficient of variance is a better measure as it includes both _____ and _____ of the data

Check Your Progress – 3 :

1. Calculate the mean and median of below six numbers :

   13, 3, 8, 10, 8, and 6.

   a. Mean = 8; Median = 10 b. Mean = 8; Median = 8

   c. Mean = 10; Median = 8 d. Mean = 10; Median = 10

2. The approximate value of the below expression :

$$f(x) = \sum_{n=1}^{n} (X_i - \overline{X})$$

   a. Zero b. Mean

   c. Variance d. Standard deviation

3. Which one of the following measures is appropriate for Nominal scale variables ?
   a. The average age of students in a class
   b. Average marks of students
   c. Gender of students
   d. Height of students
4. Calculate the variance of the below numbers; consider it as population data :
   6, 8, 2, 9, 7, 3, 1, 4
   a. 5 b. 40 c. 6 d. 7.5
5. Calculate the standard deviation of the below numbers; consider it as sample data :
   6, 8, 2, 9, 7, 3, 1, 4
   a. 2.4 b. 2.93 c. 3 d. 4.5
6. What are NOT true characteristics of mean
   a. It includes all data points
   b. It gets impacted with extreme values
   c. It depends only on the central value of an ordered list
   d. It is equivalent to the arithmetic mean
7. Mean, Median and Mode are best to represent :
   a. Shape of dataset b. Location of dataset
   c. Both d. None of the above
8. If most frequently occurring data is an outlier, then in that case :
   a. Mode is a poor measure
   b. Mode is a good measure
   c. We can't say
   d. The mode can be calculated in that case
9. Which of the following is best to compare the variability of two completely different data–set ?
   a. Standard deviation b. Variance
   c. Coefficient of variance d. Range
10. We have extracted a sample of 200 people from a city, which one of below is correct :
    a. The standard deviation of the sample is always greater than the standard deviation of the population
    b. The standard deviation of the sample is always less than the standard deviation of the population
    c. We can't say whether the Standard deviation of the sample is lesser, or the Standard deviation of the population is lesser
    d. 200 people sample is too less to calculate standard deviation

2.6 Let Us Sum Up :

1. Business analytics projects start with descriptive analytics, which includes data summarization, aggregation, descriptive statistics, and visualization
2. Analysts run queries to fetch appropriate data and consolidate required data from different sources
3. Descriptive analytics focus on information gathering from raw data which tells us what happened in the past
4. Visualization helps in storytelling through data and appropriate graphs as per the data types
5. Data scientists calculate central tendency, variation, and shape of data to understand the characteristics such as variance, central point, and skewness
6. Descriptive analytics prepare field for further in–depth analysis like diagnostic and predictive analytics

2.7 Answers to Check Your Progress :

Check Your Progress – 1 :

1. b **2.** Sample

Check Your Progress – 2 :

1. c **2.** Centrality, variation

Check Your Progress – 3 :

1. b	**2.** d	**3.** d	**4.** a	**5.** b
6. c	**7.** a	**8.** c	**9.** c	**10.** a

2.8 Glossary :

Descriptive Analytics is the simplest form of analytics; it summarises current business status in the way of narrative and innovative visualization. It emphasizes *"what is going on in the business."*

The sample is a logical subset of the population, which mimics the population.

Categorical Data : This type of data is known as discrete data. Categorical variables have a finite number of groups, e.g., payment method, gender, income group, property type, etc.

Continuous Data : This type of data can be measured on a continuous scale–like height, weight, money, time, etc. It can be divided into halves any number of times.

Nominal Data : In this type of categorical variable, there is no logical sequence among categories, which means one category is not superior or inferior to others. Categorization is just a type of segregation of data into different groups.

Ordinal Data : It is better than nominal data in terms of usage potential for any analysis. Ranked or ordered data generally come into this category.

Ratio Data : In terms of statistical relevance, this is the highest level of data which is desirable for the application of statistical tools and techniques.

Mean (Average) Value : It is the summation of all numbers divided by the total numbers.

Median Value : Median value is the midpoint of the data sets when data is sorted in ascending order. For odd data points, the median value will be $\frac{(n+1)}{2^{nd}}$ observation while for even data points, an average of middle two observations after sorting the data in ascending order.

Mode : It is the value that most often occurs in the data; it can be applicable for both numeric and categorical data.

Variance : Variance is a measure of the distance of data points from its mean.

Standard Deviation : Standard deviation is the square root of variance.

Coefficient of Variance : It is a ratio of standard deviation and means. It is unitless; hence we can compare the coefficient of variance of two completely different datasets.

2.9 Assignments :

1. Define various ways to calculate the dispersion of data, write their strengths and limitations.
2. Write down one scenario where mean is better than median and vice-versa.
3. Explain how the coefficient of variance can help in ranking important mutual funds for your customers.

2.10 Activities :

Supertronic motors have published month–wise sales data in Gujarat.

Month	Number of Units Sold
1949–01	112
1949–02	118
1949–03	132
1949–04	129
1949–05	121
1949–06	135
1949–07	148
1949–08	148
1949–09	136
1949–10	119

Calculate mean, median, mode, range, variance, standard deviation, and coefficient of variance of above data.

MEAN	129.8
Median	130.5
Mode	148.0
Variance	153.7
Standard Deviation	12.4
Range	36.0
Coefficient of Variance	0.1

2.11 Case Study :

AWQS Constructions is a premium construction company in North Gujarat. They construct villas and flats. They want to understand which unit is giving them more consistent revenue.

Monthly Revenue from Flats Unit (₹ crores)		Monthly Revenue from Villas Unit (₹ crores)	
123	128	123	102
124	123	225	97
125	129	226	129
123	121	123	121
156	154	111	102
178	174	98	215
129	122	121	205

Questions :

1. Which unit has more consistent revenue ?
2. Which unit has a better coefficient of variance ?
3. Which unit has a higher mean and median ?

2.12 Further Readings :

- "Mathematical Methods of Statistics," Princeton University Press, Carmer H. (1946)
- "Super Freakonomics," Penguin Press, Levitt S. D. and Dubner S. J. (2009)
- "An Explanation of the Persistent Doctor Mortality Association" Journal of Epidemiology and Community Hearlth, Yough F. W. (2001)

Unit 3

VISUALIZATION TECHNIQUES FOR BUSINESS ANALYTICS

: UNIT STRUCTURE :

3.0 Learning Objectives :

After learning this unit, you will be able to understand :

- Learning various types of visualization techniques
- Interpretation of visualization techniques
- Understanding the most appropriate visualization technique as per the scenario

3.1 Introduction :

In this unit we will learn various type visualization techniques and scenarios in which we use specific type of visualization techniques. At the end we will also look into the benefits and limitation of each type of graphs. We have also included few case studies so that interpretation part can also be explained in detail.

3.2 Introduction to Data Visualization :

"Visualization is the soul of business analytics; it is the language which everyone understands completely without saying much".

– Anonymous

Data Visualization : Data visualization makes descriptive analytics so important and easy to understand. It helps in taking appropriate and prompt decisions for the betterment of the business. There are various charts and graphs which assist decision–makers in extracting important insights, and hidden knowledge lies in data. The few important types of charts which we will cover in courses are Bar charts, histograms, pie charts, box–plot, radar charts, line charts, area charts and scatter plots. Data visualization is the very first step towards starting an analytical project. There are various softwares available for visualization, few most important are Microsoft Excel, Open Office Maths, tableau, Microsoft PowerBI, etc.

3.3 Histogram :

Histogram applicable for continuous data (Ratio and Interval data types), there are consecutive but non–overlapping bars where each bar shows the frequency distribution of the data. Histogram assesses the probability distribution of the data.

In MS–Excel, a histogram is not part of graphs; we have to activate the "Data Analysis" pack.

Below is an example of sales data from different customers in the last financial year. Total there are 33 sales figures from these customers, and sales amount varies from ₹ 14,786 to ₹ 29,29,278.

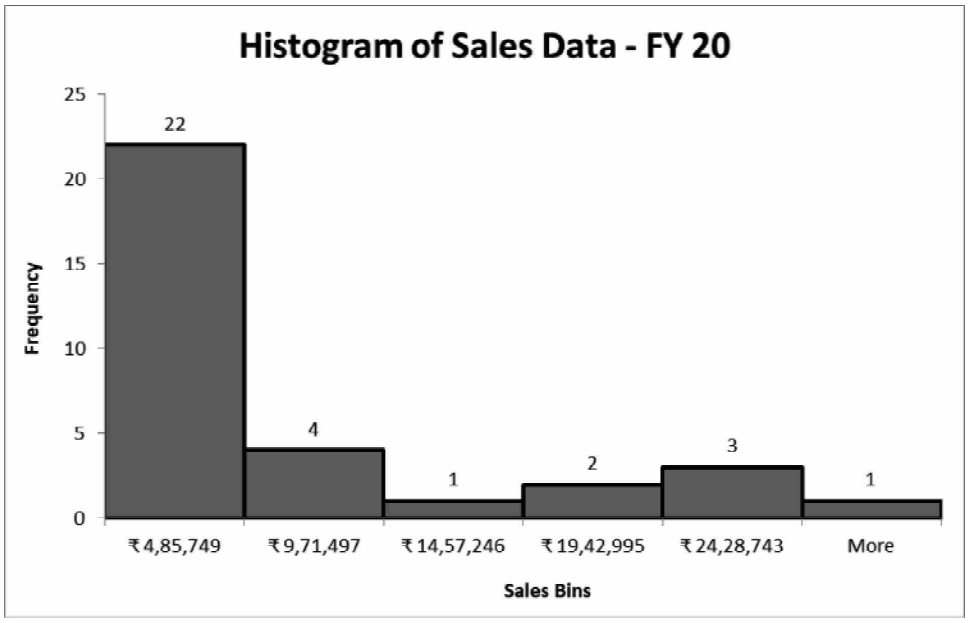


Fig 3.1 Histogram

Most statistical software automatically determines the number of bins (Intervals). Although there are formulas suggested by different statisticians by which we can decide the count of bins. So, there can be a different number of bins for the same data if we are using different statistical software. In MS Excel, we have to determine the number of bins manually. For the above histogram, six bins were supplied :

One of the most frequently used approaches to determine the number of bins is as below :

₹ 4,85,749
₹ 9,71,497
₹ 14,57,246
₹ 19,42,995
₹ 24,28,743
₹ 29,14,492

$$\text{Number of bins, } N = \frac{X_{max} - X_{min}}{W}$$

Where X_{max} is the maximum value in the dataset while X_{min} is the minimum value in the dataset, W is the width of the bins : Famous statistician Strurges (1926) suggested below equation in this research paper :

$$\text{Width of the bins (W)} = 1 + 3.332 \times \log_{10}(n)$$

Here, n is the total number of observations in the data set.

In the above example, we had n = 33 hence the number of bins suggested by the above formula is 6 (round off). Therefore,

$$\frac{(2929278 - 14786)}{6} = 485749$$

Now all data points in the range of 0–485749 will be in the first bin (interval); similarly, we will have six bins of width 485749 while the last bin will have all remaining data points also.

The histogram is one of the most important charts in statistical visualization treasure. Statisticians and business analysts use the histogram for the following reasons :

(a) It helps in identifying the outlier in the data set (if few bins are quite far from the remaining bins then it can be considered as an outlier)

(b) It helps in assessing the probability distribution of the data by judging the shape of the distribution

(c) It suggests about central tendency of the datasets like mean, median, and mode

(d) It helps in assessing the variability of the data set

(e) It also helps in assessing other important statistical measures like skewness

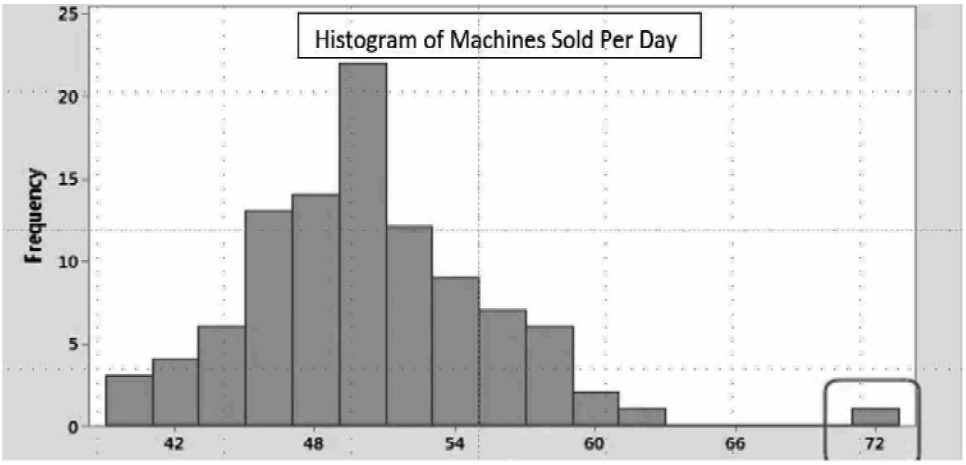


Fig 3.2 Histogram of Machines Sold Per Day

Above is another example of a histogram that depicts per day sale of machines. Here clearly, we can see 72 is an outlier. It is also suggesting that the mean is close to 50, which is also median and mode in this case. This is also telling us the probability distribution; here, data is following a normal distribution. In the second block, we will cover probability distributions in detail.

3.4 Bar Chart :

It is a chart for categorical or qualitative variables. It helps in comparing and assessing categorizes within a data set. Below is an example of the sector–wise contribution of the GDP of India.

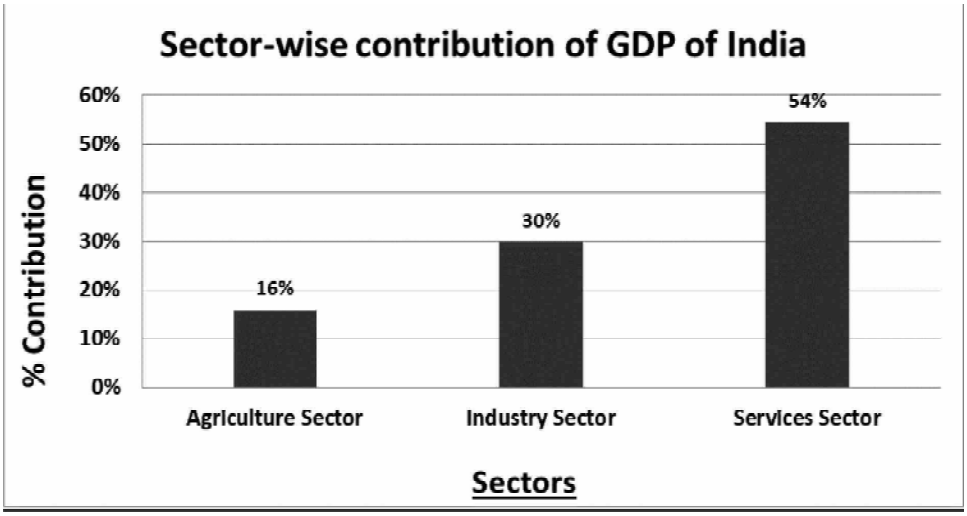


Fig 3.3 Bar Chart

Various variants for bar charts are available, we can show multiple bars for each category like each sector's revenue for the financial year 2017, the financial year 2018 and the financial year 2019, etc. We can check the Microsoft website to learn about these modifications in detail. We can showcase subcategories also in the bar chart.

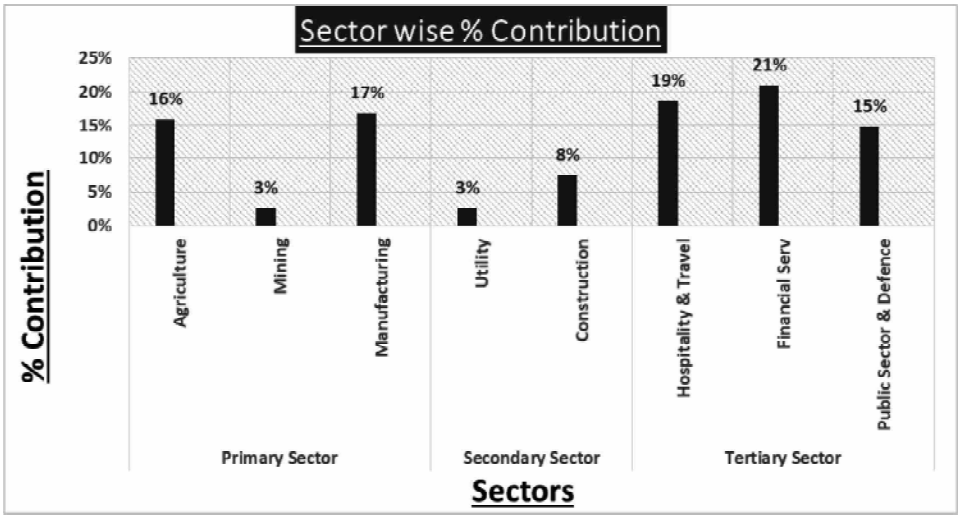


Fig 3.4 Detailed Bar Chart

All these charts are also available in the open office which is available without any license fee. A bar chart is among the frequently used charts which executives use daily to showcase various important business metrics.

3.5 Scatter Plot :

In the business world, often we need to see the relationship between various variables, for example, whether the repo rate impacts on Sensex index or whether there is a relationship between the gold price and crude oil in the international market.

Below is an example where the scatter plot is showing the relationship between the area of the property in square feet and Price in INR

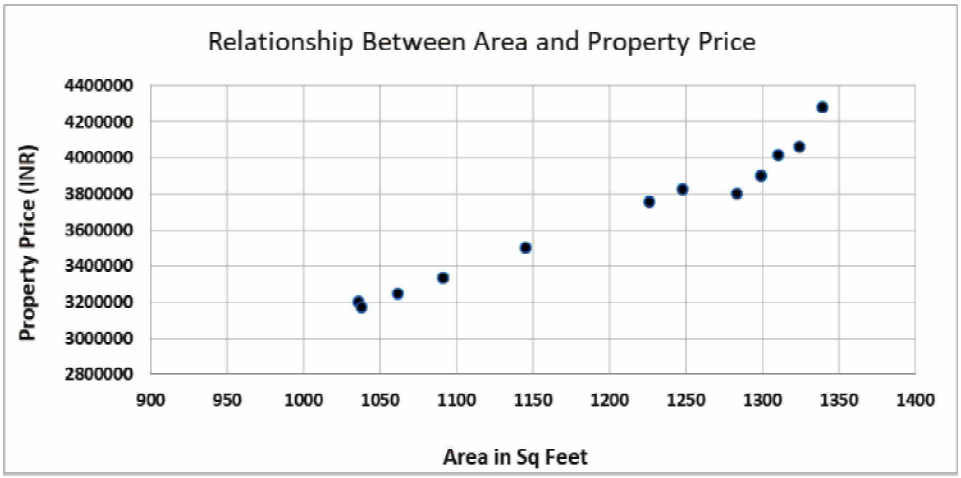


Fig 3.5 Scatter Plot

❖ **Case Study :**

Case study to analyze student performance with the help of Scatter Plot and Coefficient of Variance :

Let's say, we have a class of 25 students and the professor records their monthly test scores on a spreadsheet for the Business Analytics subject. Data is stored in the below format :

Student Roll No	Month_1	Month_2	Month_3	Month_4	Month_5	Month_6	Month_7	Month_8	Month_9	Month_10
1	86	99	97	84	88	84	87	94	100	83
2	62	73	50	50	52	50	58	57	53	51
3	64	62	61	60	73	73	74	79	75	64
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	91	84	83	84	80	97	91	89	99	89

Solution : We can calculate the average score in the last column (next to Month_10), and in the next cell, we can calculate the standard deviation. Hence, we will have centrality and variation of each student. We can create a scatter plot for these two measures on both the axis in such a way that the centre of the graph will have a midpoint of mean and standard deviation. In this way, we can segregate the entire class into four quadrants, and each quadrant will reveal important information about the progress of these students. With the help of domain knowledge, we can also name these quadrants so that we can think about customized help for each set of students.

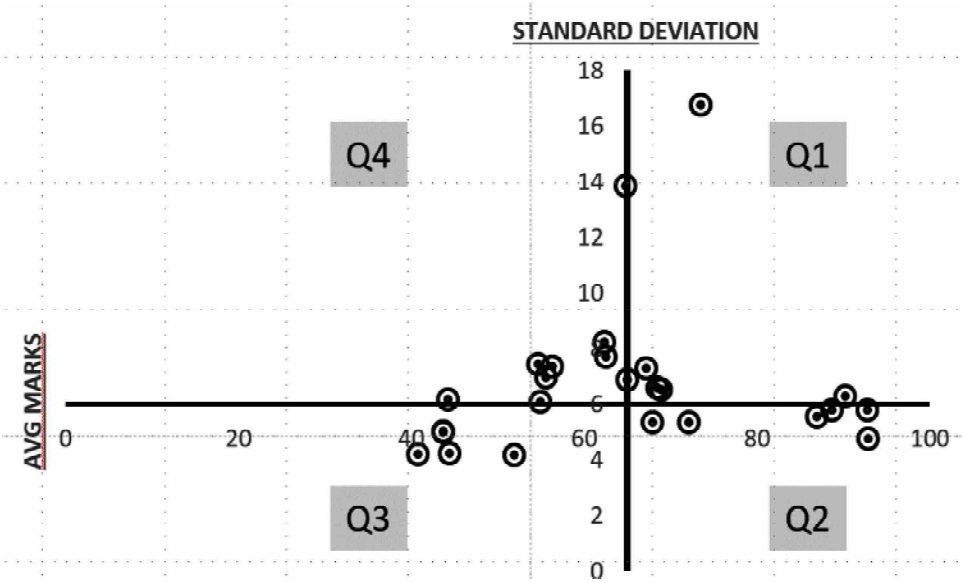


Fig 3.6 Scatter Plot – Avg Marks Vs Standard Deviation

❖ **Interpretation of Scatter Plot :**

Students fall in Q1 : Here Mean is High, which is good in this scenario while the Standard Deviation is also high, which always lesser is better. Students who fall in this quadrant are most of the time scoring above the average, but they have high fluctuation means sometimes they score very high while sometimes they score quite low. Hence teachers can deep dive into their performance so that they can identify the topics where these students are weak and arrange revision class only on those particular topics.

Students fall in Q2 : Here Mean is high, and the Standard deviation is low. Best students fall under this quadrant as they are most consistent. They score high, and there is very little fluctuation within their performance. We can benchmark these students; try to understand how they prepare for exams and what are other best practices they follow due to which their performance is consistently good.

Students fall in Q3 : Here mean is low and the standard deviation also low. These students are consistently low performers. They score poorly in most of the exams. The instructor can arrange extra classes for these students, or they can make one good, one poor student pair to clarify their poor concepts.

Students fall in Q4 : Here means is low, and the standard deviation is high. These students have the worst performance as most of the time, their scores are below average, but also, they have high fluctuations.

Conclusion : This is a very powerful technique as it provides very insightful information about the different types of student's cohorts (groups). It provides a very specific improvement area for each group without generalizing solutions like arrange extra classes etc. This study is very generic and can be replicated for share analysis, e.g. generally blue–chip shares fall under Q2 where the mean is very high, and the standard deviation is very low, in other words very consistent shares give a high and steady return. Similarly, it can be replicated for mutual fund analysis. ICC also conducts a similar analytical approach to find out the most consistent batsman, who scores high runs and has comparatively less variation among his performance (consistent performance).

Check Your Progress – 1 :

1. Histogram and scatter plots are applicable for :
   a. Continuous data b. Discrete data
   c. Both type of data
2. A Scatter plot can be used only to show a linear relationship between the variables
   a. True b. False
3. Y–axis of histograms shows _____ of the dataset

3.6 Box Plot :

Box plot is a smarter way to assess variability in the numerical data by splitting data into four quartiles. It is an effective way to identify the outliers in the data sets.

Box plot splits data into four equal quartiles (Quarter 1 quartile means 25th percentile). If 25% of data points are

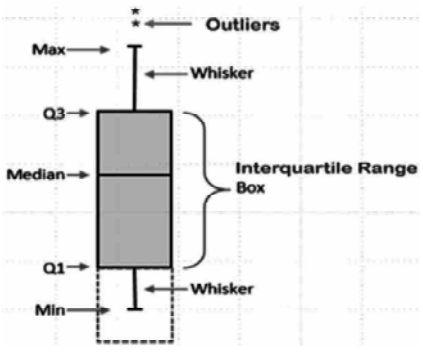


Fig 3.7 Box-Plot Graph

below a data point, then we call it the 25th percentile, or these lowest 25% data sets constitute the first quartile.

The middle part is known as the interquartile range (IQR), which is the difference between quartile 3 and quartile 1. These are 50% of data points from the middle (25% of data points are below the interquartile range box, and 25% are above it).

Any data point which is above Q3 + 1.5*IQR or Q1 – 1.5*IQR is considered as an outlier, which means these data points are not as per normal nature. There may be some special reasons for occurring these data points which need further investigation to understand the root cause.

Box plot is very important for comparing datasets. For example, if we want to compare the sales for the last three financial years or productivity of a company from its various branches etc. It is also very useful in decision making like IQR represents the middle class if we consider it as a population of any state or country. Policymakers try to draft special policy for the betterment of people fall in Q1 (low–income group), IQR (middle and upper–middle–class) and Q3 (rich or elite class).

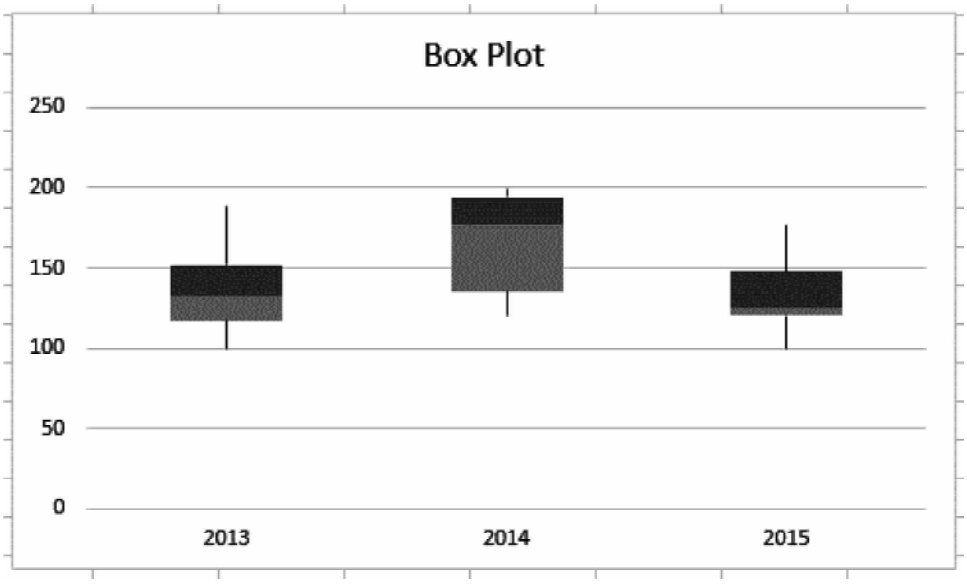


Fig 3.8 Multiple Box-Plot Graphs

In the above figure, we can see that although overall sales are high in 2014 but variability is also high. High variability reduces our capability for robust planning. It is difficult to predict sales for next year if variability in last year's data is high. Box plot is not available in Microsoft Excel or Open office graphical unit, but with a small modification to the stacked bar graph, we can create it in Excel or open office easily.

3.7 Control Charts :

A control chart is a special statistical chart that shows changes in a process over time. The X–axis of a control chart is always time, and its centreline is the **mean** of data. We can also call it a time series plot with control limits. It is also a powerful tool to identify the outliers in the data.

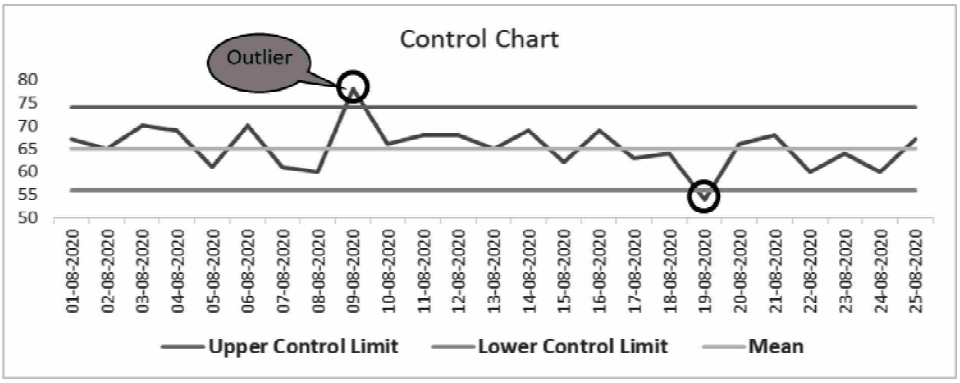


Fig 3.9 Control Charts

This is a simple control chart, also known as the Individual control chart, as all data points are individually plotted on the mean line (central line). Instead of plotting the average as a central line, we can also draw a target as a central line. There is a various advanced type of control charts like Individual – Moving Range Control Chart (I–MR Chart) or X bar Range Chart, etc. but those are not in the scope of our course.

Check Your Progress – 2 :

1. Boxplots provide us information about spread of data as well as _____ in the data.
2. Center point of box in boxplot graph is _____.

3.8 Tree Map :

A Tree map chart is very appropriate to represent hierarchical data in a tree–like structure. The Treemap is a strong visualization technique when we have to show so many categories.

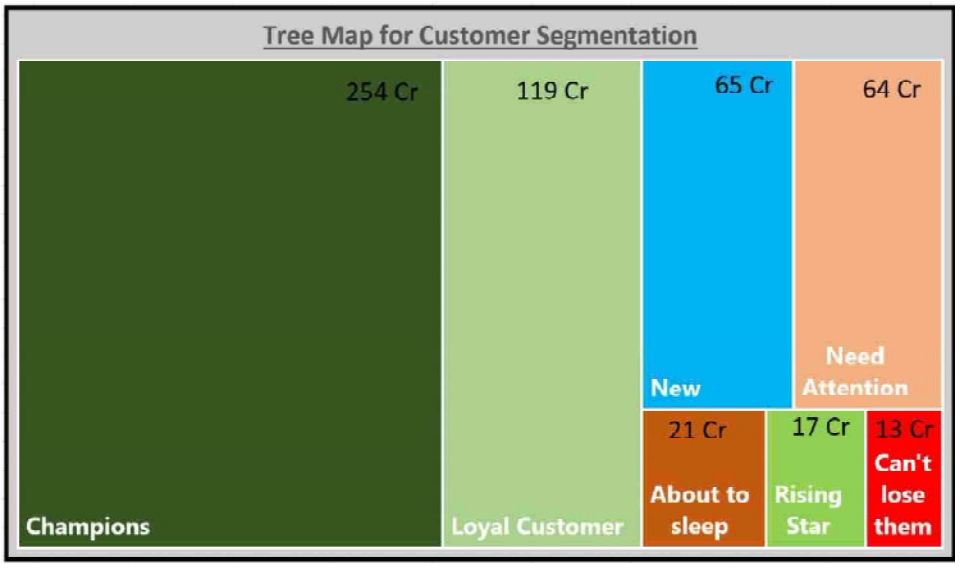


Fig 3.10 Tree Chart

Check Your Progress – 3 :

1. Any observation below Q1 – 1.5 IQR in a box plot is :
   a. Median
   b. Potential outlier
   c. Mean
   d. Minimum value

2. A bar chart is appropriate for which type of data
   a. Quantitative data b. Qualitative data
   c. Ordinal data d. Interval scale data
3. The upper control limit in the control chart is :
   a. Mean + 3 Standard deviation
   b. Median + 3 Standard deviation
   c. Mean – 3 Standard deviation
   d. Mean + 2 Standard deviation
4. Histogram can be used for which type of data
   a. Quantitative data b. Qualitative data
   c. Ordinal data d. Interval scale data
5. The main objective of the Scatter chart is :
   a. Highlight outliers in the data
   b. Showing the relationship between quantitate data sets
   c. Showing the relationship between qualitative data sets
   d. Showcase variation in the data
6. Which graph is relevant for hierarchal data ?
   a. Box plot b. Bar graph c. Histogram d. Treemap
7. Y–axis in the histogram represents :
   a. Probability b. Area under curve
   c. Frequency d. Area
8. The central line of control charts represents :
   a. Mean or target b. Standard deviation
   c. Median d. Mode
9. The central line of the IQR box of box plot represents :
   a. Mean b. Variation c. Median d. Mode
10. The formula for the coefficient of variance is :
    a. Standard deviation / Median b. Mean/standard deviation
    c. Median / Standard deviation d. Standard deviation / Mean

3.9 Let Us Sum Up :

1. Data visualization simplifies the information understanding hidden in raw data
2. Data visualization helps analysts in deciding the next step by displaying pictorial relationship among various business metrics
3. As per the data type, we should choose the right type of graph so that maximum information can be understood with minimum efforts
4. Visualization becomes more effective by adding features in data visualization like right axis scale, legends, combing important data together, or split data into sensible sub–parts

5. Histograms, boxplot, scatter plot, control charts are important graphs can be made on continuous data
6. Pie chart, bar chart, Treemap are important graphs for categorical graphs

3.10 Answers to Check Your Progress :

Check Your Progress – 1 :

1. (a) **2.** False **3.** Frequency

Check Your Progress – 2 :

1. Outliers **2.** Median

Check Your Progress – 3 :

1. b	**2.** b	**3.** a	**4.** a
5. b	**6.** d	**7.** c	**8.** a
9. c	**10.** d		

3.11 Glossary :

Bar Chart : It is a chart for categorical or qualitative variables. It helps in comparing and assessing categorizes within a data set. Here x–axis represents categories while the y–axis represents the count or percentage of data.

Histogram : It is a kind of area graph. It looks similar to a bar graph of data that buckets a range of data into columns along the x–axis while the y–axis represents the frequency (number of data count) or percentage of occurrences in the data for each column. It is applicable for continuous data only.

Scatter Plot : It is used to show representation between two continuous variables, generally we put the input variable on the x–axis while the output variable on the y–axis. It also helps in identifying outliers in the dataset.

Box–Plot : It is a smarter way to assess variability in the numerical data by splitting data into four quartiles. It is an effective way to identify the outliers in the data sets.

Control Chart : A control chart is a special statistical chart that shows changes in a process over time. The X–axis of a control chart is always time, and its centreline is the **mean** of data. We can also call it a time series plot with control limits. It is also a powerful tool to identify the outliers in the data.

Tree Map : A Tree map chart is very appropriate to represent hierarchical data in a tree–like structure. The Treemap is a strong visualization technique when we must show so many categories.

3.12 Assignments :

1. Which graph to be used if you want to show the relationship between two continuous variables, explain with an example.

2. How interquartile range helps in taking decisions, explain it with an example.
3. Write down the importance of tree diagram, mention two scenarios where you can implement a tree diagram.

3.13 Activities :

Online training firms collected the training hours for their 16 trainers and their feedback. Draw a scatter plot and see where there is the relationship between training hours and feedback scores. Also, draw a quadrant graph like provided in the chapter (fig 3.6)

Training Hours	Feedback	Training Hours	Feedback
75	8	94	5
69	5	81	5
104	5	86	9
106	4	99	3
92	8	110	7
111	2	76	4
64	6	101	4
92	4	125	4

3.14 Case Study :

Pokymon trading limited is an investment advisory organization, they have active customers across central part of India. They have listed top 40 company's stocks and their value in rupees :

292	790	354	318	345	216	362	533	256	260
233	349	467	477	50	419	565	590	522	466
247	293	273	545	566	411	470	235	247	288
364	429	569	597	248	505	309	347	353	417

(a) Make a histogram for the above dataset. List down important insights from the data

(b) Draw a box plot and see how different all four quadrants from each other, also see if there are any outlier in the data

(c) Draw a control chart of the above data and see if any trend is visible

3.15 Further Readings :

- "Mathematical Methods of Statistics," Princeton University Press, Carmer H. (1946)
- "Super Freakonomics," Penguin Press, Levitt S. D. and Dubner S. J. (2009)
- "An Explanation of the Persistent Doctor Mortality Association" Journal of Epidemiology and Community Hearlth, Yough F. W. (2001)

BLOCK SUMMARY

Business analytics has become the brain or spinal cord of businesses nowadays. Like the human body depends on brain/ spinal code for decision making and other important functions similarly entire organizational strategy is dependent on business analytics. Executives want to make all decisions based on data only. Nowadays business units are integrated with the help of enterprise–wide softwares hence businesses have a better view of the entire value chain instead of seeing opportunities or problems in individual business units. Business analytics is a very structured approach where it starts with descriptive statistics where the focus is on visualization and "What did happen" then we see all good and bad phenomenon in the business with the help of outcomes of descriptive statistics like a report card, dashboards, executive summaries, KPI metrics, etc. Then we move to diagnostics statistics to understand the root cause for all good and bad things so that we can replicate the good scenarios while minimizing the probability of happening bad things again in the business. These two types of analytics provide a solid foundation for the next level of analytics like predictive and prescriptive where focus on the future. We try to predict the future values of important business metrics and prepare a continuity plan.

BLOCK ASSIGNMENT

Short Answer Questions :

1. Explain how business analytics is a systematic approach to be established in an organization.
2. Write down few important challenges in establishing business analytics in an organization
3. Write down the important difference between descriptive and diagnostic analytics
4. Explain the importance of visualization in the business analytics
5. Why we say that "central tendency alone does not show us the complete picture, we should also consider variation also". Justify the statement with one example ?
6. Write down important components of business analytics
7. Explain the life cycle of business analytics, mention one example
8. Why should we study sample only even if we have access to entire population data ?
9. Write one scenario for the usage of mean and median as a measure of central tendency
10. Why do we call mode a weak measure of central tendency ?
11. Explain briefly the difference between bar chart and histogram
12. Write a short note on the degree of freedom

Long Answer Questions :

1. Write a note on how predictive analytics is different from prescriptive analytics. Mention important tools/ techniques under each type of analytics ?
2. Why data become so important in the last 10 years, write down few success stories from the business world about data–driven decision making ?
3. Write short note on different data measurement scales, give two examples for each data type ?
4. How control charts and boxplot help in the analysis of variance in the data. Give suitable examples ?

❖ **Enrolment No. :** ☐

1. How many hours did you need for studying the units ?

Unit No.	1	2	3
No. of Hrs.			

2. Please give your reactions to the following items based on your reading of the block :

Items	Excellent	Very Good	Good	Poor	Give specific example if any
Presentation Quality	☐	☐	☐	☐	_____
Language and Style	☐	☐	☐	☐	_____
Illustration used (Diagram, tables etc)	☐	☐	☐	☐	_____
Conceptual Clarity	☐	☐	☐	☐	_____
Check your progress Quest	☐	☐	☐	☐	_____
Feed back to CYP Question	☐	☐	☐	☐	_____

3. Any other Comments

.....

.....

.....

.....

.....

.....

.....

.....

Dr. Babasaheb Ambedkar Open University

BBAR-304

Business Analytics

BLOCK-2 STATISTICAL CONCEPTS AND HYPOTHESIS TESTING

UNIT 1

DISCRETE PROBABILITY DISTRIBUTIONS

UNIT 2

CONTINUOUS PROBABILITY DISTRIBUTIONS

UNIT 3

SAMPLING AND CONFIDENCE INTERVALS

UNIT 4

INTRODUCTION TO HYPOTHESIS TESTING

BLOCK 2 : STATISTICAL CONCEPTS AND HYPOTHESIS TESTING

Block Introduction

Business analytics try to reduce the uncertainty in the businesses so that executives can make better decisions backed by data. Descriptive and diagnostic statistics provide them important insights about business progress in terms of important business metrics and lay down the foundation blocks for the next level of analytics. Here onwards probability theory start playing an important role. Probability theory and concepts of probability distributions are essential building blocks of business analytics. These concepts help us to study the hid-den/obvious trends in the business and provide us alternatives that have fewer risks and enhance the decision-making process. Probability distributions and understanding the theory of hypothesis testing are important pillars on which predictive and prescriptive analytics models are built.

Business analysts generally avoid predicting point (single value) estimates rather they prefer to estimate the confidence interval which reduces the uncertainty e.g. instead of saying next month sales will be Rs 856 crores, analysts say sales will be in the interval of (Rs 800 Cr to 912 Cr). Confidence intervals also tell us information about population parameter instead of just sample which result into more robust decision making. Here business analysts use a very important concept called hypothesis testing which is used to validate the business as-sumptions based on sample data. It reduces the ambiguity in decision making and provides us with better opportunities.

Block Objectives

After learning this block, you will be able to understand :

- Understanding concepts and applications of the probability distribution
- Understanding important discrete probability distribution functions
- Difference between probability mass function and probability density function
- Applications of the continuous probability distributions
- Different types of sampling methodologies and techniques
- Significance of central limit theorem and its applications
- Application of confidence interval and confidence level
- Understanding the basics of hypothesis testing and its applications in decision making
- Understand the concept of significance (α) and Type 1 and Type 2 error
- Understand the significance of p-value and its application in concluding hypothesis test

Block Structure

Unit 1

DISCRETE PROBABILITY DISTRIBUTIONS

: UNIT STRUCTURE :

1.0 Learning Objectives :

- Understanding concepts and applications of the probability distributions
- Concept of Probability mass functions and cumulative density functions
- Understanding important discrete probability distribution functions
- Expected values and variance of discrete probability distributions

1.1 Introduction :

Unit Introduction : In this unit, we will study the basic concepts of probability distributions. We will see various examples of how we apply theory of these distributions in the business world to take important decisions. Further, in the unit we will see how each outcome of a random experiment is mapped to a probability distribution know as probability mass function while the cumulative distribution function is the probability that a random variable can take values less than or equal to x_t. As a part of an application, we will see the expected values and variance of these discrete probability distributions.

1.2 Random Experiments and Probability Distributions :

"Probability theory is nothing but common sense reduced to calculation." – ***Pierre–Simon Laplace***

Probability distribution consists of all possible values a random variable can take; these values tell us about the minimum and maximum

x	$P(x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$
$\Sigma P(x)$	$\frac{6}{6}$ or 1

values between which a random variable can vary. Post discussion on a range of the values; now the point is how frequently these possible values can occur within the potential range. Generally, it depends on two critical factors – mean (expected value) and standard deviation. Although there are other two factors also which are comparatively less critical are skewness and kurtosis.

Let's demonstrate it with the help of an example :

x	$count$	$\frac{count}{total} = P(x)$
1 (very dissatisfied)	5	.046
2 (dissatisfied)	10	.093
3 (neutral)	11	.102
4 (satisfied)	44	.407
5 (very satisfied)	38	.351
Σ	108	1

x is a likely outcome of the throw of dice; x = [1, 2, 3, 4, 5, 6] while P(x) is the probability of occurring 3. We represent it as $\rightarrow$ P(X = 3) = $\frac{1}{6}$. As P(x) represents the probability of any possible outcome, we call it the probability function P(x). So probability distributions represent the likelihood of an outcome depending on how frequently it is featured in the given sample space.

Let's consider one more example to understand this concept better; an online shopping company collected feedback for their packaging facility. They collected feedback from 108 customers. Feedback was collected on the Likert scare of 1–5 where 1 was 'very dissatisfied' while 5 means 'very satisfied'. We can represent these five outcomes in terms of numbers like x = 1, 2, 3, 4, 5 while P(x) represents the probability of each outcome

There are two types of probability distribution functions :

- Discrete probability distribution functions
- Continuous probability distribution functions

Let's understand these distributions in detail with the help of their definitions and applications in the business world.

1.3 Discrete Probability Distributions :

If a random variable can obtain only discrete (countable) outcomes, for example, 0, 1, 2, 3 etc but it cannot take values like 2.3 or 3.5 etc. then we say it is a discrete random variable that follows discrete probability distribution. In other words, it tells us how probabilities are assigned to every outcome of a random variable as we saw in the above two examples. As we studied earlier that probability distributions are defined by two important factors, the expected value (average) and standard deviation. For a discrete distribution, the expected value is the mean of the random variable (average of expected outcome); we represent it as follows :

$$E(X) = \mu = \Sigma x \times P(x)$$

Let's calculate the expected value of the feedback data :

x	$count$	$\frac{count}{total} = P(x)$	$xP(x)$
1	5	.046	$1 \times .046 = .046$
2	10	.093	$2 \times .093 = .186$
3	11	.102	$3 \times .102 = .306$
4	44	.407	$4 \times .407 = 1.628$
5	38	.351	$5 \times .351 = 1.755$
Σ	108	1	3.70

So we can see the expected value of the packaging facility is closer to "satisfied." Please note discrete random variables cannot obtain decimal values, but the expected value can be represented in decimals. Now let's discuss variability (how far is a random variable from its mean). For any discrete variable, the variance is calculated by the below formula :

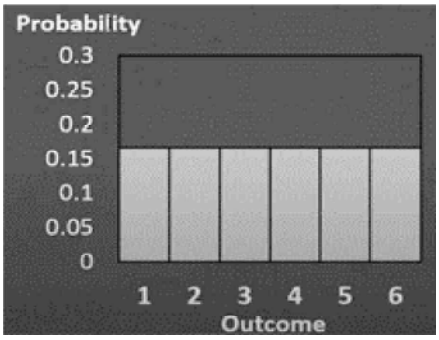
$$Var\ (x) = \sigma^2 = \Sigma(x - \mu)^2\ P(x)$$

$$\text{Standard deviation} = \sqrt{Var(x)} = \sigma = \sqrt{\left(\Sigma(x - \mu)^2\ P(x)\right)}$$

x	$P(x)$	μ	$(x-\mu)$	$(x-\mu)^2$	$(x-\mu)^2P(x)$
1	.046	3.70	$1 - 3.70 = -2.7$	7.29	$.046 \times 7.29 = .335$
2	.093	3.70	$2 - 3.70 = -1.7$	2.89	$.093 \times 2.89 = .269$
3	.102	3.70	$3 - 3.70 = -0.7$	.49	$.102 \times .49 = .05$
4	.407	3.70	$4 - 3.70 = 0.3$	.09	$.407 \times .09 = .037$
5	.351	3.70	$5 - 3.70 = 1.3$	1.69	$.351 \times 1.69 = .488$
				σ^2	1.18
				σ	1.09

Let's understand a few essential terminologies regarding the discrete probability distribution – Probability mass function (PMF) and cumulative distribution function (CDF).

Probability Mass Function (PMF) is a graphical representation of probabilities for all possible values of a discrete random variable. This is also known as a frequency function.

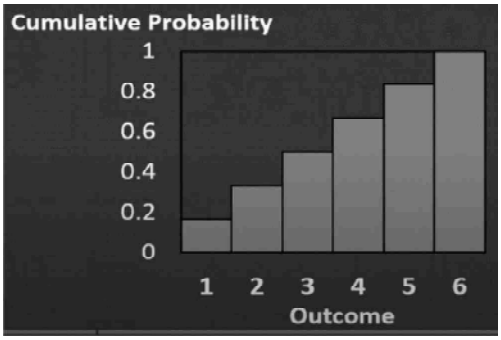


If we see an example of dice, then possible outcomes are x = [1, 2, 3, 4, 5, 6] and there is equal probability,

$$\frac{1}{6} = .167$$

There is an equal probability of each possible value of random variable x.

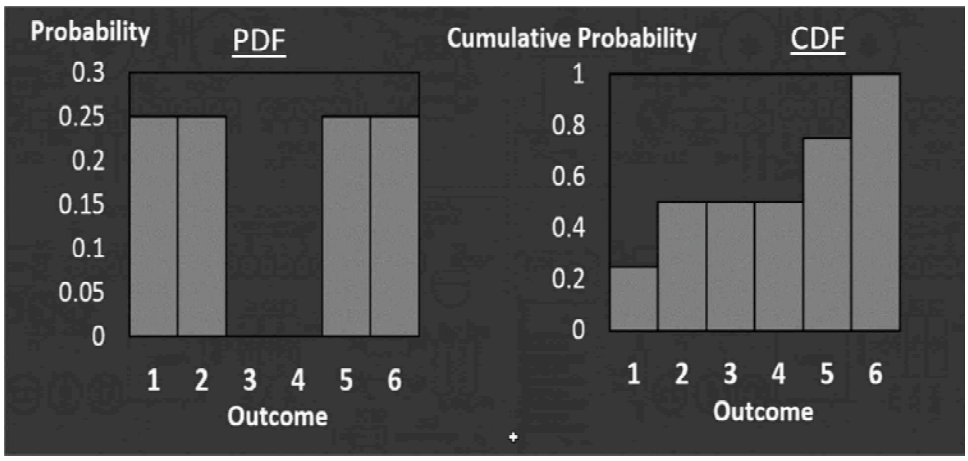
Cumulative Distribution Function (CDF) is another way to represent the distribution of a random variable, but unlike PMF, it is not limited to individual discrete variables only. At any point, it represents probability up to that point.



$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$$

Y–axis of CDF goes up to 1 as probability cannot be more than 1.

Let's consider one more example to understand this concept further, suppose I have a special dice where I can't roll 3 or 4. There x = 1, 2, 5, 6. Therefore there is blank space in PDF graphs (there is no ***"mass"*** in the probability mass function). In CDF, we can see the probability of having 2 or less is the same as 4 or less. Once we have mass at 5 in the PDF graph hence CDF value at 5 get change.



Check Your Progress – 1 :

1. Which of below are characteristics of a discrete probability distribution

   a. Mean (expected value) b. standard deviation

   c. Skewness d. Kurtosis e. All of the above

2. Probability mass function can be used to define a discrete probability function only

   a. True b. False

1.3.1 Discrete Probability Distributions – Binomial Distribution :

"Life is a school of probability" – ***Walter Bagehot***

In the binomial distribution, there are only two possible outcomes. Here prefix "bi" indicates two. For example the probability of "yes" or "no", "pass" or "fail", "heads" or "tails" etc.

There are below pre–requisites of binomial distribution :

1. Only two possible outcomes per trial
2. The probability of success or failure will remain constant across all the trials
3. There are a fixed number of trails
4. Each trial is entirely independent of each other

❖ **The General Formula for Binomial Probability :**

Probability of k out of n ways :

$$P(k \text{ out of } n) = \frac{n!}{k!(n-k)!} p^k (1-p)^{(n-k)}$$

- **Expected Value (Mean) of Binomial Distribution :** np

  Here n is the no of trials while p is the probability of success

- **Variance and Standard Deviation of Binomial distribution :** $np(1-p)$

- **Standard Deviation :** $\sqrt{np(1-p)}$

Example – 1.1 : Survey shows from Ambedkar University only 8% of students have tried paragliding. Now if we are picking randomly 10 students.

(a) What is the probability that all 10 tried paragliding

(b) What is the probability none of the students tried paragliding

(c) What is the probability exactly 2 students tried paragliding

(d) What is the probability at least 2 students tried paragliding

(e) What is the expected number of students who tried paragliding ?

(f) What is the standard deviation of the number of students who tried paragliding in the sample

Solution : Here n = 10 and p = 0.08

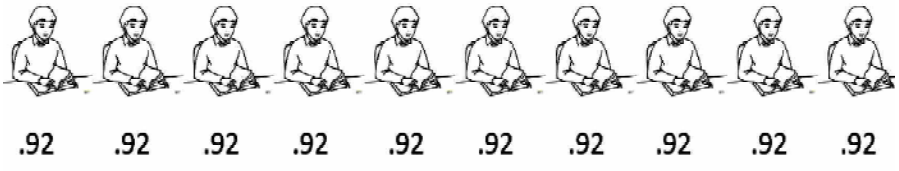
(a)

0.08 0.08 0.08 0.08 0.08 0.08 0.08 0.08 0.08 0.08

The probability of the first student tried paragliding is .08. Probability of second student also tried paragliding is .08 × .08 = .0064. Similarly, the probability of all 10 students tried paragliding is $(.08)^{10}$.

$P(X = 10) = (.08)^{10} = 1.07 \times 10^{-11}$

(b)



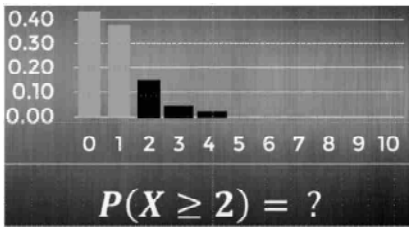
Probability of the first student NOT tried paragliding is .92. Probability of second student also not tried paragliding is .92 × .92 = .8464. Similarly, the probability of all 10 students not tried paragliding is $(.92)^{10} = P(X = 0) = .4344$

(c) It seems that the probability of 2 students tried paragliding is $(.08)^2$ while the remaining 8 students have not tried paragliding is $(.92)^8$. So the answer would be $(.08)^2 \times (.92)^8$.

But it is NOT RIGHT, because there can be the first two students who have tried paragliding or the last two students or maybe alternate one etc. Therefore there can be $^{10}C_2$ ways in which we can have 2 students out of 10, who have tried paragliding.

$P(X = 5) = {}^{10}C_2\ (.08)^2(.92)^8 = 45 \times .0000022 = .1478$

(d) At least two means, two or more. One way we can calculate all these probabilities and add them up (P(X = 2) + P(X = 3) + ... + P(X = 10). (All those black bars in the left graph).



But the easy approach can be, we can calculate P(X = 0) + P(X = 1) and subtract it from 1.

$1 - P(X \leq 1) = 1 - P(X = 1) - P(X = 0) = 1 - .378 - .434 = 0.188$

(e) The expected value of students out of a total sample of 10, who have tried paragliding

E(X) = np, n is the number of trials while p is the probability of success

Here, $E(X) = 10 \times 0.08 = .8$

Therefore .8 is the mean of the sampling distribution

(f) **Standard deviation :** $\sqrt{np(1-p)} = \sqrt{10 \times .08 \times .92} = .858$

❖ **Microsoft Excel Ealculations for Binomial Distribution :**

We calculate excel formula to calculate the probability of exact 2 students tried paragliding as we did in the last example, 2.1(c)

The formula is **BINOM.DIST(2, 10, .08, False)**, here the first argument is required individual discrete outcome (2 in this example) then the second argument is total numbers of trial, the third argument is the probability of success, and the last argument will be False for NOT cumulative while True for cumulative.

Similarly, answer to example 2.1(d) would be **1–BINOM.DIST(1, 10, .08, True) = 0.188.**

Please Note : cumulative functions always work in one direction (equal to or lower). Therefore if we want to calculate for a greater than region, then we can subtract the result from 1, as we did above.

1.3.2 Discrete Probability Distributions – Poisson Distribution :

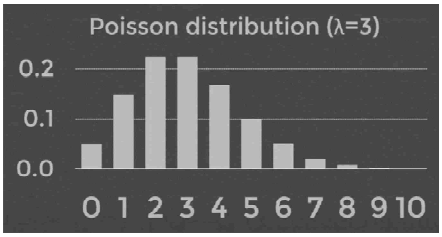
Poisson distribution is named French mathematician, **Siméon Denis Poisson.** He discovered it in 1938. In his surname "I" remain silent; hence it sounds like *Pozzon*. One of the critical assumptions of the binomial distribution is that number of events (trails) must be fixed, but in case that is not true then we use Poisson distribution.

When we cannot calculate the probability of success and failure both, we use Poisson distribution–for example, the number of potholes in a 10 km long road. Here we cannot calculate the probability of success and failure both. We can calculate only one either success or failure. For example, in a 10 km road stretch, we can only count potholes (which can be a measure of failure), but we cannot say how many potholes could be possible in this 10 km stretch (total number of trials).

Another example, number of calls received at a call centre in an hour, here theoretically there can be numerous calls of small duration are possible, but we cannot estimate the total number of possible calls, we can only calculate the average number of calls received in an hour (by analysing the historical calls data).

So Poisson distribution describes the average number of events occurring in a fixed interval of time or any continuum region of opportunity (which can be measure on a continuous scale, like length, money, distance, weight etc.). It requires only one parameter, λ or μ, which tells the average number of events in a given interval. Poisson distribution bounded by 0 and ∞. For example, there can be 0 calls in a day, or there can be an infinite number of calls in a day at a call centre.

The expected value (mean), E(X) and variance, V(X) of Poisson distribution are the same, λ.



Example : On average, three customers need a wheelchair per day (λ/μ) at a local grocery store. In the below graph, these bars go up to infinity although probability will become minuscule, so for convenience, I have put up to 10 customers only.

❖ **Important Assumptions of Poisson Distribution :**

1. Events occur at a constant rate, which means there should be equal chances of the happening number of events in a one–time interval to any other interval of time
2. The occurrence of one event must be independent of any other event (i.e. Events are independent)

Please Note : These assumptions may not hold good in reality or may not aligned with the business problem, which we are trying to solve hence we need to be sure whether it is appropriate and to apply Poisson distribution.

❖ **PMF for Poisson Distribution :**

It is the probability of each discrete outcome (height of each bar in the above graph).

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Here e is Euler's constant whose value is 2.71.

❖ **CDF for Poisson Distribution :**

$$P(X \leq x) = \frac{\Gamma(\lfloor x + 1 \rfloor, \lambda)}{\lfloor x! \rfloor}$$

We don't need to understand this formula as it may be difficult to understand for undergraduate students of this course. Instead, we will see the excel formula to calculate CDF.

❖ **Excel Formula to Calculate PMF and CDF for Poisson Distribution :**

To calculate PDF, **POISSON.DIST(X, λ, FALSE)** while calculating CDF **POISSON.DIST(X, λ, TRUE)**

Example 1.2 :

(a) Calculate the probability of arrival for 5 customers who need a wheelchair in a day, while $\lambda = 3$ is given.

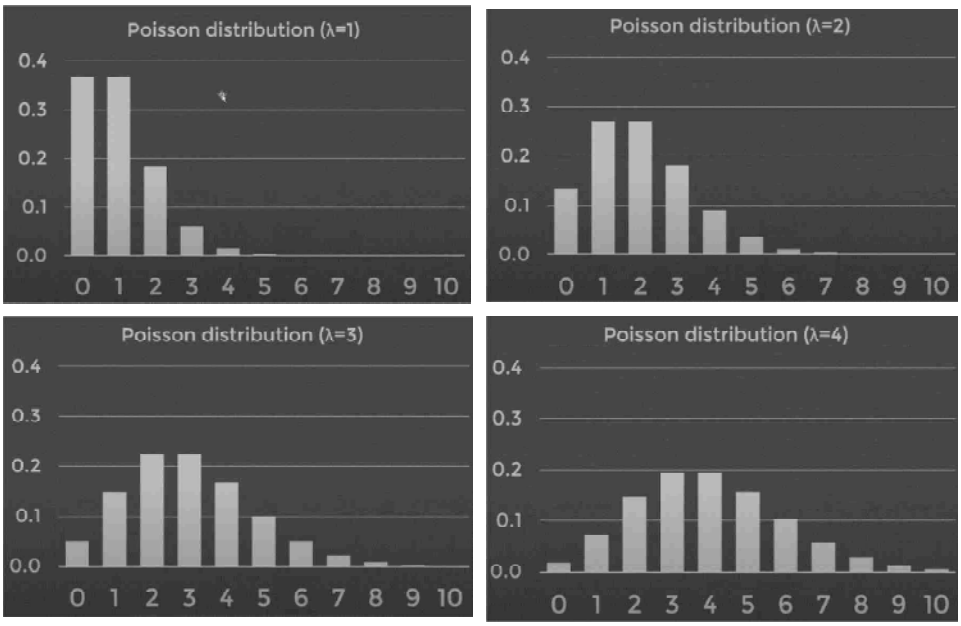
(b) Calculate the probability of arrival of at least 5 customers who need a wheel–chair in a day

Solution :

(a) $P(X = 5) = \frac{e^{-3}3^{5}}{5!} = 0.101$

(b) POISSON.DIST(5, 3, TRUE) = 0.916

❖ **Graphical Representation of Poisson Distributions for Different λ Values :**



Need to remember that λ can be a decimal value also

Example 1.3 : One watch selling company sells an average of 12 watches per day through their website.

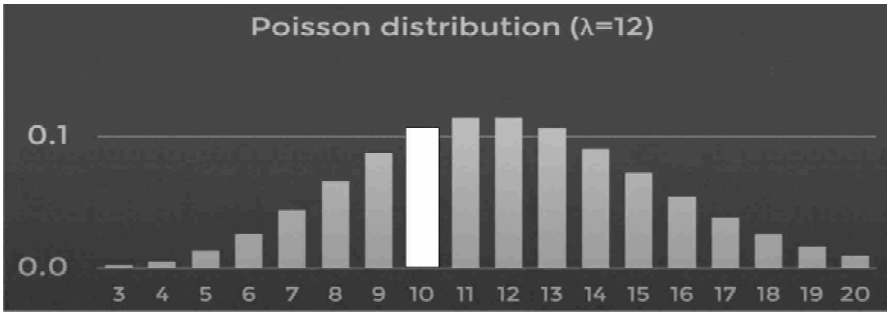
(a) Find the probability of selling 10 watches in a day

(b) Find the probability of selling at least 10 watches in a day

(c) Find the probability of selling more than 1 watch in the first hour of the day

Solution :

(a) Probability of selling exactly 10 sales in a day



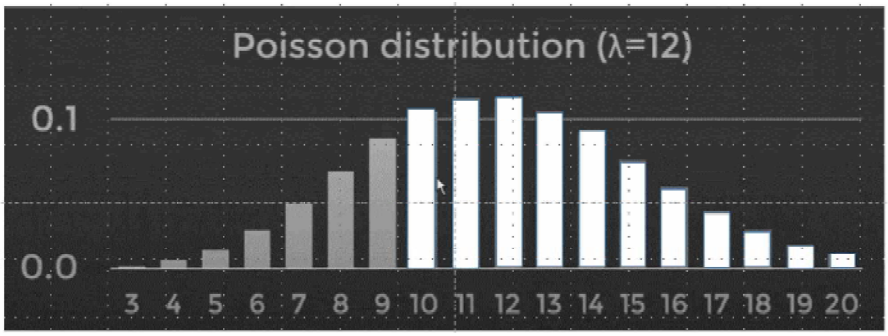
The desired bar is highlighted with white colour in the above graph.

$$P(X = 10) = \frac{e^{-12}12^{10}}{10!} = 0.105$$

POISSON.DIST(10, 12, FALSE) = 0.105

So there is approx. 10.5% probability of selling exactly 10 watches in a day

(b) Probability of selling at least 10 sales in a day, probability of 10 or more in a day



The desired bars are highlighted with white colour in the above graph.

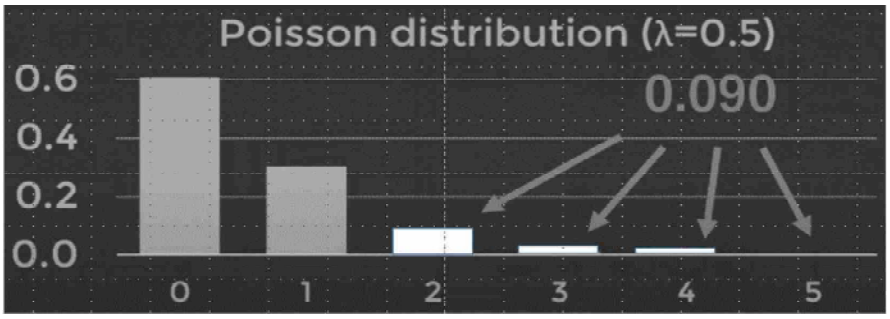
P(X > = 10) = 1 – P(X < 9) = POISSON.DIST(9, 12, TRUE) = 0.758.

So there is approx. 76% probability of selling at least 10 watches in a day

(c) probability of selling more than 1 watch in the first hour of the day

We have the value of λ in terms of sales per day while we need to calculate it in terms of sales per hour. Therefore we need to change our **λ (mean)** accordingly

$$\lambda = \frac{12}{24} = 0.5 \text{ sales per hour}$$



The desired bars are highlighted with white colour in the above graph.

1 – POISSON.DIST(1, 0.5, TRUE) = 0.090. So there is approx. a – 9% probability of selling more than 1 watch during the first hour.

Please Note : we solved the above example mathematically, that is correct but if we re–think the assumptions of Poisson distribution that there should be a constant rate of occurring events through all intervals but here in the case of online watch sales it is very unlikely that somebody will purchase a watch from late night to early morning. So in a way, it is breaching the assumption of Poisson distribution. Therefore, before applying statistical tools or techniques, we have to be very careful whether our use case fulfil all required assumptions or not. Otherwise, we will not get any error theoretically, but our inferences would be very misleading.

Poisson distribution can be an excellent approximation to the normal distribution, but only in case when the probability of success (p) must be significantly small compared to the number of trials (n). One thumb rule says Poisson distribution can be an excellent approximation to the binomial distribution, in case $n > 20$ and $np < 10$. We will cover the normal distribution detail in the next chapter.

Check Your Progress – 2 :

1. Poisson distribution can be an excellent approximation to the normal distribution, but only in case when the probability of success (p) must be _____ compared to the number of trials (n)
2. Cumulative distribution function (CDF) can be used to define a discrete probability function only

   a. True b. False

Check Your Progress – 3 :

1. Which of the following is the formula for variance for a discrete probability distribution ?

   a. Np b. $\Sigma x \times P(x)$

   c. $\Sigma(x - \mu)^2 P(x)$ d. $\sqrt{\Sigma(x - \mu)^2 P(x)}$

2. The expected value of Binomial distribution

   a. Np b. $\Sigma x \times P(x)$

   c. $\Sigma(x - \mu)^2 P(x)$ d. $\sqrt{\Sigma(x - \mu)^2 P(x)}$

3. The variance of Binomial distribution

   a. $np(1 - p)$ b. Np

   c. $\Sigma x \times P(x)$ d. $\Sigma(x - \mu)^2 P(x)$

4. Which one of the following is NOT a valid assumption of Binomial distribution

   a. Only two possible outcomes per trial

   b. The probability of success or failure will remain constant across all the trials

   c. There are an infinite number of trails possible

   d. Each trial is entirely independent of each other

5. What is the expected value of the Poisson distribution

   a. $Np(1 - p)$ b. λ(mean)

   c. $\Sigma x \times P(x)$ d. $\sqrt{\Sigma(x - \mu)^2 P(x)}$

6. Which of the following statement(s) are correct about the Poisson distribution ?

   a. Events occur at a constant rate, which means there should be equal chances of the happening number of events in a one–time interval to any other interval of time

   b. The occurrence of one event must be independent of any other event

   c. Both (a) and (b) are correct

   d. Only (a) is correct

7. Calculate the probability of arrival for 7 customers who need a wheelchair in a day, while $\lambda = 3$ given.

   a. 0.988 b. 0.082 c. 0.022 d. 0.052

8. What of following is the formula for the standard deviation for Poisson distribution

   a. λ b. np c. np(1 – p) d. $\sqrt{\lambda}$

9. One car showroom sells 60 cars per month (historical average) which following Poisson distribution, the analyst wants to calculate the probability of selling 3 cars on the first day of the month. What can be λ in this case

   a. 30

   b. 2

   c. Not sufficient information to calculate the expected value

   d. 3

10. In an exam, there are 25 multiple choice questions, each question has four probable answers but only one answer can be correct if a student is guessing every question. calculate the probability of getting 10 questions correct ?

    a. .042 b. .625

    c. 2.17 d. None of the above

1.4 Let Us Sum Up :

1. In statistics probability distribution for a given random variable explains how the probabilities are distributed over the possible values of the random variable.
2. The shape of a probability distribution depends on two critical factors – mean (expected value) and standard deviation. Although there are other two factors also which are comparatively less critical are skewness and kurtosis.
3. In the binomial distribution, there are only two possible outcomes. Here prefix "bi" indicates two. For example the probability of "yes" or "no", "pass" or "fail", "heads" or "tails" etc.

4. One of the critical assumptions of the binomial distribution is that number of events (trails) must be fixed, but in case that is not true then we use Poisson distribution.
5. When we cannot calculate the probability of success and failure both, we use Poisson distribution–for example, the number of potholes in 10 km long road
6. Poisson distribution can be an excellent approximation to the normal distribution, but only in case when the probability of success (p) must be significantly small compared to the number of trials (n)

1.5 Answers to Check Your Progress :

Check Your Progress – 1 :

1. e **2.** a

Check Your Progress – 2 :

1. Very small **2.** b

Check Your Progress – 3 :

1. c	**2.** a	**3.** a	**4.** c	**5.** b
6. c	**7.** c	**8.** a	**9.** b	**10.** a

1.6 Glossary :

Random Probability Distribution : A random probability distribution is a statistical function that explains all the possible values a random variable can take within a given minimum and maximum range.

Discrete Probability Distribution : If a random variable can obtain only discrete (countable) outcomes, for example, 0, 1, 2, 3 etc. then we say it is a discrete random variable that follows discrete probability distribution.

Probability Mass Function (PMF) : It is a graphical representation of probabilities for all possible values of a discrete random variable. This is also known as a frequency function.

Cumulative Distribution Function (CDF) : It is another way to represent the distribution of a random variable, but unlike PMF, it is not limited to discrete variables only. At any point, it represents probability up to that point.

Binomial Distribution : In the binomial distribution, there are only two possible outcomes. Here prefix "bi" indicates two. For example the probability of "yes" or "no", "pass" or "fail", "heads" or "tails" etc.

Poisson Distribution : When we cannot calculate the probability of success and failure both, we use Poisson distribution–for example, the number of potholes in a 10 km long road.

1.7 Assignments :

1. What do you mean by a probability distribution, what are different types of probability distributions ?
2. What is the generic formula of variance and standard deviation for a discrete probability distribution ?
3. What is the relationship between probability mass function and probability cumulative function, explain it with an example ?
4. Explain the MS Excel formula for Binomial distribution and its important arguments.

1.8 Activities :

1. Suppose email come to a customer care inbox follows a Poisson distribution and the average number of emails per hour is 20

(a) What is the probability that exactly 10 emails will arrive in an hour ?

(b) What is the probability of arriving more than 15 emails in an hour ?

(c) What is the probability of arriving less than 5 emails in an hour

Ans. (a) 0.006

(b) 0.84

(c) 0.000017l,

1.9 Case Study :

Myclick.com is a leading online sales website in Northern India. They have expanded themselves exponentially in the last three quarters. Post adding new segments for women and kids garments they have faced a hike in product returns. By analysing the past few months data, business analysts informed the leadership team that about 10% of their customers return the apparel because of various reasons. Three important reasons are size, colour, material. On a particular week, there were 20 customers purchased apparel from myclick.com.

Questions :

1. Calculate the probability that exactly 7 customers will return the product
2. Calculate the probability that a maximum of 12 customers will return the product
3. Calculate the probability that a minimum of 3 customers will return the product
4. The average number of customers who will return their product during this week
5. The variance and the standard deviation of the number of the returns

1.10 Further Readings :

- "Mathematical Methods of Statistics," Princeton University Press, Carmer H. (1946)
- "Super Freakonomics," Penguin Press, Levitt S. D. and Dubner S. J. (2009)
- "An Explanation of the Persistent Doctor Mortality Association" Journal of Epidemiology and Community Hearlth, Yough F. W. (2001)
- "Data Strategy : How to Profit from A World of Big Data, Analytics and The Internet of Things", O'Reilly Media, Bernard Marr
- "Predictive Analytics : The Power to Predict Who Will Click, Buy, Lie, or Die", Wiley, Eric Siegel

CONTINUOUS PROBABILITY DISTRIBUTIONS

: UNIT STRUCTURE :

2.0 Learning Objectives :

- Understanding concepts and applications of the continuous probability distribution
- Difference between probability mass function and probability density function
- Applications of the continuous probability distribution
- Important properties and assumptions for continuous probability distributions

2.1 Introduction :

In this unit, we will study the theory and application of continuous probability distributions. We will see how these distributions help us to take important business decisions about continuous business metrics like time, money, weight etc. Through various examples from the business world, we will understand the concepts of the probability density function and cumulative density functions. In the end, we will understand the assumptions and decision criteria to select the most appropriate distribution as per the experimental data.

2.2 Probability Density Function :

"I believe that we don't know anything for certain, but everything probably." – ***Christiaan Huygens***

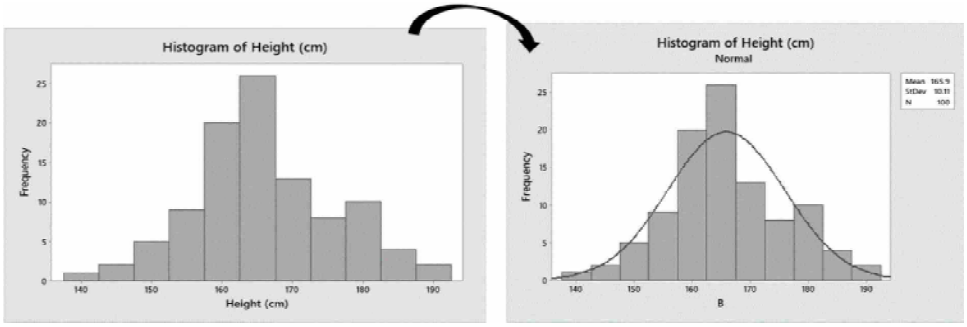
A Continuous Probability distribution can obtain any value in a given range. Unlike a discrete variable, the probability of a particular point in continuous distribution is always 0 because it is one point out of infinite possible outcomes. Hence continuous probability can't be represented in tabular form. There is always a formula or equation to express a continuous probability distribution. This equation to describe a continuous probability distribution is known as the **probability density function**. It is a counterpart of the probability mass function for the discrete probability distribution. Example of continuous variables is height, length, money, weight, distance, time, temperature, blood pressure, etc. As we can split these things into half infinite times.

2.3 The Normal Distribution :

Normal distribution is among the most important distribution in statistics. It has a wide variety of applications. Most of the naturally occurring measures in the real/ business world follow normal distributions.

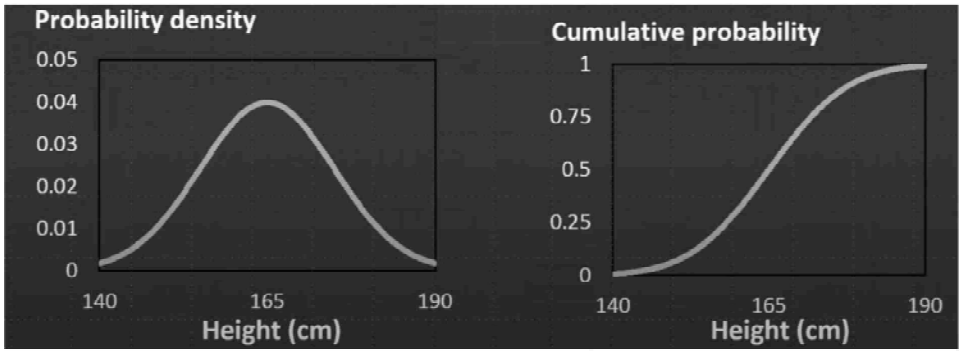
To understand the probability density function (PDF) for normal distribution, let's have sample data of the height of women working in Ahmedabad. To see the shape of the distribution, an easy way to create a histogram.

We can connect the top part of this histogram to see the approximate shape of the probability density function. This bell shape graph represents Normal Distribution.



Here we can see that mean of this height data is approx.–165 cm. The Bell shape in front of the histogram is a pictorial representation of the probability density function, but, in the probability density function graph, Y–axis is not frequency like in the histogram; instead of that Y–axis is always a probability. **A normal distribution can be defined by two parameters, μ (mean) and σ (standard deviation) of data.**

Like we draw cumulative distribution function for the discrete probability distribution; similarly, we can draw CDF in case of continuous probability distribution also.



The formula for PDF represents by f(x) and CDF represents F(x) for the normal distribution is as follows :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < +\infty$$

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt, -\infty < x < +\infty$$

We can ignore these equations due to their complexity as we can calculate these with the help of a standard normal distribution table and excel formula. But we can notice that the above equations have only two variables μ (mean) and σ (standard deviation).

We studied that any distribution can be transformed into standard normal distribution :

$$Z = \frac{X - \mu}{\sigma}$$

This is a typical normal distribution with mean = 0 and standard deviation = 1. The value of Z tells us that the calculated point is how many standard deviations away from the mean.

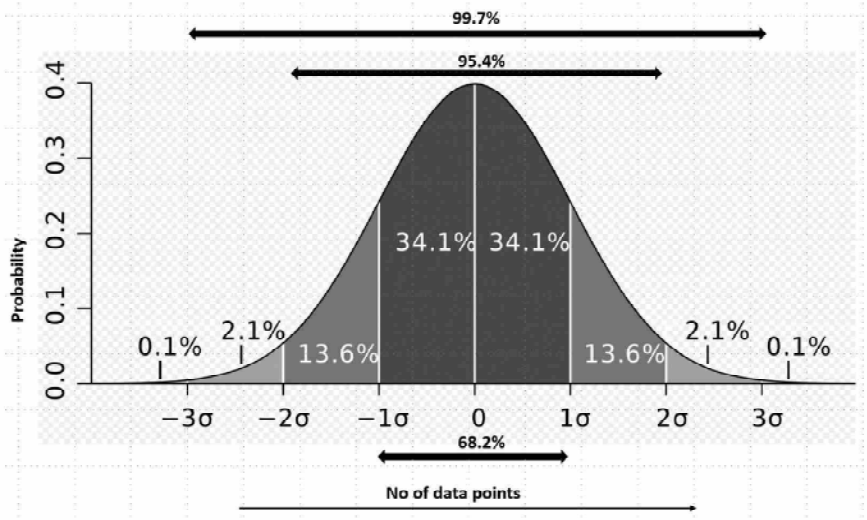
Above standard normal distribution equation can be rewritten in terms of random variable X :

$$X = \mu + \sigma Z$$

❖ **Important Properties of Normal Distribution :**

1. The total area for a normal distribution is always 1 as it is a continuous probability distribution
2. It is always symmetric bellshaped around its mean; it has a mirror image about its mean
3. Theoretically, normal distribution never touches the x–axis; it is defined from $-\infty$ to $+\infty$. This property of probability distribution is known as asymptotic

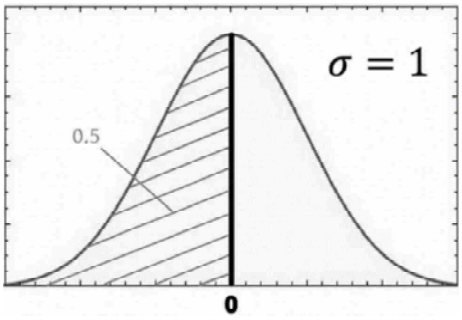
4. For any normal distribution, always area between specific values remain constant (in terms of μ and σ)
5. Linear transformation of any normal distribution is also normal distribution. If X is a normal random variable, then its linear transformation AX + B (where A and B are constants)
6. Two independent normal distributions X_1 and X_2 with means μ_1 and μ_2 and variance σ^2_1 and σ^2_2 respectively. New transformed distribution $X_1 + X_2$ will also follow the normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma^2_1 + \sigma^2_2$



7. If we extract many samples from a normal distribution and draw the distribution of their means, then this distribution of means is likely to follow the normal distribution. This property is known as the **central limit theorem**, will study the central limit theorem in unit 3.

PDF in normal distribution tell us the height of the normal distribution at different values of variable X so instead of using complex PDF formula;

An easy approach is that we can transform variable X to a standard normal distribution. Let's understand this concept in more detail.



The entire area under standard normal distribution is 1, and it is symmetric about its mean.

Therefore half area under standard normal distribution is always .50

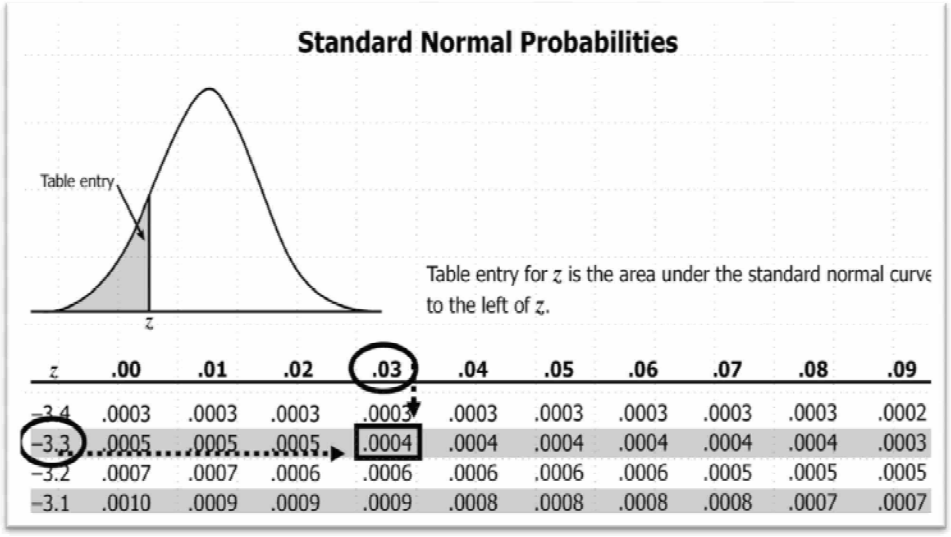
❖ **Steps to Solve Normal Distribution Problem :**

- **Step 1 :** If we have data for variable X is given, then calculate the mean and standard deviation of the data
- **Step 2 :** Draw standard normal distribution for better visualizing the problem
- **Step 3 :** Calculate the Z score

- **Step 4 :** See the value of the Z score in the Standard Normal Distribution (Z) table
- **Step 5 :** Answer will be in terms of probability, multiply it by 100 to get an answer in percentage

2.3.1 How to Check Z Score in Standard Normal Distribution Table (Z Table) :

Suppose we want to see the value of z = –3.33 in the Z table. Now we want to see the value of –3.33 in the Z table. Up to one decimal place we have to see in the left–most column (under z) so –3.3 we will see in the first column, and for the second decimal point, we have to see in the first row. Hence probability value for z = –3.33 is .0004 which means .04% data is left to z value –3.33 (area shown in below standard normal distribution is .04% only). In other words, 99.96% data points are right to z = –3.33 value (unshaded area).



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007

❖ **Excel Formula for Calculating PDF for Normal Distribution :**

As PDF for continuous distribution does not make sense, therefore, we calculate CDF only

NORM.DIST(X, μ, σ, TRUE). The formula will provide direct probability (no need to calculate z value and corresponding probability from a standard normal distribution).

Example 2.2 : Below is a sample data time taken by a bank employee to process the loan file. Bank has set a target to complete the processing of a file to 135 minutes.

Let's solve the below questions with the help of the steps mentioned above to solve a normal distribution example.

Time to process a loan file
136
127
111
122
133
111
115
118
127
140
114
140
125
134
131
126
117
137
136
118
121
114
120
133
115

(a) How many files can be processed under 125 mins ?

(b) How many files can be processed under 115 mins ?

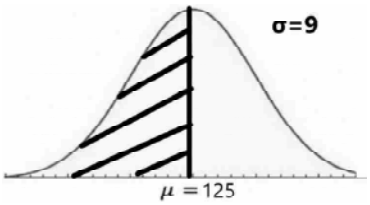
(c) How many files processes took more than 140 mins

(d) How many files took time between 120 mins to 125 mins

(e) What should be the target that they should be able to complete 95% files within the set target

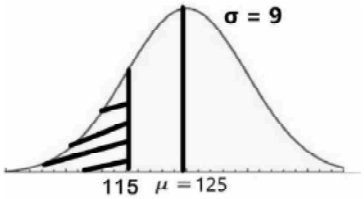
Solution : Here mean = 125 and standard deviation = 9

(a) $Z = \frac{(125 - 125)}{9} = 9$



The value of 0 in the Z table is .5 hence 50% of files are Processed within 125 mins.

(b) $Z = \frac{(115 - 115)}{9} = -1.11$

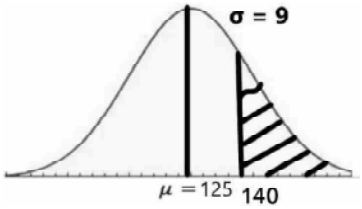


Value of –1.11 is 0.1333 hence 13.33% files processed under 115 mins

(c) P(X > = 140) = 1 – P(X < = 140)

$Z = \frac{1 - (140 - 125}{9}$

= 1 – .9525 = 0.048

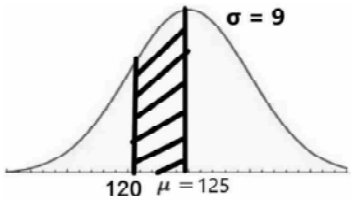


Hence 4.8% files process took more than 140 minutes.

(d) P(120 < = X > = 125)

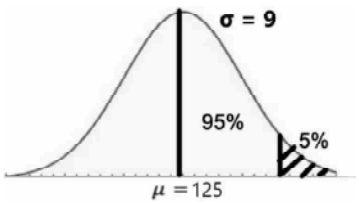
= P(X > =125) – P(X > = 120)

= .5 – .2893 = .2107



Hence 21.07% of files took time between 120 mins and 125 mins

(e) Here we know the probability beforehand hence we need to do the reverse calculation



So we can see the Z table that for which value of Z, we can see the probability .95.

We can see that at z = 1.65, we are getting approx. 0.95

$1.65 = \frac{(X - 125)}{9}$

X = 1.65 × 9 + 125 = 139.5 minutes

Therefore, if the bank will set 139.5 mins target then almost 95% of files will be processed within the set target.

Check Your Progress – 1 :

1. For Standard normal distribution mean is always _____ and standard deviation is _____.
2. Two independent normal distributions X_1 and X_2 with means μ_1 and μ_2 and variance σ^2_1 and σ^2_2 respectively. New transformed distribution $X_1 + X_2$ will also follow the normal distribution with mean _____ and variance _____.

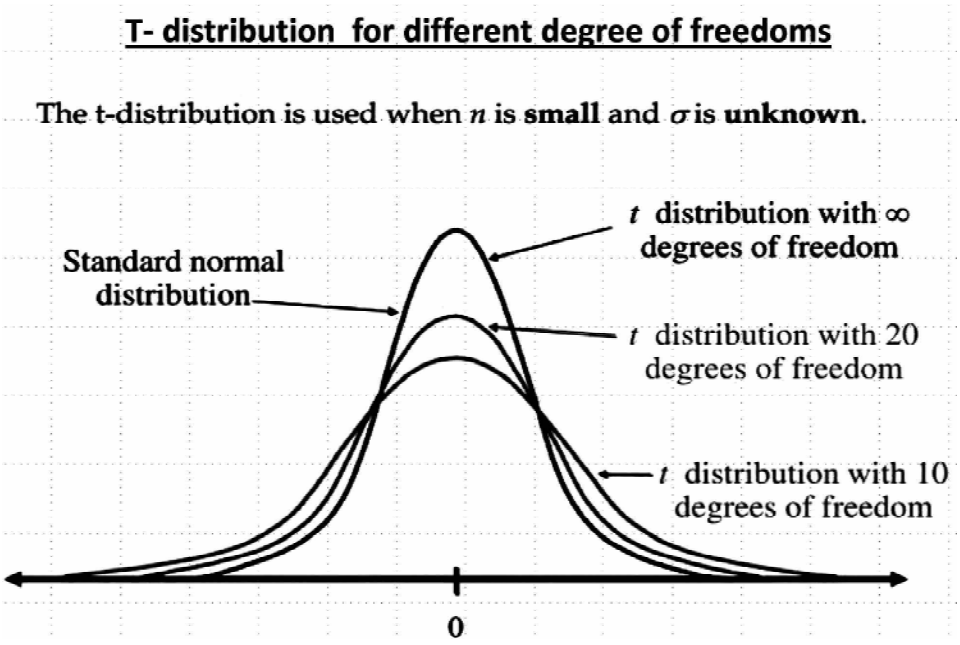
2.4 Student's t–Distribution :

One of the limitations of the normal distribution is that we should know the population standard deviation. Especially in the case of a small sample. In statistics, t–distribution plays a very significant role. For a random variable with mean $\overline{X}$ and S standard deviation of the sample. Then the random variable t can be represented as :

$$t = \frac{\left(\overline{X} - \mu\right)}{\frac{s}{\sqrt{n}}}$$

above equation follows t–distribution with n–1 degree of freedoms. Here the degree of freedom is lost due to the estimation of standard deviation from sample data.

Please Note : Degree of freedom can be calculated as the number of observations in the sample minus the number of estimates made using the data (sample). For example, in the above t–distribution formula, standard deviation needs to be estimated hence lost one degree of freedom. Similarly, for any data, if we have to estimate mean and standard deviation both then in that case degree of freedom will be n–2.



2.4.1 PDF and CDF for t–Distribution :

PDF and CDF formula for t–distribution are very complex; hence we will see only Microsoft Excel function to calculate.

MS Excel formula is **T.DIST(x, degree of freedom, true)**. Here x is the t value we have calculated with the above formula.

2.4.2 Properties of t–Distribution :

1. The mean of a t–distribution with 2 or more degrees of freedom is always 0
2. The variance of t–distribution is $\frac{n}{(n-2)}$ for 2 or more degrees of freedom
3. As we increase the degree of freedom, PDF of t–distribution will approach PDF for standard normal distribution
4. For a small sample, t–distribution has a bell curve shape, but it will be fatter than the normal distribution. But if the sample increase more than 30 observations, t–distribution start mimicking standard normal distribution
5. t–distribution play an important role in hypothesis testing of means of a population and comparing means of two populations

We will study about import properties of t–distribution in the next unit, "Hypothesis testing" and in the next two blocks, "Regression analysis" and "Time series analysis".

Check Your Progress – 2 :

1. We apply t–distribution when population _____ is unknown and sample size is _____.
2. Shape of t–distribution depends on _____.

Check Your Progress – 3 :

1. The weight of bags in a box are normally distributed with mean μ and standard deviation of 6.5 given that 20% of bags are less than 250 gms, the value of μ :

   a. 255.46 b. 250.49

   c. 244.54 d. None of the above

2. A normal distribution can be defined by :

   a. Mean

   b. Standard deviation

   c. Both

   d. Either mean or standard deviation

3. The best scenarios to use t–distribution
   a. Population standard deviation is not known
   b. The population is normally distributed
   c. The sample size is small
   d. All of the above
4. t–distribution is defined by :
   a. Mean
   b. Standard deviation
   c. Degree of freedom
   d. None of the above
5. The total area under normal distribution :
   a. It depends on observations
   b. 1
   c. .5
   d. None of the above
6. % of data points between the mean $\pm$ 2 standard deviations :
   a. Approx. 68%
   b. Approx. 95%
   c. Approx. 99.7%
   d. None of the above
7. Two standard normal distribution X_1 and X_2 with mean μ_1 and μ_2 and variance σ^2_1 and σ^2_2 respectively, then what we can say about new distribution $X_1 + X_2$:
   a. $X_1 + X_2$ will also be normally distributed
   b. Mean of $X_1 + X_2$ will be $\mu_1 + \mu_2$
   c. Variance of $X_1 + X_2$ will be $\sigma^2_1 + \sigma^2_2$
   d. All of the above
8. Mean of t–distribution with more than two degrees of freedom :
   a. Always 0
   b. Depends on the population mean
   c. Depends on the sample mean
   d. Always 1
9. The variance of t–distribution with more than two degrees of freedom :

   a. $\frac{n}{2}$ b. $\frac{n}{(n+2)}$ c. $\frac{n}{(n-2)}$ d. Always 1

10. PDF graph of a standard normal distribution will vary from :
    a. $-\infty$ to $+\infty$
    b. –3 standard deviation to +3 standard deviation
    c. 0 to ∞
    d. $-\infty$ to 0

2.5 Let Us Sum Up :

1. Unlike a discrete variable, the probability of a particular point in continuous distribution is always 0 because it is one point out of infinite possible outcomes.
2. Continuous probability distribution is defined by an equation called probability density function
3. A normal distribution can be defined by two parameters, μ (mean) and σ (standard deviation) of data.
4. The standard normal distribution has zero mean and 1 standard deviation
5. The standard normal distribution is always symmetric bell–shaped around its mean; it has a mirror image about its mean. Theoretically, normal distribution never touches the x–axis; it is defined from $-\infty$ to $+\infty$. This property of probability distribution is known as asymptotic
6. The degree of freedom can be calculated as the number of observations in the sample minus the number of estimates made using the data (sample).
7. The mean of a t–distribution with 2 or more degrees of freedom is always 0 while the variance of t–distribution is $\frac{n}{(n-2)}$ for 2 or more degrees of freedom
8. As we increase the degree of freedom, PDF of t–distribution will approach PDF for standard normal distribution

2.6 Answers to Check Your Progress :

Check Your Progress – 1 :

1. Zero, one **2.** $\mu_1 + \mu_2$, $\sigma^2_1 + \sigma^2_2$

Check Your Progress – 2 :

1. standard deviation, small **2.** Degree of freedom

Check Your Progress – 3 :

1. a	**2.** c	**3.** d	**4.** c	**5.** b
6. b	**7.** d	**8.** a	**9.** c	**10.** a

2.7 Glossary :

Probability Density Function (PDF) : PDF defines as the probability of the continuous variable coming within a given range of values. It generally tells us the height of the distribution at a given point

Cumulative Density Function (CDF) : Derivative probability density function is known as the cumulative density function. It provides us with the area from $-\infty$ to a particular point

Standard Normal Distribution : It is a special normal distribution whose mean is always 0 and the standard deviation is always 1

Degree of Freedom : Degree of freedom can be calculated as the number of observations in the sample minus the number of estimates made using the data (sample).

t–distribution : t–distribution has a bell curve shape, but it will be fatter than the normal distribution. It is defined by the degree of freedom, for a large sample size its shape start mimicking like a normal distribution

2.8 Assignments :

1. What is the probability of a particular point in a continuous probability distribution ?
2. How many data points occur between mean ± 1 standard deviation, mean ± 2 standard deviation and mean ± 3 standard deviations for a normally distributed sample data.
3. TRYC bank takes an admission test (TCAT) to hire freshers. Scores on the TCAT are normally distributed with a mean of 353 and a standard deviation of 80. What is the probability of an individual scoring above 250 in the TCAT exam ?

2.9 Activities :

The teacher surveyed 200 students to know how many hours students study Business Analytics per day. The study shows a sample mean of 68 minutes and a standard deviation of 12 minutes. Assume study hours follow the normal distribution.

(a) How many students study Business Analytics less than 90 minutes per day

(b) How many students study less than 60 minutes

(c) How many students study between 50 minutes and 80 minutes

Ans. (a) 25.46% – approx. 51 students

(b) 96.64% – approx. 193 students

(c) 77.45% – approx. 155 students

2.10 Case Study :

There are approx. 5000 tickets raised for a technical helpdesk in an Akshay software company. The technical helpdesk team resolves a ticket in 25 hours with a standard deviation of 3 hours. Let's say the technical helpdesk manager said that they would award ₹ 100 to associates whose tickets will take more than 28 hours. Calculate the monthly technical helpdesk expense due to penalty.

Question :

Now manager calculates the expense and realized that it is going out of budget, suggest to him the penalty threshold so he should not pay penalty to more than 2% of associates.

2.11 Further Readings :

- "Mathematical Methods of Statistics," Princeton University Press, Carmer H. (1946)
- "Super Freakonomics," Penguin Presss, Levitt S. D. and Dubner S. J. (2009)
- "An Explanation of the Persistent Doctor Mortality Association" Journal of Epidemiology and Community Hearlth, Yough F. W. (2001)
- "Data Strategy : How To Profit From A World Of Big Data, Analytics And The Internet Of Things", O'Reilly Media, Bernard Marr
- "Predictive Analytics : The Power to Predict Who Will Click, Buy, Lie, or Die", Wiley, Eric Siegel

SAMPLING AND CONFIDENCE INTERVALS

: UNIT STRUCTURE :

3.0 Learning Objectives :

- Understanding the requirements of sampling and its importance in business decisions
- Different types of sampling methodologies and techniques
- Significance of central limit theorem and its applications
- Understand the concept of interval estimate
- Application of confidence interval and confidence level

3.1 Introduction :

In this unit, we will study samples and their importance in the business world. Sampling helps in making decisions faster also reduce the cost of the experiment. There are different types of sampling methodologies as per the business scenarios. We will also see how estimating a range of values know as confidence intervals are better than point estimation (average, variance etc.). In the end, we will see the importance of the confidence level of a study and how it influences our study in terms of its effectiveness and efficiency.

3.2 Introduction to Sampling Process :

"Statistical analysis in cases involving small numbers can be particularly helpful because on many occasions intuition can be highly misleading." – ***Sandy Zabell***

Sampling is one of the important tools in the business world as it is directly aligned with the three most important business metrics – cost, efforts and time. In the real business world, it is very difficult to analyze entire data even if we have access to it because accessing entire data is very expensive and time–consuming. A correct sampling methodology by fulfilling all its assumptions can help us to complete endeavours within time and budget.

Sampling is a technique by which we collect only a few data points from the population to reveal information and insights about the population parameters like average, standard deviation, variance, proportion etc. Most of the time, sampling works as two edge sword where right sampling can save huge money and time, but wrong sampling can result in disastrous results. Sampling involves various steps; hence it is important we have to be very cautious at each step to get desired results.

3.2.1 Important Steps in Designing a Sampling Strategy :

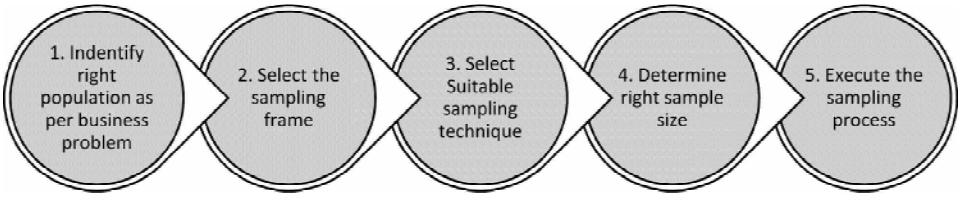


Fig 3.1 Important Steps in Sampling Process

To extract the right sample from a population, we need to follow the above five–step sequential process. Sometimes people think, a larger sample will lead to better predictions, but it is not correct. If we have not followed the right process, then a larger sample will give us more misleading results. Below are important steps in designing a robust sampling methodology :

1. **Identify the Right Population as per Business Problem :** It is utterly important that we should identify the right target population; the definition must be written and aligned with the business problem. For example, we want to understand the feedback about a restaurant from young diners. In this case term "young diners" is vague, it must be mentioned that diners between the age of 24 to 30 years old.
2. **Select the Sampling Frame :** The sampling frame describes the source from where samples have to be extracted; this can be a customer database, directory, company's employees' database, social media portals like Zomato, eat panda etc. It may also be possible that we can use more than one sampling frame for our study.

3. **Select a Suitable Sampling Technique :** The right sampling technique plays an important role in achieving the research objective. Various sampling techniques broadly can be distinguished between probabilistic and non–probabilistic techniques. We will study these techniques in the next section.

4. **Determine the Right Sample Size :** Data collection can be expensive and time–consuming; hence it is very important to calculate the right sample size, which is sufficient for achieving the research objective within budget and time frame. Sample size depends on the required level of confidence, effect size, variation and margin of error.

5. **Execute the Sampling Process :** All the above steps must be executed as the right process, and we must be agile to keep it aligned with the business problem for which we are drafting this sampling strategy. Proper execution requires a nicely identified population, correct sampling frames, appropriate sampling technique and sufficient sample size.

3.3 Sampling Methods :

There are two types of the sampling methodology, probabilistic and non–probabilistic. Which sampling method will be most appropriate always depends on the research objective, constraints etc. Below are various sampling techniques :

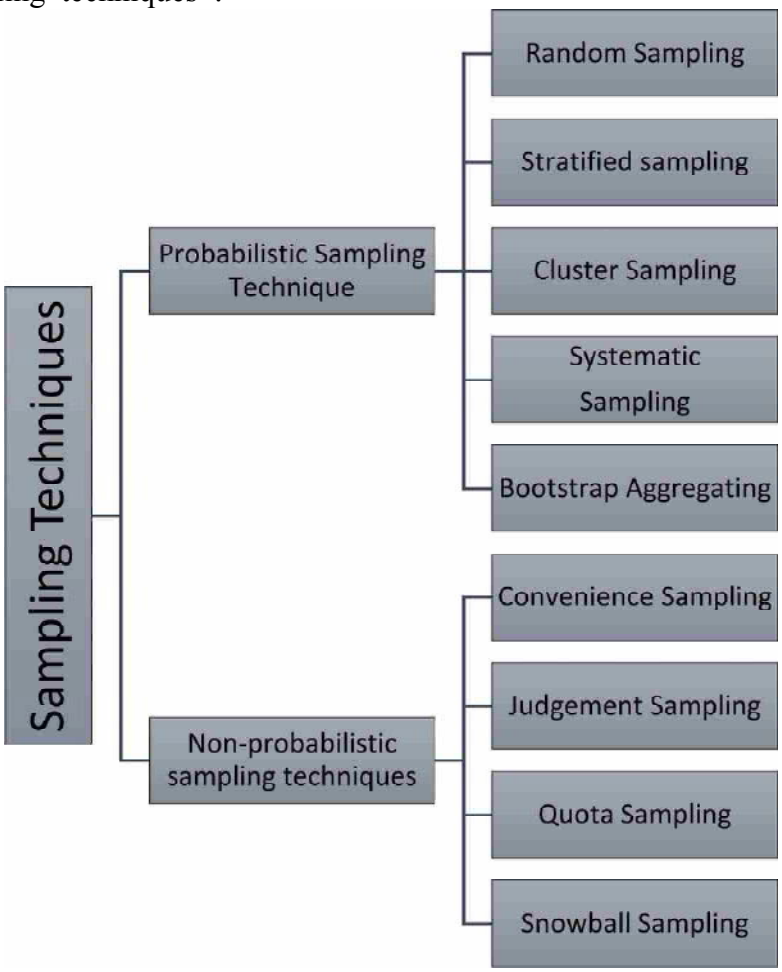


Fig 3.2 Sampling Techniques Tree

Portability based sampling techniques are based on probability distributions. Below are important probabilistic sampling techniques :

3.3.1 Probabilistic Sampling Methods :

1. **Random Sampling :** Random sampling is the most popular and frequently used sampling technique; here, each observation from the population has equal chances to get selected as part of the sample. This is a very good method when the population is homogenous means all observations are of the same kind. For example – from a bulb box of 100 bulbs, we pick up 10 bulbs to check (quality inspection). There is two type of random sampling – with replacement and without replacement. But if we have some type of group different like the population of a city, then random sampling can be a dangerous pick. Exit polls companies do not pick their sample through random sampling techniques or Mobile Phones Company also do not pick their sample customers for feedback on a random basis.
2. **Stratified Sampling :** In stratified sampling we divide the population into some mutually exclusive groups based on some factors, for example, the population of a city can be divided by urban areas, low–income group areas, slums, middle–income group areas etc. or population of a city can be a divide based on their education level, age, marital status or type of their jobs etc. The clusters formed in stratified sampling are known as strata.

   Within each stratum, then we should apply the random sampling method as now data points within each stratum are homogenously distributed. Another important point is we should see the proportion of each stratum in the population and in that proportion only we should extract the sample.

   Below are important steps in designing stratified sampling :

1. Identify the strata in the population, in the figure on the left we have considered age as a strata
2. Calculate their proportion in the population
3. Calculate the total sample size, now recalculate the sample size for each stratum in the same proportion these are in the population
4. Create the final sample by combining individual samples of each stratum

3. **Cluster Sampling :** Cluster sampling also divides the population into a few groups as we did in the case of stratified sampling. An important difference between these two types of sampling techniques is that in stratified sampling, we represent all strata in the sample while in the cluster, we can select a few

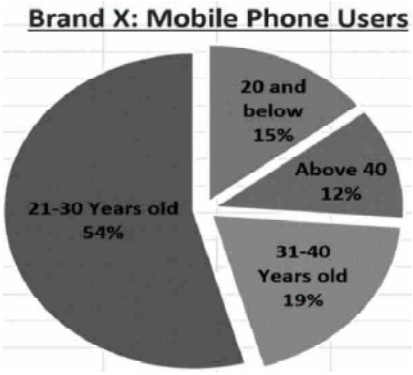


Fig 3.3 Cluster Sampling

clusters only. Second, in stratified each stratum is homogenous while in cluster sampling each cluster is generally heterogeneous (not in similar nature) hence generally we select few important clusters than all because of few reasons contribute significantly than other hence it is making sense to understand few important ones by analyzing their samples, it saves cost and very convenient.

4. **Systematic Sampling :** Here we pick up samples after a regular interval. For example, products are moving on the production belt in a manufacturing firm, and we are picking up every 20th product for quality checking.

5. **Bootstrap Aggregating Sampling (Bagging) :** It is a very useful sampling technique generally used in machine learning. Here we do sampling with replacement. Bagging, generally used in the case where we develop various models based on each sample (with replacement) and final prediction made based on voting.

3.3.2 Non–Probabilistic Sampling Methods :

1. **Convenience Sampling :** Here researcher used samples based on his/her convenience; for example, he is picking up samples from his organization or locality. It is generally a time–saving approach where the result does not get impacted due to biased sample size.

2. **Judgement Sampling :** The researcher selects the sample based on their judgement, again this is a very quick and cost–efficient approach, but as it is also a non–probabilistic approach hence it is very difficult to calculate sampling error (variation in samples). Another drawback it is not comparable with any other researcher as everybody may have a different opinion about the scope of the study.

3. **Quota Sampling :** Quota sampling is quite close to stratified sampling. Here again, the quota is selected based on a few strata like education qualification, age, gender etc. but in quota sampling researchers do not use a probabilistic approach to gather data for each stratum. The quota for each stratum is also in the same proportion these are in the population. Quota sampling is important with time and budget are constraints.

4. **Snowball Sampling :** In the case of snowball sampling, respondents are chosen as per referrals from other respondents. When a researcher wants to collect information beyond his/her domain, then researchers want to meet with somebody who has good knowledge about that subject and further leads also get from the first respondents. It is a very effective way to gather information where we collect information in terms of the survey.

Check Your Progress – 1 :

1. In which sampling technique, samples are picked up after a regular interval :

   a. Quota sampling b. Systematic sampling

   c. Cluster sampling d. Snowball sampling

2. Bootstrap sampling is a type of _____ sampling techniques while quota sampling is a type of _____ sampling technique.

3.4 Central Limit Theorem :

The central limit theorem plays a very important role in entire statistics due to its wide applications in almost every field of statistics, especially in hypothesis testing. It states that if a population is normally distributed (if the histogram is symmetric about mean then distribution said to normally distributed, we will study about various probability distributions in next chapter) and we take few samples of size n then means of these samples will always be normally distributed. The Centre limit theorem works for all types of distribution, either continuous or discrete. An important assumption for the centre limit theorem is that samples are extracted from identical distribution, and samples drawn are independent of each other. Let's see the below simulation to understand the centre limit theorem :

We want to plot the outcome of a six–phase dice throw; we are considering a very small sample size of 4. Let's say the first sample is [2, 4, 5, 5], and its mean is 4. Similarly [2, 4, 4, 5] and its average is 3.75, in the same way, we extracted means of 10 samples, below are 10 sample means :

4
3.75
5.75
4.25
3.75
3.5
3.25
1.75
2.5
4

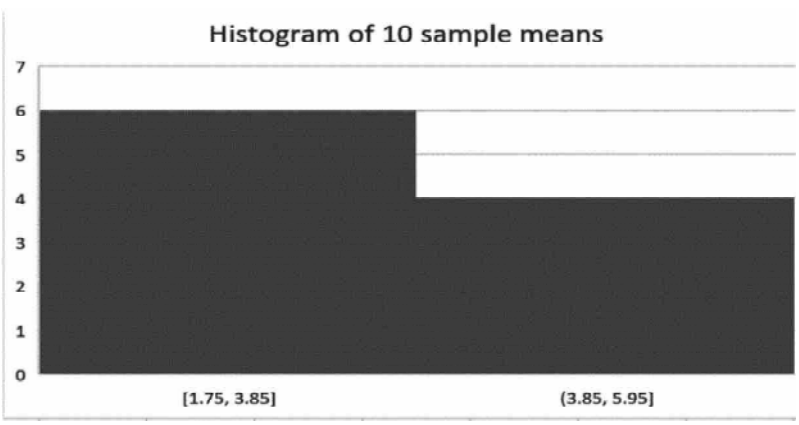


Fig 3.4 Histogram of 10 Sample Means

Now let's see the shape of distribution post 30 means

```
[3       5.75    5.75
4.25     3.75    3.5
3.25     1.75    2.5
4        2.75    2.75
4.75     3.25    3.5
4.75     4.5     2.25
3        2.25    2.75
1.25     2.75    3.75
3.75     4.25    2
3.5      3.5     3.5]
```

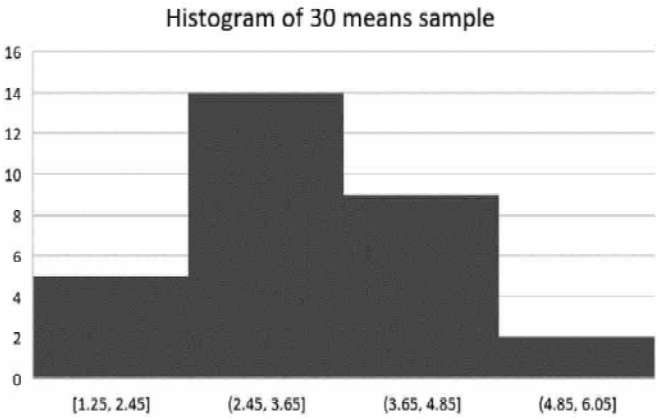


Fig 3.5 Histogram of 30 Sample Means

Now we are seeing, the shape of the distribution is shaping up like normal distribution. Below are histograms of samples 150 and 16,000.

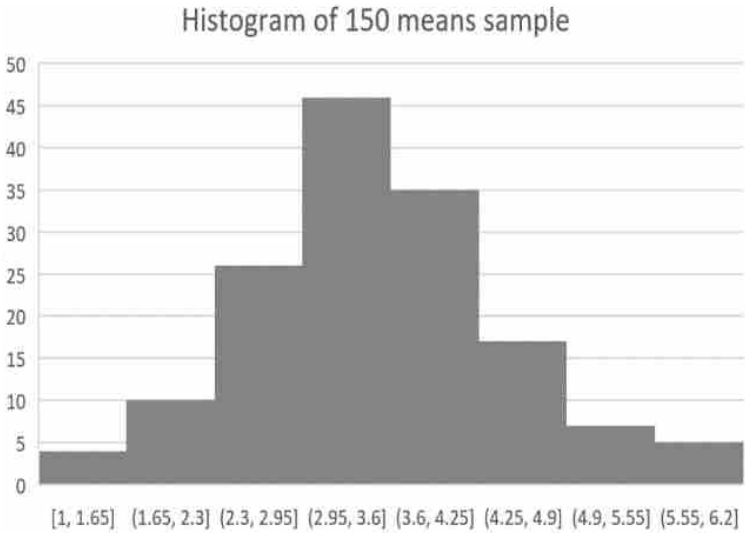


Fig 3.6 Histogram of 150 Sample Means

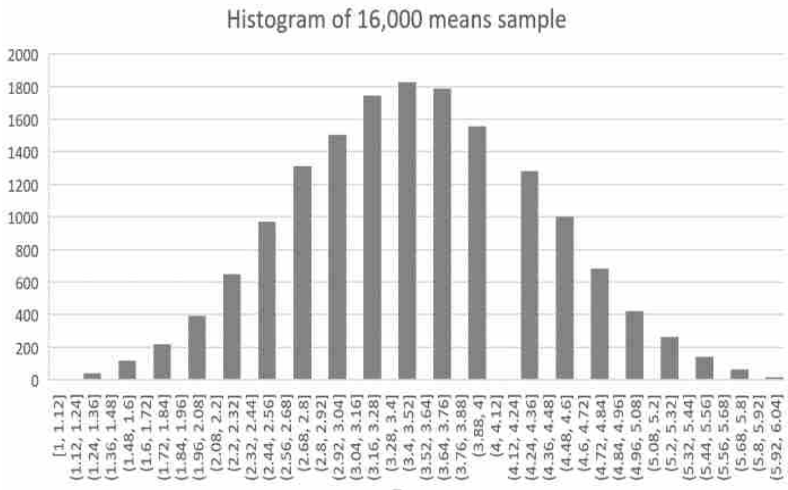


Fig 3.7 Histogram of 16,000 Sample Means

Therefore, we saw, as we are increasing the size of our sample, we are getting perfect normal distribution. This is true for any type of population distribution. There are below very important implications of the central limit theorem :

1. A bigger sample leads to a smaller spread. We saw when we had only 10 means in the sample; our bars were varying from 1 to 6 while in the case of the 150 means sample, most of the data points are varying from 2.3 to 4.9. Irrespective of the type of population distribution, a relatively large sampling distribution (n > 30) will follow the normal distribution. This distribution will have mean same as the population meanwhile standard error (standard deviation of the sample) will be $\frac{\sigma}{\sqrt{n}}$

2. The spread of sampling distribution will be lesser than the spread of population distribution (from which this sample has been drawn)

3. The variable $\frac{(X - \mu)}{\sigma}$ will always have a mean 0 and standard error = 1. This distribution of mean 0 and standard deviation 1. This distribution is known as the standard normal distribution.

	x	(x - μ) / σ
	3	-0.5
	6	2.1
	6	2.1
	4	0.6
	4	0.2
	4	-0.1
	3	-0.3
	2	-1.7
	3	-1.0
	4	0.4
	3	-0.8
	3	-0.8
	5	1.1
	3	-0.3
	4	-0.1
	5	1.1
	5	0.9
	2	-1.3
	3	-1.0
	4	0.4
	4	0.6
	4	-0.1
	3	-1.0
	3	-0.5
Mean	3.57	0.00
Standard Deviation	1.05	1.00

Fig 3.8 Stanardization

This calculation is true for any dataset. If we subtract the entire data set from its mean and divide it by standard deviation then for transformed data, the mean will always be 0, and the standard deviation always is 1. This is known as the **Standard Score**.

This helps us to compare different datasets which are measured in different units like one data is in Kgs while another is in currency.

When we compare two datasets with the help of a standard score is known as standardization as we standardize all data points. Each data is no standard deviation away from its mean. For example, – .05 means .05 standard deviation less than its mean.

3.5 Confidence Interval :

"Confidence comes not from always being right but from not fearing to be wrong." – ***Peter McIntyre***

We studied sampling and sampling error (standard deviation of the sample divided by the root of sample size) in the last section, these two things are building blocks for the formula of the confidence interval. Let's say we are studying the weight of Kesar mangoes from Saurashtra as we can't measure all mango's weight hence we take a sample and measure their weights, but the important thing to note is that we do our

study on the sample while we conclude decisions about the population. So we will extract different samples of mangoes, then every time our average weight will be different. This is called sampling error or variation due to sampling. We can't overcome the sampling error, here confidence intervals help us in this situation. Instead of mean, we should give a confidence interval of the mean. A confidence interval tells us how precise (accurate) our sample estimates are. In other words, the confidence interval states the range in which the weight of mangoes will be.

Confidence interval depends on two factors : variation in the population and sample size. If all mangoes have very little variation in their weight then our sample will also have very little variation hence our confidence interval will be narrow, which means more confidently we can say that weight of mangoes will vary in this small range. The second factor on which the width of the confidence interval depends is the sample size. If we have a small sample, then our sample will not nicely represent our population; hence we will be less confident about our inferences. If we extract different samples, then every sample will tell us quite a different average weight. In contrast, larger samples will be more similar to each other. The effect of sampling error will be reduced in the case of larger samples.

We don't know the variation of the population; hence we estimate with the help of variation in the sample. As we keep increasing our sample size, we keep getting more information about the population; hence our confidence interval keeps getting narrow, which means we are more confident about our inferences. So, if we want to be more confident (say 90% confidence level), then our confidence interval will be wider (we can say more confidently that the weight of mangoes varies from 125 gms to 175 gms. But if we want to be more confident (confidence level 95%) then our confidence interval will be even wider (let's say 110 gms to 190 gms). Somebody can say that I am 99.99% confident that the weight of mango will vary from 10 gms to 400 gms, but that kind of confidence interval does not make any business sense.

Note : Confidence level tells us how confident we want to be about our inferences (study output). It is generally taken as 90% or 95%, which means we want to be 90% or 95% confident about study inferences. 100% – confidence level is known as **α, alpha (error %)**. In other words, we are ok with a 5% error in our estimates or study output. Let's understand it with the help of confidence interval for the mean formula :

$$\text{Confidence Interval} = \overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

Here the value of t depends on sample size and confidence level; there would be a wider confidence level for bigger values of t. Here we are using $t_{\alpha/2,\, n-1}$ because of the two–tailed test, in unit 3 we will study one–tailed or two–tailed tests in detail. Below are some important t values for different confidence levels and sample sizes.

- In the above formula $\overline{X}$ is sample mean
- $t \times s\sqrt{n}$ is called the margin of error
- The confidence interval provides us with the range we are quite sure of (as per confidence level) for the mean of our population.

		Sample Size (n)				
		10	15	20	30	100
confidence	90%	1.833	1.761	1.729	1.699	1.66
	95%	2.262	2.145	2.093	2.045	1.984
	99%	3.25	2.977	2.861	2.756	2.626

- At 95% confidence level, the confidence interval does not tell us that 95% of mangoes will weigh within this range rather it will tell us that there are 95% chances that the average weight of mangoes will lie in this range.
- Therefore if we take lots of samples and create 95% confidence intervals for them, then we can be assured that 95% of them will contain the true population mean although 5% of these confidence intervals will not contain the true population mean
- The confidence interval can be calculated for any population measure (parameter) like median, standard deviation etc
- Excel formula for calculating confidence interval with the help of excel is **CONFIDENCE.T (alpha, standard deviation, size).** The value of $t_{\alpha/2,\, n-1}$ can be calculated using the excel formula **T.INV(α/2, n–1) or another excel formula T.INV.2T(α, n–1)**. There is another formula in excel **TINV(α, n–1)** while **TINV(2α, n–1)** provides value for a one–tailed T–test.

Example 3.1 : A sample of 100 students from the Business Analytics class was taken to estimate their daily study hours. The sample mean is 4.8 hours and the standard deviation of the population is given as 1.4 hours. Solve the below questions :

(a) 90% confidence interval for the population mean

(b) 95% confidence interval for the population mean

Solution :

(a) Confidence interval at 95% confidence level

Here we have n = 100, $\overline{X}$ = 4.8, s = 1.4, $t_{0.025,99}$ = 1.98

Confidence interval = $\overline{X} \pm t_{\alpha/2,n-1} \times s/\sqrt{n}$

$$= 4.8 \pm 1.98 \times 1.4/\sqrt{100} = 4.8 \pm .28$$

Therefore 95% confidence interval for the study hours is (4.52, 5.08)

(b) Confidence interval at 90% confidence level

Here we have n = 100, $\overline{X}$ = 4.8, s = 1.4, $t_{0.05,99}$ = 1.66

Confidence interval $= \overline{X} \pm t_{\alpha/2,n-1} \times s/\sqrt{n}$

$= 4.8 \pm 1.66 \times 1.4/\sqrt{100} = 4.8 \pm .23$

Therefore 95% confidence interval for the study hours is (4.57, 5.03)

An important point to observe is that we can conclude a narrow confidence interval (4.57, 5.03) at a relatively low confidence level (90%) while when we are increasing confidence level to 95% confidence interval becomes wider (4.52, 5.08).

Check Your Progress – 2 :

1. In the formula of the confidence interval, term is known as :
   a. Error %
   b. Confidence level
   c. Sampling error
   d. Margin of error
2. Confidence interval and confidence level are _____ proportional to each other.

Check Your Progress – 3 :

1. What is the main objective of the analysis of sample instead of population :
   a. Reduce cost
   b. Reduce time
   c. Reduce resources
   d. All of above
2. Which of the following is a probabilistic sampling method
   a. Convenience sampling
   b. Snowball sampling
   c. Quota sampling
   d. Stratified sampling
3. Which of the following is a non–probabilistic sampling method
   a. Random sampling
   b. Quota sampling
   c. Cluster sampling
   d. Stratified sampling
4. If there are 2000 employees in an organization and the mean height is 168 cm, then the estimated mean of the sampling distribution will be approx.
   a. 168 cm
   b. 168 / 200 cm
   c. 168 × 200 cm
   d. None of the above
5. Which of the following is not a valid assumption of the central limit theorem is
   a. Samples are drawn from the identical distribution
   b. Samples are independent of each other
   c. None of the above
   d. Both (a) and (b)

6. When we increase the confidence level, how does it impact the confidence interval
   a. Confidence interval gets narrow
   b. Confidence interval gets wider
   c. Does not impact the confidence interval
   d. None of the above is a correct statement
7. For a standard normal distribution
   a. The sample mean is always higher than the population and the standard deviation is lesser than the population
   b. Mean is always 50 and the standard deviation is always 10
   c. Mean, and standard deviation always depends on the population from where the sample extracted
   d. Mean is always ZERO while standard deviation is always ONE
8. The relation between significance (alpha) and confidence level is :
   a. Both are valid only for a two–tailed test
   b. The summation is both is always 100%
   c. Both are valid only for a one–tailed test
   d. None of the above
9. One of the below factors does not impact the width of the confidence interval :
   a. Population standard deviation b. Sample size
   c. Sample mean d. Confidence level
10. How the width of the confidence interval will change. If we increase sample size and also increase the confidence level at the same time
    a. We cannot conclude anything about the width of the confidence interval
    b. The length of the confidence interval will be increased
    c. The length of the confidence interval will be decreased
    d. The length of the confidence interval will remain the same

3.6 Let Us Sum Up :

1. It is very difficult to analyze entire data (population) even if we have access to it because accessing entire data is very expensive and time–consuming therefore we use sample instead of entire population data
2. There are two types of the sampling methodology, probabilistic and non–probabilistic. Which sampling method will be most appropriate always depends on the research objective, constraints etc.
3. If we subtract the entire data set from its mean and divide it by standard deviation then for transformed data, the mean will always

be 0, and the standard deviation always is 1. This is known as the Standard Score.

4. Point estimation of a population parameter gives us a unique value while the confidence interval provides us with the range, we are quite sure of (as per confidence level) for the mean of our population. Hence analysts prefer confidence interval for estimation

5. The confidence level tells us how confident we want to be about our inferences (study output). It is generally taken as 90% or 95%, which means we want to be 90% or 95% confident about study inferences.

3.7 Answers for Check Your Progress

Check Your Progress – 1 :

1. b **2.** Probabilistic, Non–Probabilistic

Check Your Progress – 1 :

1. d **2.** Inversely

Check Your Progress – 1 :

1. d	**2.** d	**3.** b	**4.** a	**5.** c
6. a	**7.** d	**8.** b	**9.** c	**10.** a

3.8 Glossary :

Sampling : Sampling is a technique by which we collect only a few data points from the population to reveal information and insights about the population parameters like average, standard deviation, variance, proportion etc.

Probabilistic Sampling Methods : When the researcher picks up a sample from the population based on probability theory. e.g. random sampling, stratified sampling, bootstrapping, systematic sampling, cluster sampling etc

Non–Probabilistic Sampling Methods : When the researcher picked sample without any probability–based method e.g. snowball sampling, Convenience sampling, judgement sampling etc

Center Limit Theorem : It states that if a population is normally distributed and we take few samples of size n then the means of these samples will always be normally distributed.

Standardization : When we compare two datasets with the help of a standard score, it is known as standardization

Confidence Interval : It is a measure of certainty in terms of a probability value that population parameter will fall within a range of values around the mean

3.9 Assignments :

1. Write down important steps involved in designing a sampling strategy.
2. In an organization, there is a total of 3,000 employees. A random sample of 90 engineers reveals that the average sample age is 29 years. Historically, the population ? of the age of the company's engineers is approx. 7 years. Construct a 95% confidence interval to estimate the average age of all the employees in this company.
3. What are the important differences between probabilistic and non-probabilistic sampling techniques ? Write important techniques under each of these categories.

3.10 Activities :

A sample of 70 customers from an online shopping company was taken. The sample mean is 24 transactions as per their transaction history and the standard deviation of the population is given as 35. Calculate a 95% confidence interval for the population mean.

Ans. : (23.09, 24.91)

3.11 Case Study :

ABC Bank limited is ranked 6th in an important customer satisfaction survey conducted by a market analysis firm. Bank wants to penetrate further into the semiurban and rural part of Southern India. Management put customer satisfaction as the focus strategy for next year to attract more customers in these new territories. One of the important customer concerns is the time taken to solve their queries. Analysts have extracted a sample from the database (assume the population is following a normal distribution). The variance of resolution time is 144 mins based on 750 complaints. Management wants to understand the additional requirement of manpower so that problems can be resolved within promised time to resolve a problem.

Questions :

(a) Calculate the 95% confidence interval for the resolution time.

(b) Calculate the 90% confidence interval for the resolution time.

3.12 Further Readings :

- "Mathematical Methods of Statistics," Princeton University Press, Carmer H (1946)
- "Super Freakonomics," Penguin Press, Levitt S D and Dubner S J (2009)
- "An Explanation of the Persistent Doctor Mortality Association" Journal of Epidemiology and Community Hearlth, Yough F W (2001)
- "Data Strategy : How To Profit From A World Of Big Data, Analytics And The Internet Of Things", O'Reilly Media, Bernard Marr
- "Predictive Analytics : The Power to Predict Who Will Click, Buy, Lie, or Die", Wiley, Eric Siegel

Unit 4

INTRODUCTION TO HYPOTHESIS TESTING

: UNIT STRUCTURE :

4.0 Learning Objectives :

- Understanding the basics of hypothesis testing and its applications in decision making
- Learn to set up a hypothesis test and concluding results in terms of business context
- Understand the concept of significance (α) and Type 1 and Type 2 error
- Understand the significance of p–value and its application in concluding hypothesis test

4.1 Introduction :

In this unit, we will study the concept of hypothesis testing and how it helps us to make robust decisions about the future based on a study conducted on sample data. We will see the various types of errors associated with hypothesis testing techniques. We will also understand

the concept of p–value and its application in decision–making with the help of various examples.

4.1 Life Cycle of Hypothesis Testing :

"Beware of the problem of conducting too many hypotheses as more, we torture the data, more likely they are to confess, but confession under duress may not be acceptable in the court of scientific opinion"

– Stephen Stigler

Hypothesis testing is one of the most critical tools in statistic's quiver. We can't analyze the entire population because of time, and budget constraints; so most of the time we analyze the sample only then the hypothesis test runs on those results in order to conclude whether the results will remain valid for the population or not. One of the metrics to check the authenticity of any analysis is whether results will remain valid for unseen data.

A hypothesis can be understood as a claim made by someone. The claim is generally regarding the population parameters like mean, standard deviation or proportion etc. by seeking evidence from the sample study. For example, the average time to grant a home loan is 22 days. Hypothesis testing is a process that rejects or retain this claim by analyzing the sample data.

❖ **Null and Alternate Hypothesis Statements :**

The null hypothesis, H_0 is generally the commonly accepted fact which researchers want to disprove with their research data. We are going to believe that null hypothesis statement unless it is proved wrong. Null does not mean *"nullify"*, but it means *"nothing going to change or things will remain the same"* as these were before the research. Opposite to null hypothesis (H_1/Ha) is known as alternate hypothesis statement. These two statements are mutually exclusive and exhaustive statements which mean collectively both these statements cover all possible scenarios. Below are different examples of null and alternate hypothesis statements

1. AZH company wants to check whether new training helped them in reducing call handling time significantly or not :

   H_0 : There is no change in average call handling time after a training

   H_1 : Average call handling time changed significantly after training

2. New engine helped in improving the mileage of a car brand

   H_0 : $\mu <= 15$

   H_1 : $\mu > 15$

3. A retail store wants to see whether there is a relation between time spent in–store and gender

   H_0 : There is no relation between time spent in–store and gender

   H_1 : Avg time spent in the store is significantly different for one gender than other

Writing a correct null and alternate hypothesis is an important step as it set the research objective clear and provide direction towards data collection, sampling methodology and tool and techniques applied to analyze the collected sample data. We will see it more during problem–solving in the next section.

Hypothesis testing generally plays two roles in analytics or research, it helps to find the new hypothesis for the research as we found during root cause analysis that asking and processing of unnecessary documents for a home loan cause the delay in granting the home loan to the customers. This is a new finding (which we call hypothesis in statistical term) that need to prove with the help of sample data collected. Or we already have a hypothesis like a furniture shop claim that the height of their particular model wardrobe is 175 cm.

Here the furniture shop claimed that the height of their particular model wardrobe is 175 cm. Again, we can validate this claim with the help of data collected for a few wardrobes. So how hypothesis testing helps us to validate these claims. It can be written in terms of hypothesis statements :

$H_0 : \mu = 175$

$H_1 : \mu \neq 175$

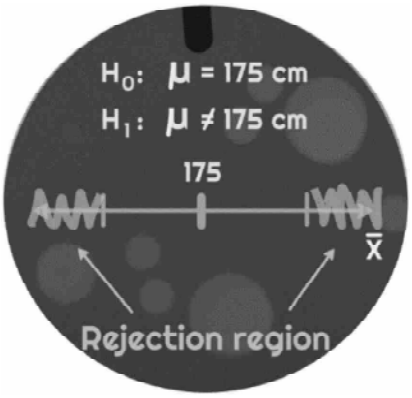


Fig 4.1 Rejection Region

If our sample mean is 170 cm, can we say that shop's claim is valid ? Or if our sample mean is 165 cm, then the shop's claim is not valid ? So, at what point we can say that our sample mean is close to the population mean (claim) or quite far away from the population mean. Hypothesis testing techniques help us to remove this ambiguity from rejecting or retaining hypothesis claims. We know that if we have extracted the right sample (by following all prerequisites of sampling theory), then our sample means should be equal to the population mean. We know there is always some random variable due to which it may not be exactly equal to the population mean, so hypothesis tests help us to draw a boundary beyond which we can say that sample mean is different (significantly) from the population mean.

4.2.1 Hypothesis Testing Process Steps :

There are eight important steps in the hypothesis testing process :

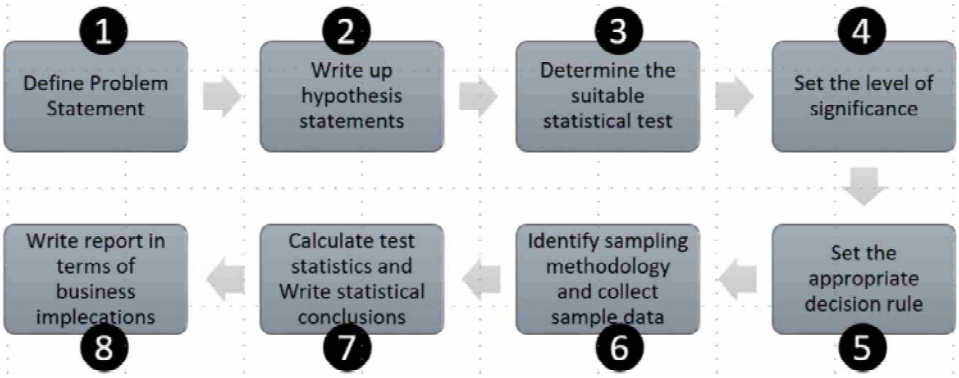


Fig 4.2 Hypothesis Testing Process Steps

1. **Define Problem Statement :** Understand the business problem and covert that business problem into a statistical problem. Describe hypothesis clearly in terms of population parameters like mean, variance, standard deviation, proportion etc. Write clear sentences to avoid ambiguity in understanding the scope of the analysis.
2. **Write Up Hypothesis Statements :** We studied that the null hypothesis is considered to be true until it is not proved wrong with the help of evidence. We should try to write hypothesis statements in terms of the equation if possible.
3. **Determine the Suitable Statistical Test :** The researcher has to identify the most suitable hypothesis test to conduct statistical analysis. Here we have to identify the test statistic to test the validity of the null hypothesis. It depends on the probability distribution that sample data follows. We will study the test statistics in the next section.
4. **Set the Level of Significance :** As we conduct statistical analysis on sample data; hence there will always be an embedded risk in our analysis's result, one type of error is the **level of significance or Alpha (α)**, which is also known as the **significance level. P** is the probability of alpha error. Alpha error is also known as **False positive**. It is the error of rejecting the null hypothesis statement when, in fact, it is TRUE. For example, we are saying a healthy person is sick.

The value of significance is always set by the researcher before collecting data. 100% – significance level is called the **confidence level**. In other words, the researcher said that I want to be 95% sure about my research result as we cannot be 100% sure from sample analysis. So it always depends on the researcher how stringent he/she wants to be sure about his analysis result.

Alpha is the highest level of **p** that we are willing to tolerate and say that a change/difference in the sample is **"Statistically significant"**.

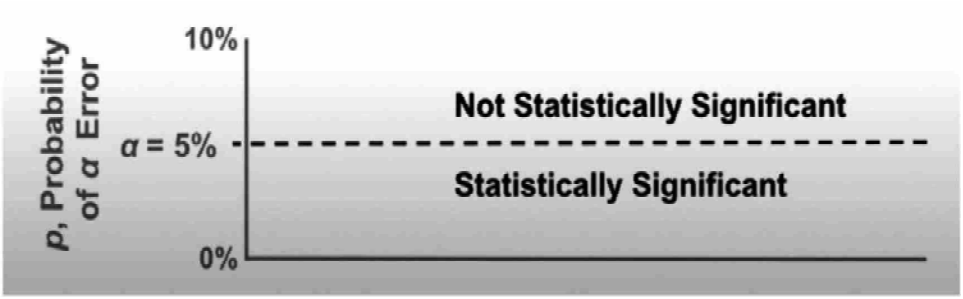


Fig 4.3 Probability of α Error

As we increase the confidence level, the required sample size will also increase exponentially. Conventionally, researchers take **confidence level at 95% or 5% alpha error**. Reducing alpha also increase the β **error**; we will study β error in the next section.

Alpha and p are cumulative probability, the area under the curve.

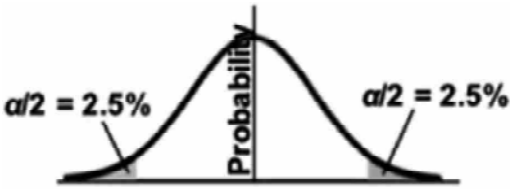


Fig 4.3 Critical Error

Left–hand side curve showing α at 5% for a two–tail test. For one tail test entire 5% will be a shaded area on one side only. This shaded area

Alpha defines the critical values for test statistics like t, z etc. The below diagram shows the relationship between critical value and confidence level.

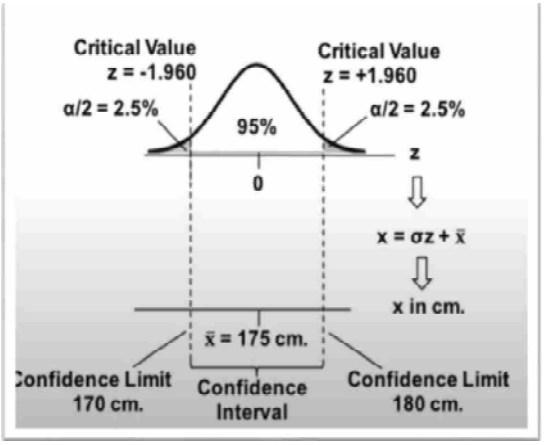


Fig 4.4 Critical Value and Confidence Limits

Let's say our sample of height has a mean of 175 cm. In the last section, we saw these calculations in detail. By converting the Z value of ± 1.960 into real value in cm, we will get the confidence interval for the mean. It is coming out 170 cm and 180 cm for the left and right sides of the confidence interval.

5. **Set the Appropriate Decision Rule :** The researcher selects the value of α, which we calculate the critical values. If the computed value of test statistics falls beyond critical values, then we reject the null hypothesis statement; otherwise, we fail to reject the null hypothesis (accepted).

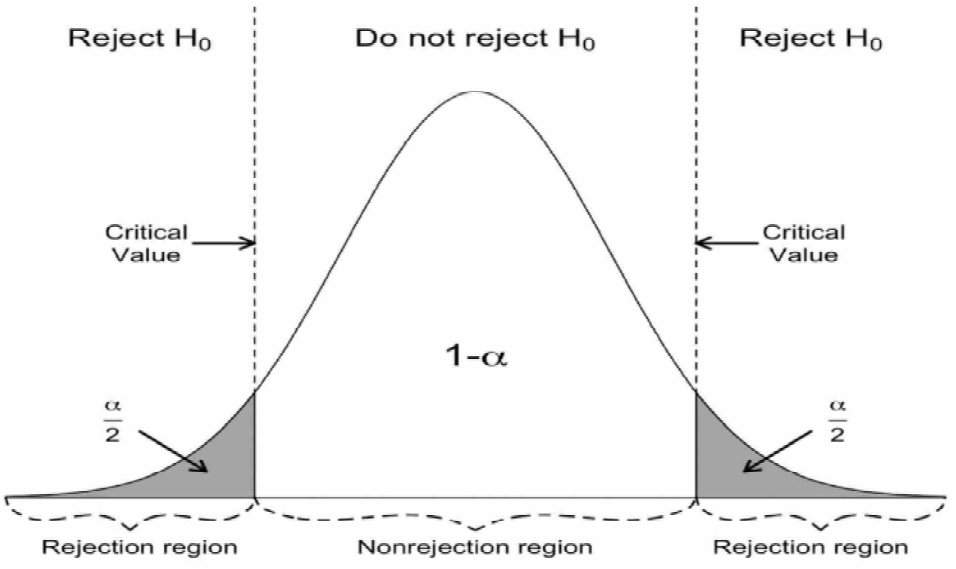


Fig 4.5 Decision Rule in Hypothesis Testing

6. **Identify Sampling Methodology and Collect Sample Data :** At this stage, we apply the appropriate sampling methodology and calculate the sample size required statistically and then collected data used for calculating the test statistic. We should have completed the first five steps before starting the data collection. We already discussed the sampling methods for data collection in the last unit.

7. **Calculate Test Statistics and Write Statistical Conclusions :** We calculate the test statistic; here important steps are selecting the appropriate probability distribution like for a sample size less than 30 we should use t–distribution instead of z–distribution.

8. **Write a Report in Terms of the Business Implications of Hypothesis Testing :** The researcher concludes based on the hypothesis testing result. We reject or fail to reject the null hypothesis testing statement and convert this statistical statement into business language and draw the predictions accordingly.

4.3 Hypothesis Test Statistics :

We create a probability density function for sample difference consider the null hypothesis is true (? = 0). Its mean is always zero and standard deviation = 1 as it is a standardized difference between the estimated value of parameter being tested calculated from the sample and the hypothesis value. It is a standardized difference between $\overline{X}$ and μ if we are testing the mean).

A test statistic is a standardized value used for calculating the p–value (probability) in support of the null hypothesis. We set the value of ? as per our risk–taking ability; then we draw the critical value beyond which we have rejection region(s). Here in the below graph, we considered α = 5%. As below is the two–tailed Test; hence we will calculate two critical values one for the left side and the other for the right side. We know each side there is a .025% region and if we see the value of .025

in the Z table, then the .025 value comes for Z = – 1.96. Similarly .9750 comes for + 1.96. Therefore ± 1.96 are critical values at both sides, and beyond these rejections, the region starts.

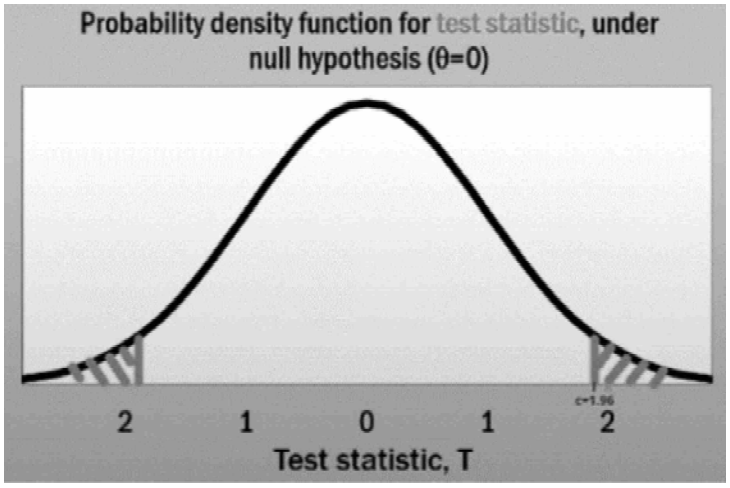


Fig 4.6 Test Statistic

Test statistics has an associated probability distribution. There are four commonly used test statistics : z, t, F and χ^2 (Chi–Square). We will z and t and in this chapter in upcoming sections. A higher value of test statistic indicates that the sample is likely to be more accurate as a representative of the population.

In case the calculated value of the test statistic is greater than the critical value (test statistic value lies in the rejection region), we conclude that there is a statistically significant difference. We reject the Null Hypothesis statement.

4.4 Two–Tailed and One–Tailed Hypothesis Test :

A researcher wants to check the hypothesis that the productivity of a Surat factory of a toy manufacturing company is different from their Vapi factory. μSurat and μVapi are the average productivity of these two locations. In this scenario, the rejection region can be either side of the mean. As the rejection region is both sides of the distribution; hence, it is a two–tailed test. Hypothesis testing statements can be written as below :

$$H_0 : \mu_{Surat} = \mu_{Vapi}$$

$$H_1 : \mu_{Surat} \neq \mu_{Vapi}$$

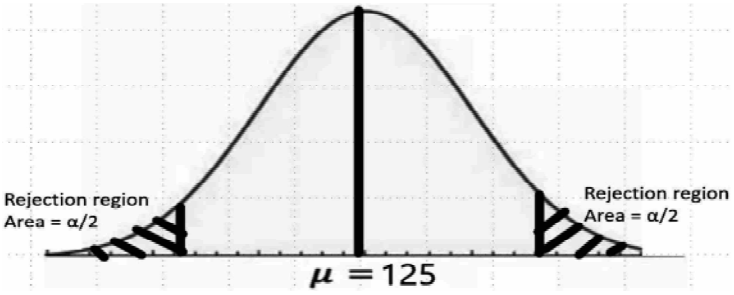


Fig 4.7 Two–Tailed Test

In scenarios where the rejection region lies only on one side of the distribution, we called it a one–tail test. If the rejection region lies on the left side, then we call it a left–tailed Test otherwise if the rejection region lies only on the right side then we call it a right–tailed test. Below are the examples :

For a right–tailed test, the average salary of a business analyst in an organization is at least ₹ 65,000

$$H_0 : \mu_{salary} <= 65000$$

$$H_1 : \mu_{salary} > 65000$$

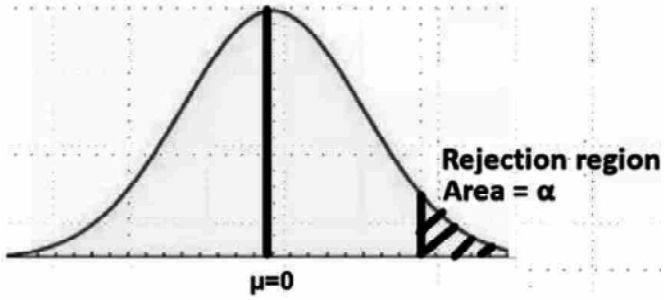


Fig 4.7 Right Tailed Test

For the left–tailed Test, the average salary of a business analyst in an organization is lesser than ₹ 65,000

$$H_0 : \mu_{salary} >= 65000$$

$$H_1 : \mu_{salary} < 65000$$

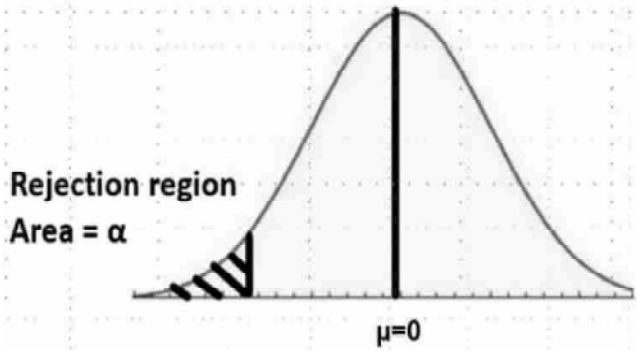


Fig 4.8 Left Tailed Test

In Microsoft Excel **NORM.S.INV(α)** can be used to find the critical value for a left–tailed test while **NORM.S.INV(1– α)** for the right–tailed Test. **NORM.S.INV(α/2)** and **NORM.S.INV(1 – α/2)** for a two–tailed test.

4.5 Concept of p–Value :

In layman terms, the p–value is the probability value that indicates how likely it is that a result occurred by chance alone. The p–value is a conditional probability of observing the sample statistic value when the null hypothesis is true. P–value (probability) is evidence in support of the null hypothesis statement.

Let's try to understand it with the help of an example. The average salary of young analysts in the Analytics and Insight department of company Circa Ltd. is at least ₹ 55,000.

Here, H_0 : μ <= 55000

He has collected sample salary information from the Human resource department and found that the sample mean ($\overline{X}$) is ₹ 55,000. Suppose the standard deviation of the population is known, and the standard error of the sampling distributions is $2500\left(\frac{\sigma}{\sqrt{n}} = 2500\right)$.

The standardized difference between the estimated sample mean value and hypothesized salary is :

$$\frac{(55000 - 50000)}{2500} = 2$$

Now we will find the probability of calculating this sample mean if the null hypothesis is true. Remember, a large standardized distance between hypothesized mean and sample mean will result in a low p–value. This is a right–tailed test. Therefore, we will see Z value in the Z table (.9772) which is an area up to Z = 2 from left most point, but we are interested in calculating area beyond Z = 2 hence our interested p–value is 1–.9772 = .0228

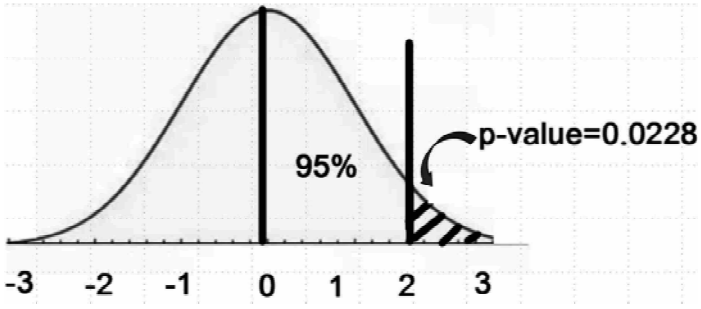


Fig 4.9 Decision Criteria

Important to note that we never say "Null hypothesis accepted" as we cannot prove it in all possible scenarios. We just find evidence where the alternate hypothesis is not valid (it may be possible that the alternate hypothesis statement is not valid in a few scenarios while it may be valid in other scenarios). As soon as we find enough evidence against the alternate hypothesis statement, we declare that we fail to reject the null hypothesis.

4.6 Type I, Type II Error and Power of the Hypothesis Test :

Hypothesis testing results in below two decisions :

1. Reject the null hypothesis statement
2. Retain the null hypothesis (fail to reject)

Below two types of error can happen during concluding hypothesis testing :

1. **Type I Error :** It is a conditional probability of incorrectly rejecting the null hypothesis testing. Here we incorrectly support the claims made by the alternate hypothesis; therefore, it is also known as False Positive. The significance value α is the probability of type I error.

   Type I Error (α) = P (Rejecting null hypothesis | H_0 is true)

   Sometimes students are confused between significance value (α) and p–value. Significance value (α) is an error due to repetitive sampling, while p–value is the evidence for the null hypothesis. Type 1 error is also known as the producer's risk as a quality team rejects a good product or services indicated by the null hypothesis.

2. **Type II Error :** It is a conditional probability of retaining a null hypothesis when the alternate hypothesis is true; therefore, it is also known as False Negative. It is represented by β.

   Type II Error (β) = P (Retain null hypothesis | H_0 is false)

   The power of the hypothesis test can be calculated as **1 – β**.

   Type II error is also known as consumer's risk as a quality team pass a faulty product. Later this faulty product will be purchased by the consumer.

Decision Matrix

		DECISION	
		Reject H_o	*Fail to Reject H_o*
ACTUAL	*H_o True*	Type I Error *Producer Risk* α-Risk False Positive	**Correct Decision** Confidence Interval = 1- α
	H_a True	**Correct Decision** **Power = 1- β**	Type II Error *Consumer Risk* β-Risk False Negative

H_o: Null Hypothesis H_a: Alternative Hypothesis

Fig 4.10 Decision Matrix

Check Your Progress – 1 :

1. In Hypothesis testing result is called "statistically significant" when :

   a. The null hypothesis statement is true

   b. The alternate hypothesis statement is true

   c. P–value is less than or equal to the significant level

   d. P–value is greater than the significant level

2. As per hypothesis testing theory significance test based on a small sample may not provide a significant output even if the correct value differs significantly from the null hypothesis statement value. This type of incorrect result is known as :

   a. Alpha value (the significance level of the test)

   b. $1 - \beta$ (the power of the hypothesis test)

   c. a Type 1 error

   d. a Type 2 error

4.7 Hypothesis Testing for a Population Mean with Known Population Variance : Z–Test :

Z–Test is also known as the one–sample Z test because this hypothesis is carried out with one sample only. In Z–Test, we claim population parameters like mean or proportion when the population variance is known. We studied in the last unit that according to the central limit theorem (CLT) a large sample extracted from a normally distributed population follow the same mean μ as the population mean and standard deviation $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population.

For Z–test, a test statistic is as below :

$$Z - \text{Statistics} = \overline{X} - \frac{\mu}{\frac{\sigma}{\sqrt{n}}}$$

The critical value will always depend on the level of significance α.

Example 4.1 : A company claims that they complete the full and final settlement of an employee within 30 days of quitting the company. The researcher collected data for 40 employees. Assume population standard deviation is 12.5 days. Conduct an appropriate hypothesis test at significance level $\alpha = .05$ to verify the company's claim.

28	26	38	24	16	23	23	41	27	21
16	16	30	37	25	22	19	35	27	32
22	25	24	32	33	28	31	18	29	28
34	28	24	35	24	21	32	29	24	35

Solution :

Hypothesis statements are as below :

$H_0 : \mu >= 30$

$H_1 : \mu < 30$

From the above data points, we calculated the sample mean,

$\overline{X}$ = 27.05 days

Standard deviation of sampling distribution $= \frac{\sigma}{\sqrt{n}} = \frac{12.5}{\sqrt{40}} = 1.974$

$$\text{Z–Statistics} = \overline{X} - \frac{\mu}{\frac{\sigma}{\sqrt{n}}} = 27.05 - \frac{30}{\frac{12.5}{\sqrt{40}}} = -1.4926$$

The critical value of left–tailed test for α = 0.05 is –1.644. Since the critical value is less than the Z–Statistic value, we fail to reject the null hypothesis. P–value for Z = – 1.4926 is 0.06777, which is greater than the value of α, so retain the null hypothesis.

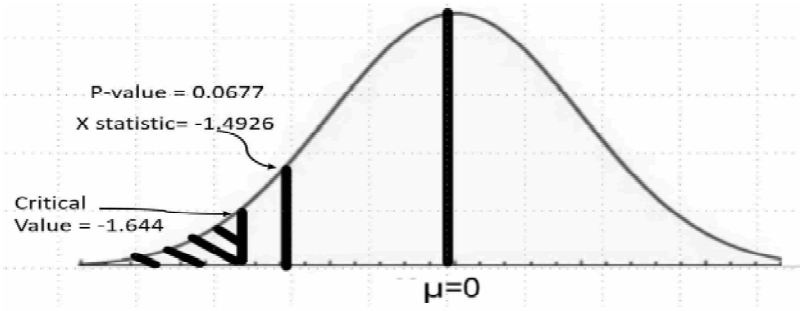


Fig 4.11 Critical Value of One Tailed Test

4.8 Hypothesis Testing for a Population Mean with Known Population Variance : t-Test :

We studied in the second unit that if we extracted a small sample (less than 30) from a normally distributed population with an unknown standard deviation, then sample distribution follows t–distribution with n–1 degree of freedoms. In the case of t–distribution, we have to estimate the variance using the sample itself. Suppose the standard deviation of the sample is S.

$$t - \text{Statistics} = \overline{X} - \frac{\mu}{\frac{S}{\sqrt{n}}}$$

Here n is the sample size; as we are estimating the value of standard deviation, hence we lost one degree of freedom. The above t–statistic formula follows t–distribution with n–1 degree of freedom.

Example 4.2 : An insurance company wants to understand the claim data for one of their products X. They believed that on average there are 500 claims every year. Assume claim data follows the normal distribution. Conduct an appropriate hypothesis test at α = 0.05 to check company's belief about an annual claim is correct.

632	457	335	252	667	636	286	444	636	292
601	627	330	364	562	353	583	254	528	470
762	439	599	708	530	402	729	593	601	408
125	60	101	110	60	252	281	227	484	402

Solution :

Here n = 40, S = 195.0337 and $\overline{X}$ = 429.55

Hypothesis statements are as below :

H_0 : μ <= 500

H_1 : μ < 500

Test statics calculation will be as follows :

$$t-\text{Statistics} = \overline{X} - \frac{\mu}{\frac{S}{\sqrt{n}}} = \frac{429.55 - 500}{\frac{195.0337}{\sqrt{40}}} = -2.2845$$

As this is a one–tailed test (right–tailed) and the critical t–value at α = 0.05 is 1.6848

[In Microsoft Excel, we can calculate the critical value, TINV(2α, degree of freedom) $\rightarrow$ TINV(.1,39) = 1.6848]. Since the t–statistics value is less than the critical t–value, we will retain the null hypothesis.

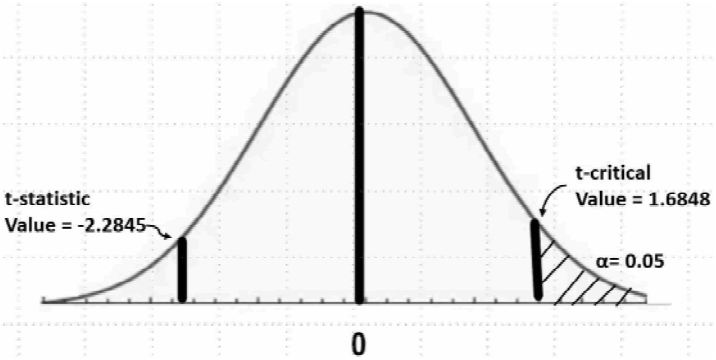


Fig 4.11 Critical Value for a Two Tailed Test

Check Your Progress – 2 :

1. Which one is the correct Null and alternate hypothesis statement for the below sentence :

   The average age of an Indian to get married in 25 years

   a. H_0 : μ > 25; Ha : $\mu \neq 25$ b. H0 : μ = 25; Ha : μ> = 25

   c. H0 : μ> = 34; Ha : $\mu \neq 34$ d. H0 : μ = 25; Ha : $\mu \neq 25$

2. **Sentence 1 – Type I Error :** We conclude that the average age at the time of marriage is NOT 25 years when it is 25 years

   Sentence 2 – Type II Error : We conclude that the average age at the time of marriage is 25 years when in fact it is NOT 25 years

a. Only sentence 1 is correct b. Only sentence 2 is correct

c. Both sentences are correct d. None of the sentences is correct

Check Your Progress – 3 :

1. A claim made about the population parameter for analysis purpose is called

   a. Test–Statistic b. Statistic

   c. Hypothesis d. Level of Significance

2. A researcher is conducting hypothesis testing at $\alpha = 0.10$. P–value is 0.06 then what is the most appropriate answer :

   a. Reject the null hypothesis

   b. Fail to reject the null hypothesis

   c. Information is incomplete to make a decision

   d. α value is too high

3. The null hypothesis gets rejected beyond a point. What we call this point

   a. Significant Value b. Critical Value

   c. Rejection Value d. Acceptance Value

4. If the critical region is split into two parts then which option is correct

   a. The Test is two–tailed b. The Test is one–tailed

   c. The Test is zero–tailed d. The Test is the three–tailed Test

5. When will type I error occurred

   a. We accept H_0 if it is True b. We reject H_0 if it is False

   c. We accept H_0 if it is False d. We reject H_0 if it is True

6. Power of Test can be calculated as

   a. α b. β c. $1 - \alpha$ d. $1 - \beta$

7. Confidence level can be calculated as

   a. $\alpha \times 100\%$ b. $\beta \times 100\%$

   c. $(1 - \alpha) \times 100\%$ d. $(1 - \beta) \times 100\%$

8. What can be another name of the alternate hypothesis

   a. Composite hypothesis b. Simple Hypothesis

   c. Null Hypothesis d. Research Hypothesis

9. Another name for type II error

   a. Producer's risk b. P–value

   c. Consumer's risk d. Confidence interval

10. $\mu = 30$, this statement can be considered as
    a. Alternate hypothesis
    b. Null hypothesis
    c. None of the above
    d. It depends on the scenario; it can be either a null or alternate hypothesis

4.9 Let Us Sum Up :

1. Hypothesis means assumption and validating these assumptions with the help of sample data is known as hypothesis testing
2. Most of the time we do analysis based on sample data, but we want results to be validated for the entire population therefore hypothesis testing techniques help us in this process
3. Hypothesis testing is an important part of various predictive analytics techniques like linear regression, logistic regression etc
4. The central limit theorem is used to calculate the test statistics in the case of t–test and z–test
5. P–value plays important role in concluding the hypothesis testing process, it is the evidence in support of null hypothesis statement
6. We retain or reject the null hypothesis statement by comparing p–value with significance value (α)
7. Rejecting a null hypothesis when in fact it is true is known as type I error or producer risk
8. Retaining the null hypothesis when it is false is called type II error or consumer's risk. It is also denoted by β.
9. we never say "Null hypothesis accepted" as we cannot prove it in all possible scenarios. We just find evidence where the alternate hypothesis is not valid (it may be possible that the alternate hypothesis statement is not valid in a few scenarios while it may be valid in other scenarios). As soon as we find enough evidence against the alternate hypothesis statement, we declare that we fail to reject the null hypothesis.
10. The value of $1-\beta$ is also known as the power of test which means how sensitive is our hypothesis testing in rejecting the null hypothesis when it is false

4.10 Answers for Check Your Progress

Check Your Progress – 1 :

1. c **2.** d

Check Your Progress – 2 :

1. d **2.** c

Check Your Progress – 3 :

1. c	**2.** a	**3.** b	**4.** a	**5.** b
6. d	**7.** c	**8.** d	**9.** c	**10.** b

4.11 Glossary :

Hypothesis Testing : It is a process of validating a claim (hypothesis) based on analysis of sample data

Null Hypothesis Statement : It is a claim about research that is assumed to be true unless it is proved incorrect with the help of collected sample data

Alternate Hypothesis Statement : It is opposite to the null hypothesis statement. It is the interested claim of a researcher which he wants to prove correct with the help of collected data

P–Value : The p–value is the probability value that indicates how likely it is that a result occurred by chance alone. The p–value is a conditional probability of observing the sample statistic value when the null hypothesis is true. P–value (probability) is evidence in support of the null hypothesis statement.

Type–I Error : When we reject a null hypothesis when in fact it is true

Type–II Error : When we fail to reject (accept) null hypothesis when in fact it is false

Power of Test : The value of $1-\beta$ is known as the power of test which means how sensitive is our hypothesis testing in rejecting the null hypothesis when it is false

4.12 Assignments :

1. Define Hypothesis testing, α risk, null and alternate hypothesis testing statements and p–value with example.
2. What is the difference between one–tail and two-tail hypothesis tests, explain with an example.
3. Why type–I error in hypothesis testing is known as producer's error. Explain with an example.
4. Write down null and alternate hypothesis testing statements for the below scenarios :
   a. A company claims that they complete the full and final settlement of an employee within 30 days of quitting the company.
   b. The passport office claimed that they issue passports within 30 days after submission of all required documents.

4.13 Activities :

Peacock training institute invented new technological ways to teach programming to their students. They picked up 20 students and checked their monthly study hours before and after training as per new technology. Below is the table that represents study hours before and after. Conduct a t–test to see whether new technology is motivating students to study more hours. Assume $\alpha = 0.05$

Before New Technology	After New Technology
349	335
449	344
378	318
359	492
469	531
329	417
389	358
497	391
493	398
268	394
445	508
287	399
338	345
271	341
412	326
335	467
470	408
354	439
496	321
351	437

Ans. : The value of test statistics is 0.5375 and the critical value of t–test when $\alpha = 0.05$ and degree of freedom 19 is 1.7291. since the t–statistics value is less than the critical value hence we will retain the null hypothesis and the difference is studying hours is not significant.

4.14 Case Study :

ABC limited company has appointed a new analytics and insight department head to streng–then its presence in India and the Asian market. He wants to analyze whether the marketing cost for units A and B for the financial year 2018 and financial year 2019 is significantly different. He asked the MIS team to provide him with sales data and marketing costs in the last 10 years :

Year	Marketing Cost for Unit A	Marketing Cost for Unit B
B2009	41	31
2010	56	36
2011	28	37
2012	42	35
2013	31	38
2014	38	41
2015	39	44
2016	39	45
2017	51	47
2018	43	49

Answer the below Questions :

a. Draw scatter plot of above data

b. Conduct hypothesis testing at a 95% confidence level and conclude your result

c. Will, there be any change in your recommendations if you will run the hypothesis testing at a 90% confidence level

1.15 Further Readings :

- "Mathematical Methods of Statistics," Princeton University Press, Carmer H (1946)
- "Super Freakonomics," Penguin Press, Levitt S D and Dubner S J (2009)
- "An Explanation of the Persistent Doctor Mortality Association" Journal of Epidemiology and Community Hearlth, Yough F W (2001)

BLOCK SUMMARY

In the business world, we have various business metrics which tells about the health of our organization. We should see these metrics over some time so that we can see the trends. Business wants to predict the values with the help of probability theory. One such tool is a probability distribution. Based on the nature of business metrics we classify them among continuous or discrete metrics. We saw how we apply the goodness of fit test to check whether data follows a specific distribution or not. This is done with the help of hypothesis testing. On the other hand, most of the time we do not want to predict the point estimation (e.g. future share price of our organization) instead of that we want to estimate the interval range in which this value can fall. The study of confidence interval helps us to calculate it more precisely and also advise businesses on how they can be ready for extreme situations. Businesses can plan continuity measures in case things will be extremely good or not so good.

In the business world, time always remains a critical resource. Most of the time we want to conclude our research fast so that results can be beneficial for businesses and their end customers. Therefore, we conduct the sample study at the end of the research whether it is meeting with the business objective and there is no adverse effect. This can be done with the help of hypothesis techniques. It helps us to conclude whether the sample result will hold in the future also. Hypothesis testing is the soul of inferential statistics, it plays important role in concluding research outputs.

BLOCK ASSIGNMENT

Short Answer Questions :

1. Write short notes on PDF, PMF and CDF
2. Explain the difference between the binomial and Poisson probability distribution
3. Explain the difference between the normal and student–t probability distribution
4. Write important assumption of Poisson distribution
5. Why do we say that we need to be very cautious about applying Poisson distribution to solve a real–life business problem ?
6. Write a short note on "How to check Z score in Standard Normal Distribution Table (Z Table)"
7. Why confidence interval is better than point estimation, explain with an example
8. Write a short note on the margin of error in the calculation of the confidence interval
9. Is there any relationship between sample size and confidence level, explain with an example ?
10. Why do we need to define null and alternate hypothesis statements ?
11. Write short notes on critical value and rejection region in the hypothesis testing process
12. Write differences between type I and type II error, draw a decision matrix to explain the concept

Long Answer Questions :

1. Write down the basic difference between discrete and continuous probability distributions. Write examples with their formulae
2. Explain important properties of normal distribution, draw diagram wherever possible to highlight the important characteristic of each property
3. Write important properties of t–distribution
4. Explain important steps in designing a sampling strategy
5. Explain two important types of sampling techniques, write down important sampling techniques under each category
6. Write down Hypothesis Testing Process Steps, explain these steps briefly

❖ **Enrolment No. :** ☐

1. How many hours did you need for studying the units ?

Unit No.	1	2	3	4
No. of Hrs.				

2. Please give your reactions to the following items based on your reading of the block :

Items	Excellent	Very Good	Good	Poor	Give specific example if any
Presentation Quality	☐	☐	☐	☐	_____
Language and Style	☐	☐	☐	☐	_____
Illustration used (Diagram, tables etc)	☐	☐	☐	☐	_____
Conceptual Clarity	☐	☐	☐	☐	_____
Check your progress Quest	☐	☐	☐	☐	_____
Feed back to CYP Question	☐	☐	☐	☐	_____

3. Any other Comments

.....

.....

.....

.....

.....

.....

.....

.....

Dr. Babasaheb Ambedkar OpenUniversity

BBAR-304

Business Analytics

BLOCK-3 CORRELATION AND REGRESSION

UNIT 1

COVARIANCE AND CORRELATION ANALYSIS

UNIT 2

SIMPLE LINEAR REGRESSION

UNIT 3

MULTIPLE LINEAR REGRESSION

BLOCK 3 : CORRELATION AND REGRESSION

Block Introduction

Finding the relationship among various business metrics and helping businesses in predicting these metrics for the next period is utterly essential for robust business decisions. Correlation and regression provide this strength to analysts. Regression is the soul of business analytics as it remains the foundation stone for various vital tools and techniques in the analytics world. In the business world, things do not happen in isolation; there is always a story behind all critical incidents; things depend on various factors. Correlation helps in establishing a relationship among business incidents, while regression helps in providing mathematical relationships among these incidents. If we know the value of one metric, then we can predict the value of others. These techniques allow us to find crucial business questions like :

- If the marketing budget is X, then what can be sales in the near future
- Which service model will attract the customers most
- What will be the sales in next month/quarter/year
- Which customer is likely to churn
- Who will be the probable customer for email marketing
- Which age group will be the right customer segment for the upcoming product

Various predictive/ prescriptive analytical models are required to answers the above questions; these models directly or indirectly depend on correlation and regression techniques. In today's era of digitization, businesses collect various data points due to regulatory and industry requirements. Correlation and regression select the most critical variables out of this plethora of data.

Block Objectives

After learning this block, you will be able to understand :

- Understand the concept of covariance and correlation
- Applications of covariance and correlation in the business world
- Mathematical interpretation of covariance and correlation
- Difference between correlation and causation
- Visual interpretation of correlation analysis
- Spearman rank correlation
- Understand the concept of simple linear regression and its mathematical interpretation
- Various stages in the regression model building
- Essential assumptions of linear regression
- Learn Ordinary–Least–Square method for estimating regression parameters
- Validation techniques for a regression model
- Application of simple linear regression in machine learning and predictive analytics

Block Structure

Unit 1

COVARIANCE AND CORRELATION ANALYSIS

: UNIT STRUCTURE :

1.0 Learning Objectives :

After learning this unit, you will be able to understand :

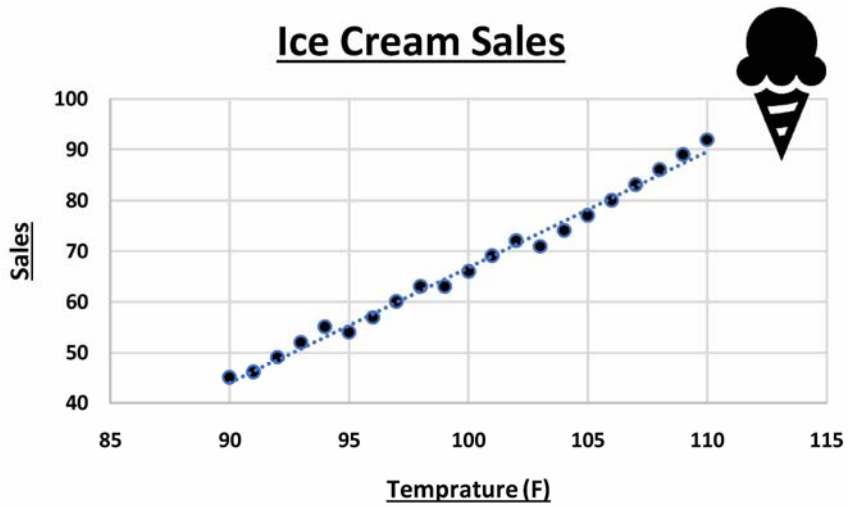
- Understand the concept of covariance and correlation
- Mathematical interpretation of covariance and correlation
- Applications of covariance and correlation in the business world
- Difference between correlation and causation
- Visual interpretation of correlation analysis
- Spearman rank correlation

1.1 Introduction :

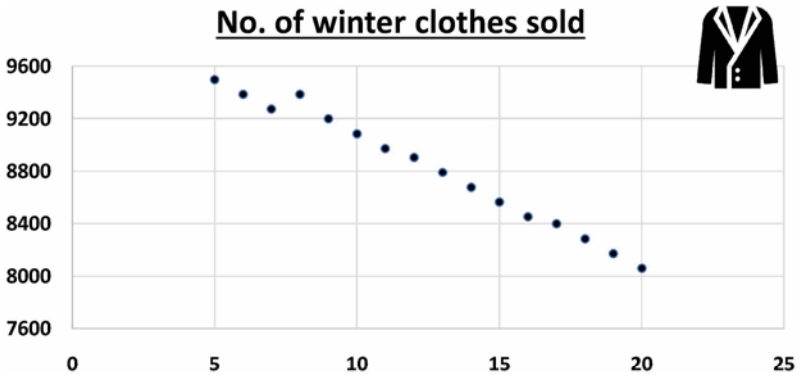
In this unit, we study the statistical relationship between continuous variables in terms of covariance and correlation. We will discuss the different scenarios in which one of these techniques will be more appropriate than another. We will see the limitation of these measures and their visualization techniques. In the end, we will also touch upon the correlation of a ranked data.

1.2 Covariance : Statistical Relationship between Variables :

Covariance is one of the techniques we use in the business world to measure the linear relationship between two variables. Other measures from the same family are correlation and linear regression, which we will study in this block later. Covariance means – "Co–Vary," variables that behave in pairs.

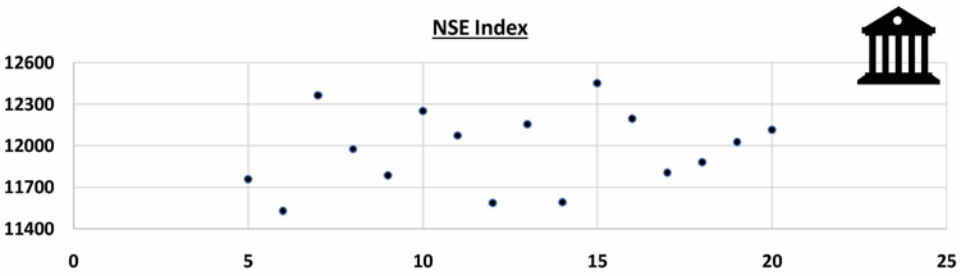


This behaviour can be positive (covariance value is a positive number) means if the value of one variable increases, other variable's value will also increase, e.g., temperature and sales of ice cream, we know in summer sales of ice cream goes up, or temperature and electricity bill in North and West India.



Covariance can be negative (covariance value is a negative number) means if the value of one variable increases other decreases, e.g., temperature and sales of woollen clothes or electricity bills in North and West India.

Or there can be no covariance (covariance value is very close to zero), e.g., movement of the stock market when the temperature goes up. E.g. Relationship between temperature and NIFTY or BSE index. In this case, if one variable increase then second remain constant or significantly less change.



1.2.1 Mathematical Interpretation of the Covariance :

Let's try to understand the mathematical formula for covariance with the help of an example of total study hours in a week and study hours for a Business analytics subject.

Week No.	Total Study Hours	Hours for Business Analytics
1	30	5
2	35	8
3	40	8
4	25	4
5	35	6
Mean	**33**	**6**

As we have studied in the last section, the objective of covariance is to find out the relationship between these study activities. There is one fundamental thing to observe that one column has comparatively high numbers than the second column. So it is not very intuitive to compare these in this form. One way to tackle this problem is to find the deviation of these two columns from their mean.

Col A	Col B	Col C	Col D	Col E	Col F
Week No.	Total Study Hours	Hours for Business Analytics	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	30	5	-3	-1	3
2	35	8	2	2	4
3	40	8	7	2	14
4	25	4	-8	-2	16
5	35	6	2	-1	-2
Mean	**33**	**6**	**0**	**0**	**35**

Now we can compare numbers in columns D and E. It is quite a trial to understand that if both numbers in these columns have the same sign, then these variables are positively correlated. While if there would have been different signs in lots of rows, then that means both variables are negatively correlated. The same thing also revealed from column F that if the number (product of column D and column E) is positive, then these variables are positively correlated, while if there are negative numbers, then negatively correlated.

The same can be derived from the mathematical formula for covariance:

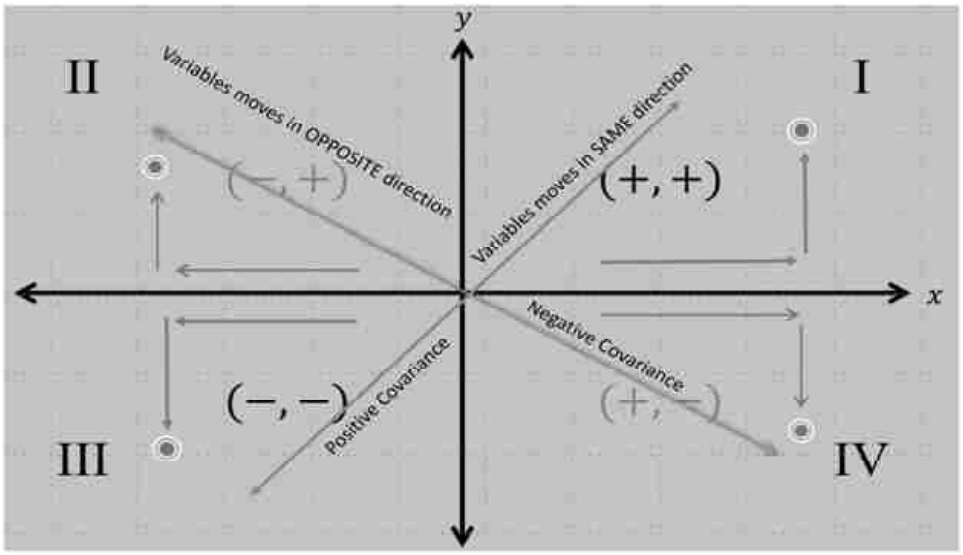
$$COV(x, y) = \sigma_{xy} = \frac{\Sigma\left((x_i - \bar{x})\,(y_i - \bar{y})\right)}{n - 1}$$

Please Note : In the above formula, we have n – 1 in the denominator as we lost one degree of freedom for estimating the mean to the sample.

$$COV(x, y) = \sigma_{xy} = \frac{35}{4} = 8.75$$

So our both variables, "total study hours" and "study hours for business analytics," are positively correlated. Here the vital point to note is that a covariance score of 8.75 does not tell us the strength of a relationship. It depends on the variable range. If we had both variables in lacs, then the value of the covariance coefficient would also be in lacks. So we cannot compare the covariance coefficient of two different studies. It just tells us whether there is positive covariance, negative covariance, or no covariance at all.

Suppose there would have been lots of negative values in column F of the above spreadsheet. In that case, the covariance value might also be negative, which indicates a negative relationship between the variables.



1.2.2 Relationship between Covariance and Variance :

In the last block, we studied variance. Below is the formula:

$$VAR(x) = \sigma_x^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

We can notice that formula for covariance and variance is quite similar; we can also rewrite the variance formula as below:

$$VAR(x) = \sigma_x^2 = \frac{\Sigma\left((x_i - \bar{x})\,(x_i - \bar{x})\right)}{n - 1}$$

if we replace x_j with y_i and $\bar{x}$ in second them with $\bar{y}$ then the formula of variance will turn into the formula of covariance. In other words, covariance is nothing but a variance formula with two variables x and y.

$$COV(x, y) = \sigma_{xy} = \frac{\Sigma\left((x_i - \bar{x})\,(y_i - \bar{y})\right)}{n - 1}$$

Check Your Progress – 1 :

1. Which statement about covariance is NOT correct

   a. If variables x and y increase simultaneously, then covariance between x and y is positive

   b. Zero covariance means we can not calculate the covariance of x and y

   c. Zero covariance means both variables are not related

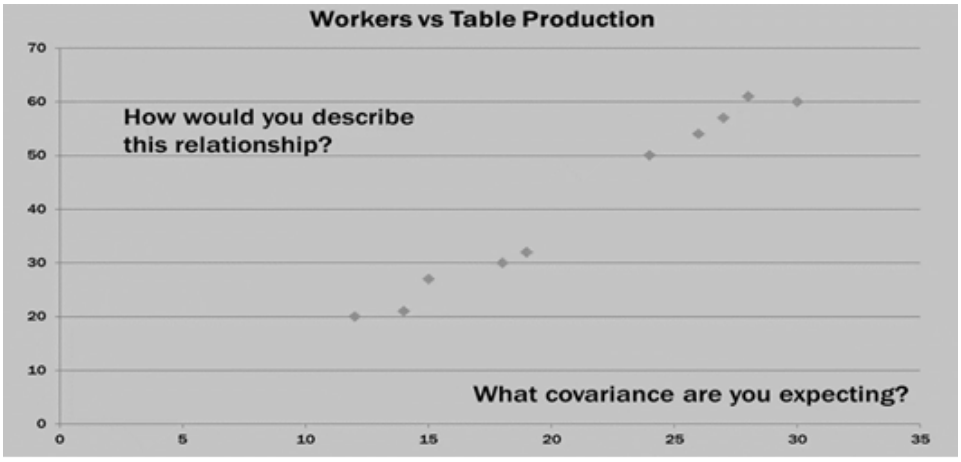
   d. If variable x increase when y decrease, then covariance between x and y is Negative

2. The covariance between rain in centimetres and occupancy percentage of movie theatres is –0.76. this indecases _____ relationship between these two variables.

Example – 1.1 : ABC company wishes to study the relationship between the number of employees (x) and the number of cardboard manufactured (y). Below is the sample data, each one hour in length from the production floor.

x	y
12	20
30	60
15	27
24	50
14	21
18	30
28	61
26	54
19	32
27	57

Solution : It is always advisable to make a scatter plot to see the relationship visually.



We can calculate measures to put in covariance formula.

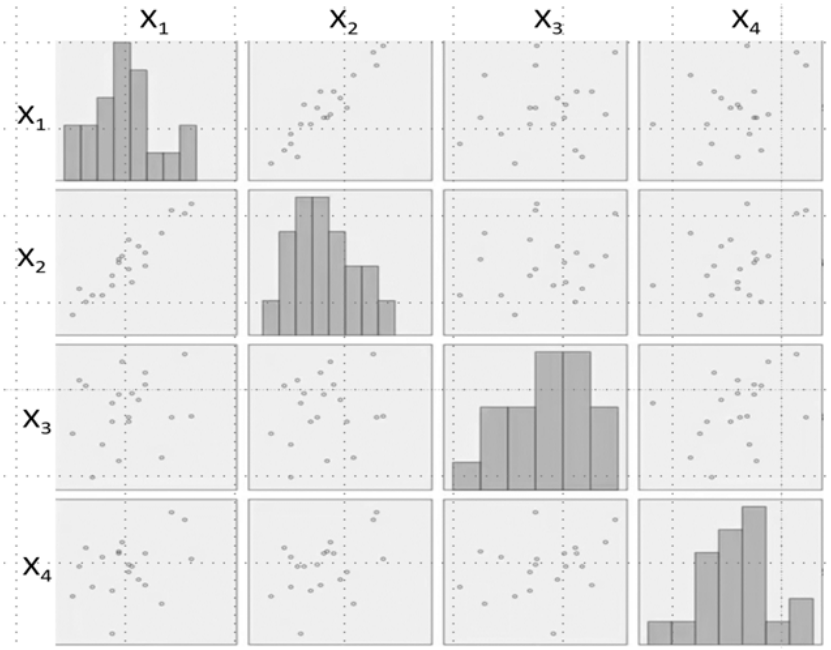
x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
12	20	-9.3	-21.2	197.16
30	60	8.7	18.8	163.56
15	27	-6.3	-14.2	89.46
24	50	2.7	8.8	23.76
14	21	-7.3	-20.2	147.46
18	30	-3.3	-11.2	36.96
28	61	6.7	19.8	132.66
26	54	4.7	12.8	60.16
19	32	-2.3	-9.2	21.16
27	57	5.7	15.8	90.06
$\bar{x} = 21.3$	$\bar{y} = 41.2$			$\Sigma = 962.4$

$$\text{COV}(x, y) = \sigma_{xy} = \frac{962.4}{9} = 106.93$$

Here it is a positive covariance between the number of employees and cardboard manufactured, so if the company wants to increase productivity, then they should increase the workforce.

1.3 Covariance Matrix :

In case we have various variables, then we can create a covariance matrix; we calculate covariance for all possible pairs of variables. Below is the covariance matrix for four variables x_1, x_2, x_3 and x_4



Below is the covariance matrix, where we have covariance for all possible pairs of variables. Diagonally, there is the covariance of each variable with itself, which is a variance of each variable.

	x1	x2	x3	x4
x1	1.008	.895	.634	.545
x2	.895	.918	.490	.652
x3	.634	.490	9.392	1.592
x4	.545	.652	1.592	2.282

The covariance matrix shows us a comprehensive picture of all variables included in the study. The lower diagonal and upper diagonal of the covariance matrix is always identical as the covariance of variable x and y is the same as the covariance of variable y and x.

Check Your Progress – 2 :

1. Which statement about the covariance matrix is NOT correct
   a. Covariance is advisable if we have more than two variables in the study
   b. Off diagonal elements provides covariance between each pair of variables
   c. In the covariance matrix, lower triangular and upper triangular elements are the same
   d. Values always vary between 1 and 5
2. Diagonal elements of covariance matix shows _____ of each variable

1.4 Relationship between Covariance and Correlation :

Correlation values are independent of the scale of data, which means whether we have two variable's values in millions, still, their correlation will always vary between –1 and +1 because correlation is the standardized score by the standard deviation of both the variables. The symbol r represents correlation; the below formula represents it:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Therefore correlation is covariance divided by the standard deviation of both variables x and y.

So if we put values covariance and standard deviation of x and y from the above formula:

$$r = \frac{8.75}{5.70 \times 1.79} = 0.86$$

Correlation also tells us the strength of the relationship as it always varies between –1 and 1. Where –1 indicates a strong negative correlation, while +1 indicates a strong positive correlation between variables. It is always advisable to see the scatter plot of variables before calculating the value of correlation coefficients. As correlation is only applicable for a linear relationship if the scatter plot is showing curvilinear or any other non–linear trend, then we should not use the above formula to calculate the correlation. It may be possible that the value of the correlation coefficient is close to zero for a non–linear relationship, while they may be a strong relationship between both variables.

Please Note : Covariance tells only the direction (no covariance, positive covariance, and negative covariance) of the linear relationship between two variables. In contrast, correlation tells us both direction and strength. The sign of the correlation coefficient is the same as the sign of covariance.

The above correlation formula can be rewritten as below:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\Rightarrow \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Statistician Karl Pearson discovered the correlation coefficient; hence it is also known as Pearson's coefficient of correlation. In Microsoft Excel, the formula for the correlation coefficient is CORREL(array1, array2).

If the correlation between two variables is .7, then it can say that these two variables are explaining a total of 49% variation in the study. In other words, 49% variation is explained by these two variables, while the remaining 51% variation is due to different variables that are not included in the study.

Example – 1.2 : In a manufacturing firm, the maintenance team collected data for the number of batches runs on each machine daily, and the number of the month the machine works without breakdown.

No. of batches per day (x)	25	35	10	40	85	75	60	45	50
Machine life in months (y)	63	68	72	62	65	46	51	60	55

Check whether there is a correlation between the number of batches and machines life.

Solution : We can create a table to calculate all components of the coefficient of the correlation formula:

Sr No	No of batches per day (x)	Machine life in months (y)	X^2	Y^2	XY
1	25	63	625	3969	1575
2	35	68	1225	4624	2380
3	10	72	100	5184	720
4	40	62	1600	3844	2480
5	85	65	7225	4225	5525
6	75	46	5625	2116	3450
7	60	51	3600	2601	3060
8	45	60	2025	3600	2700
9	50	55	2500	3025	2750
Sum	425	542	24525	33188	24640

Hence there is 61% that the number of batches influences the life of the machine before breakdown.

$$r_{xy} = \frac{N\Sigma x_i y_i - \Sigma x_i y_i}{\sqrt{N\Sigma x_i^2 - (\Sigma x_i)^2}\ \sqrt{N\Sigma y_i^2 - (\Sigma y_i)^2}}$$

$$= \frac{9.24640 - 425.542}{\sqrt{9.24525 - 180625}\ \sqrt{9.33188 - 293764}}$$

$$= \frac{-8590}{\sqrt{40100}\ \sqrt{4928}} = -0.61$$

1.5 Spearman Rank Correlation :

Pearson correlation is more appropriate when variables show a linear relationship and both from either interval or ratio scale. When variables show the non–linear relationship or there are outliers in the data or variables are of an ordinal scale.

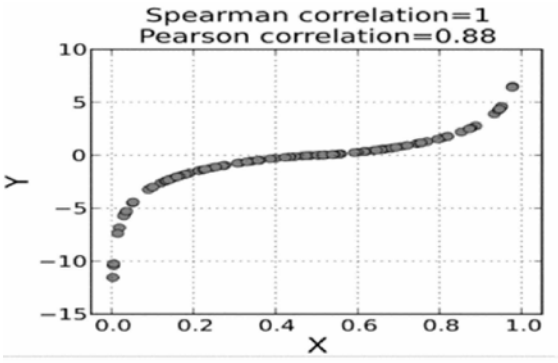


Fig: X and Y are showing non linear relationship

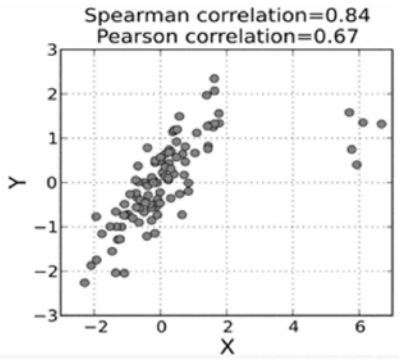


Fig: Scatter plot showing Outliers in the data

Spearman rank correlation is a better choice ρ_s, or r_s denote it:

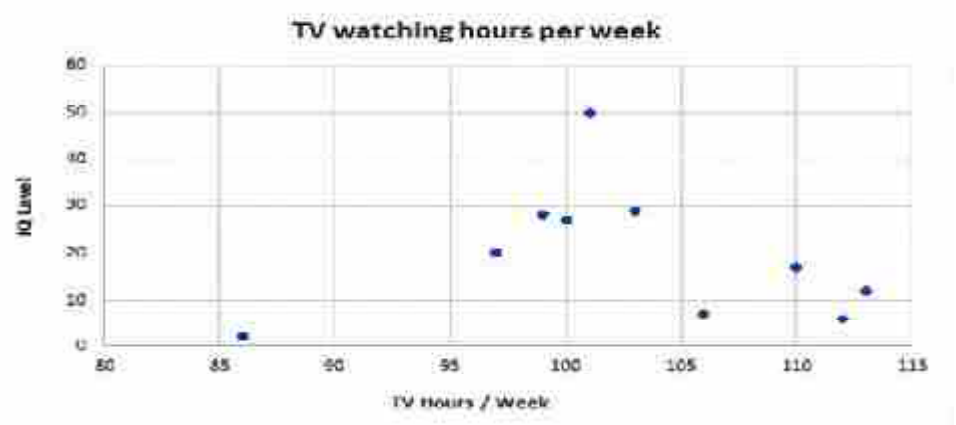
$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)}$$

Here, d_i is the difference between the rank of x_i and y_i

Example – 1.3 : An advertisement company collected data for 'Number of hours TV watch' and 'IQ level.' Below is the data–calculated Spearman rank correlation to check whether there is a relationship between TV watching hours and IQ level.

IQ	106	100	86	101	99	103	97	113	112	110
TV Watching hours / Week	7	27	2	50	28	29	20	12	6	17

Solution : Let's make a scatter plot to see the relationship



The first thing we have to calculate a rank for both x and y variables.

IQ	TV Watching hours/ Week	Rank IQ (Rank X_i)	Rank TV (Rank Y_i)	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Here $\Sigma d_i^2 = 194$ and $n = 10$

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)}$$

to give

$$\rho = 1 - \frac{6 \times 194}{10(10^2 - 1)}$$

Finally, $\rho = \frac{-29}{165} = -0.17576$

The value is very close to zero, which indicates no correlation between "TV–watching hours" and "IQ level."

Check Your Progress – 3 :

1. The covariance of two numeric variables varies:
   a. –1 to +1
   b. –50 to + 50
   c. It depends on the median of data
   d. It depends on the distribution of data
2. Which statement is NOT correct regarding covariance of data
   a. Covariance depends on the sample size
   b. It is applicable for numeric data only
   c. It always varies between –1 and +1
   d. All of the above
3. Which of the following is NOT correct about the correlation
   a. Correlation value always depends on the scale of variables
   b. The correlation value is independent of the scale of the variables
   c. It always varies between –1 and +1
   d. All of the above
4. If the correlation between two variables is .8, then how much variance explained by these two variables
   a. 16%
   b. 40%
   c. 64%
   d. Can't calculate variance if the correlation is provided between two variables
5. Rajesh found a covariance level of 21.98 between the two variables. What we can say about this:
   a. It means there is a strong negative between two variables
   b. It means there is a strong positive correlation between the two variables
   c. It means there is no correlation between two variables
   d. It means there is a weak positive correlation between two variables
6. Mahesh created a scatter plot diagram for two variables for which we want to assess the relationship; he could see a few outliers in his graph. What would you like to suggest as the most appropriate technique to analyze the data:

a. He should calculate the covariance

b. He should calculate spearman rank correlation

c. He should calculate the variance of both variables individually and then compare

d. He should calculate the correlation coefficient for both the variables

7. Which one is the most appropriate option regarding the covariance matrix

a. We should fill all the cells of the covariance matrix, including the upper triangular and lower triangular matrix

b. We should create a covariance matrix only in case when we have more than two variables in our study

c. The covariance matrix is vital for the non–linear relationship

d. We should always calculate the variance of each variable individually before creating the covariance matrix

8. If the correlation coefficient between study hours and the final score is .9, how much variance in the final score is not accounted for by study hours

a. 19% b. 18% c. 45% d. 9%

9. Which statement about covariance is NOT correct:

a. Covariance can be defined as an unstandardized version of the correlation coefficient

b. It is a measure of the linear relationship between two variables

c. It is a synonym of the correlation coefficient

d. It is dependent on units of measurement of the variables

10. **Statement 1 :** Covariance tells about the direction of the linear relationship between two variables

Statement 2 : Correlation tells about the direction and strength of the linear relationship between two variables

a. Only statement 1 is true b. Only statement 2 is true

c. Both statements are true d. None of the statements is true

1.6 Let Us Sum Up :

1. It is always recommended to see the relationship between the variables with the help of a scatter plot before applying a mathematical formula.
2. Covariance explains the linear relationship between two continuous variables
3. Covariance tells us only about direction (whether a relationship is positive, no relation, or negative), but it does not tell anything about the strength of the relationship

4. Covraince score depends on the unit of variables; therefore, we can not compare the covariance score of two different studies
5. Correlation is a standardized relationship score (covariance standardized by the standard deviation of both variables); therefore, its value is always between –1 to +1. We can compare the correlation score of two different studies
6. Correlation explains the only association of relationship but not guarantees the causal relationship between the variables
7. Pearson correlation is appropriate to apply if both variables are continuous (either interval or ratio scale). In the case, variables are in ordinal scale or if there are outlier or relationship is non–linear then Spearman rank correlation would be a better choice

1.7 Answers for Check Your Progress :

Check Your Progress – 1 :

1. b **2.** Negative

Check Your Progress – 2 :

1. d **2.** Variance

Check Your Progress – 3 :

1. d	**2.** c	**3.** a	**4.** c	**5.** d
6. b	**7.** b	**8.** a	**9.** c	**10.** c

1.8 Glossary :

Covariance : It is a linear relationship between continuous variables; it explains only direction, not the strength of the relationship. Covariance scores always depend on unit of variables; hence we can not compare covariance scores of two independent studies

Pearson Correlation Coefficient : It is a standardized linear relationship between two variables; therefore, values always come in between –1 and +1. Correlation does not guarantee causation

Covariance Matrix : In case there are more than two variables in the study, then it is always advisable to see the linear relationship between all possible pairs of variables. This combination of the relationship among various variables is known as a covariance matrix.

Spearman Rank Correlation : If variables are on an ordinal scale, or there are some outliers, or the relationship is not non–linear, the Spearman rank correlation is more appropriate than Pearson's correlation. Here we calculate the correlation between the ranks of the variable instead of raw form.

1.9 Assignment :

1. Why do we call covariance means co–vary, explain it with an example ?
2. What is the significance of covariance metrics visualization ?
3. If two variables are measured in different scales why correlation is better to measure for the relationship than covariance.

1.10 Activities :

1. In the below table, there are dividends provided by two banks for the last five years. Calculate covariance and correlation for the dividends of these two banks. Describe the relationship between the shares.

	Year 1	Year 2	Year 3	Year 4	Year 5
Bank A	5%	7%	2%	–5%	3%
Bank B	–1%	0%	5%	–1%	3%

2. Use the following tabular data to answer the below questions

Variable A	Variable B	Variable C	Variable D
10	110	92	930
11	120	94	900
12	115	97	1020
13	128	98	990
11	137	100	1100
10	145	102	1050
9	150	104	1150
10	130	105	1120
11	120	105	1130
14	115	107	1200

   a. Make covariance matrix of the above data

   b. Calculate the correlation between variable A and variable B

Ans. : **1.** Covariance = 1.075 Correlation = 0.0904

2. (b) –0.51

1.11 Case Study :

ABC limited company has appointed a new analytics and insight department head to strengthen its presence in India and the Asian market. He wants to analyze sales and advertisement costs to examine whether different advertising channels are providing enough sales or not. He asked the MIS team to provide him with sales data and marketing costs in the last 10 years :

Year	Sales (in a million rupees)	Marketing Cost (in a million rupees)
2009	2064	31
2010	2389	36
2011	2418	37
2012	2509	35
2013	2608	38
2014	2706	41
2015	2802	44
2016	2905	45
2017	3056	47
2018	3189	49

Answer the below Questions :

a. Draw scatter plot of above data

b. Calculate the variance of Sales and Marketing Cost data

c. Calculate covariance and correlation of sales and marketing cost

d. Suggest whether Pearson correlation or Pearson rank correlation would be better in this case, justify your answer

1.12 Further Readings :

- "Mathematical Methods of Statistics," Princeton University Press, Carmer H (1946)
- "Super Freakonomics," Penguin Presss, Levitt S D and Dubner S J (2009)
- "An Explanation of the Persistent Doctor Mortality Association" Journal of Epidemiology and Community Hearlth, Yough F W (2001)

SIMPLE LINEAR REGRESSION

: UNIT STRUCTURE :

2.0 Learning Objectives :

After learning this unit, you will be able to understand :

- Understand the concept of simple linear regression and its mathematical interpretation
- Various stages in the regression model building
- Essential assumptions of linear regression
- Learn Ordinary–Least–Square method for estimating regression parameters
- Validation techniques for the regression model

- Application of simple linear regression in machine learning and predictive analytics

2.1 Introduction

Unit : In this unit, we will study Simple Linear regression that explains the relationship between two continuous variables and how it is more effective than correlation and covariance. There are various ways to calculate linear relationships, we will touch upon the most fundamental technique know as least square optimization. We will also touch upon various validating techniques for linear regression. In the end, we will see various examples where linear regression helps us to make better decisions and empowered us with predicting capabilities.

2.2 Essence of Simple Linear Regression :

In the last unit, we discussed covariance and correlation as a measure of the linear relationship between variables. On similar lines, simple linear regression is a statistical technique for finding a relationship between variables, but unlike correlation, it also provides us with a mathematical equation between these two variables. If the value of one variable is known, then we can estimate the value of other variables. Organizations use various metrics to measure their performance, e.g., sales revenue, employee tenure, feedback score, cost of goods sold, growth rate, market share, return on investment, etc. Performance of these matrics depends on various influential factors such as:

- Price of product
- Competitor's price
- Market share by competition
- Marketing budget
- Marketing channels
- New product introduction
- Promotional strategy
- Macroeconomic variables such as GDP, unemployment index, inflation rate, inflation, etc.

In order to make robust business decisions, organizations want to understand the relationship of their essential business metrics with these influencing factors.

2.2.1 Introduction to Simple Linear Regression :

Regression analysis is the statistical technique of developing a statistical model, which can be used to predict the value of one variable (output variable) with the help of another variable (input variable). The output variable is also known as the dependent variable as the value of the output variable depend on the value of the input variable, generally, we denote output by alphabet Y. In the business world, there is always a cause behind every business outcome. The variable on which our output

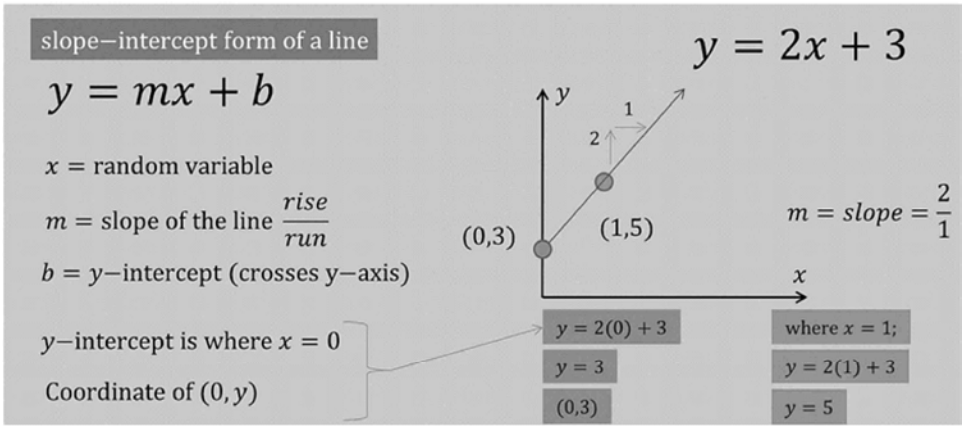
variable (metric) depends is known as the input variable. Regression analysis helps us to calculate the impact of this input variable on the output variable. To see the clear impact of input variable on output variable, regression has an important assumption that input variables should not have any relationship among themselves otherwise we will not be able to calculate the influence of each input variable on output variable. As input variables do not show any relationship, therefore, these are also known as independent variables. In regression analysis, the dependent variable is also known as regressed or explained or response variable. At the same time, the independent variable is also be defined as a regressor or predictor or explanatory or input variable. In simple linear regression analysis, a mathematical relationship between two variables is explained by a straight line, for example, predicting sales revenue based on marketing expenditure or the number of years of experience of employees and salary package or temperature and water consumption, etc. The relationship between one dependent variable and various independent variables is known as multiple linear regression; we will study multiple regression in detail in the next unit.

2.2.2 Determining the Equation of the Linear Regression Line :

Simple linear regression is established on the slope–intercept equation of a line. We have studied the equation of line as:

$$y = b + mx$$

Here, m is the slope (angle with x–axis) of the line, and b is the y–intercept of the line (the point where the best–fit line crosses the y–axis).



But this is a mathematical expression that can be defined for a finite dataset. In case we want to write it for a sample, then there is a score for random error. In that case, instead of a mathematical expression (also known as a deterministic model), we write a statistical expression (also known as the probabilistic model). This can be expressed as below:

$$y = b + mx + \text{random error}$$

Here we can think as y is a function of x: y = f(x);

If the value of x (independent/input variable) is known, then we can estimate the value of y (dependent/ output variable).

Simple linear regression can be expressed in a functional form as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

For a dataset with n observations, we can write the above functional form for each of these observations as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where

Y_i is the ith observation (data) of the dependent (output) variable in the dataset

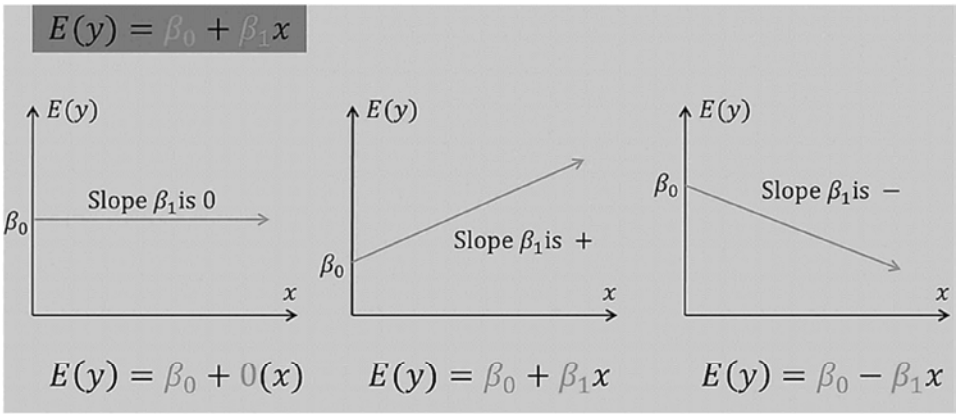
X_i is the ith observation of the independent variable in the dataset

ε_i is the random error

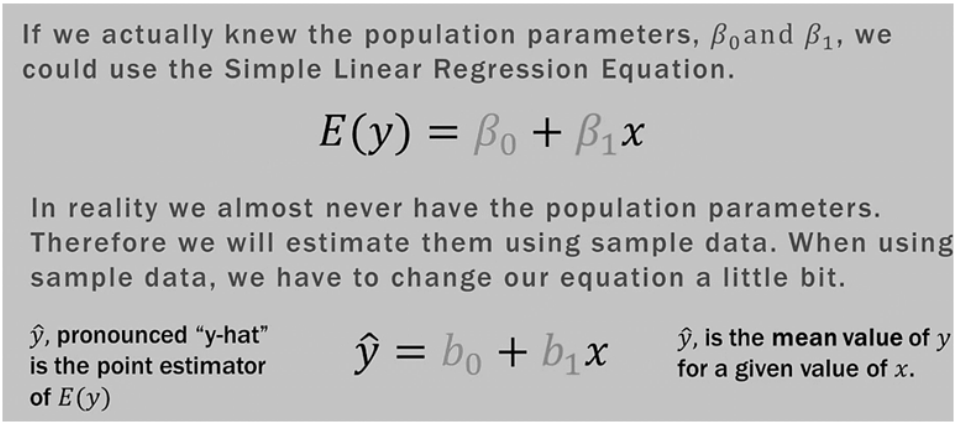
β_0 and β_1 are the regression parameters (or regression coefficient)

The expected value of simple linear regression is as follows:

$$E(y) = \beta_0 + \beta_1 x$$



So the value of β_1 determines the slope of the relationship between variables.



Check Your Progress – 1 :

1. In simple linear regression there is _____ output variable and _____ input variable.
2. Regression is important technique in _____ analytics.
3. Linear regression implies that the relationship between output variable and input variable is _____.

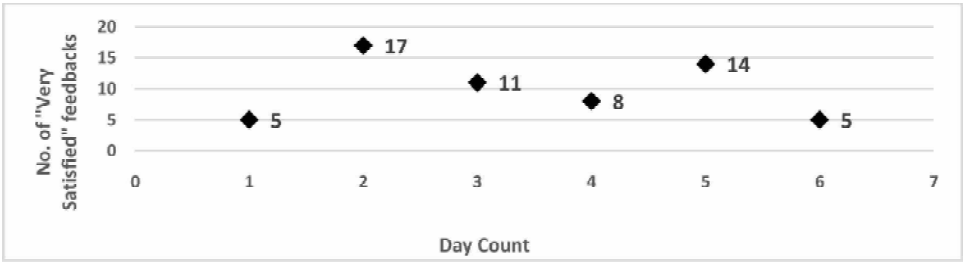
2.3 Baseline Prediction Model :

A restaurant chain collected feedback scores for its customers. They have received a "Very satisfied" rating for six days but suppose they have not counted the total number of feedbacks received. Now data for "Very Satisfied" customers are as below:

Day Count	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Count of Very Satisfied Feedback	5	17	11	8	14	5

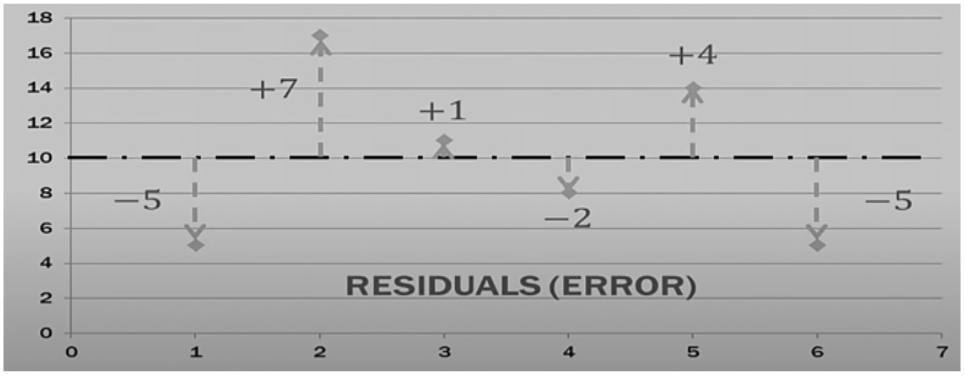
Here we do not have the input variable (total number of feedbacks). Now can we predict the count of "Very Satisfied" feedback on the 7th day?

Let's visualize our data with the help of a simple line graph.



So one way to predict the "Very Satisfied" feedback score is considering the average of these six days feedback count. The average of six days of "Very Satisfied" feedback is 10. So for one variable best prediction can be average.

The next step is to validate how good is this prediction as we can see, all feedback counts are either more than 10 or less than 10. Let's see the difference of each day count from this predicted count of 10 feedbacks. This difference is an error; in statistics, it is known as residuals.



So if we sum up all the errors in our prediction model, then it will be zero.

Day #	Residual	Residual²
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

So to get rid out of these negative differences, we can square all these residuals. It helps in two ways:

1. It makes all differences positive
2. It emphasizes larger deviations

If we sum all these squared residuals, then this term is known as the sum of squared errors (SSE), which is 120 in this case.

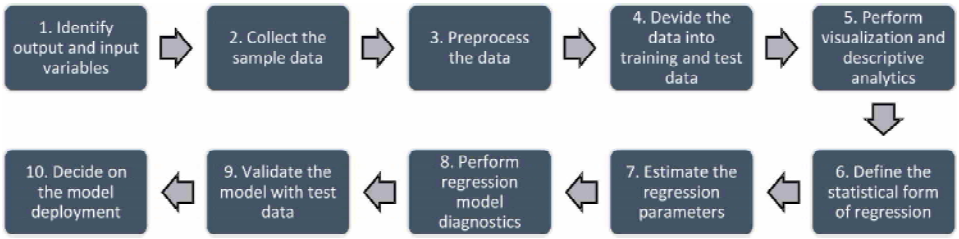
So we have established a basic prediction model and calculate the error in that model. Now the goal of simple linear regression is to create a model that can minimize this sum of squared error (SSE). The regression model will try to minimize this error by introducing an independent variable in the model and draw a best fit line, so that difference between the dependent variable value and line best is minimum; this error explained by the regression model is known as Regression error.

Check Your Progress – 2 :

1. If we have only dependent variable then we can predict the next value by taking _____ of all the observeations.
2. Difference between predicting line and each observation is know as _____.
3. If we add all residuals, then its sum will be very close to _____.

2.4 Simple Linear Regression Model Building :

To understand the statistical relationship between output and input variables, we develop a simple linear regression model. A very high–level framework for linear regression model building, it generally works for the type of industries irrespective of their size.



Regression models can lead to previously unknown relationships, therefore leading to a new hypothesis. Therefore it is a very systematic approach where we see the relationship between important business metrics and important factors that influence those metrics. By control those factors, we can improve the performance of those business metrics.

2.5 Ordinary Least Square Method to Estimate Parameters :

We understood in the earlier section that the overall objective of the regression model is the minimize the residuals by estimating the best fit–line.

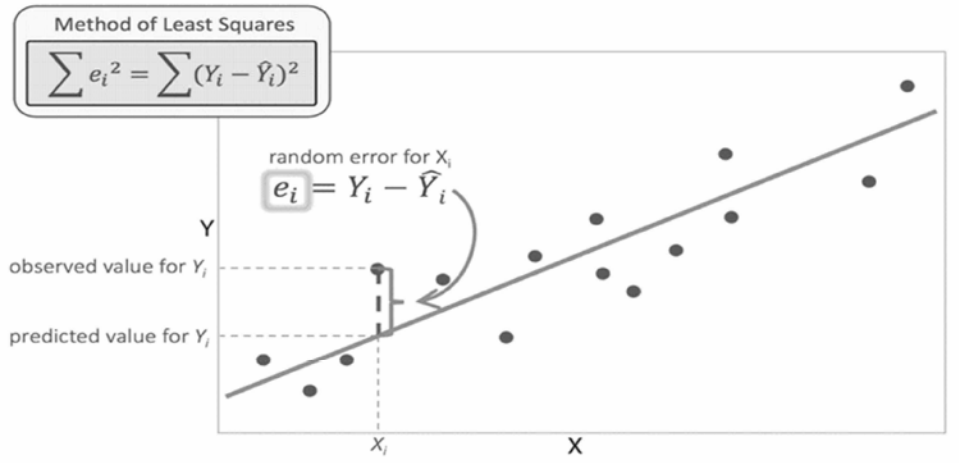
$$\min \sum (y_i - \hat{y}_i)^2$$

Where

Y_i = observed value of dependent variable

$\hat{Y}_i$ = predicted (estimated) value of the dependent variable

Difference between Yi and $\hat{Y}_i$ is residual. There can be infinite lines that can go through these data points; our task is to find the best fit which yields a minimum sum of residuals squares.



2.5.1 Calculation of Regression Parameters :

It is relatively easy to calculate regression parameters if we can put them all in the form of a table.

Intercept $\hat{y}_i = b_0 + b_1 x_i$ Slope

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$\bar{x}$ = mean of the independent variable x_i = value of independent variable

$\bar{y}$ = mean of the dependent variable y_i = value of dependent variable

	Total Feedback	# "Very Satisfied" Feedback	Total Feedback Deviation	"Very Satisfied" Feedback Deviation	Deviation Product	Total Feedback Deviation Squared
	x	y	$x_i - \dot{x}$	$y_i - \bar{y}$	$(x_i - \dot{x})(y_i - \bar{y})$	$(x_i - \dot{x})^2$
1	34	5	-40	-5	200	1600
2	108	17	34	7	238	1156
3	64	11	-10	1	-10	100
4	88	8	14	-2	-28	196
5	99	14	25	4	100	625
6	51	5	-23	-5	115	529
	$\dot{x} = 74$	$\bar{y} = 10$			$\sum = 615$	$\sum = 4206$

$$b_1 = \frac{615}{4200} = 0.1462 \rightarrow \text{slope of line}$$

$$b_0 = \bar{y} - b_1 \times \bar{x} = 10 - 0.1462 \times 74 = -0.8188 \rightarrow \text{y-intercept}$$

so the regression equation is as below:

$$\hat{y}_i = 0.1462x - 0.8188$$

Please Note : Best–fit line always cross from the centroid (intersection point of the mean of x and mean of y)

2.5.2 Interpretation of Regression Equation :

$$\hat{y}_i = 0.1462x - 0.8188$$

The above regression equation tells us that for every new addition in feedback, there will be a .1462 addition to the "Very satisfied" category. In other words, for any new feedback, there are 15% chances of having it in the "Very Satisfied" category. If there is zero feedback, then there are –.8188 "Very satisfied" feedback, which does not make any sense; therefore, intercept may or may not make sense in terms of its business value. So the important point is how the dependent variable changes in relation to one unit change in the independent variable.

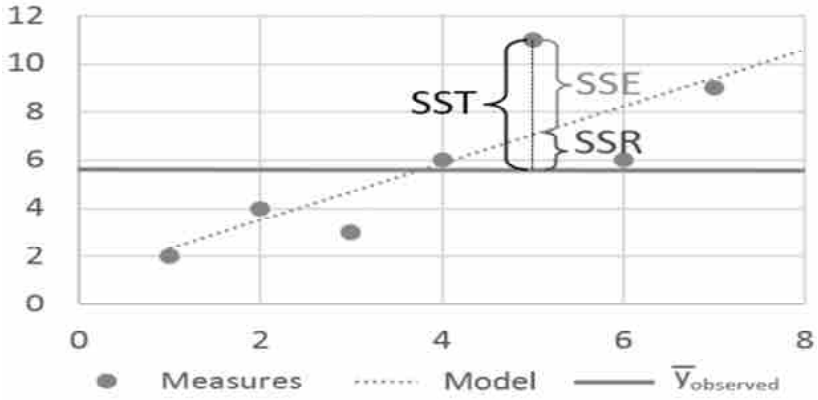
In this way, the regression model tells us the relationship between important business metrics and factors that influence that metric. An important point to remember is that the regression model can only predict the output variable only in the range of input variables supplied. For example, in the above example, input variable rage from 34 to 108; hence technically, we can predict the number of "Very Satisfied" feedback only if total feedbacks is in the range of 34 to 108. If the total number of feedbacks is significantly high or lower than this range, then the regression model may give us misleading predictions.

2.6 Measures of Variation :

Still, one question which is unanswered yet that how good is this regression model? We saw in the earlier section that in the case of baseline model total error, SSE, which was 120. In that case, it was also a total sum of error (SST). Similarly, we have to measure in case of the regression model where instead of the average line, we have the best–fit line. Again we will calculate the distance of each observation from this best–fit line, and we will square those distances to get rid out of negative numbers and also to emphasize large deviations.

Total variation can be segregated into two parts: variation due to relationship between x and y variable, it is referred to as regression sum of squares (SSR), other is due to unexplained variation due to factors which are not part of the regression model and is referred to as error of sum of squares (SSE).

Total sum of squares (SST) = Regression sum of squares (SSR) + Error sum of squares (SSE)



- Total sum of squares (SST) is the difference between the mean of y and each observed value (y_i)

$$\textbf{Total Sum of Squares = SST = } \Sigma(y_i - \bar{y})^2$$

- The regression sum of squares (SSR) is the difference between the mean of y and each predicted value ($\hat{y}_i$)

$$\textbf{Regression Sum of Squares = SSR = } \Sigma(\hat{y}_i - \bar{y})^2$$

- Error sum of squares (SSE) is the difference between each observed value of y and each predicted value ($\hat{y}_i$)

$$\textbf{Error Sum of Squares = SSE = } \Sigma(y_i - \hat{y}_i)^2$$

We can calculate $\hat{y}_i$ value by substituting the value of each observation in the above regression equation as follows:

$\hat{y}_i = 0.1462x - 0.8188$ $\hat{y}_i$

	x	y		
1	34	5	$\hat{y}_i = 0.1462(34) - 0.8188$	4.1505
2	108	17	$\hat{y}_i = 0.1462(108) - 0.8188$	14.9693
3	64	11	$\hat{y}_i = 0.1462(64) - 0.8188$	8.5365
4	88	8	$\hat{y}_i = 0.1462(88) - 0.8188$	12.0453
5	99	14	$\hat{y}_i = 0.1462(99) - 0.8188$	13.6535
6	51	5	$\hat{y}_i = 0.1462(51) - 0.8188$	6.6359
	$\bar{x} = 74$	$\bar{y} = 10$	Observed vs. Predicted	

We can calculate SSE by substituting their corresponding values in the above equations

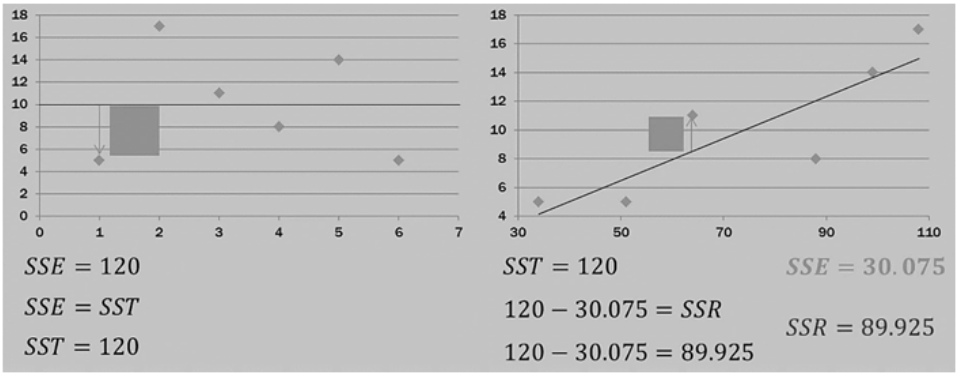
	Total Feedback	# "Very Satisfied" Feedback	$\hat{y}_i$	Error $(y-\hat{y}_i)$	Squared Error $(y-\hat{y}_i)^2$
	x	y			
1	34	5	4.1505	0.8495	0.7217
2	108	17	14.9693	2.0307	4.1237
3	64	11	8.5365	2.4635	6.0688
4	88	8	12.0453	−4.0453	16.3645
5	99	14	13.6535	0.3465	0.1201
6	51	5	6.6359	−1.6359	2.6762
	$\bar{x} = 74$	$\bar{y} = 10$		SSE= $\sum$	= 30.075

So when we conducted regression, the SSE decreased from 120 to 30.075. That is, 30.075 of the sum of squares was unexplained by the regression model or allocated to error.

The 89.925 is the sum of squares error due to regression.

2.6.1 Comparison of Two Models :

Below is the performance of two models, the first one without regression (with the help of dependent variable only) and the second one by applying regression calculations.



2.6.2 Coefficient of Determination :

Coefficient of determination measures that how well the regression equation fits our data

We saw in the earlier section that SST further splits into SSE and SSR. The coefficient quantifies this ratio as a percentage.

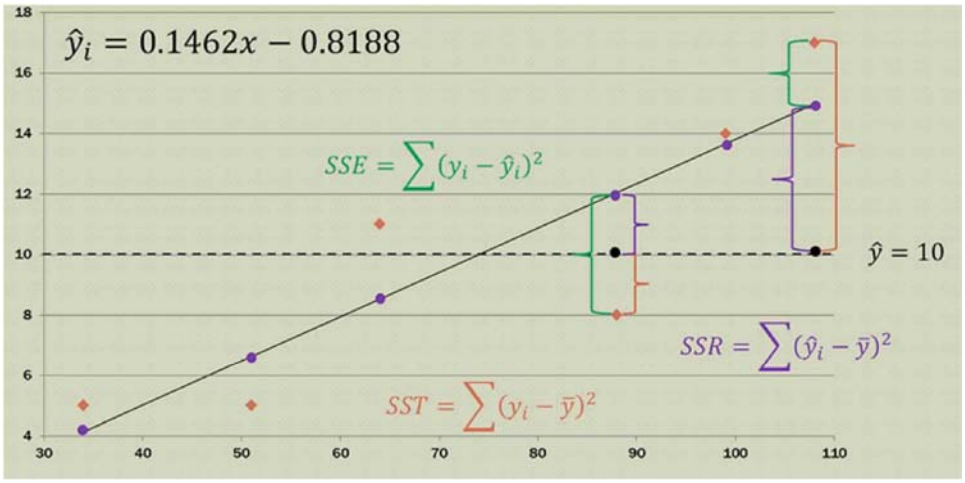
$$\text{Coefficient of determination} = r^2 = \frac{SSR}{SST}$$

Coefficient of determination for the above feedback example:

$$\text{Coefficient of determination} = r^2 = \frac{89.925}{120} = 0.7493 \text{ or } 74.93\%$$

We can conclude that 74.93% of the total sum of squares can be explained by using the estimated regression equation to predict the "Very Satisfied" feedback. The remainder is an error. For simple linear regression, r^2 is the square of correlation coefficient r. In the above survey feedback

example, the correlation coefficient is .866; if we square it, then it will be equal to the coefficient of determination r^2, which we calculated through the sum of squares method.



2.6.3 Mean Square Error and Root Mean Square Error (Standard Error) :

MSE, s^2 is an estimate of σ^2 the variance of the error, ε. In other words, how spread out the data points are from the regression line. MSE is SSE divided by its degree of freedom; in the above example, it will be n–2 as we lost two degrees of freedom as we estimated slope and intercept. In simple linear regression, it is always n–2, while in the case of multiple linear regression, it will be different; we will study that in the next unit.

$$\text{MSE} = s^2 = \frac{\text{SSE}}{n - 2}$$

This is the reason MSE is not a simple average of residuals. But in case we are using a population instead of a sample then we will use n in the denominator, which is a simple average of residuals.

The standard error of the estimates σ (also known as the standard error) is the standard deviation of the error term, ε.

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n - 2}}$$

If we calculate standard error for the above feedback example:

$$\mathbf{S = \sqrt{7.5187} = 2.742}$$

So the average distance of the data points from the fitted line is about 2.74 feedbacks. We can think of s as a measure of how well the regression model makes the predictions.

Check Your Progress – 3 :

1. A total sum of squares is the summation of sum of squares regression and _____.
2. Cofficient of determination is the ratio of _____ and _____.
3. If SSE is 100 and there are 27 observations in the sample then standard error would be _____.

2.7 Simple Linear Regression in MS Excel :

Microsoft Excel provides an option to conduct simple linear regression,

- In MS Excel, under the "Data" tab → "Data Analysis" package (if it is not visible then go to File → Options → Activate data analysis package
- Select the input and output data range
- Select all graphs options available towards the lower side of the pop–up window

Detailed videos "How to perform a regression analysis in MS Excel" are available on Microsoft's website.

Below is MS Excel output for regression analysis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.866
R Square	0.749
Adjusted R Square	0.687
Standard Error	2.742
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	89.9251	89.9251	11.9602	0.0259
Residual	4	30.0749	7.5187		
Total	5	120			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.820	3.323	-0.247	0.817	-10.046	8.406	-10.046	8.406
Total feedbacks	0.146	0.042	3.458	0.026	0.029	0.264	0.029	0.264

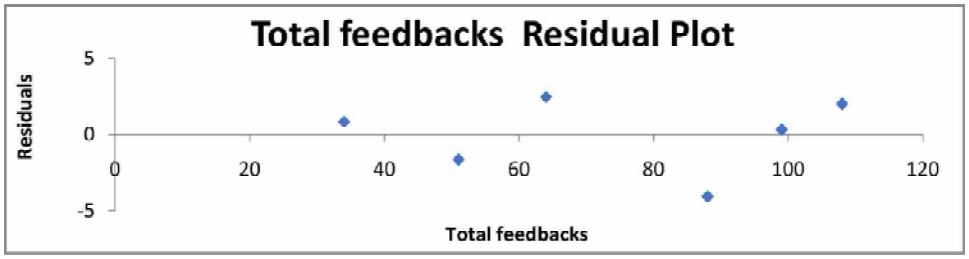
Below is the explanation for the essential components of the regression output:

1. The correlation coefficient for both the variables
2. The coefficient of determination (R^2) $\rightarrow \frac{SSR}{SST} \rightarrow .749 \rightarrow 74.9\%$ variation in output variable due to input variable
3. The standard error $\rightarrow \sqrt{\left(\frac{SSE}{n-2}\right)} \rightarrow 2.742$
4. Regression sum of squares (SSR) $\rightarrow$ 89.925
5. Error sum of squares (SSE) $\rightarrow$ 30.749
6. Total sum of squares (SST) $\rightarrow$ 120
7. The overall significance of the model, if the value is less than .05 at 95% confidence level, then the overall model showing a significant relationship between input and output variables
8. y–intercept $\rightarrow$ –0.820
9. Slope $\rightarrow$ 0.146
10. The P–value for each variable, if it is less than .05 at 95% CL, then the input variable is significant in the model

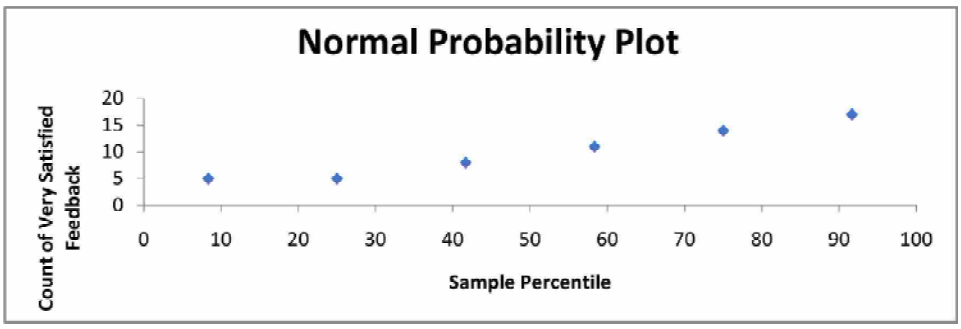
2.7.1 Residual Analysis to Test The Regression Assumptions :

Residual analysis is mainly used to test the assumptions of the regression model. Below two are important residual analyses provided by MS Excel:

1. **The Linearity of the Regression Model :** The linearity of the regression model can be obtained by plotting the residual on the vertical axis against the corresponding xi values of the independent variable on the horizontal axis. There should not be any apparent pattern in the plot for a fit regression model.



2. **Normality of Error :** The normality probability plot of the residuals should roughly follow a straight line for meeting the assumption of normality. A straight line correcting all the residuals indicates that the residuals are normally distributed. It means that residuals are random in nature, and there is not any hidden trend. A curve in the tail is an indication of skewness.



There are some more advanced assumption validations, which are not part of MS Excel like **Constant error variance (Homoscedasticity),** which means constant error variance. **Independence of Error,** which tells us that there is no relationship between error and independent variable; this effect is known as autocorrelation.

Check Your Progress – 4 :

1. The _____ is to mean as the standard error is to the regression line.
2. If SSR is 90 and SST is 130, coefficient of determination will be _____.
3. In a regression output if P–value is 0.029 at 95% confidence level, we can say that input variable is _____ for the study.

Check Your Progress – 5 :

1. In a simple linear regression, if the input variable will change by one unit, how much the output variable will be changed (impacted)

   a. By intercept b. By its slope c. No change d. By one unit
2. In a regression equation, $Y = \beta_0 + \beta_1 X + \varepsilon$. β_0 and β_1 refer to:

   a. Intercept and slope b. Error and Intercept

   c. Slope and Error d. Intercept and Error
3. In simple linear regression, how many coefficients do we need to estimate

   a. 1 b. 2 c. 0 d. Can't say
4. To evaluate a regression model, which of the following metrics can be used:

   a. R squared b. Standard Error

   c. Residual analysis d. All of the above
5. In simple linear regression, what does the slope represents

   a. The value of Y when the input variable is zero

   b. The estimated change in the input variable per unit change in the output variable

   c. The estimated change in output variable per unit change in the input variable

   d. Variation around the best–fit line

6. How can we explain the standard error
   a. It is defined as a variation around the regression line
   b. Total error in regression model divided by its standard deviation
   c. Total error in the regression model
   d. Total variation in output and input variable together
7. How can we define the coefficient of determination in simple linear regression
   a. SSR/SST
   b. Square of the coefficient of correlation
   c. The proportion of variation in the output variable explained by the input variable
   d. All of the above
8. How can we define residual in a regression model
   a. Difference between the variation of input and output variable
   b. Difference between the predicted value and observed value
   c. Difference between the input and output variable
   d. All of the above
9. Which of the following statements are correct
   a. Residual in a regression model must be distributed normally
   b. Residuals must have constant variance
   c. There should not be any significant relation between the input variable and residuals
   d. All of the above
10. Which of the below option is correct about the below two statements:

    Statement 1 : Constant variance among residuals is known as homoscedasticity

    Statement 2 : Relationship between the input variable and residuals is known as autocorrelation

    a. Only 1 is correct b. Only 2 is correct
    c. Both are correct d. Both are incorrect

2.8 Let Us Sum Up :

- Regression is an essential technique in the business world to predict business metrics
- Simple linear regression helps to understand the relationship between input and output variables and also provide mathematical equation so that by supplying a value of input variable, we can estimate the value of the output variable
- Regression parameters can be estimated using the method of ordinary least squares under the assumption that the residuals follow a normal

distribution. The method of ordinary least squares gives the best linear unbiased estimate.

- The efficacy of the regression model can be validated by measures like R^2, standard error, significance F value, etc.
- Residual analysis is performed to check whether the model satisfies assumptions such as normality of residuals, homoscedasticity, autocorrelation, etc.

2.9 Answers for Check Your Progress :

Check Your Progress – 1 :

1. one, one **2.** predictive **3.** linear

Check Your Progress – 2 :

1. mean **2.** residuals **3.** zero

Check Your Progress – 3 :

1. Sum of squares error (SSE)

2. SSR, SST **3.** 2

Check Your Progress – 4 :

1. Standard deviation **2.** 0.6923 **3.** significant

Check Your Progress – 5 :

1. b **2.** a **3.** b **4.** d **5.** c

6. a **7.** d **8.** b **9.** d **10.** c

2.10 Glossary :

Simple Linear Regression : It is an important statistical technique for finding the linear (straight line) relationship between an input and output variable. Regression gives us a mathematical expression by which we can predict the value of the output variable by supplying a value of the input variable.

Total Sum of Squares (SST) : Total sum of squares in a model can be measured as the summation of total variation explained by the regression model (SSR) and unexplained variation due to random error (SSE)

Coefficient of Determination (R^2) : It is a measure of fit of the regression model. It is a ratio of SSR to SST. The value of the coefficient of determination varies between 0 and 1.

Standard Error : Standard error can be understood as a standard deviation around the regression line. A large standard error indicates a large amount of variation or scatters around the regression line

Residual Analysis of Regression Model : Residual or error analysis is a vital step to check whether the assumptions of regression models have been satisfied

Homoscedasticity : The assumption of homoscedasticity means constant error variance across the range of input variable

Autocorrelation : The input variable must be independent of error terms. The Durbin–Watson statistic can measure the autocorrelation effect.

2.11 Assignment :

1. What is residual in simple linear regression.
2. What do we mean by the baseline prediction model in simple linear regression ?
3. Write down important steps in simple linear regression model calculation.
4. In simple linear regression equation, $Y_i = 23x - 7$. Explain the relationship between the x and y variable.

2.12 Activities :

Company Nirja Plc has produced its expenses (input variable) and revenue (output variable) for the last 10 years:

Year	Expenses (In million $)	Revenue (In million $)
2009	3506	4108
2010	3518	4190
2011	3609	4278
2012	3689	4295
2013	3712	4307
2014	3754	4389
2015	3798	4401
2016	3803	4456
2017	3874	4497
2018	3967	4501
2019	3994	4508

a. Make a scatter plot and check whether it is showing the linear relationship, and are there a few outliers ?

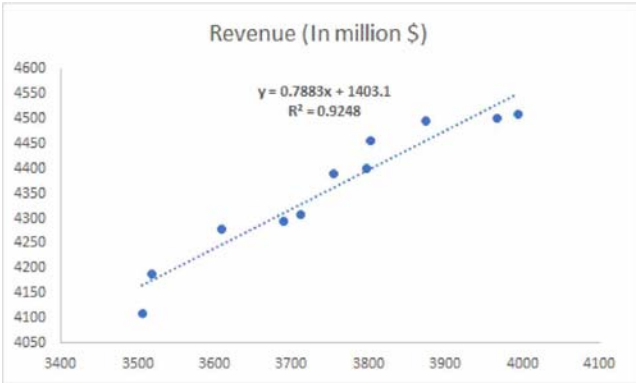
b. Calculate the Pearson's correlation coefficient and coefficient of determination (r^2)

c. Calculate the regression equation

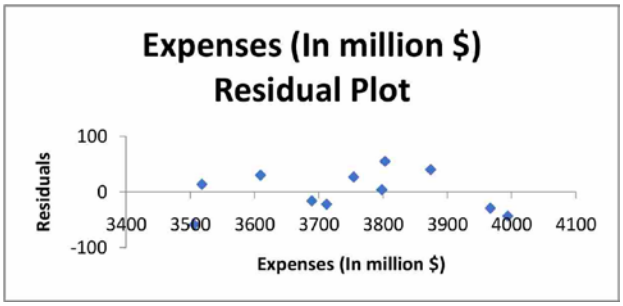
d. Calculate standard error, the total sum of squares, the sum of square error, and a regression sum of squares

e. Validate the regression equation through residual analysis

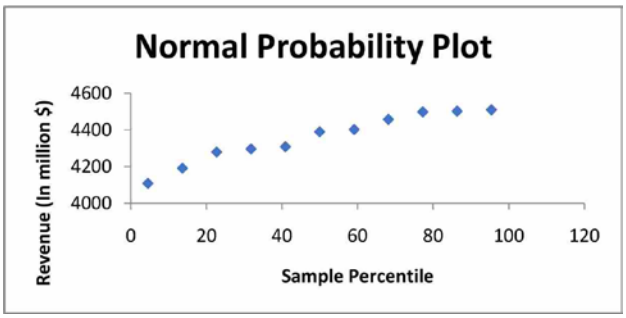
- Yes, the scatter plot is showing a linear relationship. There are not any outliers visible in the scatter plot



- Pearson's correlation coefficient = .9617 and coefficient of determination (r^2) = 0.9248
- Revenue = 1403.10 + 0.788 × expenses
- Standard error = 38.60, SST = 178372.18, SSE = 13412.05, SSR = 164960.13



The residual plot is not showing any visible trend



Errors are in a straight line with a slight curve, need to understand the reason for the curve in the residual plot

2.13 Case Study :

ABC limited company has appointed a new analytics and insight department head to strengthen its presence in India and the Asian market. He wants to analyze sales and advertisement costs to examine whether different advertising channels are providing enough sales or not. He asked the MIS team to provide him with sales data and marketing costs in the last 10 years:

Year	Sales (in a million rupees)	Marketing Cost (in a million rupees)
2009	2064	31
2010	2389	36
2011	2418	37
2012	2509	35
2013	2608	38
2014	2706	41
2015	2802	44
2016	2905	45
2017	3056	47
2018	3189	49

Answer the below Questions :

a. Make a scatter plot and check whether it is showing the linear relationship, and are there a few outliers ?

b. Calculate the Pearson's correlation coefficient and coefficient of determination (r^2)

c. Calculate the regression equation

d. Calculate standard error, the total sum of squares, the sum of square error, and a regression sum of squares

e. Validate the regression equation through residual analysis

2.14 Further Readings :

- Ken Black, "Business Statistics: Contemporary Decision Making", John Wiley and Sons (2009)
- U Dinesh Kumar, "Business Analytics: The Science of Data–Driven Decision Making", Wiley (2017)
- Naval Bajpai, "Business Research Methods", Pearson (2013)

Unit 3

MULTIPLE LINEAR REGRESSION

: UNIT STRUCTURE :

3.0 Learning Objectives :

After learning this unit, you will be able to understand :

- Understand the concept of multiple linear regression and its mathematical interpretation
- Understand the concept of coefficient of multiple determination and adjusted coefficient of multiple determination
- Interpretation of multiple regression parameters and the concept of partial regression coefficients
- Understand the effect of multicollinearity and auto–correlation on the regression model
- Residual analysis for multiple regression model
- Running multiple regression on MS Excel and interpretation of the model output

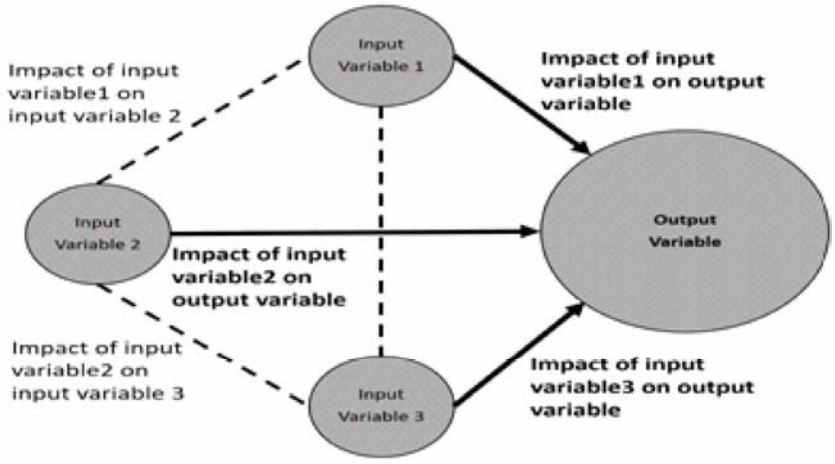
3.1 Introduction :

In this unit, we will study multiple linear regression where we see the relationship between one output variable and multiple input variables. It resembles business problems that we encounter often in the corporate

world. We will see various techniques to validate the multiple linear regression. In the end, we will see various examples to see the effectiveness of multiple linear and its influence in decision making.

3.2 Essence of Multiple Linear Regression :

Multiple linear regression is an important statistical technique for finding the relationship between one dependent variable and multiple independent variables. In business scenarios generally, any performance metric depends on various factors; hence multiple regression is used most often than simple linear regression. Besides finding the relationship, it also helps us to find the most important factors influencing the business metric so that instead of controlling various factors, we can concentrate on vital few factors. For example, the percentage salary hike of an employee may depend on his/her performance rating, working hours, office environment, technology, salary, career growth options etc. or mileage of a car can be dependent on the age of the car, engine power, fuel type, engine technology, city/countryside driving, driver proficiency etc. For multiple regression model output variable should be continuous (ratio or interval) and independent variables can be either continuous or discrete (nominal or ordinal scale).



Ideally, in the regression model, there should not be any relationship between independent variables because if there is a relationship between independent variables, then we can not calculate their actual impact on the dependent variable. When dependent variables show a relationship, then we call it multicollinearity. We will discuss the ways to tackle multicollinearity in the next section.

On the other hand, too many independent variables explain the higher variance of the dependent variable than actually, it holds but also increases the risk of over–fitting of the regression model. This means the model will work very nicely on training data but will not work fine on the test and unknown data. We will discuss the over–fitting problem in detail in the next section. Here point to be noted is that having more variables will not solve the purpose instead few variables with a significant relationship with dependent variables will constitute a better regression model.

3.2.1 Introduction to Multiple Linear Regression :

In simple linear regression, we discussed the probabilistic regression model as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In the case of multiple regression as we have more than one independent variable. Hence above probabilistic model can be extended to **multiple probabilistic regression models** as:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + ... + \beta_k X_k + \varepsilon_i$$

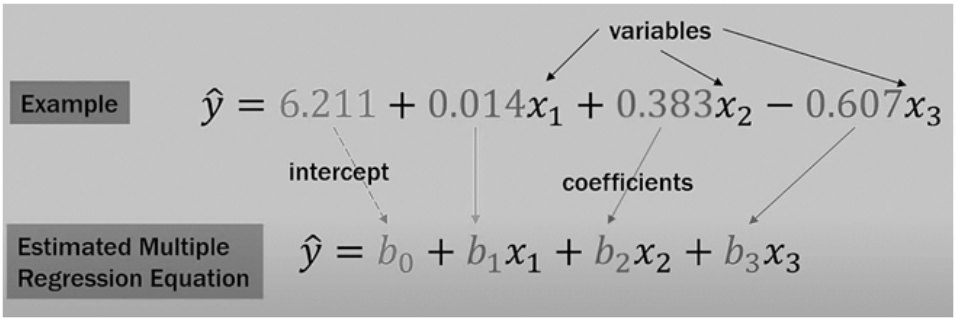
Where, Y_i is the i^{th} value of the dependent variable, β_0 the y–intercept, β_1 the slope of y with input variable X_1 by keeping X_2, X_3, X_4……, X_k constant. Similarly, β_k the slope of y with input variable X_k by keeping X_1, X_2, X_3……, X_{k-1} constant. ε_i is a random error in y for ith observation.

In multiple regression analysis, β_i is the slope of y with independent variable x_i keeping all other independent variables constant. This is also referred to as the partial regression coefficient for the independent variable x_i.

To predict the value of y, we have to calculate the value of β_0, β_1, β_2,…… β_k. But this can be done only if we have access to the entire population, which is not possible most of the time in the business world. Hence we use sample data to predict the value of these regression coefficients. So the equation for estimative y with the sample information is given as:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_k x_k$$

For example,



So in the above example if x_1 will increase by one unit, then y will increase by .014 units provided x_2 and x_3 are constant.

Check Your Progress – 1 :

1. Multiple linear regression has_____ independent variables and _____ dependent variable.
2. If we have large number of independent variables then it may lead to the _____ problem.
3. If there is a significant relationship between two independent variables, it is known as _____.

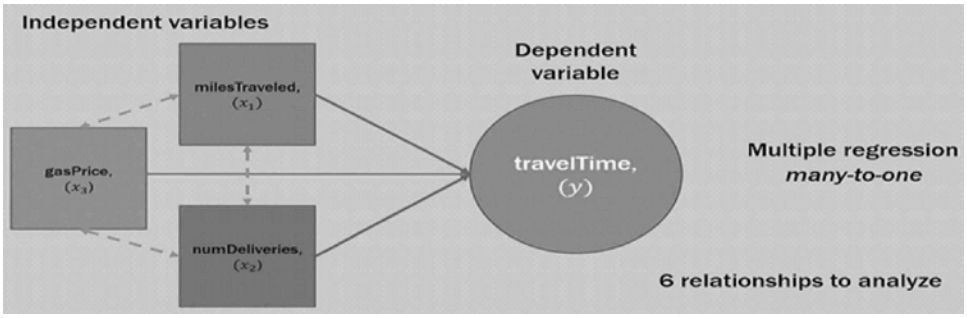
3.3 Understanding the Concept of Multiple Linear Regression with a Worked Example :

Let's consider an example of a supply chain of an online retail company; their logistic arm tries to deliver the packages on the same day. They try to optimize their delivery associate's trips by strategizing trips with the help of city maps to reduce time and fuel costs. Below is the data for 10 random trips, each trip has three important pieces of information – Total miles travelled, number of deliveries, daily gas price and a total time of travel in hours.

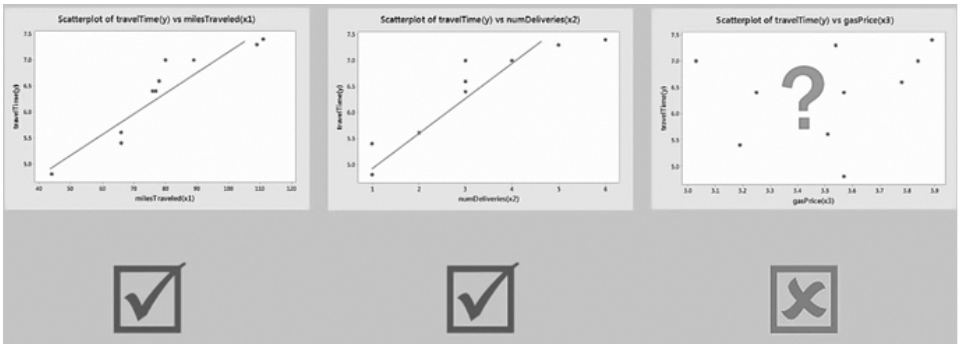
As an analyst, you want to estimate how long delivery will take (dependent variable) based on two inputs – total distance, number of deliveries and daily gas (fuel) price (independent variables).

MileTraveled	numDeliveries	gasPrice	travelTime(hrs)
89	4	3.84	7
66	1	3.19	5.4
78	3	3.78	6.6
111	6	3.89	7.4
44	1	3.57	4.8
77	3	3.57	6.4
80	3	3.03	7
66	2	3.51	5.6
109	5	3.54	7.3
76	3	3.25	6.4

Below is the pictorial view of the relationship between the dependent and independent variables.



Before starting calculations, it is always recommended to use the scatter plot to visualize the relationship of every independent variable with the dependent variable separately.

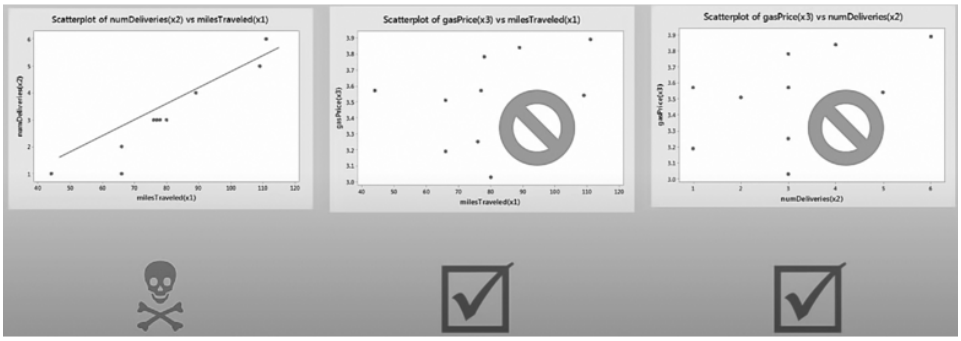


Here we saw miles travelled and the number of deliveries showing a significant relationship while gas price does not show any relationship with the dependent variable (travel time).

As gas price does not appear to be correlated with travel time, we should not use it in a multiple regression model.

But we are keeping it for the learning process so that we can see the consequences of having a non–significant variable in the model.

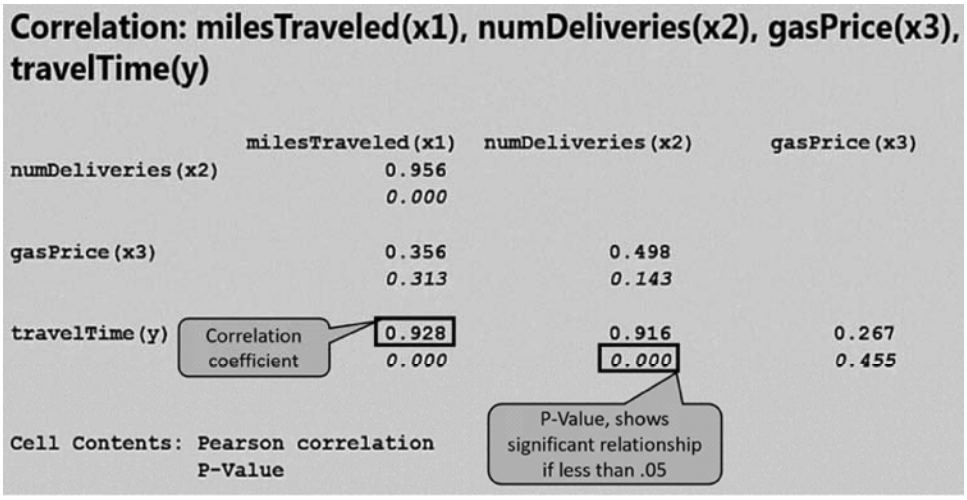
Now we should also see the relationship among independent variables.



So miles travelled, and the number of deliveries is showing a strong relationship; hence it is a problem (multicollinearity). Ideally, we should not use BOTH in the regression model; instead, we should use either of them, which makes more sense from the business point of view.

But for learning purposes, let's keep both in the regression model. We will remove it in the later stage of the model building.

It is recommended that we should also see the correlation coefficient between these variables in case it is not visible in the graph then we can rely on the numerical coefficient better.



So correlation coefficients tell us clearly, which variables have to include in the model and which one has to remove.

Check Your Progress – 2 :

1. In multiple linear regression, there is _____ relationship between dependent and independent variables
2. For a good multiple linear regression Scatter plot should show_____ relationship between each individual variable and dependent variable.
3. If an independent variable is found to be highly significant in a regression model, we can conclude that the changes in _____ variable are associted to the changes in _____ variable.

3.4 The Correlation Coefficient for Multiple Linear Regression :

In the case of simple linear regression, we calculated the relationship between one independent and one dependent variable. In the case of multiple linear regression, we see the relationship between the dependent variable and the cumulative impact of all independent variables. Here we see the relationship between the dependent variable (y) and predicted variable ($\hat{y}_i$). Let's say below are the coefficient of the regression model.

So it is a correlation between column D (y) and column F ($\hat{y}_i$) :

A	B	C	D	E	F
Mile Traveled	num Deliveries	gas Price	Time travel	travelTime_Predicted	travelTime Predicted
89	4	3.84	7	6.21+0.01*A3+0.38*B3–0.61*C3	6.67
66	1	3.19	5.4	6.21+0.01*A4+0.38*B4–0.61*C4	5.59
78	3	3.78	6.6	6.21+0.01*A5+0.38*B5–0.61*C5	6.17
111	6	3.89	7.4	6.21+0.01*A6+0.38*B6–0.61*C6	7.72
44	1	3.57	4.8	6.21+0.01*A7+0.38*B7–0.61*C7	5.05
77	3	3.57	6.4	6.21+0.01*A8+0.38*B8–0.61*C8	6.28
80	3	3.03	7	6.21+0.01*A9+0.38*B9–0.61*C9	6.65
66	2	3.51	5.6	6.21+0.01*A10+0.38*B10–0.61*C10	5.78
109	5	3.54	7.3	6.21+0.01*A11+0.38*B11–0.61*C11	7.52
76	3	3.25	6.4	6.21+0.01*A12+0.38*B12–0.61*C12	6.46

Correlation Coefficient	travelTime(hrs) (y)	travelTime Predicted
travelTime(hrs) (y)	1	
travelTime Predicted	0.9459	1

Therefore correlation coefficient between independent variables and the dependent variable is .9459. If we calculate the square of it, then it will become a coefficient of determination (R^2), which is 0.8947 in the above example.

3.5 Coefficient of Coefficient (R^2), Adjusted R^2, and Standard Error :

In the last unit, we calculated SSR, SSE and SST for simple linear regression. Calculations remain the same in the case of multiple linear regression also:

- Total sum of squares (SST) is the difference between the mean of y and each observed value (yi)

$$\textbf{Total Sum of Squares = SST = } \Sigma(y_i - \bar{y})^2$$

- The regression sum of squares (SSR) is the difference between the mean of y and each predicted value ($\hat{y}_i$)

$$\textbf{Regression Sum of Squares = SSR = } \Sigma(\hat{y}_i - \bar{y})^2$$

- Error sum of squares (SSE) is the difference between each observed value of y and each predicted value ($\hat{y}_i$)

$$\textbf{Error Sum of Squares = SSE = } \Sigma(y_i - \hat{y}_i)^2$$

Mile Traveled	num Deliveries	gas Price	travel Time (y)	$(y_i - \bar{y})$	$(yi - \bar{y})^2$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
89	4	3.84	7	0.61	0.3721	0.2817	0.0793	0.3283	0.1078
66	1	3.19	5.4	-0.99	0.9801	-0.7983	0.6373	-0.1917	0.0367
78	3	3.78	6.6	0.21	0.0441	-0.2204	0.0486	0.4304	0.1853
111	6	3.89	7.4	1.01	1.0201	1.3283	1.7644	-0.3183	0.1013
44	1	3.57	4.8	-1.59	2.5281	-1.3395	1.7943	-0.2505	0.0627
77	3	3.57	6.4	0.01	0.0001	-0.1072	0.0115	0.1172	0.0137
80	3	3.03	7	0.61	0.3721	0.2627	0.0690	0.3473	0.1206
66	2	3.51	5.6	-0.79	0.6241	-0.6093	0.3712	-0.1807	0.0327
109	5	3.54	7.3	0.91	0.8281	1.1292	1.2751	-0.2192	0.0481
76	3	3.25	6.4	0.01	0.0001	0.0728	0.0053	-0.0628	0.0039
	Average		**6.39**	**Sum**	**6.769**		**6.0561**		**0.7129**

- **Total Sum of Squares = SST = $\Sigma(y_i - \bar{y})^2 = 6.769$**
- **Regression Sum of Squares = SSR = $\Sigma(\hat{y}_i - \bar{y})^2 = 6.0561$**
- **Error Sum of Squares = SSE = $\Sigma(y_i - \hat{y}_i)^2 = 0.7129$**

In the case of simple linear regression, R^2 explains the total variation in output (dependent) variable explained by an independent

variable. In the case of multiple linear regression, R2 is the variation in the dependent variable explained by the combination of independent variables.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{6.0561}{6.769} = 1 - \frac{.7129}{6.769} = 0.8947$$

In the last section, we calculated the multiple correlation coefficient = .9459. Therefore if we square this multiple correlation coefficient, then it is 0.8947.

Adding any new variable in the regression model does not change the total sum of squares (SST) while it is likely to increase the SSR. Sometimes despite the newly added variable not making the regression model better but still it increases the R^2 value, hence considering the R^2 value along as efficacy of the regression model is wrong.

This problem can be tackled by another measure called adjusted R^2; it considers additional information added by new variable(s) and changed in the degree of freedom.

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

$$= 1 - \frac{.7129/(10-3-1)}{6.769/(10-1)} = 0.8420$$

Where n is the sample size taken and k is the number of independent variables.

If more insignificant variables are added to the regression model, the difference between R^2 and adjusted R^2 will be larger. In the case of adjusted R^2 total variation in the dependent variable is adjusted by the sample size and number of independent variables.

The standard error is the standard deviation of residuals around the regression line. In the case of multiple linear regression, there is a slightly different formula for standard error which is adjusted by sample size (number of data points and number of independent variables in the model) and number of independent variables.

$$\text{Standard error} = \sqrt{SSE/(n-k-1)}$$

$$= \sqrt{.7129/(10-3-1)} = 0.3447$$

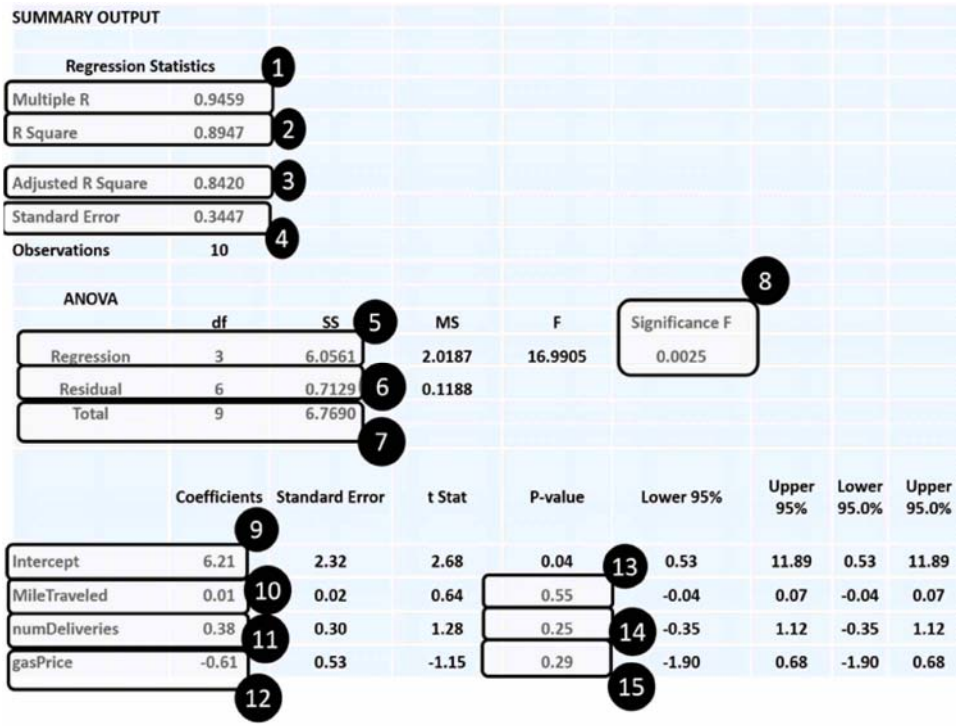
Check Your Progress – 3 :

1. In multiple linear regression, correlation coefficient is calculated between dependent variable and _____.
2. Adjusted R^2 is always _____ than R^2 value.
3. Standard error in multiple linear regression is adjusted by _____ and _____.

3.6 Multiple Linear Regression in MS Excel :

For the same example, estimate how long delivery will take (dependent variable) based on three inputs – total distance, number of deliveries and daily gas (fuel) price (independent variables).

Below is the output of MS Excel for the above example:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9459
R Square	0.8947
Adjusted R Square	0.8420
Standard Error	0.3447
Observations	10

ANOVA

	df	SS	MS	F	Significance F
Regression	3	6.0561	2.0187	16.9905	0.0025
Residual	6	0.7129	0.1188		
Total	9	6.7690			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6.21	2.32	2.68	0.04	0.53	11.89	0.53	11.89
MileTraveled	0.01	0.02	0.64	0.55	-0.04	0.07	-0.04	0.07
numDeliveries	0.38	0.30	1.28	0.25	-0.35	1.12	-0.35	1.12
gasPrice	-0.61	0.53	-1.15	0.29	-1.90	0.68	-1.90	0.68

Below is the explanation for the essential components of the regression output:

1. The correlation coefficient for the independent variable and all independent variables, .9459
2. The coefficient of determination $(R^2) \rightarrow \frac{SSR}{SST} \rightarrow .8947$ ' 89.47% variation in output variable due to input variables
3. Adjusted $R^2 \rightarrow 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{.7129/(10-3-1)}{6.769/(10-1)} = 0.8420$
4. The standard error $\rightarrow \sqrt{SSE/(n-k-1)} = \sqrt{.7129/(10-3-1)} = 0.3447$
5. Regression sum of squares (SSR) $\rightarrow$ 6.0561

6. Error sum of squares (SSE) → 0.7129
7. Total sum of squares (SST) → 6.6769
8. The overall significance of the model, if the value is less than .05 at 95% confidence level, then the overall model showing a significant relationship between input and output variables
9. y–intercept → 6.21
10. Cofficient for first independent variable (MileTraveled), $x_1 \rightarrow 0.01$
11. Cofficient for second independent variable (numDeliveries), $x_2 \rightarrow 0.3831$
12. Cofficient for third independent variable (gasPrice), $x_3 \rightarrow -0.606$
13. The P–value for the first independent variable (MileTraveled), is less than .05 at 95% CL, which indicates input variable (MileTraveled) is NOT significant in the model
14. The P–value for the second independent variable (numDeliveries), is less than .05 at 95% CL, which indicates the input variable (MileTraveled) is NOT significant in the model
15. The P–value for the third independent variable (numDeliveries), is less than .05 at 95% CL, which indicates the input variable (gasPrice) is NOT significant in the model

We have observed that the overall model is significant (point number 8) while all three independent variables are not significant. It is a problem due to multicollinearity. Below are two issues with the above model:

1. The first two independent variables are highly correlated
2. The third variable is not correlated with the dependent variable

Check Your Progress – 4 :

1. In multiple linear regression, F statistics tells us about _____ of overall model.
2. At 95% confidence level, p–value of each independent variable should be less than _____ .
3. At 95% confidence level, if overall model has p–value less than .05 but each independent variable has p–value more than .05. In that scenario, there can be an issue of _____ in the regression model.

3.6.1 The Modified Regression Model in Excel :

It is always recommending that we should apply the above remedies to improve the model in a stepwise manner, so let's remove the third variable (gasPrice) from the model and rerun to see how it is performing:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.9335**
R Square	**0.8714**
Adjusted R Square	**0.8347**
Standard Error	**0.3526**
Observations	10.0000

ANOVA

	df	*SS*	*MS*	*F*	***Significance F***
Regression	2	5.899	2.949	23.716	**0.001**
Residual	7	0.870	0.124		
Total	9	6.769			

	Coefficients	*Standard Error*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0*
Intercept	3.732	0.887	4.208	0.004	1.635	5.830	1.635	5.830
Mile Traveled	0.026	0.020	1.310	**0.232**	-0.021	0.074	-0.021	0.074
num Deliveries	0.184	0.251	0.733	**0.487**	-0.409	0.777	-0.409	0.777

Still, we are observing that the overall model is significant, but both variables are non–significant while individually, they are positively correlated with the dependent variable.

So now we have to go with a single linear regression and check the model for each variable in order to determine the best predicting independent variable. The below table shows a comparable study for all significant combinations of independent variables.

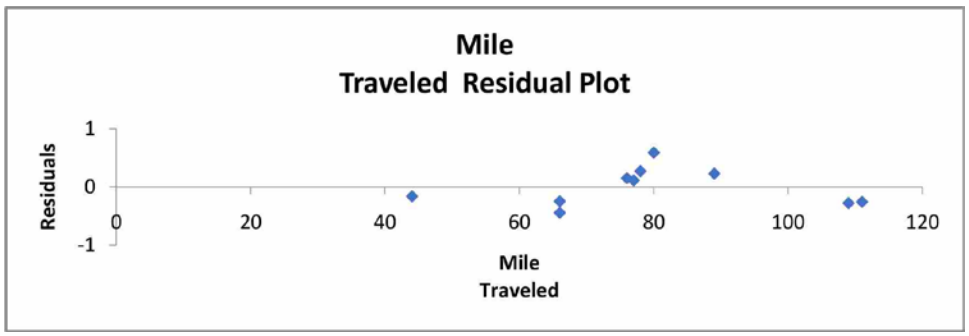
F Value	P-Value	Standard Error	R^2 Adjusted	MileTraveled Significant	numDeliveries Significant	gasPrice Significant	Model Selected
49.77	<.001	.3423	84.42%	v			v v
41.96	<.001	.3680	81.99%		v		v
0.62	.455	.8864	0.00%			X	X
23.72	<.001	.3526	59.95%	X	X		X
22.63	<.001	.3599	68.11%	X		X	X
27.63	<.001	.3297	71.76%		X	X	X
16.99	.002	.3447	57.49%	X	X	X	X

So, in this case, simple linear regression with MileTraveled as an independent variable makes the most significant model. So it is not always necessary that for the same dependent variable, multiple linear regression will make a better model than simple linear regression.

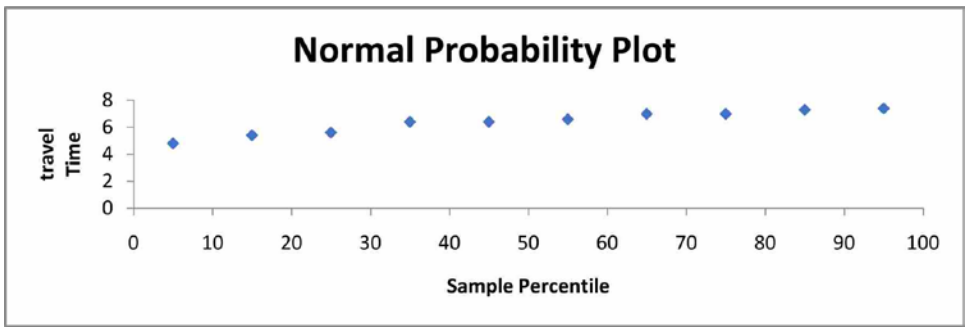
3.6.2 Residual Analysis to Test the Regression Assumptions :

Residual analysis is mainly used to test the assumptions of the regression model. Below two are important residual analyses provided by MS Excel:

1. **The Linearity of the Regression Model :** The linearity of the regression model can be obtained by plotting the residual on the vertical axis against the corresponding xi values of the independent variable on the horizontal axis. There should not be any apparent pattern in the plot for a fit regression model.



2. **Normality of Error :** The normality probability plot of the residuals should roughly follow a straight line for meeting the assumption of normality. A straight line correcting all the residuals indicates that the residuals are normally distributed. It means that residuals are random in nature, and there is not any hidden trend. A curve in the tail is an indication of skewness.



There are some more advanced assumption validations, which are not part of MS Excel like **Constant error variance (Homoscedasticity),** which means constant error variance. **Independence of Error,** which tells us that there is no relationship between error and independent variable; this effect is known as autocorrelation.

Check Your Progress – 5 :

Question 1 – 5 : Use below data to answer the questions

Analytical Score	Hours Study	Final Score
125	30	100
104	40	95
110	25	92
105	20	90
100	20	85
100	20	80
95	15	78
95	10	75
85	0	72
90	5	65

1. The value of R^2 for the above data is:
   a. 0.9515 b. 0.9053 c. 0.8782 d. 3.8751
2. The value of SSR and SST for the above data is:
   a. 1004.484 and 1109.6 b. 105.1158 and 1109.6
   c. 502.2421 and 1004.484 d. 33.4459 and 1109.6
3. Value of Standard error for the above data is:
   a. 0.9515 b. 0.9053 c. 0.8782 d. 3.8751
4. Value of Adjusted R^2 for the above data is:
   a. 0.9515 b. 0.9053 c. 0.8782 d. 3.8751
5. By visualizing scatter plot and regression output, we can conclude that:
   a. Data have multicollinearity issue
   b. Data has overfitting issue
   c. Data has both multicollinearity and overfitting issues
   d. Data has neither multicollinearity nor overfitting issue
6. If there is multicollinearity in the model, it may result in:
   a. The regression coefficient may be in the opposite sign
   b. May add a new variable that is not significant
   c. Removing a significant variable from the data
   d. All of the above
7. If we have added a new variable in the model then:
   Statement 1 : R^2 will always increase
   Statement 2: Adjusted R^2 will always increase
   a. Statement 1 is always true
   b. Statement 2 is always true
   c. a and b both statements are correct
   d. a and b both statements are incorrect
8. In the case of multiple linear regression, the correlation coefficient is:
   a. Correlation between the dependent variable and predicted variable
   b. The square root of the coefficient of determination
   c. a and b both options are correct
   d. a and b both options are incorrect

9. In the below formula of standard error, n and k represent:

Standard error $\rightarrow = \sqrt{SSE/(n-k-1)}$

a. n and k are sample size and number of independent variables respectively

b. n and k are the number of independent variables respectively and sample size respectively

c. n is the degree of freedom and k is the sample size

d. None of the above

10. In the case of two independent variables:

a. Running two separate simple linear regression is always better than multiple linear regression

b. It depends on the relationship of each independent variable and dependent variable

c. Running multiple linear regression would always be better than two separate simple linear regression

d. None of the above

3.7 Let Us Sum Up :

- Multiple linear regression is a statistical extension of simple linear regression where two or more independent variables used to predict variance in one dependent variable.
- In a multiple regression model, each coefficient is interpreted as the estimated change in the dependent variable (y) corresponding to a one–unit change in the independent variable when all other variables are held constant.
- Too many independent variables may cause to overfitting of the regression model as these variables keep increasing coefficient determination R^2.
- The correlation coefficient in the case of multiple linear regression is calculated between the dependent variable and predicted variable.
- R^2 adjusted helps in determining whether the newly added variable is adding value to the model or not; it is a better measure than R^2 alone.
- Multicollinearity happens when some independent variables are correlated with each other.
- In the case of multiple linear regression, R^2 explains the total variation in output (dependent) variable explained by an independent variable while in the case of adjusted R^2, the total variation in the dependent variable is adjusted by the sample size and number of independent variables.
- In multiple linear regression, the overall model significance is calculated by F–test.

- If their a multicollinearity issue in the model then there are chances to remove the significant independent variable.
- Residual normality plot should show a straight line while there should not be any relationship between residuals and independent variables.

3.8 Answers for Check Your Progress :

Check Your Progress – 1 :

1. Two or more, one **2.** overfitting

3. multicollinearity

Check Your Progress – 2 :

1. Many to one **2.** Linear

3. Independent, dependent

Check Your Progress – 3 :

1. Predicted variable **2.** less

3. Sample size, number of independent variables

Check Your Progress – 4 :

1. Significance **2.** 0.05 **3.** multicollinearity

Check Your Progress – 5 :

1. b	**2.** a	**3.** d	**4.** c	**5.** d
6. c	**7.** a	**8.** c	**9.** a	**10.** b

3.9 Glossary :

Multiple Linear Regression : Multiple linear regression is an important statistical technique for finding the relationship between one dependent variable and multiple independent variables.

Multiple R (Multiple Correlation Coefficient) : In multiple linear regression, correlation coefficient represents the relationship between the dependent variable and predicted dependent variable.

Overfitting : In multiple linear regression if there are too many significant variables in the model, then there would be a higher fictitious coefficient of determination. Also, if we have not split and validated the model on test data, then also model may be overfitted on train data.

Multicollinearity : If independent variables show a strong relationship among themselves, then it is difficult to measure the actual impact of an independent variable on the dependent variable as an independent variable also has an impact on another independent variable.

Residual Analysis : Residual analysis validates the regression assumptions like residuals are random in nature, and there is not any hidden trend. Another assumption is the residuals do not show any relation with the independent variable.

Adjusted R^2 : Coefficient of determination (R^2) increase whenever we add a new variable in the model irrespective of its significance, hence sometimes R^2 tells us fictitious model accuracy. Adjusted R^2 is a measure where R^2 get adjusted by the degree of freedom (sample size and the number of independent variables).

Regression Equation : It is a mathematical expression that represents the mathematical relationship between the dependent variable and independent variables. If we have values of all independent variables then with the help of the regression equation, we can predict the value of the dependent variable.

3.10 Assignment :

1. Why it is always recommended to see a scatter plot before applying multiple linear regression. Explain with an example ?
2. What do we mean by non-linear regression, explain with an example ?
3. What do we mean by y-intercept, explain its role in the interpretation of regression equation ?

3.11 Activities :

Below is the data for car sales. Here "Price" is the dependent variable, while all others are independent variables. Calculate SST, SSR, SSE, standard error, R^2 and adjusted R^2, regression equation for the below data. Also, justify the difference between R^2 and adjusted R^2.

engine_s	horsepower	wheelbase	width	length	curb_wgt	fuel_cap	mpg	price
1.8	140	101.2	67.3	172.4	2.639	13.2	28	21.5
3.2	225	106.9	70.6	192	3.47	17.2	26	27.3
3.2	225	108.1	70.3	192.9	3.517	17.2	25	28.4
3.5	210	114.6	71.4	196.6	3.85	18	22	42
1.8	150	102.6	68.2	178	2.998	16.4	27	23.99
2.8	200	108.7	76.1	192	3.561	18.5	22	33.95
4.2	310	113	74	198.2	3.902	23.7	21	62
2.5	170	107.3	68.4	176	3.179	16.6	26.1	26.99
2.8	193	107.3	68.5	176	3.197	16.6	24	33.4
2.8	193	111.4	70.9	188	3.472	18.5	24.8	38.9
3.1	175	109	72.7	194.6	3.368	17.5	25	21.975
3.8	240	109	72.7	196.2	3.543	17.5	23	25.3
3.8	205	112.2	73.5	200	3.591	17.5	25	27.885
3.8	205	113.8	74.7	206.8	3.778	18.5	24	31.965
3	200	107.4	70.3	194.8	3.77	18	22	31.01
4.6	275	108	75.5	200.6	3.843	19	22	39.665
4.6	275	115.3	74.5	207.2	3.978	18.5	22	39.895
5.7	255	117.5	77	201.2	5.572	30	15	46.225
1	55	93.1	62.6	149.4	1.895	10.3	45	9.235
1.8	120	97.1	66.7	174.3	2.398	13.2	33	13.96

Ans. :

Regression Statistics					
Multiple R	0.9460				
R Square	0.8950				
Adjusted R Square	0.8187				
Standard Error	4.9892				
Observations	20				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	2334.10	291.76	11.72	0.00
Residual	11	273.81	24.89		
Total	19	2607.91			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-116.159	78.407	-1.481	0.167	-288.731
engine_s	-6.055	4.886	-1.239	0.241	-16.809
horsepow	0.203	0.065	3.133	0.010	0.060
wheelbas	1.310	0.520	2.521	0.028	0.166
width	0.261	0.954	0.274	0.789	-1.840
length	-0.359	0.287	-1.250	0.237	-0.990
curb_wgt	-3.136	10.581	-0.296	0.772	-26.425
fuel_cap	1.831	1.203	1.521	0.156	-0.818
mpg	0.475	0.661	0.718	0.488	-0.980

3.12 Case Study :

Ranco automobile has established a business for car repair and service in Ahmedabad. They are receiving a few random complaints about car mileage. To understand the factors affecting car mileage (mpg), they collected data for 20 cars for various attributes. Below is the data:

engine_s	horsepower	wheelbase	width	length	curb_wgt	fuel_cap	mpg
2.5	163	103.7	69.7	190.9	2.967	15.9	24
2.5	168	106	69.2	193	3.332	16	24
2.7	200	113	74.4	209.1	3.452	17	26
3.5	253	113	74.4	207.7	3.564	17	23
3.5	253	113	74.4	197.8	3.567	17	23
2	132	105	74.4	174.4	2.567	12.5	29
2.5	163	103.7	69.1	190.2	2.879	15.9	24
2.5	168	108	71	186	3.058	16	24
8	450	96.2	75.7	176.7	3.375	19	16
2.4	150	113.3	76.8	186.3	3.533	20	24
2.5	120	131	71.5	215	3.557	22	19
3.9	175	109.6	78.8	192.6	4.245	32	15
3.9	175	127.2	78.8	208.5	4.298	32	16
5.2	230	115.7	71.7	193.5	4.394	25	17
5.2	230	138.7	79.3	224.2	4.47	26	17
2	110	98.4	67	174.7	2.468	12.7	30
2	107	103	66.9	174.8	2.564	13.2	30
2.5	170	106.5	69.1	184.6	2.769	15	25
2.5	119	117.5	69.4	200.7	3.086	20	23
3.8	190	101.3	73.1	183.2	3.203	15.7	24

Consider mpg (mileage per gallon) as predicted variable and answers for the below questions:

1. Calculate the correlation coefficient between the dependent variable (mpg) and independent variables
2. Calculate R2 and adjusted R2
3. Calculate SSR, SSE, SST and standard error
4. Calculate the regression equation
5. Scatter plot between the dependent variable and top 3 independent variables

3.13 Further Readings :

- Ken Black, "Business Statistics: Contemporary Decision Making", John Wiley and Sons (2009)
- U Dinesh Kumar, "Business Analytics: The Science of Data–Driven Decision Making", Wiley (2017)
- Naval Bajpai, "Business Research Methods", Pearson (2013)

BLOCK SUMMARY

Correlation plays a vital role in the corporate world as decision making about any business metric always depends on various factors. Therefore we can not take decisions in silos; we need to make a robust strategy about all positive and negative consequences of our decision. At one side correlation validates all these relationships while we take the help of simple and multiple linear regression to establish mathematical relations between these business metrics (dependent and independent variables). Regression models help us to predict the value of output metrics like sales, revenue, productivity, time etc. based on various factors (independent variables) like manpower requirement, marketing budget, number of sales pitches etc.

We have to be cautious at the time of doing regression analysis as model accuracy and validity depends on selecting the right independent variables. We should not unnecessary try to add independent variables as one side; it rises the problem of overfitting another side practically it is difficult to control too many input variables in the model. Regression also has two critical assumptions about independent variables – firstly, there should be a linear relationship between the variables (relationship should not be curvilinear or zig–zag); secondly, independent variables should not have a significant relationship among themselves (multicollinearity problem). Regression is also very sensitive towards residual analysis, therefore, it is utterly crucial to observe and analyze residuals as it may lead to autocorrelation (the relationship between residual and independent variables) or heteroscedasticity (error variance is not constant).

We should not supply the entire data into the regression model as it leads to overfitting. The model works very well for trained data but behaves poorly for unobserved data. Therefore, it is vital to split the entire data into training and test set (80% training and 20% test). Before exposing the model to the external world, it should be validated on the test dataset. Regression is also an important foundation for various important tools and techniques in predictive and prescriptive analytics. Currently, most machine learning algorithms use regression as a base methodology.

BLOCK ASSIGNMENT

Short Questions

1. Explain the fundamental difference between covariance and correlation
2. Write down three scenarios where Spearman rank correlation can be better than the Pearson correlation coefficient
3. Regression analysis plays a vital role in decision making. Explain this statement
4. Explain the concept of SST, SSE and SSR in a regression model
5. Explain the importance of the coefficient of multiple determination (R^2) in interpreting the multiple regression output
6. Why adjusted R^2 is better than the coefficient of multiple determination (R^2), explain with an example

Long Questions

1. Explain the relationship between variance and covariance also derive the Pearson correlation formula from the covariance formula
2. Write down the assumptions of simple linear regression analysis, write down the consequences of these assumptions that are not validated by the model
3. Explain the concept of standard error and coefficient of determination and their significance in the regression model
4. Explain the importance of residual analysis in the regression model, define different types of residual assumptions
5. Explain the steps of carrying out multiple linear regression and the significance of each step in decision making

❖ **Enrolment No. :** ☐

1. How many hours did you need for studying the units ?

Unit No.	1	2	3
No. of Hrs.			

2. Please give your reactions to the following items based on your reading of the block :

Items	Excellent	Very Good	Good	Poor	Give specific example if any
Presentation Quality	☐	☐	☐	☐	_____
Language and Style	☐	☐	☐	☐	_____
Illustration used (Diagram, tables etc)	☐	☐	☐	☐	_____
Conceptual Clarity	☐	☐	☐	☐	_____
Check your progress Quest	☐	☐	☐	☐	_____
Feed back to CYP Question	☐	☐	☐	☐	_____

3. Any other Comments

.....

.....

.....

.....

.....

.....

.....

.....

Dr. Babasaheb Ambedkar OpenUniversity

BBAR-304

Business Analytics

BLOCK-4 TIME SERIES ANALYSIS

UNIT 1

INTRODUCTION TO FORECASTING TECHNIQUES

UNIT 2

MOVING AVERAGE AND SINGLE EXPONENTIAL SMOOTHING TECHNIQUES

UNIT 3

REGRESSION METHODS FOR FORECASTING

UNIT 4

AUTO–REGRESSION (AR) AND MOVING AVERAGE (MA) FORECASTING MODELS

BLOCK 4 : TIME SERIES ANALYSIS

Block Introduction

One of the main objectives of business analytics is to forecast important business metrics like sales, raw material, manpower, budget etc. Therefore, forecasting is among the most important and fre–quently addressed in analytics. Forecasting is the process of predicting the future values of any business metric based on historic values for the same metric. For example, forecasting of sales data for next month based on historic sales data for the last 24 months. For forecasting, we use Time Series Analysis techniques. In the current era of globalization, the supply chain plays the most important role in the success of any organization and forecasting or time series analysis is the backbone of supply chain analytics which helps the organizations in predicting resources at each milestone of the entire supply chain (organizational value chain). In case the forecasting of an organization is not up to the mark then it will have a severe impact on the organization's top–line (revenue earned) and bottom–line (cost incurred in running the businesses). It can be understood with the help of an example like if forecasting is not correct and we have predicted less demand than actual then either we will not have enough goods in the market hence it would impact on the organization revenue or in case we have predicted more than required demand then we unnecessary we would to purchase extra raw material, deployed extra manpower and utilized extra space in warehouses, all this will impact on the cost of the organization. Hence it is utterly important to forecast the demand for a service/ product as accurately as possible.

All organizations despite their size and industry prepare a short term (weekly, monthly or quarterly) and long–term planning (half–yearly, yearly or five–yearly etc.) and forecasting remain the most important integral part of all these planning activities. It directly impacts the revenue and cost of the organization.

Block Objectives

After learning this block, you will be able to understand :

- Understand the significance of forecasting and how it influences decision making
- Understand various important components of time series analysis
- Learn various forecasting accuracy techniques
- Learning smoothing techniques like naïve forecasting models, average models and exponential smoothing techniques
- Application of regression theory in time series analysis
- Forecasting time series data in influence with seasonal variation
- Learn the concept of autocorrelation and how it impacts the time series analysis
- Understanding of Auto–regressive (AR) forecasting models

Block Structure

Unit 1

INTRODUCTION TO FORECASTING TECHNIQUES

: UNIT STRUCTURE :

1.0 Learning Objectives :

- Understand the significance of forecasting and how it influences decision making
- How forecasting is an integral part of supply chain management and the overall performance of the organization
- Understand various important components of time series analysis
- Time series data modelling techniques – additive methods and multiplicative methods
- Learn various forecasting accuracy techniques such as mean absolute error, mean absolute percentage error, mean square error and root mean square error

1.0 Introduction :

In this unit, we will study the basic concepts of time series analysis and its role in decision making. We will understand the impact of time

series analysis in an organization with the help of various industrial examples. Later we will study various components of time series analysis and how these components influence the overall performance of an organization. In the end, we will see important techniques to measure the accuracy of a time series analysis technique and how to determine which forecasting modelling is best for a given scenario.

1.2 Forecasting : Magical Crystal Ball of Statisticians :

"Those who know don't predict. Those who predict don't have knowledge" – ***Lao Tzu***

Forecasting plays important role in all major departments in an organization including man–power hiring, sales, production, quality control, supply chain, finance etc. Therefore, time series analysis holds a very strong role in the predictive analytics tool kit. Time series analysis impacts both the revenue side (top–line performance) and expenditure side (bottom–line performance) of an organization. Following are time series forecasting examples from the business world:

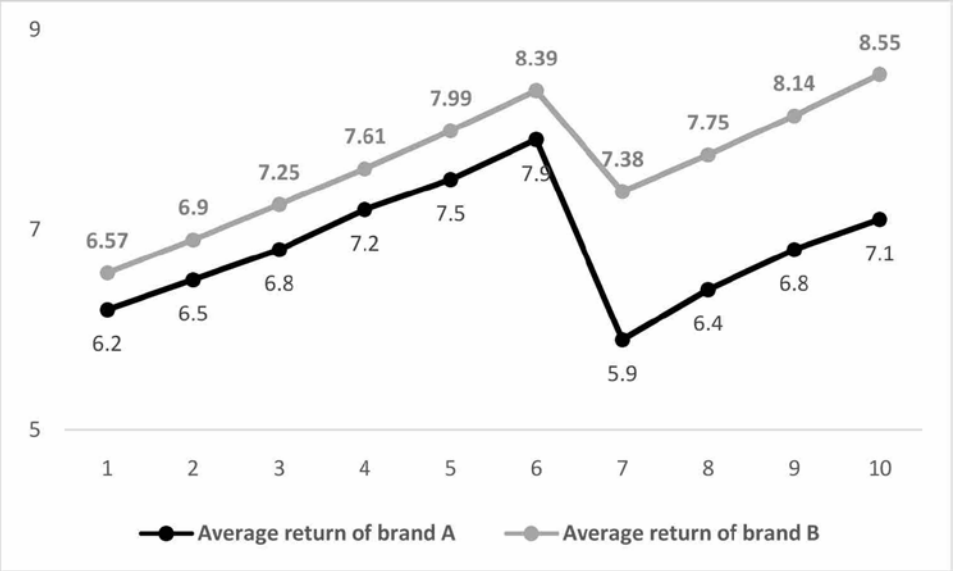
- Organizations want to predict the manpower requirement for the upcoming financial year
- The finance department wants to estimate the budget for procuring hardware and software units for the new plant
- Municipality predicts the water shortage in Ahmedabad during May and June month
- The weather department predicts normal rainfall in northern India
- The aviation industry estimates a downfall in demand due to the corona pandemic in 2022 also
- There will be a significant increment in demand for the electric vehicle due to income tax relaxation by the government
- The life insurance market will grow at the rate of 5% in the next five years

Besides severe application in the business world, time series also plays important role in personal life. For example, somebody wants to invest in a mutual fund, and he must choose between two brands. Below is the data for the last ten years return of two renowned mutual funds:

Year	Average return of brand A	Average return of brand B
2011	6.2	6.57
2012	6.5	6.90
2013	6.8	7.25
2014	7.2	7.61
2015	7.5	7.99
2016	7.9	8.39
2017	5.9	7.38
2018	6.4	7.75
2019	6.8	8.14
2020	7.1	8.55

Selection will depend on the prediction about their returns in the upcoming years which is an application of time series analysis. As studied in block 1, time–series measure the centrality and variance in the data to estimate the future value but it is not limited to these two factors only, we will see it in detail in upcoming topics and units.

Time series data is always collected over a period which means the x–axis of a time series data is always time (day, date, month etc). Line charts are the most preferred graph to visualize time–series data.

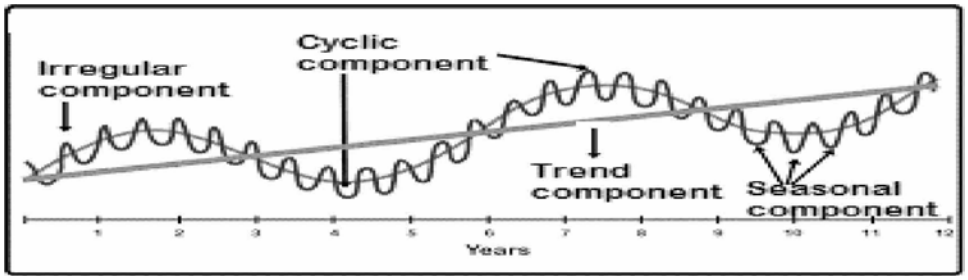


1.3 Time–Series Data and Components of Time–Series Data :

In time–series data, the x–axis always denotes time while the y–axis represents the response variable, Y_t, such as share price, productivity, revenue, cost etc at different time points t. The variable Y_t is a random variable and generally, it is collected at regular intervals of time and arranged in chronological order. If time–series data contains data only for one variable, then it is called univariate time series for example daily gold price or the number of employees who join an organization every month. While if time–series involves more than one variable then it is known as multivariate time–series data, for example per employee revenue earned, per employee cost, per employee logged in hours etc collected for each month.

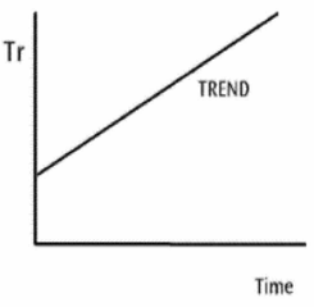
Month	Per employee revenue	Per employee cost	Per employee logged in hours
Jan'20	1,70,000	1,15,000	176
Feb'20	1,72,000	1,17,000	176
Mar'20	1,65,000	1,16,000	172
Apr'20	1,83,000	1,19,000	175
May'20	1,79,000	1,18,000	173
Jun'20	1,84,000	1,21,000	179

Most statisticians believed that time–series data from a forecasting perspective can be categorized into four components: trend component, seasonal component, cyclic component, and irregular component. Most of the time all these components are not present in time–series data.

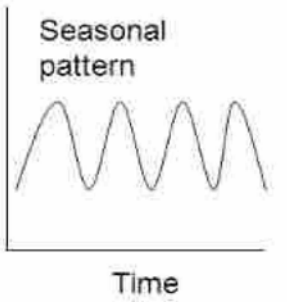


Below is the description of each of these components:

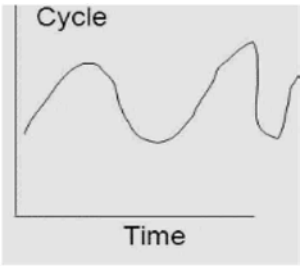
1. **Trend Component (T_t) :** The trend component depicts long term overall downward or upward movement of time–series data. Generally, the trend does not repeat in time–series data, it shows the overall slope of the time–series data. There are two main ways to find a trend in the data either by simply plot the data e.g. line plot, run chart or control chart etc. another way is the decomposition of time–series data but this is not included in the scope of this text. For example, in Gold price in the last three years, there are many fluctuations in the daily price but overall, there is an increasing trend.



2. **Seasonal Component (S_t) :** When time–series data shows repetitive downward or upward movement/ fluctuations from the trend then we say there is a seasonal influence in our data. These repetitions are generally shorter (less than a year) for example, daily, weekly, monthly, quarterly etc. But the frequency of this repetition is less than a year most of the time. There can be various reasons for seasonality in the time–series data like a weekly footfall in the shopping mall or theatres, school holidays, festivals or end of season sales etc. The time after which seasonal trend repeats is known as the periodicity of seasonal variation and it repeats over a period.

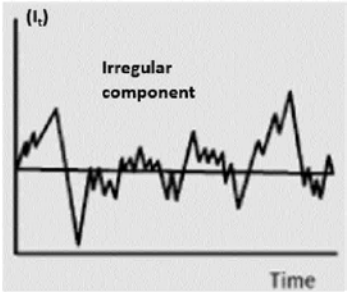


3. **Cyclical Component (C_t) :** Cyclical component is the movement over a longer period (generally more than a year). These movements or fluctuations are because of macro economic changes while in the case of seasonal component movement observed within a year and depend on micro–economic factors like festivals



and customs that depend on society, the end of season sale etc. Another big difference between cyclical and seasonal components is that cyclical movement has not fixed time between fluctuations, it is random in nature while seasonal component has a fixed time period within a year generally. In another world periodicity of cyclical fluctuations is not constant while it is constant for a seasonal component.

4. **Irregular Component (I_t) :** Irregular component consists of random movements in the time series and follow a normal distribution with mean 0 and constant variance. Distribution with mean zero and constant variance is known as white noise. From time–series data if we remove the trend, seasonal and cycle components then residual time–series data is an irregular component.



1.4 Time–Series Data Modelling Techniques :

Data modelling plays important role in business analytics. Modelling means if we have Time–series data as an input then we can forecast future value for this Time–series, this predicting future capability based on input data is known as **modelling** in statistics. There are two important techniques to model Time–series data either through the addition of all or few Time–series components or product of these components. Let's discuss these in brief:

1.4.1 Additive Model of Time–Series Modelling :

In the Additive model we add these components like below

$$Y_t = T_t + C_t + S_t + I_t$$

The basic assumption behind the additive model is that cyclical and seasonal components are completely independent of the trend component. Which is generally not the case in real–life scenarios hence additive models are not very popular for forecasting purpose. An additive model can be un–derstood with an example, there are few extra weekend classes in an institute and there are fixed students enrolled in those classes hence extra food demand in the canteen can be forecasted through an additive model.

1.4.2 Multiplicative Model of Time–Series Modelling :

Multiplicative models are more common in business scenarios as they do not have the assumption that seasonal component should be uncorrelated with trend component. The general structure of the multiplicative model is as below:

$$Y_t = T_t \times C_t \times S_t \times I_t$$

The cyclical component needs long term data to be used in the forecasting model, which is difficult to have all the time. In case long

term (few years data) is not available then we use a modified version of the above equation:

$$Y_t = T_t \times S_t$$

A multiplicative model better fits the data that's why it is more common than the additive model.

1.5 Measuring Forecasting Accuracy Techniques :

There are several forecasting techniques, which analysts can use to predict the future value of their desired KPI/ business measure. It is relatively a difficult question to select the most appropriate technique as it depends on various factors and business situations, we will cover these techniques in detail in upcoming topics and units.

One of the major factors of choosing the right forecasting technique is calculating the accuracy measure. We calculate accuracy measures for different forecasting techniques and compare the forecasting error, whichever method gives the least error become the strong contender for choosing forecasting technique, although this approach can be computationally expensive and time consuming for very large data sets. Besides accuracy measure, business objective, analysis software capability and analysts command on these techniques also plays important role in selecting the most relevant forecasting technique for a given business scenario.

Forecasting error or deviation is the difference between forecasted value and actual value of a Time–series. It is also known as residuals.

$$e_t = x_t - F_t$$

Where

e_t = Forecasted error/ deviation for given time t

x_t = Actual value at time t

F_t = Forecasted value at time t

There are four important forecasting accuracy measures used frequently:

1. Mean Absolute Error (MAE) / Mean Absolute Deviation (MAD)
2. Mean Absolute Percentage Error (MAPE)
3. Mean Square Error (MSE)
4. Root Mean Square Error (RMSE)

1.5.1 Mean Absolute Error (MAE) / Mean Absolute Deviation (MAD):

This is the average error in forecast without considering the sing of the error. The formula is as below:

$$MAE = \sum_{t=1}^{n} \frac{|x_t - F_t|}{n}$$

We saw during the calculation of variance, some of the deviations from the mean line were positive and the rest were negative hence

variances of the different signs were cancelling each other to overcome that we squared all variance terms. Another way to overcome this sign (direction) problem is taking absolute value instead of squaring these terms. It will tell us the overall error in the forecasted values. Let's see the examples of the forecasted value of an organization for the last 10 years.

Year	Sales	Forecast	Error Value	Absolute error
1	1336	-		
2	1392	1477	-85	85
3	1487	1532	-45	45
4	1547	1460	87	87
5	1610	1626	-16	16
6	1689	1729	-40	40
7	1741	1669	72	72
8	1798	1760	38	38
9	1760	1752	8	8
10	1714	1733	-19	19
Sum			**0**	**410**

$$\text{MAD} = \frac{410}{9} = 2.878$$

Here n = 9 (forecasting is available only for 9 observations).

This example explains that some of the forecasted errors are positive while others are negative. If add all these error values, then most of the time outcome will be either 0 or very close to 0 as these negative and positive errors are cancelling each other. By taking the absolute value we can resolve this problem. In the above example, the summation of all error terms is zero but it is not mandatory that we will get error summation all the time zero but for sure it will remain close to zero all the time until there are lots of outliers in our data.

One of the important limitations of mean absolute error is that we cannot compare the MAD score of two different studies as it directly depends on the absolute value of the Time–series data. If one time series in hundreds, share prices of an organization is in range of ₹ 600 to ₹ 850 while the share price of another organization is in range of ₹ 45,000 to ₹ 56,000 then MAD in case of the second organization will be higher as their absolute values are higher. Another accuracy measure means absolute percentage error resolve this problem.

1.5.2 Mean Absolute Percentage Error (MAPE) :

Mean absolute percentage error consider the average of absolute percentage error instead of the average of just absolute value of the error terms. The formula is as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|x_t - F_t|}{x_t}$$

As MAPE is dimensionless hence it can be used to compare two different Time–series irrespective of the magnitude of their values.

Year	Sales	Forecast	Error Value	Absolute values of errors divided by actual values
1	1336	-		
2	1392	1477	-85	0.061
3	1487	1532	-45	0.030
4	1547	1460	87	0.056
5	1610	1626	-16	0.010
6	1689	1729	-40	0.024
7	1741	1669	72	0.041
8	1798	1760	38	0.021
9	1760	1752	8	0.005
10	1714	1733	-19	0.011
Sum			**0**	**0.259**

$$\text{MAPE} = \frac{.259}{9} \times 100 = 2.878$$

Here n = 9 (forecasting is available only for 9 observations).

1.5.3 Mean Square Error (MSE) :

When we calculate error terms in Time–series forecasting, we had already seen the method of taking the absolute value to get rid out of negative values. Another way to do the same thing is to calculate the square of error terms.

$$\text{MSE} = \sum_{t=1}^{n} \frac{(x_t - F_t)^2}{n}$$

Year	Sales	Forecast	Error Value	Square of Error Terms
1	1336	-		
2	1392	1477	-85	7225
3	1487	1532	-45	2025
4	1547	1460	87	7569
5	1610	1626	-16	256
6	1689	1729	-40	1600
7	1741	1669	72	5184
8	1798	1760	38	1444
9	1760	1752	8	64
10	1714	1733	-19	361
Sum			**0**	**25728**

$$\text{MSE} = \frac{25728}{9} = 2859$$

Here n = 9 (forecasting is available only for 9 observations).

1.5.4 Root Mean Square Error (RMSE) :

Square root of the mean square error is known as root mean square error (RMSE). It is the standard deviation of errors.

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(x_t - F_t)^2}{n}}$$

Year	Sales	Forecast	Error Value	Square of Error Terms
1	1336	-		
2	1392	1477	-85	7225
3	1487	1532	-45	2025
4	1547	1460	87	7569
5	1610	1626	-16	256
6	1689	1729	-40	1600
7	1741	1669	72	5184
8	1798	1760	38	1444
9	1760	1752	8	64
10	1714	1733	-19	361
Sum			**0**	**25728**

$$RMSE = \sqrt{\frac{25728}{9}} = \sqrt{2859} = 53.47$$

RMSE and MAPE are the two most popular forecasting accuracy measures. Generally, the organization put it as a target.

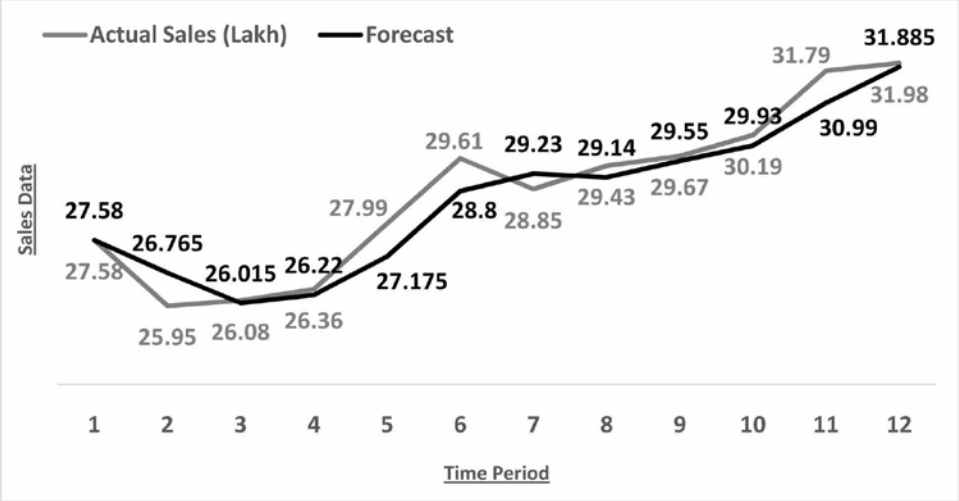
Worked Example : An organization has below sales and forecasting figures. Calculate MAD, MAPE, MSE and RMSE.

Year	Actual	Forecast
1	27.580	27.580
2	25.950	26.765
3	26.080	26.015
4	26.360	26.220
5	27.990	27.175
6	29.610	28.800
7	28.850	29.230
8	29.430	29.140
9	29.670	29.550
10	30.190	29.930
11	31.790	30.990
12	31.980	31.885

Solution :

In the earlier section, we have seen the formula and significance of different forecasting accuracy measures. Here Time–series data is available for 12 data periods. Generally, we do not have forecasting figures available for the first period or sometimes we consider actual

sales itself forecasting number. Generally, a line chart is preferred to visualise time–series data.



Year	Actual	Forecast	Error Value $x_t - F_t$	Absolute Value |xt - Ft|	Square of Error Terms $(x_t - F_t)^2$	Absolute error value divided by actual values $\|x_t - F_t\|/x_t$
1	27.580	27.580	0.000	0	0.000	0.000
2	25.950	26.765	-0.815	0.815	0.664	0.031
3	26.080	26.015	0.065	0.065	0.004	0.002
4	26.360	26.220	0.140	0.14	0.020	0.005
5	27.990	27.175	0.815	0.815	0.664	0.029
6	29.610	28.800	0.810	0.81	0.656	0.027
7	28.850	29.230	-0.380	0.38	0.144	0.013
8	29.430	29.140	0.290	0.29	0.084	0.010
9	29.670	29.550	0.120	0.12	0.014	0.004
10	30.190	29.930	0.260	0.26	0.068	0.009
11	31.790	30.990	0.800	0.8	0.640	0.025
12	31.980	31.885	0.095	0.095	0.009	0.003
Sum			**2.2**	**4.59**	**2.9679**	**0.160**

Here n = 12

$$\text{MAD/MAE} = \sum_{t=1}^{n} \frac{|x_t - F_t|}{n} = \frac{4.59}{12} = 0.383$$

$$\text{MSE} = \sum_{t=1}^{n} \frac{(x_t - F_t)^2}{n} = \frac{2.968}{12} = 0.247$$

$$\text{RMSE} = \sqrt{\sum_{t=1}^{n} \frac{(x_t - F_t)^2}{n}} = \sqrt{\frac{2.968}{12}} = 0.497$$

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|x_t - F_t|}{x_t} = \frac{.160}{12} \times 100 = 1.329$$

1.6 Factors Affecting Forecasting Accuracy :

Forecasting always remain challenging as there are so many factors that influence it, we apply various fixes to improve the forecasting accuracy. Below are few important factors:

A. The high volume of data helps in better forecasting accuracy : High volumes always help in attaining better forecasting accuracy as it covers various possible business scenarios also smoothes the impact of natural variance in the data. Therefore, forecasting a superstore is always easier than a local confectionary or grocery store. High volume is the prerequisite for most of the data science techniques as large data explains its variation well.

B. Aggregation helps in improving forecasting accuracy : Forecasting at an individual product/ service level is difficult but if we aggregate data into a few logical groups then it becomes easy. For example, forecasting at each course in a university may be difficult but if we try to forecast graduate students in science streams by aggregating individual specialization in the science department then it would be relatively easy.

C. Short period forecasting works better than long term : A long term period consists of various business scenarios which are not so regular therefore longer forecasting has relatively lower accuracy.

D. Forecasting is relatively easy for a stable business : In a start–up or new business, the unit has unstable demand which makes forecasting challenging while on the other hand, a mature business has regular customers and a stronger brand which helps in superior accuracy.

Check Your Progress :

1. In Time–series data y–axis denotes the business measures while x–axis denotes _____.
2. Seasonality component is relatively easier to estimate while cyclical component is _____ to estimate.
3. Root mean square error and _____ forecasting accuracy measures can be compared across the industries.
4. Mean absolute error is also known as _____.

❖ **Multiple Choice Questions :**

1. Cyclic components in a Time–series is caused due to:
   a. Random events occur in local geography
   b. Festivals in society
   c. Macroeconomic changes
   d. Changes in customer behaviour

2. Which option is NOT a valid component of Time–Series data ?
   a. Irregular component b. Seasonal component
   c. Trend component d. Noise
3. White noise is:
   a. Errors with mean and the standard deviation both 1
   b. Uncorrelated errors with a mean value of 0 and constant standard deviation
   c. Errors with constant mean and standard deviation
   d. Completely random errors
4. Which forecasting accuracy measures can be measured across different industries
   a. MSE and RMSE b. MAD and MAPE
   c. Only MAPE d. RMSE and MAPE
5. Which forecasting model will be more appropriate if seasonality is completely independent of the trend
   a. Additive model
   b. Multiplicative model
   c. Both additive and multiplicative can be used
   d. Neither additive nor multiplicative can be used
6. Which Time–Series component depicts long term upward and downward movement of the data
   a. Seasonality b. Cyclic c. Irregular d. Trend
7. Which component depends on local festivals and events in the society
   a. Trend b. Seasonality c. Cyclic d. Irregular
8. Which forecasting accuracy measure is the standard deviation of errors
   a. MAPE b. MAD c. RMSE d. MSE
9. Which of the below factors help in better forecasting ?
   a. Large data set
   b. Predicting for a shorter period
   c. Forecasting for a stable business
   d. All the above
10. What is modelling in statistics ?
   a. Predicting output variable based on input variables
   b. Predicting input variables
   c. Predicting error terms
   d. To check errors following a normal distribution or not

1.7 Let Us Sum Up :

1. Time–series analysis plays an important role in descriptive analytics as it shows timely movement (fluctuations) of important KPIs over the period
2. Time–series forecasting is the backbone of predictive analytics as all organizations need to forecast their important business KPIs (key performance indicators)
3. Time–series data always have an x–axis as time while the y–axis is the output variable, for which we want to show the movement over the time
4. Time–series has four important components, trend, seasonality, cyclical and irregular
5. Trends showcase overall upward or downward movement of data throughout the time period
6. When time–series data shows repetitive downward or upward movement / fluctuations from the trend then we say there is a seasonal influence in our data. These repetitions are generally shorter (less than a year) for example, daily, weekly, monthly, quarterly etc.
7. A cyclical component is a movement over a longer period (generally more than a year).
8. An irregular component consists of random movements in the time series and follow a normal distribution with mean 0 and constant variance.
9. One of the major factors of choosing the right forecasting technique is calculating the accuracy measure, technique which shows the least error generally become the choice of forecasting, but it is not the only factor. Businesses consider several other factors also to identify the right forecasting technique
10. There are four important forecasting accuracy measures used frequently, these are Mean Absolute Error (MAE) / Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Root Mean Square Error (RMSE)
11. The value of MAD and MSE depends on the magnitude of data hence these cannot be used for benchmarking across the department or businesses
12. MAPE and RMSE can be used as a target for benchmarking as these are represented as ratio and percentage respectively
13. Forecasting is relatively easy for large data, stable businesses, aggregated data and shorter period

1.8 Answers for Check Your Progress :

Check Your Progress :

1. Time **2.** Difficult

3. MAPE **4.** Mean absolute deviation

❖ **Multiple Choice Questions :**

1. c	**2.** c	**3.** b	**4.** d
5. a	**6.** d	**7.** b	**8.** c
9. d	**10.** a		

1.9 Glossary :

Time–Series : Time–series analysis represents the output variable on the y–axis while the x–axis always remains time (generally time interval remain constant).

Forecasting Technique : Predicting future value of output variable based on historic data is known as a forecasting technique.

Component of Time–Series : Time series can be decomposed into four important parts namely, trend, seasonality, cyclical and irregular. These parts of time–series are known as components of time series.

Forecasting Error : The difference between the actual value and forecasting value is known as a forecasting error.

Forecasting Accuracy Measure : The approach for calculating the forecasting error in a time series is known as the forecasting accuracy measure.

Mean Absolute Error (MAE) / Mean Absolute Deviation (MAD): This is the average error in forecast without considering the sing of the error.

Mean Absolute Percentage Error (MAPE) : Mean absolute percentage error consider the average of absolute percentage error instead of the average of just absolute value of the error terms.

Mean Square Error (MSE) : It is the average of the square of the error term.

Root Mean Square Error (RMSE) : Square root of the mean square error is known as root mean square error (RMSE).

1.10 Assignments :

1. Why forecasting is known as one of the most important functions of business analytics.
2. Write down a few examples of the application of forecasting techniques in personal life.
3. Write down various components of time series analysis and their importance in brief.
4. Write important scenarios where additive and multiplicative forecasting models can be used. Explain with a few examples.

1.11 Activities :

One of the leading organizations in the retail store provided their west India store's revenue (in crores rupees) of last 12 months. Calculate the MAD, MSE, RMSE and MAPE for this data.

Year	Actual	Forecast
1	26.580	26.580
2	24.950	25.765
3	25.080	25.015
4	25.360	25.220
5	26.990	26.175
6	28.610	27.800
7	27.850	28.230
8	28.430	28.140
9	28.670	28.550
10	29.190	28.930
11	30.790	29.990
12	30.980	30.885

Ans. MAD – 0.383, MSE – 0.247, MAPE – 1.377, RMSE – 0.497

1.12 Case Study :

Pokymon construction company uses more than thousands of different small and big machinery parts which range from ₹ 7 to ₹ 5 million. They supply ready to mix and other intermediate materials require in the construction business. Consistent supply of these parts are required for their customers as in absence of these parts, they may face severe losses on the other hand if the Pokemon industry keeps surpassing inventory then it blocks their working capital as well as there is always a risk of part become obsolete. Therefore, forecasting these parts is very critical for the Pokymon industry.

Below is the actual quantity and forecasted of one of the small parts:

Year	Actual	Forecast
1	321	398
2	345	382
3	543	375
4	424	408
5	390	411
6	321	407
7	422	390
8	362	396
9	516	389
10	462	415
11	462	424
12	459	432

Questions :

1. Suggest the most appropriate forecasting accuracy measure
2. Plot the actual quantity with trend line
3. Which forecasting accuracy measures, Pokymon industry can be compared with their competitors

1.13 Further Readings :

- "Time series analysis and Control," Holden Day, Box and Jenkins (1970)
- "How to get a better forecast, Harvard Business Review", Parker G, Segura E (1971)
- "An introduction to Time series analysis and forecasting with applications of SAS and SPSS", Management science journal, Yaffee R, McGee M (2000)

Unit 2

MOVING AVERAGE AND SINGLE EXPONENTIAL SMOOTHING TECHNIQUES

: UNIT STRUCTURE :

2.0 Learning Objectives :

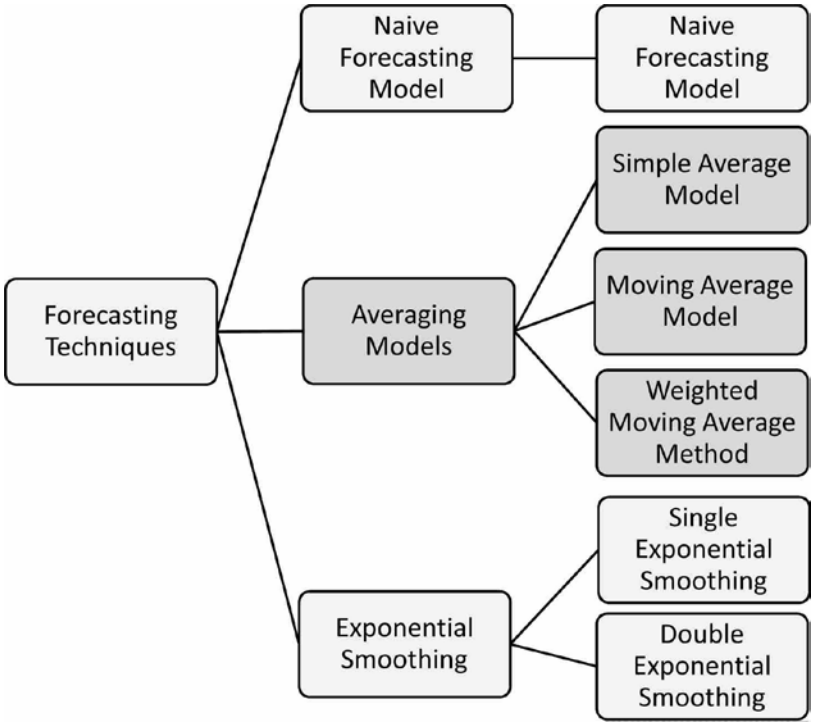
- Understand smoothing techniques for time–series data
- Graphical representation of smoothing techniques
- Learning smoothing techniques like naïve forecasting models, average models and exponential smoothing techniques
- Learning optimization of smoothing constants
- Establishing the relationship between smoothing constant and accuracy measures

2.1 Introduction :

In this unit, we will study various forecasting techniques also known as smoothing techniques as these techniques smooth out the irregular fluctuation effects in the time–series data. We will analyse the graphical representation of these smoothing techniques and will see the ways to optimize smoothing constants. In the end, we will see the relationship between smoothing constant and accuracy measures.

2.2 Parts of Forecasting Techniques :

Forecasting techniques for stationary time series data (no significant level of seasonal, trend, cyclical component) are also known as smoothing techniques as smooth (minimize) the fluctuation due to irregular effects in the time–series. Below are important forecasting techniques:



2.3 Naïve Forecasting Models :

These are simple methods where we consider the most recent periods of data to represent the forecast or prediction of future outcomes. Here we do not consider anything like trend or seasonality. Shorter frequency data like hourly, daily or weekly naïve model generally works better. The simplest form of a naïve model is as follows:

$$F_t = x_{t-1}$$

Where,

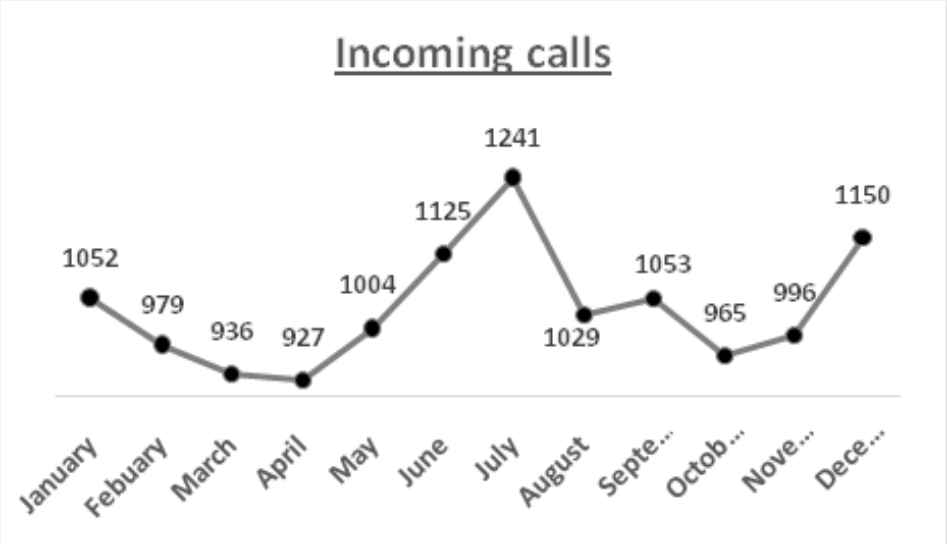
x_{t-1} = The actual value at time t–1

F_t = Forecast value for time t

So, in simple words, if last month incoming call volume for a help desk was 1150 then the naïve forecasting model would forecast that the helpdesk will receive 1150 calls this week also.

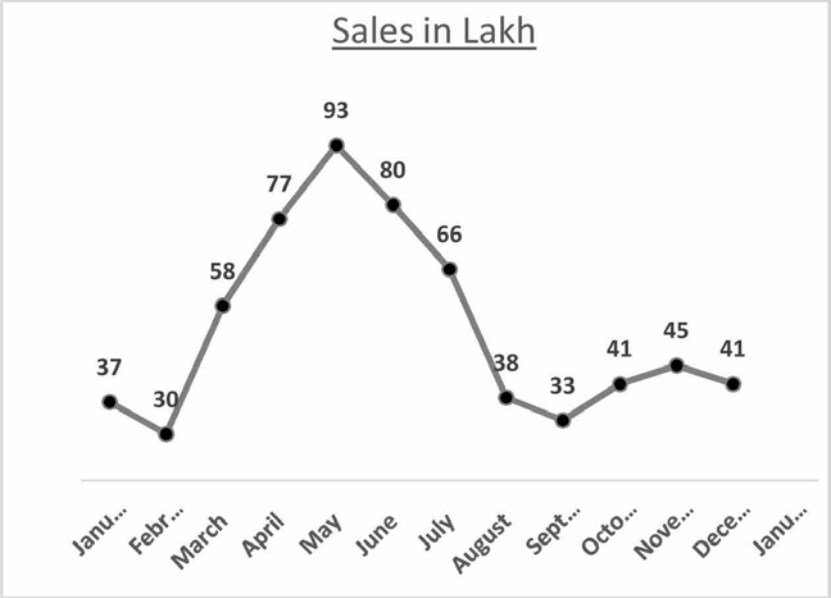
This data is for the 12^{th} month, now if we want to predict incoming call volume for the 13^{th} month then it will remain the same as for the 12^{th} month, 1150 calls.

Month	Incoming calls
January	1052
February	979
March	936
April	927
May	1004
June	1125
July	1241
August	1029
September	1053
October	965
November	996
December	1150
January	???



Another way to use the naïve forecasting model is that we can use the value of last year during the same week on month. For example, if last year January sale was ₹ 37 lakh then this year January we will consider sales of ₹ 37 lakh only (here we will not consider the sale of last month, December).

Month	Sales in Lakh
January	37
February	30
March	58
April	77
May	93
June	80
July	66
August	38
September	33
October	41
November	45
December	41
January	



2.4 Averaging Models :

Averaging models smooths the fluctuation by taking an average of several data points in order to calculate the forecast for the next time period. There are various ways to calculate the averaging models, below three are the most important ones:

2.4.1 Simple Averages :

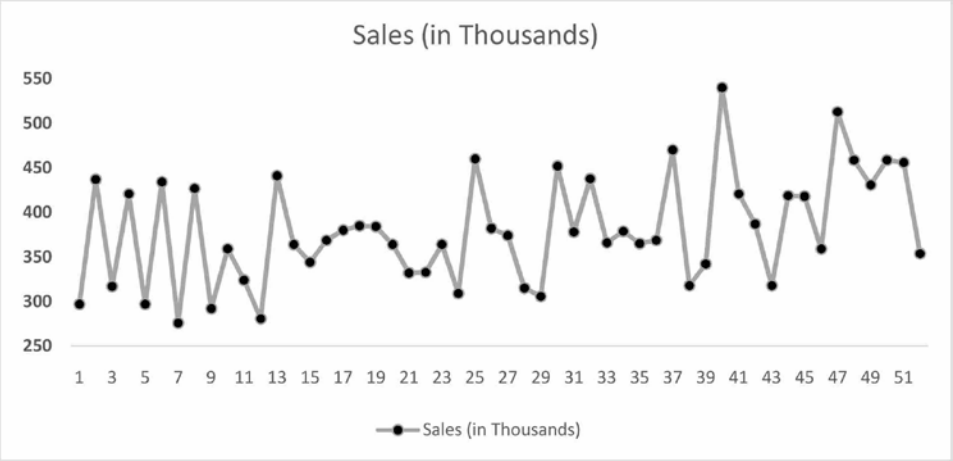
Most fundamental averaging model is a simple average model. Here forecast for time t is calculated by averaging the value for a given number of previous time periods. Below is the formula for the same:

$$F_t = \frac{x_{t-1} + x_{t-2} + ... + x_{t-n}}{n}$$

Below is the data for a year, given in the form of weekly sales data.

Week	Sales (in Thousands)	Week	Sales (in Thousands)
1	297	**27**	374
2	437	**28**	315
3	317	**29**	306
4	421	**30**	452
5	297	**31**	378
6	434	**32**	438
7	276	**33**	366
8	427	**34**	379
9	292	**35**	365
10	359	**36**	369
11	324	**37**	470
12	281	**38**	318
13	441	**39**	342
14	364	**40**	540

15	344		**41**	421
16	369		**42**	387
17	380		**43**	318
18	385		**44**	419
19	384		**45**	418
20	364		**46**	359
21	332		**47**	513
22	333		**48**	459
23	364		**49**	431
24	309		**50**	459
25	460		**51**	456
26	382		**52**	354



Here if want to calculate the forecast for the 53rd week then we can calculate the average of all 52 weeks.

$$F_{53} = \frac{x_{52} + x_{51} + ... + x_1}{52} = 380$$

In case we have data for several months, for example, 30 months then instead of calculating the average of all 30 months we can also calculate the average of 12 months. Generally, statisticians take decisions by consulting with domain experts.

2.4.2 Moving Averages :

Moving average is another simple and interesting way to forecast the time series data. It forecast the future value of a time series using an average of the last 'N' observations. Below is the formula for calculating a simple moving average:

$$F_{t+1} = \frac{1}{N} \sum_{k=t+1-N}^{t} Y_k$$

Here we take the average of the last N observations.

Worked Example : Below is the sales data of an organization for the last 12 months. Use a 4–month moving average to calculate the forecast for Jan 2021. Also calculate the MAD, MAPE, MSE and RMSE.

Year	Sales (in Thousands)
1	297
2	437
3	317
4	421
5	297
6	434
7	276
8	427
9	292
10	359
11	324
12	281

Solution :

We can calculate the moving average by calculating the average of 4 consecutive terms, first forecasting term will be $\frac{(297 + 437 + 317 + 421)}{4} = 368$ similarly second forecasting term will be $\frac{(437 + 317 + 421 + 297)}{4} = 368$ so on so for.

Month	Sales (in Thousands)	Moving Average	Error	Absolute error value	Square value of Error	Absolute values of Errors Divided by Actual values
Jan'20	297					
Feb'20	437					
Mar'20	317					
Apr'20	421					
May'20	297	368	-71	71	5041	0.239
Jun'20	434	368	66	66	4356	0.152
Jul'20	276	367	-91	91	8327	0.331
Aug'20	427	357	70	70	4900	0.164
Sep'20	292	359	-67	67	4422	0.228
Oct'20	359	357	2	2	3	0.005
Nov'20	324	339	-15	15	210	0.045
Dec'20	281	351	-70	70	4830	0.247
Jan'21		314				
Average				**56**	**4011**	**0.176**

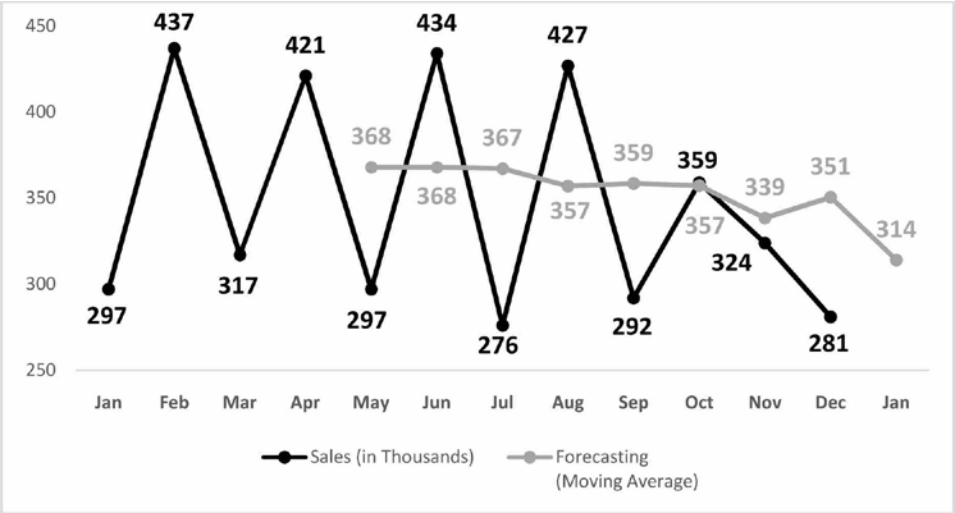
Forecasted value (Jan'21) $= \frac{\mathbf{292 + 359 + 324 + 281}}{\mathbf{4}} = \mathbf{314}$

MAD – 56

MSE – 4011

RMSE – 22

MAPE – 0.176



Besides so many benefits of moving averages, there are few disadvantages also as determining the optimal length of time periods for moving average is always challenging also moving average does not adjust the effects of trend, cycles or seasonality.

2.4.3 Weighted Moving Averages :

When we calculate moving average, we give equal weights to all the terms but in the weighted moving average, we assign varying weights. Generally, we give higher weights to most recent terms and relatively lower weights to older data points (observations). The general formula is as follows:

$$F_{t+1} = \sum_{k=t+1-N}^{t} W_k \times Y_k$$

Where W_k is the weight given to the value of Y at time K (Yk)

and $\sum_{k=t+1-N}^{t} W_k = 1.$

The summation of all weights is equal to 1.

There are two ways to keep the summation of all weights to 1 either we keep all the weights in decimal in such a manner that summation of all these weights in decimal must be equal to 1 or we can multiply different data points by different weights and divided by the summation of all weights. Let's see both with the help of the examples.

Scenario 1 – weights are in decimals and their summation is 1.

$$F_{t+1} = .4(F_t) + .3(F_{t-1}) + .2(F_{t-2}) + .1(F_{t-3})$$

Scenario 2 – weights are in integers and we divide the entire term by summation of all weights

$$F_{t+1} = \frac{5 \times F_t + 3 \times F_{t-1} + 2 \times F_{t-2} + 1 \times F_{t-3}}{11}$$

Worked Example : Calculate a 3–month weighted moving average for a retail firm. Using weights of 3 for the recent month, 2 for the month prior and 1 for the month before that. Calculate forecast for Jan 2021 also calculate MAD, MSE, RMSE and MAPE. 12 months data as below:

Month	Sales (in Thousands)
Jan'20	496
Feb'20	339
Mar'20	427
Apr'20	421
May'20	398
Jun'20	434
Jul'20	276
Aug'20	427
Sep'20	397
Oct'20	359
Nov'20	424
Dec'20	361

Solution :

First moving average will be calculated on the first three values (Jan, Feb and Mar data).

$$\frac{(3 \times 496 + 2 \times 337 + 1 \times 427)}{6} = 432$$

Month	Sales (in Thousands)	Forecasting (Moving Average)	Error	Absolute error value	Square value of Error	Absolute values of Errors Divided by Actual values
Jan'20	496					
Feb'20	339					
Mar'20	427					
Apr'20	421	432	-11	11	125	0.027
May'20	398	382	16	16	256	0.040
Jun'20	434	420	14	14	191	0.032
Jul'20	276	416	-140	140	19460	0.505
Aug'20	427	390	37	37	1394	0.087
Sep'20	397	380	17	17	283	0.042
Oct'20	359	347	13	13	156	0.035
Nov'20	424	406	18	18	336	0.043
Dec'20	361	389	-28	28	775	0.077
Jan'21		381				
Average				35	2856	0.108

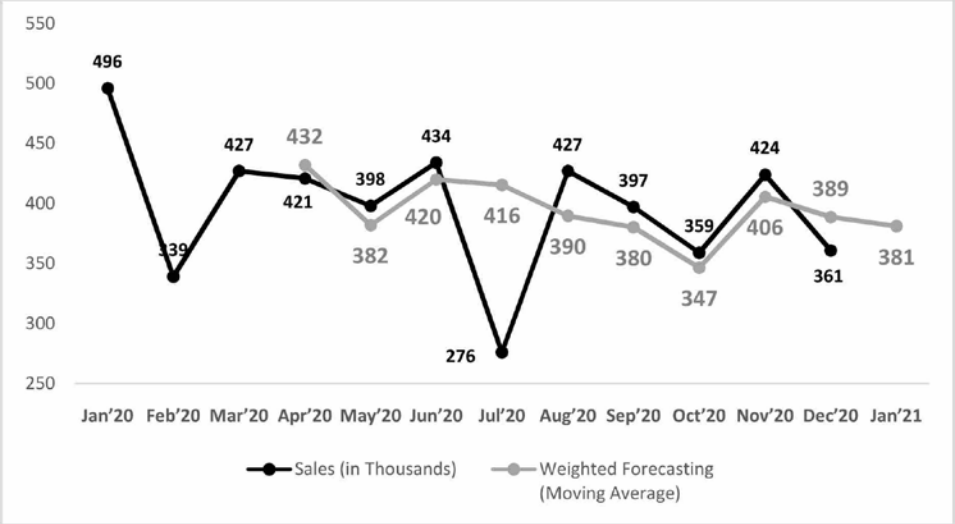
$$\textbf{Forecasted value (Jan'21)} = \frac{3 \times 359 + 2 \times 424 + 1 \times 361}{6} = 381$$

MAD – 35

MSE – 2856

RMSE – 18

MAPE – 0.108



Deciding on the optimal length of moving average remain a challenge for weighted moving average also besides that deciding weights is also very difficult and remain subjective.

2.5 Single Exponential Smoothing Forecasting Technique :

Exponential smoothing techniques are quite like the weighted moving average technique; the only difference is that here weights decrease their importance exponentially. Here we have a smoothing constant α, generally a value between 0 and 1. The general form of single exponential smoothing forecasting equation is:

$$F_{t+1} = \alpha \times x_t + (1-\alpha) \times F_t$$

Where

x_t = The actual value at time t

F_t = Forecast value for time t

F_{t+1} = Forecast value for time t+1

α = Exponential smoothing constant, a value between 0 and 1.

Here the value of α is decided by the analyst. They may require consulting with domain experts.

As we have a general equation for a single exponential smoothing equation, we can write an equation for F_t as below:

$$F_t = \alpha \times x_{t-1} + (1-\alpha) \times F_{t-1}$$

We can substitute the value for Ft in the above equation for F_{t+1}

$$F_{t+1} = \alpha \times x_t + (1-\alpha) \times \{\alpha \times x_{t-1} + (1-\alpha) \times F_{t-1}\}$$

Above equation can be simplified as follows:

$$F_{t+1} = \alpha \times x_t + \alpha\,(1-\alpha) \times x_{t-1} + (1-\alpha)^2 \times F_{t-1}$$

Similarly, if we can substitute the value of $F_{t-1} = \alpha \times x_{t-2} + (1-\alpha) \times F_{t-2}$ in the above equation, it can be further expanded as below:

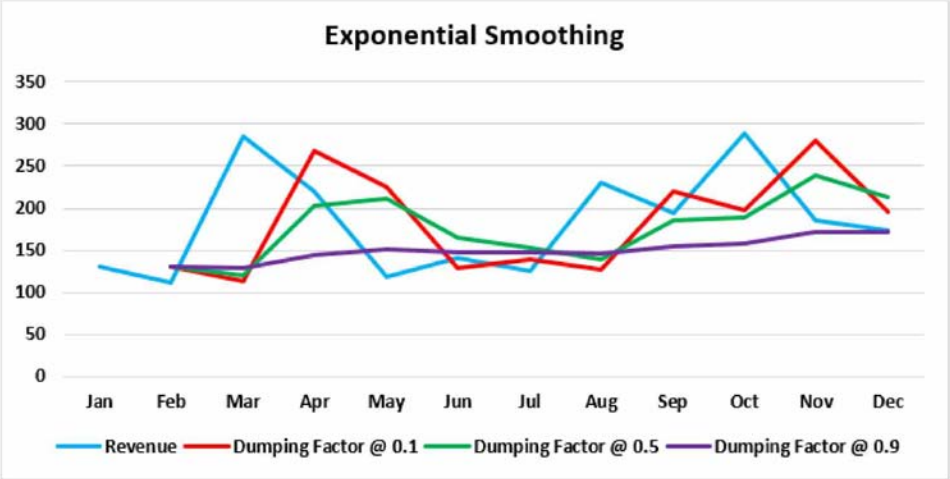
$$F_{t+1} = \alpha \times x_t + \alpha\,(1-\alpha) \times x_{t-1} + (1-\alpha)^2 \times \{\alpha \times x_{t-2} + (1-\alpha) \times F_{t-2}\}$$

$$\mathbf{F_{t+1} = \alpha \times x_t + \alpha\,(1-\alpha) \times x_{t-1} + \alpha\,(1-\alpha)^2 \times x_{t-2} + (1-\alpha)^3 \times F_{t-2}}$$

Continuing the above process explains the weights on the older time period and forecasts include $(1- \alpha)^n$ (exponential values). Because of exponential values impacts higher emphasis on recent values and rapidly decreasing emphasis on older time periods.

The basic idea about exponential smoothing techniques is that the forecasted value of every next time period is a combination of current actual and forecasted value. If the value of α is chosen less than .5 then there would be less weight on actual values and more weight on forecasted value while if the value of α is more than .5 then there would be more weight on actual value and less weight on forecasted values.

α is known as the smoothing constant and $(1-\alpha)$ is known as the dumping factor.



Worked Example : 12 months sales figures are given in the table below. Apply exponential smoothing to predict the values of each month. Calculate forecast for January 2021 using α = .3, .5 and .9

Month	Sales (in Thousands)
Jan'20	496
Feb'20	339
Mar'20	427
Apr'20	421
May'20	398
Jun'20	434
Jul'20	276
Aug'20	427
Sep'20	397
Oct'20	359
Nov'20	424
Dec'20	361

Solution :

We can create a table with forecast and error values for each alpha value. Generally, the forecast is not available for the very first data point hence we cannot calculate a forecasted value for the second time period. Therefore, we use the actual value of the first time period as the forecasted value for the second period and start the forecasting process. Applying the exponential smoothing formula to calculate the forecasted value for third and fourth data points (for α = .3) :

F3 = .3(339) + .7(496) = 449

F4 = .3(427) + .7(449) = 442

F5 = .3(421) + .7(442) = 436

F6 = .3(398) + .7(436) = 425

Similarly, we can calculate forecasted values with other α values also. The below table has all calculated values.

Month	Sales (in Thousands)	0.3						0.5						0.9					
		Forecast(α=.3)	Error	Abs Error	Squared Error	Abs(Error)/Actual value		Forecast(α=.5)	Error	Abs Error	Squared Error	Abs(Error)/Actual value		Forecast(α=.9)	Error	Abs Error	Squared Error	Abs(Error)/Actual value	
Jan'20	496																		
Feb'20	339	496	-157	157	24649	0.463		496	-157	157	24649	0.463		496	-157	157	24649	0.463	
Mar'20	427	449	-22	22	480	0.051		418	10	10	90	0.022		355	72	72	5227	0.169	
Apr'20	421	442	-21	21	455	0.051		438	-17	17	287	0.040		429	-8	8	67	0.019	
May'20	398	436	-38	38	1439	0.095		432	-34	34	1133	0.085		423	-25	25	632	0.063	
Jun'20	434	425	9	9	89	0.022		417	17	17	290	0.039		402	32	32	1037	0.074	
Jul'20	276	427	-151	151	22918	0.549		429	-153	153	23493	0.555		433	-157	157	24666	0.569	
Aug'20	427	382	45	45	2028	0.105		352	75	75	5671	0.176		291	136	136	18458	0.318	
Sep'20	397	395	2	2	2	0.004		404	-7	7	56	0.019		422	-25	25	650	0.064	
Oct'20	359	396	-37	37	1364	0.103		396	-37	37	1387	0.104		397	-38	38	1432	0.105	
Nov'20	424	385	39	39	1532	0.092		377	47	47	2165	0.110		363	61	61	3758	0.145	
Dec'20	361	397	-36	36	1267	0.099		404	-43	43	1886	0.120		420	-59	59	3491	0.164	
Jan'21		386						379		0				365		0			
Average		418	-33	51	5111	0.149		412	-27	54	5555	0.158		400	-15	70	7643	0.196	

Conclusion :

The smaller value of ? (larger the damping factor), the smoothers the entire time–series. On the other hand, the larger ? (smaller the damping factor), forecasted values follow time–series data.

Benefits and Drawbacks of Single Smoothing Forecasting Techniques:

Benefits :

The single exponential smoothing technique performs very well in most scenarios. Its important benefits are as follows:

1. It is simple in calculation and works well with the data if there is no strong trend or seasonal patterns in the time series
2. This technique uses all historic data unlike moving average forecasting techniques
3. It assigns progressively decreasing weights which ensure more importance to recent data

Drawbacks :

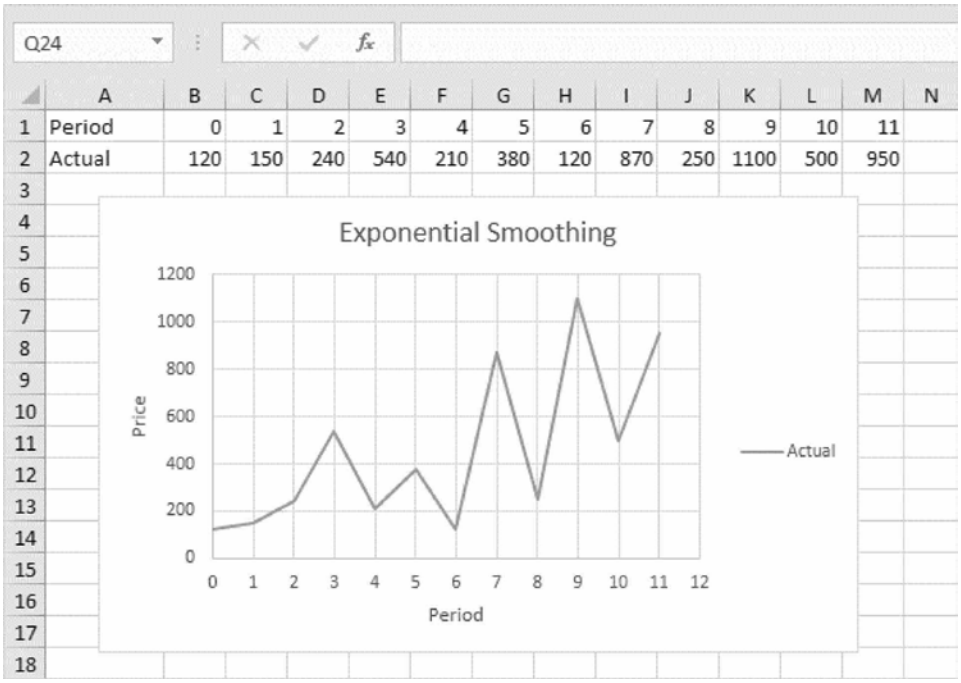
1. For every large data set, the forecast becomes less sensitive to changes in the data
2. It always lags the trend and it is based on past observations. The longer the time period n, the greater the last as it is slow to recognize the shifts in the level of the data points
3. It doesn't work well if data has trend and seasonality influence

2.6 Single Exponential Smoothing Forecasting Technique in MS Excel :

In the single exponential smoothing forecasting technique, we have only one smoothing constant that's why we call it single exponential

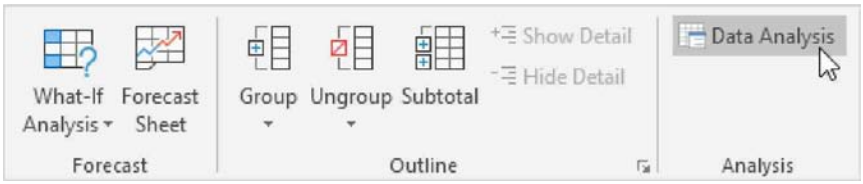
smoothing. There are advanced versions of smoothing techniques like the double exponential smoothing technique and triple exponential smoothing technique. The single exponential smoothing technique can also be carried out in MS Excel. Below are the important steps:

Step – 1 : We can create a simple line graph to visualise the fluctuations in the time series.

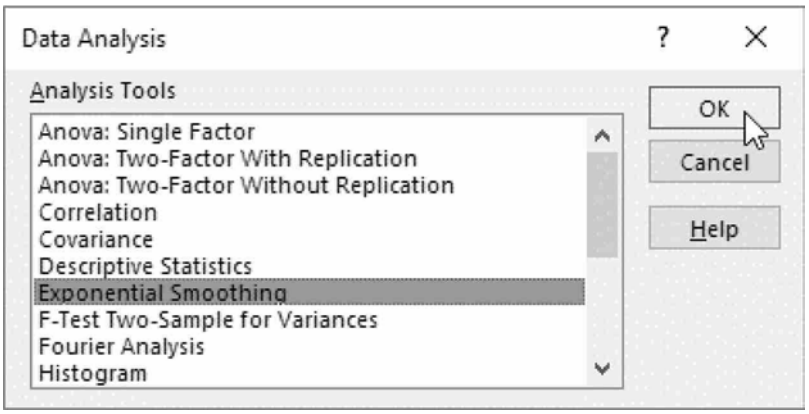


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Period	0	1	2	3	4	5	6	7	8	9	10	11
2	Actual	120	150	240	540	210	380	120	870	250	1100	500	950

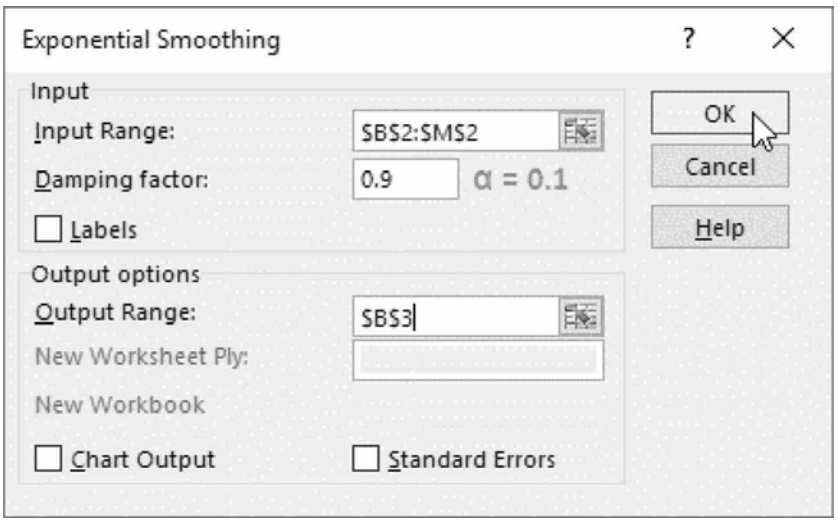
Step – 2 : Select the "Data Analysis" pack under the Data tab



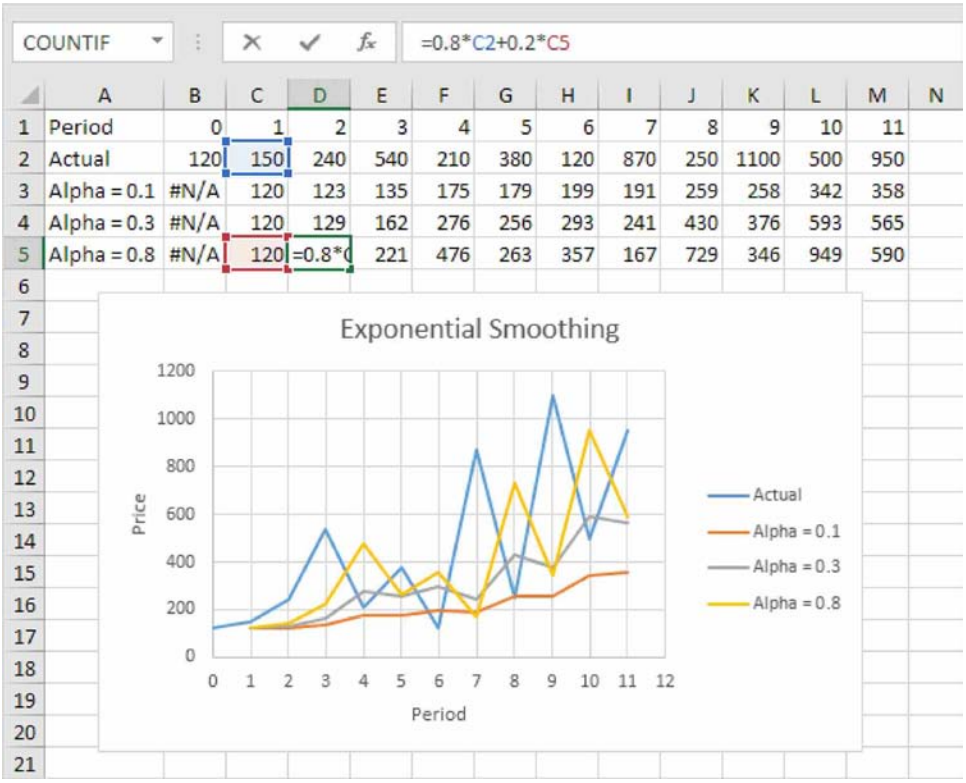
Step – 3 : Select the option, "Exponential smoothing" and click on the OK button



Step – 4 : Provide the reference of the input range, enter damping factor $(1-\alpha)$, output cell reference and click on the OK button



Step – 5 : Calculate the forecasted values for different values of α.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Period	0	1	2	3	4	5	6	7	8	9	10	11
2	Actual	120	150	240	540	210	380	120	870	250	1100	500	950
3	Alpha = 0.1	#N/A	120	123	135	175	179	199	191	259	258	342	358
4	Alpha = 0.3	#N/A	120	129	162	276	256	293	241	430	376	593	565
5	Alpha = 0.8	#N/A	120	=0.8*C	221	476	263	357	167	729	346	949	590

Notice that there is always a lag because forecasts depend on the last observations. Lower the alpha value smoother the forecast.

Check Your Progress :

1. Naïve forecasting methods works well with _____ frequency time period
2. In simple moving average, there are _____ weights assigned for all data period
3. Summation of all weights in weighted moving average is always _____.
4. Value of smoothing constant in single exponential smoothing constant is generally between _____ and _____.

5. If α is the smoothing constant in exponential smoothing time series forecasting technique, then $1-\alpha$ is known as _____.

❖ **Multiple Choice Questions :**

1. If smoothing constant in single smoothing is less than .5 then it means, there will be more emphasis on
   a. Actual value
   b. Residual or error term
   c. Forecasted value
   d. The absolute value of the error term
2. If α is small, then the value of the dumping factor will be
   a. Small
   b. Large
   c. It doesn't matter
   d. None of above
3. How smoothing techniques transform the time series
   a. It doesn't change anything to the time series data
   b. It makes time series more fluctuating
   c. It reduces the peak part of the time series
   d. It smooths the fluctuations of a time–series data
4. Single smoothing exponential technique has:
   a. Single smoothing constant
   b. Single trend line
   c. Both options are correct
   d. None of the above
5. The value of the smoothing constant always lies between
   a. –1 to +1
   b. 0 and the average value
   c. 0 and 10
   d. 0 and 1
6. One of the main differences between moving average and single exponential smoothing technique is:
   a. single exponential smoothing technique uses a constant value always more than .5
   b. single exponential smoothing technique uses all the historic data
   c. Both above options are correct
   d. None of the above
7. Which forecasting technique uses progressively decreasing weights to historic data
   a. Naïve forecasting technique
   b. The single exponential smoothing technique
   c. Moving average
   d. Weighted moving average

8. What is the main motive behind the progressively decreasing weights in the single exponential smoothing technique ?
   a. Recent actual values will have higher weights
   b. Recent actual values will have lower weights
   c. Time series will smooth all the fluctuations
   d. Forecasting accuracy measure will be better
9. One of the parameters to select the appropriate forecasting technique is
   a. Forecasting accuracy measure
   b. Business domain
   c. Knowledge of statistician about the forecasting technique
   d. None of the above
10. One of the drawbacks of moving average is
   a. It is computationally very expensive
   b. It provides the same importance to all data points
   c. It is difficult to choose the optimal length of moving average
   d. None of the above

2.7 Let Us Sum Up :

1. Forecasting techniques for stationary time series data are also known as smoothing techniques as smooth (minimize) the fluctuation due to irregular effects in the time series.
2. Naïve forecasting techniques are simple methods where we consider the most recent periods of data to represent the forecast or prediction of future outcomes
3. In the simple average forecasting method forecast for time t is calculated by averaging the value for a given number of previous time periods.
4. Moving average is a simple and interesting way to forecast the time series data. It forecast the future value of a time series using an average of the last 'N' observations.
5. In the weighted moving average forecasting technique, we assign varying weights. Generally, we give higher weights to most recent terms and relatively lower weights to older data points (observations).
6. In the exponential smoothing technique, weights decrease their importance exponentially. Here we have a smoothing constant ?, generally a value between 0 and 1.
7. Single exponential smoothing technique uses all historic data unlike moving average forecasting techniques, it assigns progressively decreasing weights which ensure more importance to recent data

8. Single exponential smoothing technique, there is always a lag because forecasts depend on the last observations. Lower the alpha value smoother the forecast.

2.8 Answers for Check Your Progress :

Check Your Progress :

1. Shorter **2.** Same **3.** One

4. 0 and 1 **5.** Damping factor

❖ **Multiple Choice Questions :**

1. c **2.** b **3.** d **4.** a

5. d **6.** b **7.** b **8.** a

9. a **10.** c

2.9 Glossary :

Forecasting Technique : Predicting future value of output variable based on historic data is known as a forecasting technique

Moving Average Forecasting Techniques : Moving average is a forecasting technique that forecasts the future value of a time series data using average (or weighted average) of past 'N' observations

Single Exponential Smoothing Forecasting Technique : This technique uses exponentially reducing weights with the help of a smoothing constant alpha. Unlike moving average methods, it uses all historic data

Damping Factor : It is used to smooth the time series and progressively assign weights to historic data. Its value is $1-\alpha$

2.10 Assignment :

1. What is the basic difference between naïve, averaging models and exponential smoothing forecasting techniques ? Write down the important type of forecasting techniques under these categories.
2. "It is not necessary that more complex mathematical forecasting technique will always give better forecasting accuracy". Explain this statement with appropriate examples.
3. What is the basic difference between moving average and weighted moving average forecasting techniques ?

2.11 Activities :

One of the leading organizations in the retail store provided their west India store's revenue (in crores rupees) of last 12 months. Calculate the forecast using a 3–month moving average and single exponential smoothing technique (use α = .3). Also calculate MAD, MSE, RMSE and MAPE for this data.

Year	Revenue
1	26.580
2	24.950
3	25.080
4	25.360
5	26.990
6	28.610
7	27.850
8	28.430
9	28.670
10	29.190
11	30.790
12	30.980

2.12 Case Study :

"Delicious bowl" is a famous restaurant chain in North–East India. They receive their raw materials from various parts of Assam and Sikkim. It is the area where rain is very high compared to the rest of India. Sometimes heavy rain impacts their sales hence they always prepare their dishes as per the weather forecasts and procurement also goes with the same logic. Below is the table where their restaurant chain daily revenue and rain in cms.

Year	Actual	Forecast
1	321	398
2	345	382
3	543	375
4	424	408
5	390	411
6	321	407
7	422	390
8	362	396
9	516	389
10	462	415
11	462	424
12	459	432

Questions :

1. Develop a forecasting model using 5–days and 7 days moving average and single exponential smoothing technique for alpha value .3 and .8
2. Visualise the forecasted values
3. Calculate the MAD, MSE, RMSE and MAPE also suggest which measure is most appropriate for their restaurant chain
4. Which forecasted model will you recommend for their business

2.13 Further Readings :

- "Time series analysis and Control," Holden Day, Box and Jenkins (1970)
- "How to get a better forecast, Harvard Business Review", Praker G, Segura E (1971)
- "Time Series Based Predictive Analytics Modelling: Using MS Excel", Glyn Davis, Branko Pecar; 1st edition (2016)
- "An introduction to Time series analysis and forecasting with applications of SAS and SPSS", Management science journal, Yaffee R, McGee M (2000)

REGRESSION METHODS FOR FORECASTING

: UNIT STRUCTURE :

3.0 Learning Objectives :

- Application of regression theory in time series analysis
- Forecasting time series data in influence with seasonal variation
- Industry examples to understand the interpretation of regression–based forecasting outputs
- Application of regression methods in business decisions

3.1 Introduction :

In this unit, we will study the forecasting methods when we can have extra information about time series. For example, besides sales data, we also have information about marketing expenses, competition information, consumer's demographic information etc. Here we can also include information about seasonality variation. In the last, we will study a few worked examples to understand how regression–based forecasting methods help in robust decision making in various industries.

3.2 Forecasting Techniques with a Trend :

Forecasting techniques discussed in unit 2 have an assumption that trend is not included in the time–series data. In this unit, we will discuss the possible ways to determine a trend in the data. One of the most effective ways to determine trend components in time–series data is

regression analysis. Here we consider the output variable as the metric we are determining on a time scale or the metric we want to forecast while the x–axis represents time.

Regression is a more powerful forecasting technique when the time–series has values of various independent variables also besides the dependent variable Y_t. The forecasting techniques we discussed in the last units have only output variables that's why their application in the real world is comparatively less as most of the time we have various independent variables which influence output variables. Therefore, forecasting only based on output variables is less effective. For example, if we want to predict the sales price alone then it may not be so effective but if we also include information like marketing cost, whether a promotion was going on, nearby festival etc then our forecasting will be more accurate. The general form of a regression equation is as follows:

$$F_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_n X_{3t} + \varepsilon_t$$

Where

F_t = The forecasted value at time t

X_{1t}, X_{2t} etc are the predictor variables measures at time t

ε_t = Error or Irregular component at time t

Worked Example : Below are crude oil per barrel prices in the Indian market (IN USD). See if a trend is visible, can we predict the price for 17th August 2021

Day	Crude Oil Price per Barrel in India
1	$73.62
2	$73.95
3	$71.26
4	$70.56
5	$68.15
6	$69.09
7	$68.28
8	$66.48
9	$68.29
10	$69.25
11	$69.09
12	$68.44
13	$67.29

Solution : We can use regression analysis under MS Excel "Data Analysis" tab

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.76187804
R Square	0.580458147
Adjusted R Square	0.542317979
Standard Error	1.528988275
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	35.58	35.58	15.22	0.00
Residual	11	25.72	2.34		
Total	12	61.30			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	72.6142	0.8996	80.720	0.000	70.634	74.594	70.63	74.59
Day	-0.4421	0.1133	-3.9012	0.003	-0.6916	-0.1927	-0.691	-0.192

$$F_t = 72.6142 - 0.4421 \times X_t$$

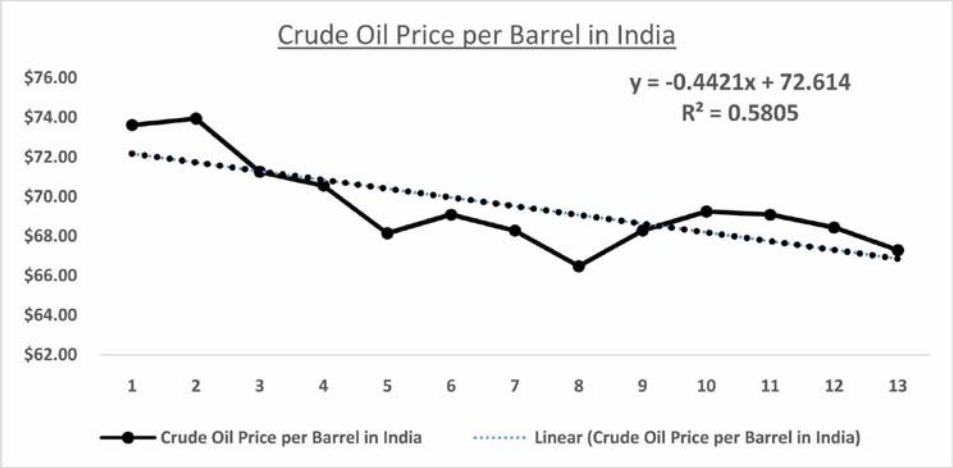
Here, F_t is forecasted per barrel price of crude oil and X_t is the time period

So, if we want to forecast the sales for the next time period then the value will be:

$$F_{14} = 72.6142 - 0.4421 \times X_{14} = 72.6142 - 0.4421 \times 14 = \$66.42$$

Here can interpret it like every unit increase in the time period X_t, per barrel price of crude oil will decrease by 0.442 USD. Y–intercept 72.6142 indicates that the time period before the very first time period (at Day 0) per barrel price of crude oil was $ 72.61.

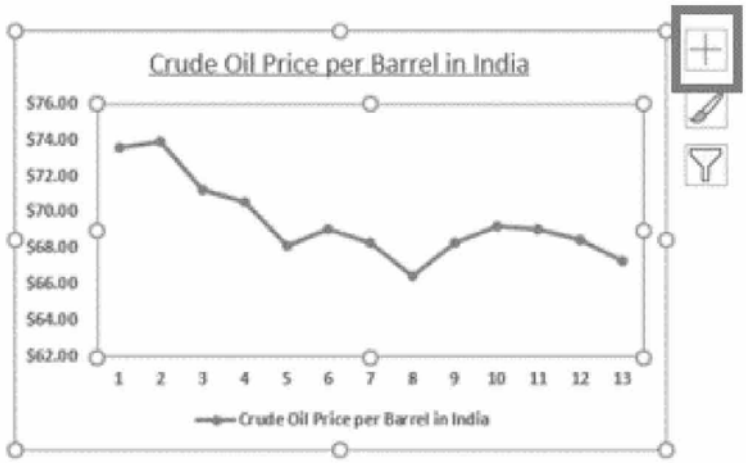
The P–value for the independent variable "Day" is 0.0025 represents a significant relationship between input and output variables. Below is the visual representation of the trend.



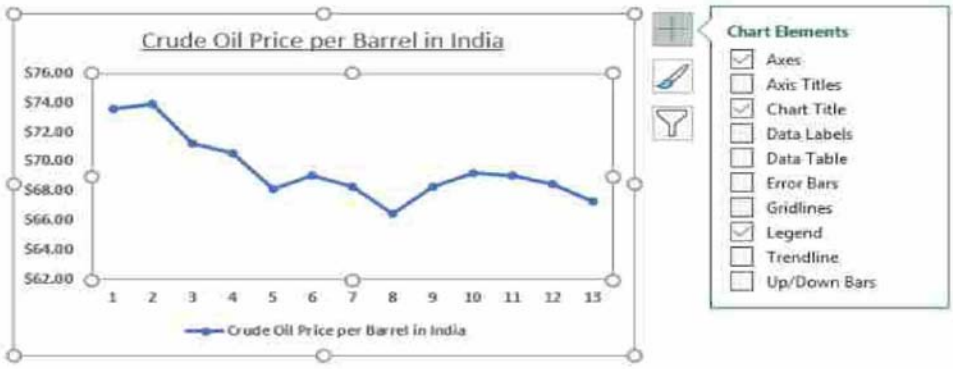
3.3 How to Draw Trendline and Regression Equation in Time–Series Graph :

We can add a trend line to our time–series graph to visualize a trend. Below are the important steps to create a trend line and publish a regression equation, R square value etc.

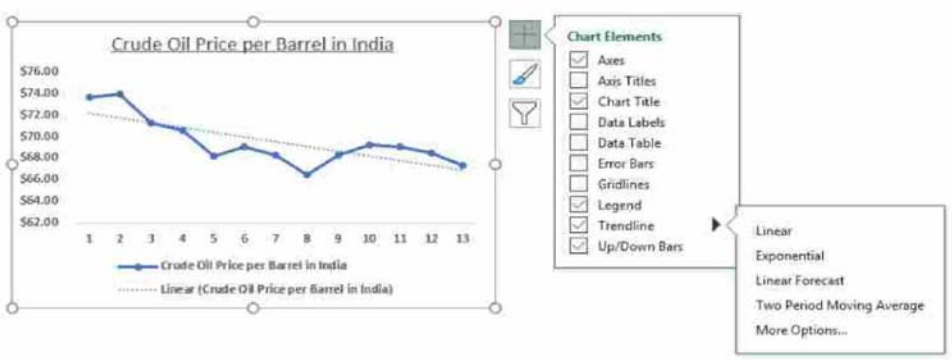
1. Select the dependent and independent variables and create a line graph, time variable must be on X–axis and the output variable on Y–axis
2. Select the graph, right hand side three features will appear. Select the + icon on the right–hand top side of the graph



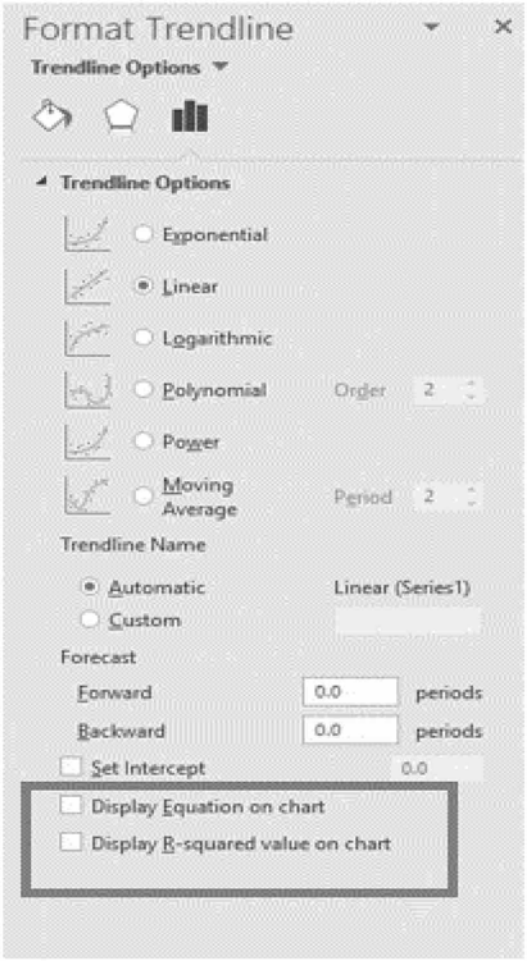
3. On clicking on the + icon, one menu bar will be open



4. Select the second last option, "Trendline", which will create a trend line as per the data. There is various type of trendline like linear (straight line), exponential, Linear forecast etc. In this course, we have included only Linear, but others are also quite intuitive, please refer to the Microsoft excel help website to learn about these different trendlines. By default, Excel will create a Linear trendline only (the very first option).

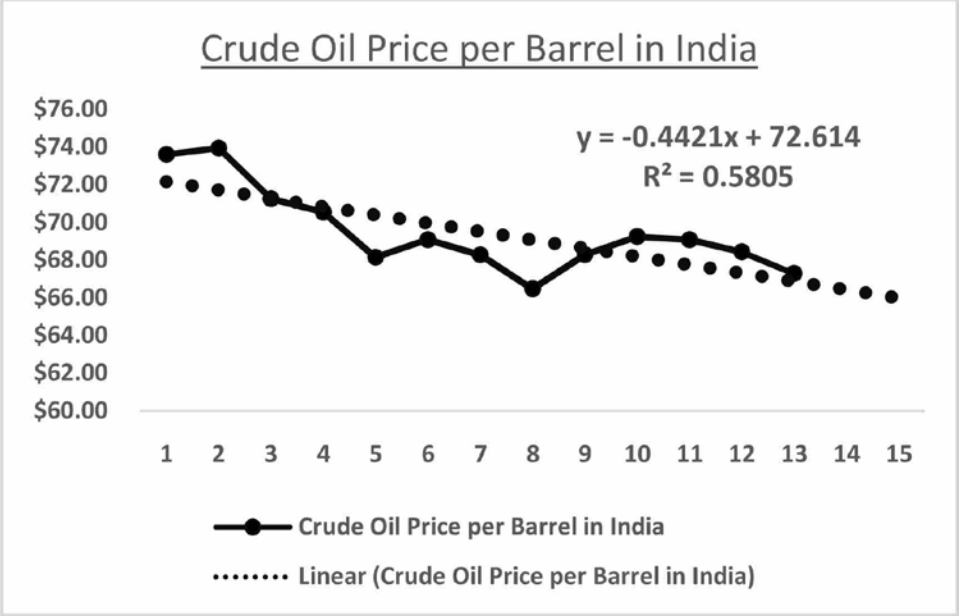


5. Select the trendline and right–click, to choose the "format trendline option"



Here we can select a different types of trendlines and select the options provided below, "Display Equation on chart" or "Display R–squared value on chart".

6. We can also extend the trendline either forward or backwards to forecast the data points



3.4 Double Exponential Smoothing Constant Technique for Forecasting :

One of the drawbacks of single exponential smoothing is that the model does not do well if a trend is visible in the data. This can be improved by introducing an additional equation for capturing the trend in the time–series data. Double exponential smoothing uses two equations to forecast the future values of the time series, one for forecasting the level (short term average value) and another for capturing the trend. Below are the two equations:

Level (or Intercept) equation (L_t) :

$$L_t = \alpha \times Y_t + (1-\alpha) \times F_t$$

Equation for the trend (T_t) :

$$T_t = \beta \times (L_t - L_{t-1}) + (1-\beta) \times T_{t-1}$$

Where

α and β are smoothing constant for level and trend, respectively and $0 < \alpha < 1$ and $0 < \beta < 1$.

The forecast at time $t + 1$ is given by:

$F_{t+1} = L_t + T_t$

$F_{t+n} = L_t + (n)T_t$

This technique does not work very well if the seasonality component is also available besides the trend. In that situation, we use the triple exponential smoothing technique which is not in the scope of this text.

Worked Example : Below is the sales data of ABC Limited for 42 months. Use double exponential smoothing technique to forecast sales for the next 6 months.

Month	Sales		Month	Sales
Apr-17	9.75		Jan-19	11.46
May-17	10.09		Feb-19	11.06
Jun-17	10.27		Mar-19	10.76
Jul-17	8.98		Apr-19	10.32
Aug-17	8.79		May-19	9.34
Sep-17	8.23		Jun-19	8.39
Oct-17	8.24		Jul-19	7.01
Nov-17	8.97		Aug-19	6.71
Dec-17	9.03		Sep-19	5.93
Jan-18	9.31		Oct-19	5.07
Feb-18	10.25		Nov-19	5.21
Mar-18	9.5		Dec-19	6.46
Apr-18	10.05		Jan-20	6.35
May-18	9.73		Feb-20	5.36
Jun-18	9.48		Mar-20	5.25
Jul-18	9.38		Apr-20	4.94
Aug-18	9.44		May-20	5.13
Sep-18	10.02		Jun-20	5.63
Oct-18	10.15		Jul-20	5.35
Nov-18	10.55		Aug-20	5.74
Dec-18	11.13		Sep-20	5.26

Solution :

Step 1 : Let's assume value for smoothing constants, $\alpha = 0.7$ and $\beta = 0.6$. Later we will see how we can optimize these values with the help of the "Solver" functionality of MS Excel.

Step 2 : We must initialize the very first values of Level, trend and forecast. One of the popular ways to initialize the level value is considering the same actual value for that period While the trend is the difference between the current and last actual value.

Month	Sales	Level	Trend	Forecast	Error
Apr-17	9.75				
May-17	10.09	10.09	0.34		
Jun-17	10.27			10.43	-0.16
Jul-17	8.98				
Aug-17	8.79				
Sep-17	8.23				
Oct-17	8.24				

The initial value of forecast value is the summation of Level and trend while Error is the difference between the actual value and forecasted value.

Step 3 : Now we must calculate the level value for the row "Jun–17" $\rightarrow \alpha \times Y_t + (1-\alpha) \times F_t$.

$\text{Level}_{\text{Jun–17}} = 0.7 \times 10.27 + (1 - 0.7) \times (10.09 + 0.34) = 10.318$

Step 4 : Now we must calculate the trend value for the row "Jun–17" $\rightarrow \beta \times (L_t - L_{t-1}) + (1-\beta) \times T_{t-1}$

$\text{Trend}_{\text{Jun–17}} = 0.6 \times (10.318 - 10.090) + (1 - 0.6) \times (0.340) = 0.273$

Step 5 : Copy the same formula for entire column of "Level" and "Trend" (up to the cells actual values are available). Also copy the formula for "Forecast" and "Error" column

Month	Sales	Level	Trend	Forecast	Error
Apr-17	9.75				
May-17	10.09	10.090	0.340		
Jun-17	10.27	10.318	0.273	10.43	-0.16
Jul-17	8.98	9.463	-0.404	10.59	-1.61
Aug-17	8.79	8.871	-0.517	9.06	-0.27
Sep-17	8.23	8.267	-0.569	8.35	-0.12
Oct-17	8.24	8.077	-0.341	7.70	0.54
Nov-17	8.97	8.600	0.177	7.74	1.23
Dec-17	9.03	8.954	0.283	8.78	0.25
Jan-18	9.31	9.288	0.314	9.24	0.07
Feb-18	10.25	10.056	0.586	9.60	0.65
Mar-18	9.5	9.842	0.107	10.64	-1.14
Apr-18	10.05	10.020	0.149	9.95	0.10
May-18	9.73	9.862	-0.035	10.17	-0.44
Jun-18	9.48	9.584	-0.181	9.83	-0.35
Jul-18	9.38	9.387	-0.190	9.40	-0.02
Aug-18	9.44	9.367	-0.088	9.20	0.24
Sep-18	10.02	9.798	0.223	9.28	0.74
Oct-18	10.15	10.111	0.277	10.02	0.13
Nov-18	10.55	10.502	0.345	10.39	0.16
Dec-18	11.13	11.045	0.464	10.85	0.28
Jan-19	11.46	11.475	0.443	11.51	-0.05
Feb-19	11.06	11.317	0.083	11.92	-0.86
Mar-19	10.76	10.952	-0.186	11.40	-0.64
Apr-19	10.32	10.454	-0.373	10.77	-0.45
May-19	9.34	9.562	-0.684	10.08	-0.74

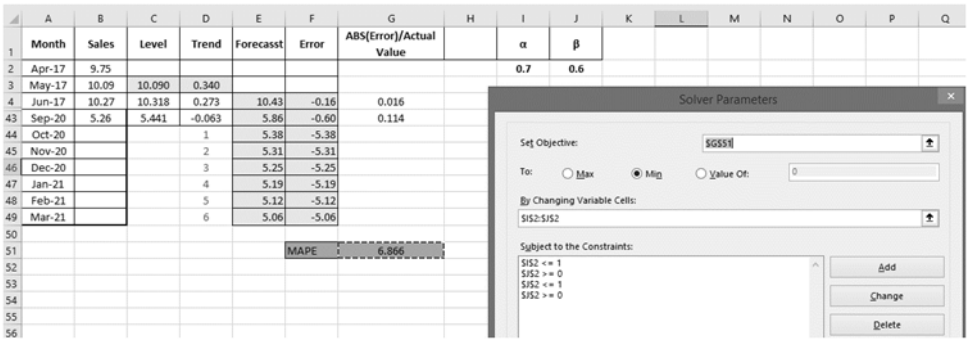
Jun-19	8.39	8.536	-0.889	8.88	-0.49
Jul-19	7.01	7.201	-1.157	7.65	-0.64
Aug-19	6.71	6.510	-0.877	6.04	0.67
Sep-19	5.93	5.841	-0.753	5.63	0.30
Oct-19	5.07	5.076	-0.760	5.09	-0.02
Nov-19	5.21	4.942	-0.384	4.32	0.89
Dec-19	6.46	5.889	0.415	4.56	1.90
Jan-20	6.35	6.336	0.434	6.30	0.05
Feb-20	5.36	5.783	-0.158	6.77	-1.41
Mar-20	5.25	5.362	-0.316	5.62	-0.37
Apr-20	4.94	4.972	-0.361	5.05	-0.11
May-20	5.13	4.974	-0.143	4.61	0.52
Jun-20	5.63	5.391	0.193	4.83	0.80
Jul-20	5.35	5.420	0.095	5.58	-0.23
Aug-20	5.74	5.672	0.189	5.51	0.23
Sep-20	5.26	5.441	-0.063	5.86	-0.60

Step 6 : Enter the numbers of the forecast horizon. It is just the number under the "Trend" column that starts with 1. Here we use the same last value (fixed excel cell) of Level and Trend for all next forecasted values, formula for the forecast is $\rightarrow F_{t+n} = L_t + (n)T_t$. Here n is the forecast horizon.

	A	B	C	D	E	F
1	Month	Sales	Level	Trend	Forecasst	Error
40	Jun-20	5.63	5.391	0.193	4.83	0.80
41	Jul-20	5.35	5.420	0.095	5.58	-0.23
42	Aug-20	5.74	5.672	0.189	5.51	0.23
43	Sep-20	5.26	5.441	-0.063	5.86	-0.60
44	Oct-20			1	5.38	-5.38
45	Nov-20			2	=C43+D45*D43	
46	Dec-20			3	5.25	-5.25
47	Jan-21			4	5.19	-5.19
48	Feb-21			5	5.12	-5.12
49	Mar-21			6	5.06	-5.06

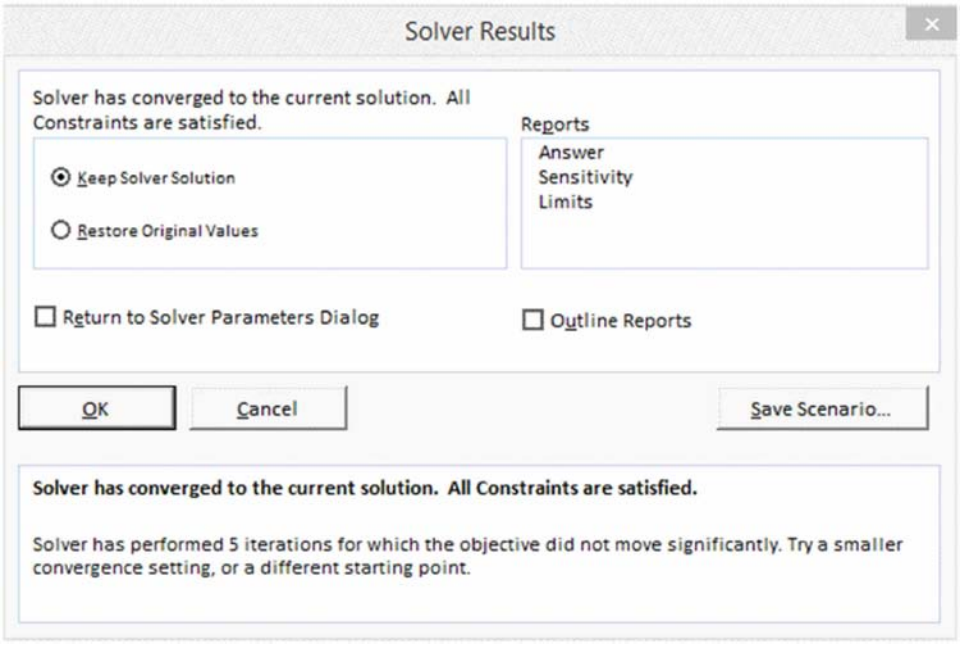
Step 7 : Calculate the forecasted accuracy like MAPE and RMSE. MAPE value for the above data is 6.86.

Step 8 : Optimize the error value by adjusting different values of α and β. It can be done using MS Excel functionality, solver. It is available under the "Data" tab.



Month	Sales	Level	Trend	Forecasst	Error	ABS(Error)/Actual Value		α	β
Apr-17	9.75							0.7	0.6
May-17	10.09	10.090	0.340						
Jun-17	10.27	10.318	0.273	10.43	-0.16	0.016			
Sep-20	5.26	5.441	-0.063	5.86	-0.60	0.114			
Oct-20			1	5.38	-5.38				
Nov-20			2	5.31	-5.31				
Dec-20			3	5.25	-5.25				
Jan-21			4	5.19	-5.19				
Feb-21			5	5.12	-5.12				
Mar-21			6	5.06	-5.06				
					MAPE	6.866			

Here we want to reduce the MAPE value (we selected as a minimum) by changing the values of α and β. We also put the constraint that these values can't be less than 0 and more than 1. Click the "Solve" button and excel using a linear programming approach to find the minimum value MAPE.



Select "Keep solver solution" and click on the OK button to see the improved MAPE value and changed values of α and β smoothing constants.

Trend	Forecast	Error	ABS(Error)/Actual Value		α	β
					0.906461	0.360152
0.340						
0.288	10.43	-0.16	0.016			
-0.069	5.82	-0.56	0.106			
1	5.24	-5.24				
2	5.17	-5.17				
3	5.11	-5.11				
4	5.04	-5.04				
5	4.97	-4.97				
6	4.90	-4.90				

MAPE	6.332

Here we can see MAPE has been reduced to 6.332 by adjusting α = .906461 and β = .360152.

Check Your Progress :

1. Regression based forecasting techniques are advisable if clear _____ is visible in the time series data
2. In regression equation, $F_t = \beta_0 + \beta_1 X_{1t}$; β_0 is known as _____ while β_1 is known as _____.
3. If a significant value under the ANOVA table in regression output is 0 then _____ shows significant relationship between input and output variables.
4. Double smoothing forecasting techniques has two smoothing constants, one for _____ and other for _____.
5. Smoothing constants have values between _____ and _____.

❖ **Multiple Choice Questions :**

1. Double smoothing forecasting techniques are applicable if the below effect is NOT visible:

   a. Trend b. Irregular or random

   c. Seasonality d. Level

2. Which of the below statements are true ?

I. Both smoothing constants must have the same value

II. Both smoothing constants can have a value between 0 and .5

   a. Only statement 1 is correct b. Only statement 2 is correct

   c. Both statements are correct d. Both statements are incorrect

3. Complete the Level (or Intercept) equation (L_t) : $L_t = \alpha \times Y_t +$ _____ from below options:

   a. $(1-\alpha) \times F_t$ b. $\alpha \times F_t$

   c. $(1-\beta) \times F_t$ d. $\beta \times F_t$

4. If the R–Square value is .72 in an MS Excel output of regression analysis, then it means:

   a. Input variables justify 72% variation in an output variable

   b. The researcher needs more input variables in the study

   c. Regression output cannot be trusted

   d. The researcher can use a regression model at a 72% confidence level only

5. Complete the trend equation (T_t) : $T_t = \beta \times (L_t - L_{t-1}) +$ _____ from below options:

   a. $(1-\alpha) \times F_t$ b. $\alpha \times F_t$

   c. $(1-\beta) \times F_t$ d. $\beta \times F_t$

6. As per unit content, MAPE/ RMSE value can be minimized by using the below functionality in MS Excel :
   a. Pivot table b. Solver
   c. Both the functionality d. None of the above
7. Solver optimises MAPE/ RMSE to :
   a. Maximum b. More than 1
   c. Minimum d. Less than 0

3.5 Let Us Sum Up :

1. Regression is a more powerful forecasting technique when the time–series has values of various independent variables also besides the dependent variable Y_t.
2. Sometimes simple forecasting techniques results better than complex techniques hence various forecasting models must be tried before selecting the final model.
3. The effectiveness of regression is generally more than other forecasting techniques as it includes both input and output variables.
4. MS Excel help us to visualize the trend line clearly, it also shows the regression equation and R–Square value in the graph itself.
5. The residual analysis must be performed before selecting the regression model.
6. Double exponential smoothing uses two equations to forecast the future values of the time series, one for forecasting the level (short term average value) and another for capturing the trend.
7. The double exponential smoothing technique uses two smoothing constants, α (for level/ intercept) and β (for trend).
8. The final forecast comes as a summation of trend and level components both.
9. By using MS Excel Solver functionality, the forecasting error can be reduced.
10. Solver adjusts the values of smoothing constants in order to optimize the forecasting error.

3.6 Answers for Check Your Progress :

Check Your Progress :

1. Trend **2.** Intercept, slope **3.** Overall model
4. Level, trend **5.** 0, 1

❖ **Multiple Choice Questions :**

1. c **2.** d **3.** a **4.** a
5. c **6.** b **7.** c

3.7 Glossary :

Regression Model for Forecasting : When information about input and output variables are available then regression–based forecasting techniques work better than other techniques which are based on output only.

Double Exponential Smoothing Forecasting Techniques : When the clear trend is visible in the time–series data then double exponential smoothing techniques generally works better, here we use two equations, one for level (intercept) and the other for trend. Final forecasts come as a summation of both equation.

Optimization of Smoothing Constants : We use linear programming techniques to adjust the values of smoothing constants α and β to minimize the forecasting error. In MS–Excel there is an optimizing technique in form of Solver.

Decomposition of Time–Series : Separating the impact of trend, seasonality and irregular components from a time series is known as decomposition of a time series.

3.8 Assignment :

1. What is the basic difference between regression-based and other forecasting techniques, why most of the time regression-based forecasting techniques works better in terms of accuracy of forecasts.
2. Write down the important steps to draw a trend line in a forecasting graph in MS Excel.
3. What is the basic difference between single and exponential forecasting techniques ? Write down the important scenarios where these techniques work better than other.

3.9 Activities :

Below are the sales (in ₹ Crore) data of Contesa Ice cream. Use double exponential smoothing technique to forecast sales for the next 5 days.

Date	Sales		Date	Sales
January 1, 2012	8.75		January 22, 2012	10.26
January 2, 2012	9.09		January 23, 2012	9.86
January 3, 2012	9.27		January 24, 2012	9.56
January 4, 2012	7.98		January 25, 2012	9.12
January 5, 2012	7.79		January 26, 2012	8.14
January 6, 2012	7.23		January 27, 2012	7.19
January 7, 2012	7.24		January 28, 2012	5.81
January 8, 2012	7.97		January 29, 2012	5.51
January 9, 2012	8.03		January 30, 2012	4.73
January 10, 2012	8.31		January 31, 2012	3.87
January 11, 2012	9.25		February 1, 2012	4.01
January 12, 2012	8.5		February 2, 2012	5.26
January 13, 2012	9.05		February 3, 2012	5.15
January 14, 2012	8.73		February 4, 2012	4.16
January 15, 2012	8.48		February 5, 2012	4.05
January 16, 2012	8.38		February 6, 2012	3.74
January 17, 2012	8.44		February 7, 2012	3.93
January 18, 2012	9.02		February 8, 2012	4.43
January 19, 2012	9.15		February 9, 2012	4.15
January 20, 2012	9.55		February 10, 2012	4.54
January 21, 2012	10.13		February 11, 2012	4.06

3.10 Case Study :

OTT based entertainment platforms become quite popular during the corona virus–based pandemic as most of the cinema halls and theatres were closed. Most of the big entertainment companies launched their OTT based platforms also big production houses and artists signed contracts with OTT platforms as these were gaining popularity during the pandemic. Below are TRP ratings of a famous web series.

Episode	TRP		Episode	TRP
1	7.98		11	8.5
2	9.8		12	9.03
3	9.54		13	9.82
4	7.28		14	9.77
5	9.62		15	10.77
6	9.8		16	9.46
7	7.9		17	9.57
8	8.26		18	9.81
9	8.17		19	9.92
10	8.36		20	9.98

Questions :

1. Is there any clear trend visible in the data, use MS Excel to visualize the trend also show R–Square and regression equation in the same graph ?
2. Which technique do you think will be more appropriate between double exponential smoothing technique or regression? Justify your selection
3. Use the initial value of both α and β constants as .6 but optimize them using solver functionality of MS Excel

3.11 Further Readings :

- "Time series analysis and Control," Holden Day, Box and Jenkins (1970)
- "How to get a better forecast, Harvard Business Review", Praker G, Segura E (1971)
- "Time Series Based Predictive Analytics Modelling: Using MS Excel", Glyn Davis, Branko Pecar; 1st edition (2016)
- "An introduction to Time series analysis and forecasting with applications of SAS and SPSS", Management science journal, Yaffee R, McGee M (2000)

Unit 4

AUTO–REGRESSION (AR) AND MOVING AVERAGE (MA) FORECASTING MODELS

: UNIT STRUCTURE :

4.0 Learning Objectives :

- Learn the concept of autocorrelation and how it impacts the time series analysis
- Understanding of Auto–regressive (AR) forecasting models
- Identification of the order of autoregression through MS Excel
- Understanding and application of moving average (MA) forecasting models

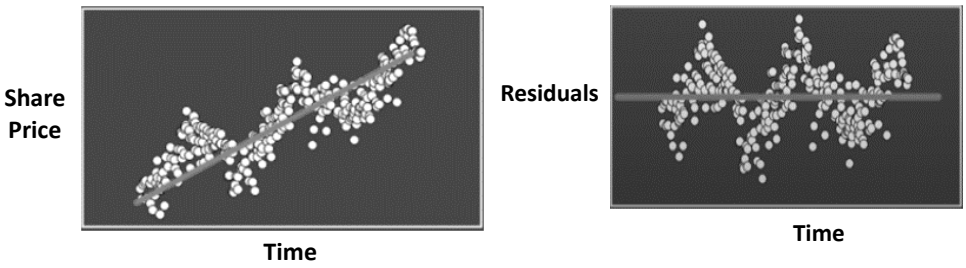
4.1 Introduction :

In this unit, we will study the concepts of autocorrelation and how autore–gression models help to overcome the problem of autocorrelation. We will see the difference between moving average based forecasting techniques studied in the last units and the moving average process of time–series modelling.

4.2 Introduction to Autocorrelation :

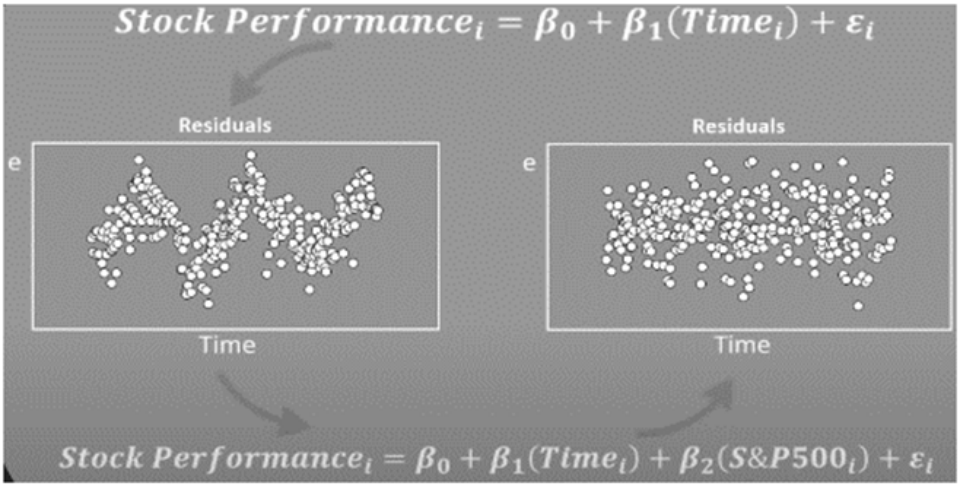
Business data generally collected in chronological order means older date's data comes first and then we have more recent data. This data cannot be completely independent of each other. Productivity of today must be inclined with yesterday's production similarly gold price, crude oil price or share prices etc. always depends on older values. Now this older impact may be of one day, two days or even much older. In other words, these metrics correlate with themselves which is a breach of the

regression assumption that there should not be any correlation among independent variables. When an independent variable shows correlation with itself then we called it autocorrelation. In the case of autocorrelation error terms of the regression model are correlated. As error terms in the case of autocorrelation are not random, this creates several problems in regression analysis. In the below example, share price shows the autocorrelation because error terms are no longer random. If we know the error term at a time t, we can predict the error term for next time t+1.



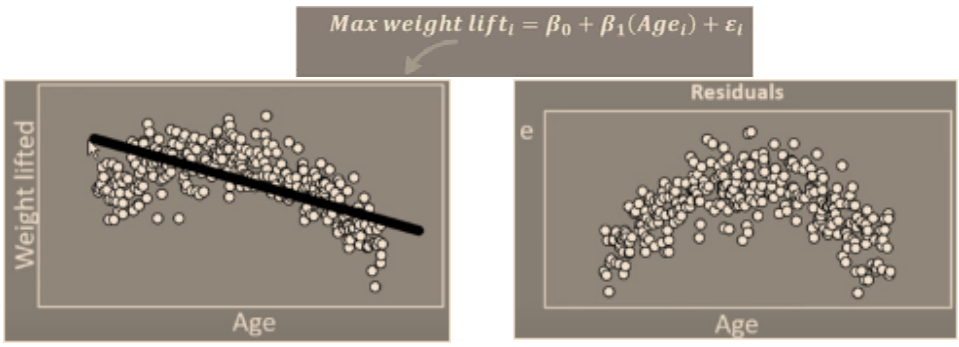
4.2.1 Reasons for Autocorrelation :

- **Missing of the Important Variable(s) from the Regression Model:** The important variable(s) is not part of the regression model then the impact of this missing variable will be seen in error terms. For example, in the above example share price also depends on the market so including that variable may reduce the impact of autocorrelation from the study

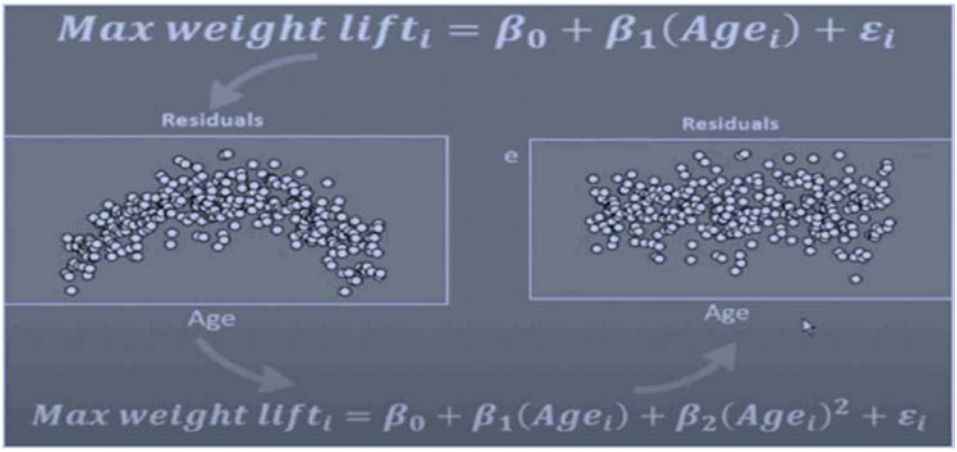


So, the inclusion of the market index (S&P500) variable must resolve the autocorrelation problem, now error terms are relatively more random

- **The Incorrect Functional form of a Regression Model :** In case, the scatter plot is not showing a linear relationship while we are writing an equation of linear regression equation then it may cause autocorrelation. For example, the maximum weight a person can lift shows a curvilinear model, young age we can lift less weight while maximum at the middle ages and again less weight during old age. But if we try to build a linear model then there may be an effect of autocorrelation.



A here linear equation is not appropriate to fit the data points, it seems that a quadratic model may fit the data better.



A quadratic regression model is fitting data well and removed the effect of autocorrelation from the model.

Autocorrelation can be a problem of non–time–series data also, few statisticians have this myth that autocorrelation problem occurs only in the case of time–series data.

4.2.2 Impact of Autocorrelation on a Regression Model :

Below are the important consequences of autocorrelation:

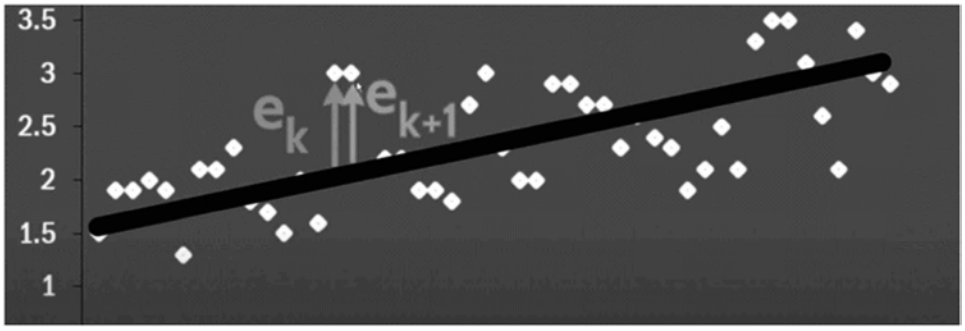
- The estimation of regression coefficients is not correct as they do not show minimum variance property
- The variance of the error term is greatly underestimated so it may look like a good regression model but, it is because of the autocorrelation problem and generally, these models work very badly for new (future) data
- The standard deviation of the regression coefficient may be calculated very less than actual
- Autocorrelations make the overall regression model unreliable (unlike multicollinearity, which does not impact the R2 value of the overall model) as confidence interval and hypothesis tests using F and t distribution have also become unreliable

Therefore, we need to correct the autocorrelation before evaluating the model performance.

4.2.3 Ways to Detect Autocorrelation : Durbin Watson Test :

There are two important ways to detect autocorrelation, first one is the "Durbin Watson test" which detects only first–order autocorrelation (only consecutive error terms are correlated) and the second one is the "Breusch–Godfrey test", which can detect autocorrelation of any order (error terms are corrected with any past error order of error terms). In this text, we will consider only the Durbin–Watson test as the second one is more complex mathematically and rarely used in day to day business analytics.

Durbin Watson Test : This test was developed by two statisticians in 1950. Here we capture the error terms in a regression model. We try to see the successive error terms if these are related.



So, if a positive error term is correlated with the very next positive error term then we call it positive autocorrelation. Similarly, if a positive error term is related to the next negative error term then it is known as negative autocorrelation. There may be the cases when error term at the k^{th} position is correlated with the $K+2^{nd}$ error term or any other error of higher–order but the Durbin–Watson test is only effective for the first order of autocorrelation.

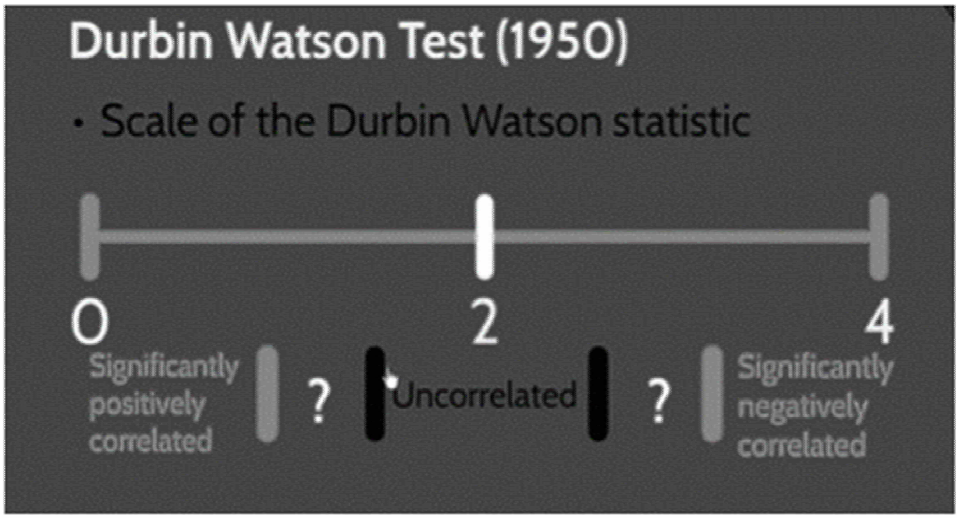
$$\text{dw statistics} = \frac{(e_2 - e_1)^2 + (e_3 - e_2)^2 + ... + (e_n - e_{n-1})^2}{e_1^2 + e_2^2 + ... e_n^2}$$

in the simple and generic form above equation can be rewritten as:

$$\text{dw statistics} = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} (e_t)^2}$$

We can observe that in numerator one term is less than the denominator.

Durbin–Watson statistics give us a value between 2 and 4. If the value is more towards 2 then it is positively autocorrelated and if it is more towards 4 then it is more negatively autocorrelated error terms.



These cut–off values of the Durbin–Watson test depend on two factors:

- Number of data points (n)
- Number of input variables in the regression model (k)

Worked Example : A leading laptop manufacturing company manufacture laptops and special circuits which require in few models. Check with the help of the Durbin–Watson test, if the data shows the autocorrelation at $\alpha = 0.05$.

Month	No of Laptops	Number of Special Ics Manufactured		Month	No of Laptops	Number of Special Ics Manufactured
1	17194	18005		13	14342	16597
2	17144	17769		14	13694	17123
3	17298	16241		15	13324	17783
4	17376	14585		16	13120	17714
5	17758	15945		17	12930	18169
6	17942	14967		18	12904	18276
7	17360	14586		19	12504	18161
8	16698	15239		20	11762	18022
9	16280	15899		21	11644	18158
10	15226	16141		22	11602	18561
11	14710	16486		23	11492	18227
12	14834	16450		24	11362	18251

Solution : Using MS Excel we can write the predicted regression equation. Below is the coefficient section of Excel regression output.

Regression	Coefficients
Intercept	23350.34888
No of Laptops	–0.436671535

Number of Special Ics Manufactured = 23350.34888 + (–0.436671535 * No. of Laptops)

For the first month, an error term can be calculated as:

Actual – Predicted (Regressed) = 18005 – 15842 = 2163

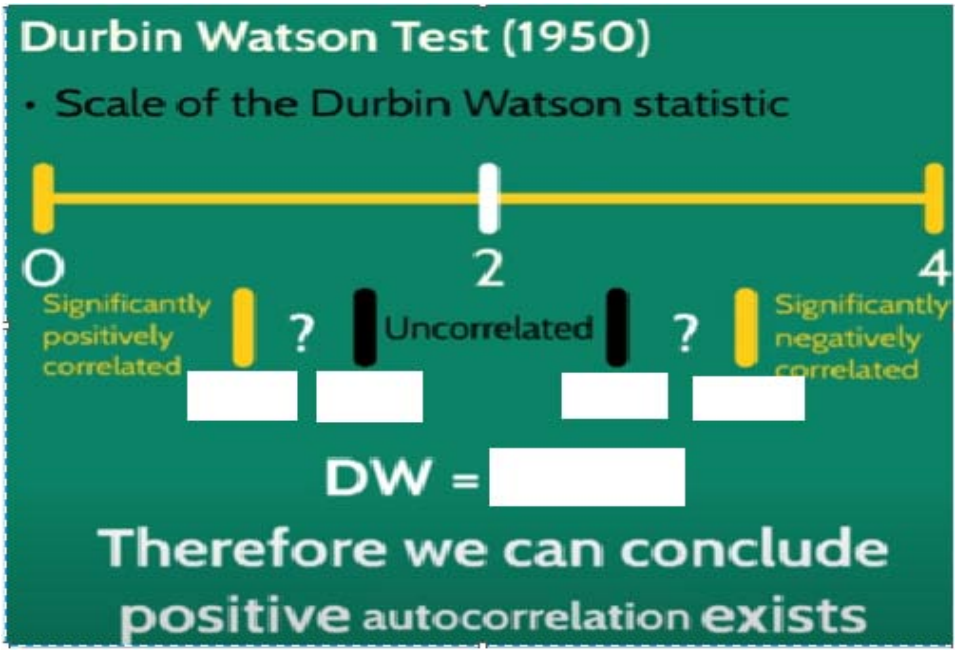
Difference of the first pair of consecutive error terms:

$e_{\text{Second month}} - e_{\text{First month}} = -258$

Similarly, we can calculate the error terms for the rest of the data points

Month	Y Predicted	e_t	e^2_t	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$
1	15842	2163	4677624	-	-
2	15864	1905	3628827	-258	66478
3	15797	444	197309	-1461	2133798
4	15763	-1178	1387082	-1622	2630688
5	15596	349	121846	1527	2331144
6	15516	-549	300949	-898	805780
7	15770	-1184	1401219	-635	403406
8	16059	-820	672084	364	132440
9	16241	-342	117194	477	227979
10	16702	-561	314259	-218	47634
11	16927	-441	194402	120	14323
12	16873	-423	178729	18	329
13	17088	-491	240694	-68	4603
14	17371	-248	61290	243	59067
15	17532	251	62932	498	248434
16	17621	93	8608	-158	24990
17	17704	465	216052	372	138408
18	17716	560	314116	96	9148
19	17890	271	73328	-290	83908
20	18214	-192	36948	-463	214379
21	18266	-108	11609	84	7136
22	18284	277	76682	385	147963
23	18332	-105	11050	-382	145950
24	18389	-138	19013	-33	1074
Sum			**14323847**		**9879058**

$$\text{dw statistics} = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} (e_t)^2} = \frac{9879058}{14323847} = 0.6897$$



Here, we have k = 1 as we have only one input variable, n = 24 (number of data points/observations) and α = 0.05 (given). We can see critical values from the Durbin–Watson table for α = 0.05, k = 1 and n = 24. The table gives us two critical values, one for the upper critical value (d_U) and the other is the lower critical value (d_L). For k = 1 and n = 24, d_L = 1.273 and d_U = 1.446. For right–hand side critical values, we can subtract these values from 4, we can get d_L = 2.554 and d_U = 2.727.

In the case of DW statistics value would have been in between d_L and d_U then we could not conclude anything. Below is the sample DW statistics table for α = .05 similar tables are available on the internet for different values of α.

Durbin-Watson Statistic: 5 Per Cent Significance Points of dL and dU

	k'=1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----
7	0.700	1.356	0.467	1.896	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----
8	0.763	1.332	0.559	1.777	0.367	2.287	----	----	----	----	----	----	----	----	----	----	----	----	----	----
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	----	----	----	----	----	----	----	----	----	----	----	----
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	----	----	----	----	----	----	----	----	----	----
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	----	----	----	----	----	----	----	----
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	----	----	----	----	----	----
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	----	----	----	----
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	----	----
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164

4.3 Autoregression : Remedy to Resolve Autocorrelation :

Autoregression resolves the autocorrelation problem by creating a multiple regression model of a variable on itself measured at different time periods (different time periods are known as lags). Here independent variables are the input variable itself and the same variable at lag1, lag2, leg3 etc. up to lag n. The optimum value of lags is determined by examining the regression output. A general equation for an autoregression model can be written as :

$$Y = \beta_0 + \beta_1 \times Y_{t-1} + \beta_2 \times Y_{t-2} + \ldots + \beta_n \times Y_n$$

Worked Example : For below "Number of special ICs" data, try to fit the third–order autoregression model if require fitting second or first–order regression model (Given α = .05). Also, forecast the value for the 25th month.

Month	Number of Special Ics Manufactured	Month	Number of Special Ics Manufactured
1	13541	13	12133
2	13305	14	12659
3	11777	15	13319
4	10121	16	13250
5	11481	17	13705
6	10503	18	13812
7	10122	19	13697
8	10775	20	13558
9	11435	21	13694
10	11677	22	14097
11	12022	23	13763
12	11986	24	13787

Solution : We can create data for lagged 1, lagged 2 and lagged 3 (as below)

Month	Number of Special Ics Manufactured	One Period Lagged Y_{t-1} (X1)	Two Period Lagged Y_{t-2} (X2)	Three Period Lagged Y_{t-3} (X3)
1	13541	-	-	-
2	13305	13541	-	-
3	11777	13305	13541	-
4	10121	11777	13305	13541
5	11481	10121	11777	13305
6	10503	11481	10121	11777
7	10122	10503	11481	10121
8	10775	10122	10503	11481
9	11435	10775	10122	10503
10	11677	11435	10775	10122
11	12022	11677	11435	10775
12	11986	12022	11677	11435
13	12133	11986	12022	11677
14	12659	12133	11986	12022
15	13319	12659	12133	11986

16	13250	13319	12659	12133
17	13705	13250	13319	12659
18	13812	13705	13250	13319
19	13697	13812	13705	13250
20	13558	13697	13812	13705
21	13694	13558	13697	13812
22	14097	13694	13558	13697
23	13763	14097	13694	13558
24	13787	13763	14097	13694
		13787	13763	14097
			13787	13763
				13787

A regression model can be developed only for the rows where data is available for input variables X_1, X_2 and X_3. Hence, we can remove the first three rows and the last three rows. We will consider the first column as the output variable and the remaining three columns for input variables. Below is the regression output:

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.8896				
R Square	0.7914				
Adjusted R Square	0.7545				
Standard Error	664.0296				
Observations	21.0000				
ANOVA					
	df	***SS***	***MS***	***F***	***Significance F***
Regression	3	28432295	9477432	21	0
Residual	17	7495900	440935		
Total	20	35928195			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	1254.631	1565.633	0.801	0.434	
X Variable 1	0.957	0.215	4.452	0.000	
X Variable 2	-0.086	0.298	-0.290	0.775	
X Variable 3	0.036	0.221	0.161	0.874	

Here the higher value of R^2 and adjusted R^2 shows decent predictability of the regression model. Here we can see that only the first X variable (One period lagged) is showing a significant relationship between output and input variable.

It is always advisable that lagged period should be chosen by consulting with domain/ business experts and visualizing past data. If data has been captured on daily basis then 7 days lagged may make sense as there may be some cyclical effect during different weekdays and weekends. Similarly, if data is captured quarterly then 4 quarter lagged data may make more sense as 4 quarters complete one yearly cycle.

4.4 Moving Average Model MA(q) :

Moving average (MA) processes are regression models in which the past residuals are used for forecasting future values of the time–series data. The moving average process is different from the moving average technique discussed in earlier units except that the regression model of the MA process can be considered as a weighted moving average of past residuals. Moving average process of lag 1, MA (1) is as follows:

$$Y_{t+1} = \omega + \theta e_{t-1} + \varepsilon_t$$

MA (1) process uses the previous residual, ε_t to forecast the next value of the time series. The reasoning behind the MA process is that the error at the current period, ε_t, and the error at the next period, ε_{t+1}, drive the next value of the time series Y_{t+1}, A moving average process with q lags, MA(q) process, is given by

$$Y_{t+1} = \omega + \theta_1 e_t + \theta_2 e_{t-1} + \ldots. + \theta_q e_{t-q+1} + \varepsilon_{t+1}$$

Check Your Progress :

1. AR (1) models depends only on past values of the time series, these input variables of past values are known as _____
2. $Y_t = \beta_0 + \beta_1 Y_{t-1} + e_t$ is an example _____ autoregression model.
3. Moving average Forecast model based on past _____
4. Range of Durbin–Watson statistics is always between _____ and _____.
5. Durbin–Watson is useful to identify only first order _____ in the time–series.

❖ **Multiple Choice Questions :**

1. The problem created by autocorrelation:
   a. One variable correlates with itself which denies regression assumptions
   b. Error terms are not random
   c. Both of above
   d. None of the above

2. What are important reasons for autocorrelation
   a. Missing of the important variable(s) from the regression model
   b. The incorrect functional form of a regression model
   c. None of the above
   d. Both options a and b are correct
3. Due to autocorrelation, we cannot estimate the regression coefficients because of:
   a. Minimum variance property
   b. Non–randomness of the error term
   c. Seasonality component in the time series
   d. None of the above
4. AR models can be used when:
   a. There is no autocorrelation effect in the data
   b. Data is stationary
   c. When there is more than one variable
   d. None of the above
5. Auto–regressive models are regression models where:
   a. Y_t is the output variable
   b. Y_{t-1}, Y_{t-2} etc are independent variable
   c. The error term is not an independent variable
   d. All the above are correct
6. For not showing autocorrelation, Durbin–Watson statistics should be around:
   a. Statistics value should be around 2
   b. Statistics value should be around 0
   c. Statistics value should be around 4
   d. Statistics value should be less than 1
7. In the Durbin–Watson test if the value of the statistics is in between d_U and d_L
   a. There is no autocorrelation
   b. Result is inconclusive
   c. There is a strong autocorrelation
   d. There is weak autocorrelation

8. In the Durbin–Watson test, which of the below statements are true:
   a. If the test statistics value is around zero, there is a positive autocorrelation
   b. If the test statistics value is around four, there is a negative autocorrelation
   c. If the test statistics value is around two, there is no autocorrelation
   d. All the above are true
9. For moving average of lag 1, MA (1) model
   a. Past residual is the independent variable
   b. Error terms correlate with one error term before
   c. Both a and b statements are wrong
   d. Both a and b options are correct
10. If data is showing an autoregressive model of third–order then which test can be used to validate it
   a. Durbin–Watson test
   b. Breusch–Godfrey test
   c. Both above test
   d. None of the tests given in options a and b

4.5 Let Us Sum Up :

1. Most of the economic phenomenon shows autocorrelation where data of the current day depend on historic data.
2. If today's data depends on yesterday's data, then it is known as first–order auto–correlation but if today's data depends on the day before yesterday's data then it is second–order autocorrelation
3. If autocorrelation is there, then we cannot rely on the regression coefficient
4. The variance of the error term is greatly underestimated so it may look like a good regres–sion model but, it is because of the autocorrelation problem and generally, these models work very badly for new (future) data
5. The standard deviation of the regression coefficient may be calculated very less than actual. The first–order autocorrelation can be detected by the Durbin–Watson test
6. Any order autocorrelation can be detected by the Breusch–Godfrey test
7. Moving average models, MA(q) use past error terms as independent variables

4.6 Answers for Check Your Progress :

Check Your Progress :

1. Lags **2.** First–order **3.** Errors

4. 1 and 4 **5.** Autocorrelation

❖ **Multiple Choice Questions :**

1. c **2.** d **3.** a **4.** b

5. d **6.** a **7.** b **8.** d

9. d **10.** b

4.7 Glossary :

Autocorrelation : Autocorrelation is a mathematical representation where a time–series data shows correlation with a lagged version of itself.

Durbin–Watson Test : This test helps to detect autocorrelation of the first order when a time series data correlate with lag 1 data of itself.

Breusch–Godfrey Test : This test helps to detect autocorrelation of any order.

Autoregressive (AR) Models : AR Models forecast time–series data using a linear combination of its lagged values. Here the output variable is Y_t while input variables are Y_{t-1}, Y_{t-2} etc.

Moving Average MA(q) Models : Moving average models use past forecast errors as input variables. MA (q) models are also known as the moving average process.

4.8 Assignments :

1. What could be important consequences of ignoring the auto-correlation problem in a forecasting technique.
2. Write down two important reasons for autocorrelation in a forecasting technique.
3. What are two important ways to detect auto-correlation, write down the appropriate scenarios where these techniques can be utilized.
4. What is the basic difference between positive and negative autocorrelation, explain with an example.

4.9 Activities :

Laburnum chemicals limited has issued their production data for one of the agricultural pesticide products for the last 26 years.

Year	Unit Produced	Year	Unit Produced
1	1196	14	2330
2	1536	15	3792
3	1726	16	2756
4	1636	17	3152
5	1682	18	2920
6	2280	19	3178
7	2430	20	3512
8	2818	21	3446
9	2526	22	3780
10	2468	23	3166
11	2892	24	3350
12	2672	25	3744
13	2486	26	3690

Check the first–order autocorrelation in the above data with the help of the Durbin–Watson test also develop a one–period (lag 1) two–period (lag 2) AR model and check if there is significant autocorrelation.

4.10 Case Study :

Vedic Fabrics Pvt Limited conducts an employee satisfaction survey every year, below is the data for the last 32 years. Below are two important factors – Salary satisfaction and Motivation at work

Year	Salary Satisfaction	Motivation at work	Year	Salary Satisfaction	Motivation at work
1989	8.5		2005		
1990	7.8		2006		
1991	4.1		2007		
1992	2.3		2008		
1993	3.7		2009		
1994	2.3		2010		
1995			2011		
1996			2012		
1997			2013		
1998			2014		
1999			2015		
2000			2016		
2001			2017		
2002			2018		
2003			2019		
2004			2020		

Questions :

1. Is there any clear trend visible in the data, use MS Excel to visualize the trend also show R–Square and regression equation in the same graph ?
2. Calculate a Durbin–Watson test to check if there is a first–order autocorrelation at $\alpha = 0.05$
3. Develop an autoregressive model with a one–period lag and then develop a model with two–period lag. Compare the results and write your observations.

4.11 Further Readings :

- "Time series analysis and Control," Holden Day, Box and Jenkins (1970)
- "How to get a better forecast, Harvard Business Review", Praker G, Segura E (1971)
- "Time Series Based Predictive Analytics Modelling: Using MS Excel", Glyn Davis, Branko Pecar; 1st edition (2016)
- "An introduction to Time series analysis and forecasting with applications of SAS and SPSS", Management science journal, Yaffee R, McGee M (2000)

BLOCK SUMMARY

Business analytics is the brain of the organization while forecasting techniques are the spinal cord. Forecasting was one of the fundamental requirements to establish business analytics as a separate department in a leading organization. Better forecasting is the key difference between successful and struggling organizations, in the last two decades supply chain management has emerged as one of the most important departments and time–series forecasting is the heart of successful supply chain management. Forecasting techniques like MAPE and RMSE can be used to benchmark different industries, these parameters have gained huge importance in most quality standard certifications like ISO, CMMI etc. Most of the fundamental features for forecasting are available in MS Excel while few advanced softwares make forecasting quite easy, for example, e–views, Minitab, SAS etc. Forecasting techniques are important for both short term and long term planning in the organization. Good forecasting helps us to reduce warehouse cost, additional manpower, additional bandwidth and server capacity. It also helps in optimizing manpower cost, budgeting, revenue management etc.

BLOCK ASSIGNMENT

Short Answer Questions :

1. What are the different components of a time series ?
2. How forecasting techniques impact both the top and bottom line of an organization
3. Write a short note on additive and multiplicative forecasting models and write down the ideal scenarios in which we can apply each of these models
4. Write down the difference between weighted average and moving average forecasting techniques
5. Write a short note on autoregression and what are important reasons for it
6. Describe the basic concept of the Durbin–Watson test to detect the first–order autocorrelation
7. Write a short note on the Moving average, MA(q) forecasting model
8. Write down different factors affecting forecasting accuracy of a time–series

Long Answer Questions :

1. Explain the benefits of forecasting in the business world with the help of few examples in the banking, retail and textile industries
2. Write down different forecasting accuracy techniques, which techniques are suitable for benchmarking across different industries
3. Explain the intuition behind single and double exponential smoothing techniques. Take data from with assignment question and use MS Excel solver the optimize the smoothing constants
4. Explain the difference between autoregressive models and the moving average MA (q) forecasting model. Write down the similarity between moving average techniques and moving average forecasting model MA (q)
5. Write down impacts of autocorrelation

❖ **Enrolment No. :** ☐

1. How many hours did you need for studying the units ?

Unit No.	1	2	3	4
No. of Hrs.				

2. Please give your reactions to the following items based on your reading of the block :

Items	Excellent	Very Good	Good	Poor	Give specific example if any
Presentation Quality	☐	☐	☐	☐	_____
Language and Style	☐	☐	☐	☐	_____
Illustration used (Diagram, tables etc)	☐	☐	☐	☐	_____
Conceptual Clarity	☐	☐	☐	☐	_____
Check your progress Quest	☐	☐	☐	☐	_____
Feed back to CYP Question	☐	☐	☐	☐	_____

3. Any other Comments

.....

.....

.....

.....

.....

.....

.....

.....

DR.BABASAHEB AMBEDKAR OPEN UNIVERSITY

'Jyotirmay' Parisar,
Sarkhej-Gandhinagar Highway, Chharodi, Ahmedabad-382 481.
Website : www.baou.edu.in