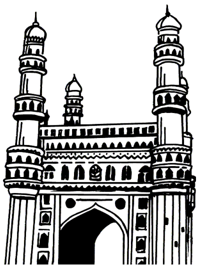
Rahul's ✔
Topper's Voice

R19
New Syllabus



JNTU (H)
MBA

II Year III Semester

Latest **2022** Edition

DATA ANALYTICS

- Study Manual
- FAQ's and Important Questions
- Short Question and Answers
- Solved Model Papers
- Solved Previous Question Papers

Price
₹. 159-00

- by -

WELL EXPERIENCED LECTURER

Rahul Publications™
Hyderabad. Ph : 66550071, 9391018098

JNTU(H) MBA

II Year III Semester

DATA ANALYTICS

Inspite of many efforts taken to present this book without errors, some errors might have crept in. Therefore we do not take any legal responsibility for such errors and omissions. However, if they are brought to our notice, they will be corrected in the next edition.

Price ₹. 159 -00

Sole Distributors : **☎ : 66550071, Cell : 9391018098**

VASU BOOK CENTRE

Shop No. 2, Beside Gokul Chat, Koti, Hyderabad.
Maternity Hospital Opp. Lane, Narayan Naik Complex, Koti, Hyderabad.
Near Andhra Bank, Subway, Sultan Bazar, Koti, Hyderabad -195.

CONTENTS

DATA ANALYTICS

STUDY MANUAL

SOLVED PREVIOUS QUESTION PAPERS

SOLVED MODEL PAPERS

SYLLABUS

UNIT - I

Introduction to Data Analytics: Introduction to Data- Importance of Analytics- Data for Business Analytics - Big Data - Business Analytics in Practice. Data Visualization - Data Visualization tools, Data queries, Statistical methods for Summarizing data, Exploring data using pivot tables.

UNIT - II

Descriptive Statistical Measures: Population and samples, Measures of location, Measures of Dispersion, Measures of variability, measures of Association. Probability distribution and Data Modeling – Discrete Probability distribution, Continuous Probability distribution, Random sampling from Probability Distribution, Data Modeling and Distribution fitting.

UNIT - III

Predictive Analytics : Karl Pearson Correlation Techniques - Multiple Correlation-Spearman's rank correlation-Simple and Multiple regression-Regression by the method of least squares - Building good regression models - Regression with categorical independent variables - Linear Discriminant Analysis - One way and Two way ANOVA.

UNIT - IV

Data Mining: Scope of Data Mining, Data Exploration and Reduction, Unsupervised learning – cluster analysis, Association rules, Supervised learning- Partition Data, Classification Accuracy, prediction Accuracy, k-nearest neighbors, Classification and regression trees, Logistics Regression.

UNIT - V

Simulation : Random Number Generation, Monte Carlo Simulation, What if Analysis, Verification and Validation, Advantages and Disadvantages of Simulation, Risk Analysis, Decision Tree Analysis.

Contents

Topic **Page No.**

UNIT - I

UNIT - II

UNIT - III

Topic **Page No.**

UNIT - IV

UNIT - V

Frequently Asked & Important Questions

UNIT - I

1. **What are the Data Visualization Techniques?**

Ans : **(Aug.-21, Nov.-20, May-19, Dec.-18, Imp.)**

Refer Unit-I, Q.No. 15.

2. **Discuss briefly about role of business analytics in current business environment.**

Ans : **(Aug.-21, Nov.-20, May-19, Dec.-18, Imp.)**

Refer Unit-I, Q.No. 9.

3. **What is Big Data? Explain the various technologies in big data?**

Ans : **(Nov.-20)**

Refer Unit-I, Q.No. 5.

4. **Explain the different statistical methods for summerizing data.**

Ans : **(Aug.-21, Dec.-19)**

Refer Unit-I, Q.No. 18.

5. **How to explore data in pivot tables?**

Ans : **(Nov.-20, Dec.-19, Imp.)**

Refer Unit-I, Q.No. 19.

UNIT - II

1. **Explain briefly about Measures of Association?**

Ans : **(Nov.-20, Dec.-18, Imp.)**

Refer Unit-II, Q.No. 5.

2. **Explain about Data Modelling?**

Ans : **(Nov.-20, Dec.-19)**

Refer Unit-II, Q.No. 14.

3. **Explain the various ways of Measure of variability?**

Ans : **(Nov.-20)**

Refer Unit-II, Q.No. 4.

4. **What is probability distribution? Explain the two classes of probability distribution?**

Ans : **(Dec.-19)**

Refer Unit-II, Q.No. 6.

5. **What is continuous probability distribution explain with its example.**

Ans : **(Dec.-19)**

Refer Unit-II, Q.No. 12.

UNIT - III

1. **What is regression by the method of lease square?**

Ans : **(Nov.-20, Dec.-19, Imp.)**

Refer Unit-III, Q.No. 15.

2. **What is regression with categorical independent variables?**

Ans : **(Dec.-19, Imp.)**

Refer Unit-III, Q.No. 19.

3. **Explain about linear discriminant analysis.**

Ans : **(Dec.-18)**

Refer Unit-III, Q.No. 20.

4. Discuss briefly about simple and multiple regression.

Ans : **(Dec.-19, May-19)**

Refer Unit-III, Q.No. 12, 14.

5. Explain One Way ANOVA with an example?

Ans : **(Imp.)**

Refer Unit-III, Q.No. 22.

UNIT - IV

1. Explain about cluster analysis with an example.

Ans : **(Dec.-19, May-19)**

Refer Unit-IV, Q.No. 6.

2. What is data exploration ? Explain the Steps of Data Exploration and Preparation

Ans : **(Aug.-21, May-19, Imp)**

Refer Unit-IV, Q.No. 3.

3. What is data mining. Explain the scope of data mining.

Ans : **(May-19, Imp)**

Refer Unit-IV, Q.No. 1, 2.

4. What is data reduction and explain data reduction techniques?

Ans : **(Dec.-18, Imp.)**

Refer Unit-IV, Q.No. 4.

5. What is supervised learning explain the steps involves to solve a given problem of supervised learning?

Ans : **(Imp.)**

Refer Unit-IV, Q.No. 9.

UNIT - V

1. Explain the advantages and dis advantages of simulation?

Ans : **(Aug.-21, Nov.-20, Dec.-19, Imp.)**

Refer Unit-V, Q.No. 14.

2. Explain about Monte Carlo simulation and explain how it works.

Ans : **(May-19, Dec.-18, Imp.)**

Refer Unit-V, Q.No. 6.

3. Explain Decision tree analysis with an example.

Ans : **(Dec.-19, Imp.)**

Refer Unit-V, Q.No. 18.

4. Discuss in detail various techniques of risk analysis.

Ans : **(Dec.-18)**

Refer Unit-V, Q.No. 17.

5. Explain about risk analysis with its benefits and uses?

Ans : **(Imp.)**

Refer Unit-V, Q.No. 15.

UNIT I

Introduction to Data Analytics: Introduction to Data- Importance of Analytics- Data for Business Analytics - Big Data - Business Analytics in Practice. Data Visualization - Data Visualization tools, Data queries, Statistical methods for Summarizing data, Exploring data using pivot tables.

1.1 Introduction to Data

Q1. What is Data?

Ans :

Definition of data

According to Hicks [1993: 668] quoted by Checkland and Holwell [1998]

Data

A representation of facts, concepts or instructions in a formalised manner suitable for communication, interpretation, or processing by humans or by automatic means.

Three aspects of data can be identified.

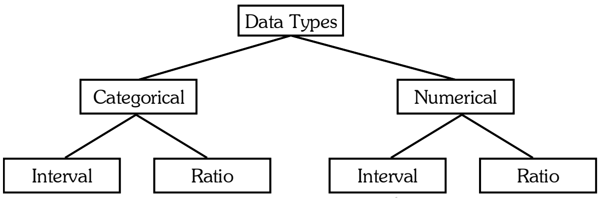
- Record objective facts which will be understood in exactly the same way by everyone;
- Record absolutely any type of concept, with no guarantees as to its accuracy or validity,

Which will be interpreted in all sorts of different ways by individuals?

- Use agreed structures and conventions for representing information, recording it and Transmi-tting it, all in order to communicate it.

Types of Data

Data Types are an important concept of statistics, which needs to be understood, to correctly apply statistical measurements to your data and therefore to correctly conclude certain assumptions about it.



1. Categorical Data

Categorical data represents characteristics. Therefore it can represent things like a person's gender, language etc. Categorical data can also take on numerical values

It has two types, They are :

i) Nominal Data

ii) Ordinal Data

i) Nominal Data

Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as "labels". Note that nominal data that has no order.

ii) Ordinal Data

Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters.

2. Numerical Data

It has two types. They are:

i) Interval data

ii) Ratio data

i) **Interval Data:** Interval values represent ordered units that have the same difference. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values.

ii) **Ratio Data :** Ratio values are also ordered units that have the same difference. Ratio values are the same as interval values, with the difference that they do have an absolute zero. Good examples are height, weight, length etc.

Q2. What is data analytics? Explain the different types of data analytics application?

Ans :

Definition

Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements. Data analytics is also known as data analysis.

- **Exploratory data analysis:** At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to find patterns and relationships in data, and
- **Confirmatory data analysis (CDA),** which applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work. CDA is akin to the work of a judge or jury during a court trial - a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis
- **Quantitative data**: Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically.
- **Qualitative approach:** It is more interpretive - it focuses on understanding the content of non-numerical data like text, images, audio and video, including common phrases, themes and points of view.

1.2 Importance of Analytics

Q3. Explain the Importance of Analytics.

Ans :

As “Data Analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain.” ... Let’s dig more deep into the conceptual understanding of DataAnalytics and how it is important from a business perspective.

- **Important for your business:** To the extent that making good decisions is. The practice of analytics is all about supporting decision making by providing the relevant facts that will allow you to make a better decision. And allows you to make decisions on a scale that can hardly be believed.
- **Data increasing:** It increases at a rapid speed and the rate of growth of information is very high. Data generation occurs through many users, industries, and businesses. It is crucial to amalgamate this data that have been generated through the business. If it gets wasted, lots of valuable information will be lost. Previously, skilled analysts were required for processing the data but these days there are tools used for high-speed data and this helps in incorporating the data analytics at the time of making decisions
- **Improving Efficiency:** All the data collected by the business is not only related to the individuals external to the organization. Most of the data collected by the businesses are analyzed internally. Market Understanding With the development of algorithms nowadays, huge datasets can be collated and analyzed. This process of analysis is called Mining. Regarding the other kinds of physical resources, data collection is done in raw form and thereafter refined. Industry Knowledge Industry knowledge can be comprehended and it can show how a business can run in

the near future. Also, it can tell you what kind of economy is already available for business expansion purpose.

- **Cost Reduction**: Big data technologies like cloud-based analytics and Hadoop can bring huge cost advantages if it relates to storage of large data. They can also identify the efficient ways to do business. Faster and Better Decision-Making The high-speed in-memory analytics and Hadoop in combination with the ability for analyzing the new data sources, businesses can analyze the information almost instantly. New Products/ Services With the power of Data Analytics, the needs and satisfaction of the customers are met in a better way. This helps one to make sure that the product/service aligns with the values of the target audience

- **Importance of Data Analytics is truly changing the world**: Whether it is the sports, the business field, or just the day-to-day activities of the human life, data analytics have changed the way people used to act. It now, not plays a major role in business, but too, is used in developing artificial intelligence, track diseases, understand consumer behavior and mark the weaknesses of the opponent contenders in sports or politics. This is the new age of data and it has unlimited potential.

Q4. What is data for business analytics ?

Ans :

Data is used to gain insights that inform business decisions and can be used to automate and optimize business processes. Data-driven companies treat their data as a corporate asset and leverage it for a competitive advantage. Successful business analytics depends on data quality, skilled analysts who understand the technologies and the business, and an organizational commitment to data-driven decision-making.

Business Analytics Techniques

Business analytics techniques break down into two main areas.

1. **The first is basic business intelligence**

   This involves examining historical data to get a sense of how a business department, team or staff member performed over a particular time. This is a mature practice that most enterprises are fairly accomplished at using.

2. **The second area of business analytics involves deeper statistical analysis**

   This may mean doing predictive analytics by applying statistical algorithms to historical data to make a prediction about future performance of a product, service or website design change. Or, it could mean using other advanced analytics techniques, like cluster analysis, to group customers based on similarities across several data points. This can be helpful in targeted marketing campaigns, for example.

   - **Descriptive analytics,** which tracks key performance indicators to understand the present state of a business;

   - **Predictive analytics,** which analyzes trend data to assess the likelihood of future outcomes; and

   - **Advanced areas of business analytics**: It can start to resemble data science, but there is a distinction. Even when advanced statistical algorithms are applied to data sets, it doesn't necessarily mean data science is involved. There are a host of business analytics tools that can perform these kinds of functions automatically, requiring few of the special skills involved in data science.

   - **True data science:** It involves more custom coding and more open-ended questions. Data scientists generally don't set out to solve a specific question, as most business analysts do. Rather, they will explore data using advanced statistical methods and allow the features in the data to guide their analysis.

Business analytics tools come in several different varieties

1. Data visualization tools
2. Business intelligence reporting software
3. Self-service analytics platforms
4. Statistical analysis tools
5. Big data platforms

- **Self-service** : It has become a major trend among business analytics tools. Users now demand software that is easy to use and doesn't require specialized training. This has led to the rise of simple-to-use tools from companies such as Tableau and Qlik, among others. These tools can be installed on a single computer for small applications or in server environments for enterprise-wide deployments. Once they are up and running, business analysts and others with less specialized training can use them to generate reports, charts and web portals that track specific metrics in data sets
- **Data Acquisition :** Once the business goal of the analysis is determined, an analysis methodology is selected and data is acquired to support the analysis. Data acquisition often involves extraction from one or more business systems, data cleansing and integration into a single repository, such as a data warehouse or data mart. The analysis is typically performed against a smaller sample set of data.
- **Analytics Tools**: Range from spreadsheets with statistical functions to complex data mining and predictive modeling applications. As patterns and relationships in the data are uncovered, new questions are asked, and the analytical process iterates until the business goal is met.
- **Deployment of Predictive Models**: It involves scoring data records - typically in a database - and using the scores to optimize real-time decisions within applications and business processes. BA also supports tactical decision-making in response to unforeseen events. And, in many cases, the decision-making is automated to support real-time responses.

1.3 Big Data

Q5. What is Big Data? Explain the various technologies in big data?

Ans :

Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.

Big data is a term that is used to describe data that is high volume, high velocity, and/ or high variety; requires new technolo- gies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

SAS describes Big Data as "a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis." What's important to keep in mind about Big Data is that the amount of data is not as important to an organization as the analytics that accompany it. When companies analyze Big Data, they are using Business Analytics to get the insights required for making better business decisions and strategic moves.

Technologies in big data

We can categorise them into two (storage and Querying/Analysis).

- Apache Hadoop. Apache Hadoop is a java based free software framework that can effectively store large amount of data in a cluster. ...
- Microsoft HDInsight. ...
- NoSQL. ...
- Hive. ...
- Sqoop. ...
- PolyBase. ...
- Big data in EXCEL. ...
- Presto

- Some of the most common of those big data challenges include the following:
- Dealing with data growth. ...
- Generating insights in a timely manner. ...
- Recruiting and retaining big data talent. ...
- Integrating disparate data sources. ...
- Validating data. ...
- Securing big data. ...
- Organizational resistance

Q6. Explain the evolution of Big Data.

Ans :

The evolution of big data is discussed below,

(i) 1970s and before

(ii) 1980s and 1990s

(iii) 2000s and beyond

(i) 1970s and before : The data generation and storage of 1970s and before is fundamentally primitive and structured. This era is termed as the era of mainframes, as it stores the basic data.

(ii) 1980s and 1990s : In 1980s and 1990s the evolution of relational data bases took please. The relational data utilization is complex and thus this era comprises of data intensive applications.

(iii) 2000s and beyond : The World Wide Web (www) and the Internet of Things (IOT) have an aggression of structured, unstructured and multimedia data. The data driven is complex and unstructured.

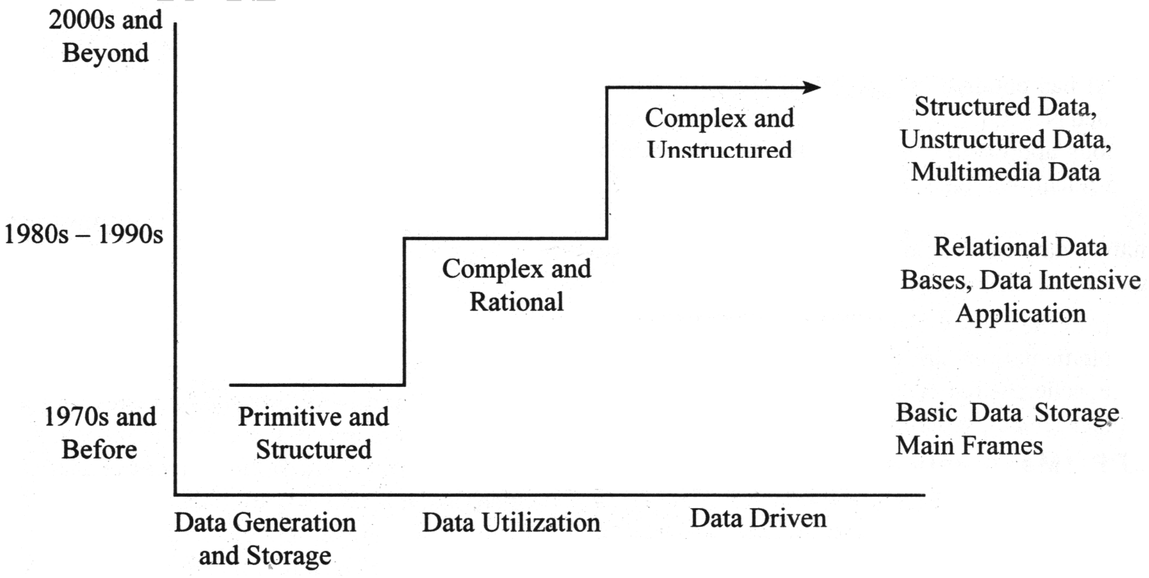

2000s and Beyond
1980s – 1990s
1970s and Before
Complex and Unstructured
Complex and Rational
Primitive and Structured
Structured Data, Unstructured Data, Multimedia Data
Relational Data Bases, Data Intensive Application
Basic Data Storage Main Frames
Data Generation and Storage
Data Utilization
Data Driven

Fig.: The Evolution of Big Data

Q6. Explain the dimensions of big data?

Ans :

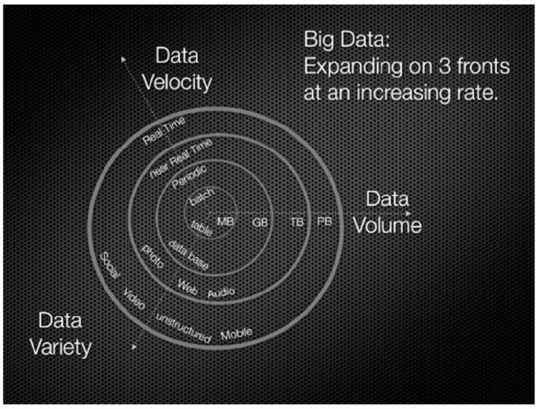
Big data refers to datasets whose size is beyond the ability of typical database software tool to capture, store, managed and analyze. Big data is data that goes beyond the traditional limits of data along four dimensions:

i) Volume

ii) Variety

iii) Velocity

iv) Variability



i) **Data Volume:** Data Volume can be measured by quality of transactions, events and amount of history. Big Data isn't just a description of raw volume. The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That's where Big Data analytics become necessary.

ii) **Data Variety :** It is the assortment of data. Traditionally data, especially operational data, is structured as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.). Wide variety of data: Internet data(Social media ,Social Network-Twitter, Face book),

iii) **DataVelocity** : It is the measure of how fast the data is coming in. Remember our Facebook example. 250 billion images may seem like a lot. But if you want your mind blown, consider this: Facebook users upload more than 900 million photos a day. So that 250 billion number from last year will seem like a drop in the bucket in a few months. Facebook has to handle a tsunami of photographs every day. It has to ingest it all, process it, file it, and somehow, later, be able to retrieve it.

iv) **Variability** : The increase in the range of values typical of a large data set - and value, which addresses the need for valuation of enterprise data.

Q7. Explain how data can be collected for research ?

Ans :

Primary Research (Surveys, Experiences, Observations),

Secondary Research: (Competitive and Market place data, Industry reports, Consumer data, Business data), Location data (Mobile device data,Geospatial data), Image data (Video, Satellite image, Surveillance), Supply Chain data (vendor Catalogs, Pricing etc), Device data (Sensor data, RF device, Telemetry)

Structured Data: They have predefined data model and fit into relational database. Especially, operational data is ¯structured as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.)

Semi-structured data: These are data that do not fit into a formal structure of data models. Semi-structured data is often a combination of different types of data that has some pattern or structure that is not as strictly defined as structured data. Semi-structured data contain tags that separate semantic elements which includes the capability to enforce hierarchies within the data.

Unstructured data : Do not have a predefined data model and /or do not fit into a relational database. Oftentimes, text, audio, video, image, geospatial, and Internet data (including click streams and log files) are considered unstructured data.

Q8. Explain the relationship of big data with other areas?

Ans :

Big data models Improve Operational Efficiencies Increase Revenues Achieve Competitive Differentiation Reduce risks and costs Sell to microtrends Offer new services Save time Enable self service Seize market share Lower complexity Improve customer experience Incubate new ventures Enable self service Detect fraud

- Digital Marketing.
- Financial Services
- Big data and Advances in health care
- Pioneering New Frontiers in medicine
- Advertising

I) Digital Marketing

Introduction Database Marketers, Pioneers of Big Data Big Data & New School of Marketing Cross Channel Life cycle Marketing Social and Affiliate Marketing Empowering marketing with Social Intelligence Introduction Digital marketing encompasses using any sort of online media (profit or non-profit) for driving people to a website, a mobile app etc. and retaining them and interacting with them to understand what consumers really want.

Digital \marketing is easy when consumers interact with corporate' primary platform (ie. The one the corporate own) because corporate get good information about them. But corporate get very little information once people start interacting with other platforms (eg. Face book, Twitter, Google +). One of the problems that have to be dealt with in a digital existence, is that corporate do not have access to data for all consumers (ie. There is very little visibility about people when they interact in social networking sites). So corporate lose control of their ability to access the data they need, in order to make smart, timely decisions. Big data on the web will completely transform a company's ability to understand the effectiveness of its marketing and the ability to understand how its competitors are behaving. Rather than a few people making big decisions, the organizations will be able to make hundreds and thousands of smart decisions every day

II) Financial Services

Fraud & Big Data - Fraud is intentional deception made for personal gain or to damage another individual. - One of the most common forms of fraudulent activity is credit card fraud. - Social media and mobile phones are forming new frontiers fraud. - Capegemini financial services team believes that due to the nature of data streams and processing required BIG Data Technologies provide an optimal technology solution based on the following three Vs : 1. High volume: Years of consumer records and transactions (150 billion + 2. records per year). 3. High velocity: Dynamic transactions and social media info. 4. High variety: Social media plus other unstructured data such as customer E-mails, call center conversations as well as transactional structured data.

III) Big data and Healthcare

Big data promises enormous revolution in healthcare, with the advancements in everything from the management of chronic disease to the delivery of personalized medicine. In addition to saving and improving lives, Big Data has the potential to transform the entire healthcare system by replacing guesswork and institution with objective data-driven science.-The healthcare industry now has huge amount of data: from biological data such as gene expression, Special Needs Plans (SNPs), proteomics, metabolomics, and next-generation gene sequence data etc. The exponential growth in data is further accelerated by the digitization of patient level data stored in Electronic Medical Records (EMRs) or Electronic Health Records (EHRs) and Health Information Exchanges (HIEs) enhanced with data from imaging and test results, medical and prescription claims and personal health devices. - In addition to saving and improving lives, Big Data has the potential to transform the entire health care system by replacing guesswork and intuition with objective, data-driven science

IV) Advertising and Big Data

Big Data is changing the way advertisers address three related needs. (i) How much to spend on advertisements. (ii) How to allocate amount across all the marketing communication touch points. (iii) How to optimize advertising effectiveness. Given these needs advertisers need to measure their advertising end to end in terms of Reach, Response & Reaction. Reach, Resonance, and Reaction Reach: First part of reach is to identify the people who are most volumetrically responsive to their advertising and then answer questions such as what do those people watch? What do they do online? How to develop media plan against intended audience. The second part of reach is delivering advertisements to the right audience. That is, to understand if we are actually reaching our desired audience. If we think about the online world, it's a world where we can deliver 100 million impressions but we never really know for sure who our campaign was actually delivered to. If our intended audience is women aged 18 to 35, of our 100 million impressions, what percentage of impressions were actually delivered to the intended audience? What was the reach, what was the frequency, what was the delivery against the intended audience? Resonance: If we know whom we want to reach and we're reaching them efficiently with your media spend, the next question is, are our ads breaking through? Do people know they're from our brand? Are they changing attitudes? Are they making consumers more likely to want to buy our brand? This is what is called resonance. Reaction: Advertising must drive a behavioral reaction or it isn't really working. We have to measure the actual behavioral impact.

1.4 Business Analytics

Q9. Discuss briefly about role of business analytics in current business environment.

Ans :

1. Financial Analytics

Organizations use predictive models for forecasting future financial performance for constructing financial instruments like derivatives and assessing the risk involved in investment projects and portfolios. They also use prescriptive models for creating optimal capital budgeting plans for constructing optimal portfolios of investments and allocating assets. Addition to this, simulation is also used for ascertaining risk in the financial sector.

Example : GE Asset Management utilizes optimization models of analytics to make investment decisions of cash received from various sources. The approximate benefit obtained from using optimization models over a five-year period was $ 75 million.

2. Marketing Analytics

Business analytics is used in marketing for obtaining a better understanding of consumer behaviours by using the scanned data and social networking data. It leads to efficient use of advertising budgets, improved demand forecasting, effective pricing strategies, increased product line management and improved customer loyalty and satisfaction. Marketing analytics has gained much interest due to the data generated from social media.

Example : NBC Universal utilizes a predictive model every year to aid the annual up front market. An upfront market is a period in ending of May when every TV network sells most of the on-air advertisements for the upcoming season of television. The results of forecasting model are utilized by more than 200 NBC sales for supporting sales and pricing decisions.

3. Human Resource (HR) Analytics

HR function utilizes analytics to ensure that the organization consists of the employees with required skills to meet its needs, to ensure that it achieves its diversity goals and to ensure that it is hiring talent of the highest quality and also offering an environment which retains it.

Example : Sears Holding Corporation (SHC) owners of Roebuck Company, retailers Kmart and Sears. They made a team of HR analytics inside the corporate HR function. They apply predictive and descriptive analytics for tracking and influencing retention of employees and for supporting employee hiring.

4. Health Care Analytics

Health care organizations utilize analytics to provide more effective treatment and control cost. They use descriptive, prescriptive and predictive analytics for improving patient flow, staff and facility scheduling, purchasing and control of inventory. However, prescriptive analytics is specially used for the purpose of treatment and diagnosis. It is the most important proven utility of analytics.

Example : Memorial Sloan-Kettering Cancer Center along with Georgia Institute of Technology created a real-time prescriptive model for determining the optimal placement of radio active seeds for prostate cancer treatment. The results led to requirements of20-30% lesser seeds and less invasive and faster procedure.

5. Supply Chain Analytics

Analytics is used by logistics and supply chain management to achieve efficiency. The entire spectrum of analytics is utilized by them. Various organizations such as=UPS and FedEx apply analytics for efficient delivery of goods. Analytics helps them in optimal sorting of goods, staff and vehicle scheduling and vehicle routing, which helps in increasing the profitability. Analytics enable better processing control, inventory and more effective supply chains.

Example : ConAgra Foods utilized the prescriptive and predictive analysis for a better plan capacity utilization by incorporation of inherent uncertainty in pricing of commodities. ConAgra Foods attained a 100% return on investment in just three months.

6. Analytics for Government and Non-Profit Organizations

Government and non-profit organizations apply analytics for driving out inefficiencies and increasing the accountability and effectiveness of programs. During the period of Word War II, advanced analytics was first applied by the English and U.S. military. Analytics applicability is very extensive in government agencies from elections to tax collections. Non-profit organizations utilize analytics for ensuring the accountability and effectiveness to their clients and donors.

Examples : The New York State Department incorporated with IBM for using prescriptive analytics in developing a more efficient tax collection approach.

Catholic Relief Services (CRS) is a non-profit organization which is the official international humanitarian agency of the U.S. Catholic community. This offer helps to the victims of both human-made and natural disasters. It also offers various other services through its agricultural, educational and health programs. It utilizes analytical spread sheet model for helping in the annual budget allocation based on the effects of its relief programs and efforts in various countries.

7. Sports Analytics

Analytical applicability in area of sports became popular when a renowned author Michael Lewis published Money ball in the year 2003. The book explained how the athletics of Oakland applied an analytical approach for evaluating players for assembling a competitive team with a limited budget. Analytics is used for evaluation of on-field strategy which is a common thing in professional sports. Analytics is also used in off-the-field decisions to ensure customer satisfaction.

Example : Professional sports teams utilize analytics for assessing players for the amateur drafts and for decision making of contract negotiations offered to the players. Various franchises across many major sports utilize prescriptive analytics for adjusting the ticket prices throughout the season for reflecting the potential demand and relative attractiveness for every game.

8. **Web Analytics**

The analysis of internet activity including visits of users to social media sites like Linkedln and Facebook and other websites is called web analytics. It plays a vital role in sales or promotions of products and services. Through internet various leading companies utilize advanced and descriptive analytics by applying them to the data gathered from online experiments for determining the best way for configuration of ads, websites and proper utilization of social networks to promote products and services. Online experimentation is exposing various group of visitors to varied versions of a website and then tracking the results. These experiments can be conducted without risking the overall business disruption of the company due to several number of internet users. But these experiments have proven to be invaluable as they enable the company for using trial-and-error method of determine statistically the reasons for differences in the sales and website traffic.

Q10. Explain differences between Business Intelligence and Business Analytics.

Ans :

BUSINESS INTELLIGENCE	BUSINESS ANALYTICS
Business Intelligence is the process comprising of technologies and strategies incorporated by the enterprise industries to analyze the existing business data which provides past (historical), current and predictive events of the business operations.	Business Analytics is the process of technologies and strategies used to continuously exploring and to extract the insights and performance from the past business information to drive the successful future business planning.
Business Intelligence uses past and present available data to drive the present business successfully. Business Intelligence maintains, operates, streamlines and increases the productivity of the on-going businesses.	Business Analytics uses past data to drive current business planning successfully. Business Analytics gathers and analyses the data by using predictive analytics method and provides rich visual reports to the viewers about the current business operations and its' operations efficiency. so let us see the Difference between Business Intelligence and Business Analytics in detail.

Q11. Explain the various Challenges in Business Analytics

Ans :

Penn State University's John Jordan described the challenges with Business Analytics: there is "a greater potential for privacy invasion, greater financial exposure in fast-moving markets, greater potential for mistaking noise for true insight, and a greater risk of spending lots of money and time chasing poorly defined problems or opportunities." Other challenges with developing and implementing Business Analytics include…

- **Executive Ownership:** Business Analytics requires buy-in from senior leadership and a clear corporate strategy for integrating predictive models.
- **IT Involvement:** Technology infrastructure and tools must be able to handle the data and Business Analytics processes.

- **Available Production Data vs. Cleansed Modeling Data :** Watch for technology infrastructure that restrict available data for historical modeling, and know the difference between historical data for model development and real-time data in production.
- **Project Management Office (PMO)** : The correct project management structure must be in place in order to implement predictive models and adopt an agile approach.
- **End user Involvement and Buy-In :** End users should be involved in adopting Business Analytics and have a stake in the predictive model.
- **Change Management :** Organizations should be prepared for the changes that Business Analytics bring to current business and technology operations.
- **Explainability vs. the "Perfect Lift"**: Balance building precise statistical models with being able to explain the model and how it will produce results.

Q12. Explain the Business Analytics Best Practices.

Ans :

Adopting and implementing Business Analytics is not something a company can do overnight. But, if a company follows some best practices for Business Analytics, they will get the levels of insight they seek and become more compe titive and successful. We list some of the most important best practices for Business Analytics here, though your organization will need to determine which best practices are most fitting for there needs.

- Know the objective for using Business Analytics. Define the business use case and the goal ahead of time.
- Define the criteria for success and failure.
- Select the methodology and be sure to know the data and relevant internal and external factors
- Validate models using to predefined success and failure criteria.

Business Analytics is critical for remaining competitive and achieving success. When they get BA best practices in place and get buy-in from all stakeholders, the organization will benefit from data-driven decision making.

1.5 Data Visualization

Q13. What is Data visualization?

Ans :

It is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

- Today's data visualization tools go beyond the standard charts and graphs used in Microsoft Excel spreadsheets, displaying data in more sophisticated ways such as infographics, dials and gauges, geographic maps, sparklines, heat maps, and detailed bar, pie and fever charts. The images may include interactive capabilities, enabling users to manipulate them or drill into the data for querying and analysis. Indicators designed to alert users when data has been updated or predefined conditions occur can also be included.

Importance of data visualization

- **Modern Business Intelligence**

  Data visualization has become the de facto standard for modern business intelligence (BI). The success of the two leading vendors in the BI space, Tableau and Qlik — both of which heavily emphasize visualization - has moved other vendors toward a more visual approach in their software. Virtually all BI software has strong data visualization functionality.

- **Democratizing Data**

  Data visualization tools have been important in democratizing data and analytics and making data-driven insights available to workers throughout an organization. They are typically easier to operate than traditional statistical analysis software or earlier versions of BI software. This has d to a rise in lines of business implementing data visualization tools on their own, without support from IT.

- **Data Visualization Software**

  It plays an important role in big data and advanced analyticsprojects. As businesses accumulated massive troves of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Visualization tools were a natural fit.

- **Advanced Analytics**

  Visualization is central for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

Q14. Explain the various Data visulization.

Ans :

Data visualization tools can be used in a variety of ways. The most common use today is as a BI reporting tool. Users can set up visualization tools to generate automatic dashboardsthat track company performance across key performance indicators and visually interpret the results.

Many business departments implement data visualization software to track their own initiatives. For example, a marketing team might implement the software to monitor the performance of an email.

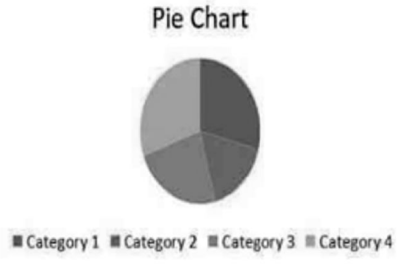
VIDI: The VIDI tool lets you create a visualization of your data for free. All you have to do is upload your data, select type, do a little customization and you are good to go.	**Microsoft Excel:** Yes you can even use Microsoft Excel to visualize your data. You can know about the whole topic if you do some search about the data visualization tools excel.
Qlik Sense Desktop: The Qlick Sense desktop tool lets you interactive data reports visualization for free.	**Microsoft PowerPoint:** Just like the Microsoft Excel Microsoft PowerPoint is also one of the great tools to do the job. You can easily visualize your data Microsoft PowerPoint and so on
Microsoft BI Platform: The Microsoft BI platform allows you to update your data from different sources and makes a report out of it.	**Fusion Charts:** From the basic charts (line, column, pie etc. – 2D & 3D) to the most complex ones (waterfall, gantt, candlestick, zoomline etc.). One of the most exhaustive collection of javascript charts, widgets & maps in the industry.
Google Fusion Tables: The google fusion tables is one of the simple tools to visualize the data. You can simply upload a file and choose how to display it. You can display your file as a map, table, line chart, or pie chart. It's highly customizable and user friendly.	**Data Wrapper:** Known to have a simple, clear interface easy to work with. Provides option to upload csv data and create straightforward charts, and also maps, that can quickly be embedded into reports.

Q15. What are the Data Visualization Techniques?

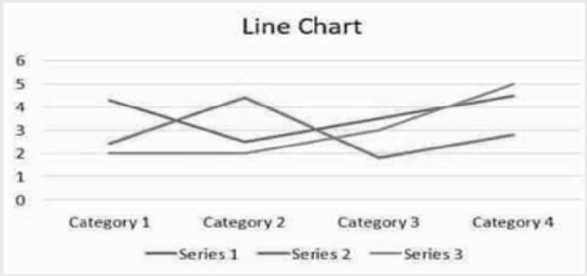
Ans :

The data visualization techniques are Diagrams, charts, graphs.

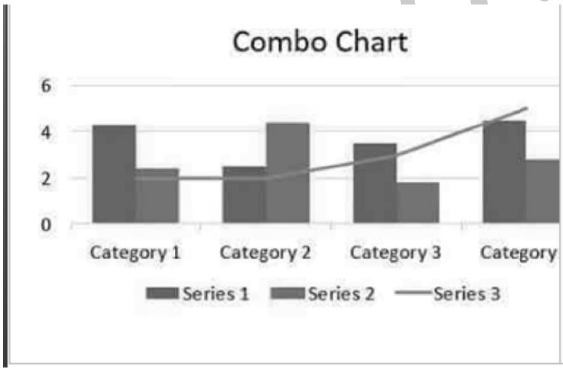
Most widely used forms of data visualisation are presented below:



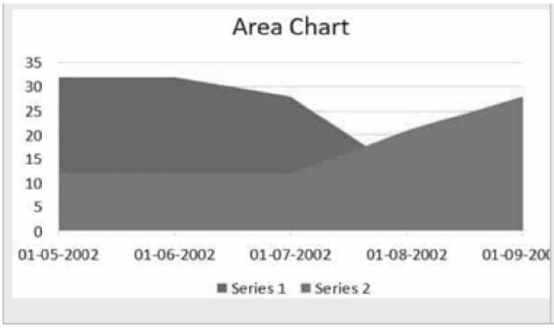
Pie Charts: Pie Charts are one of the common yet popular techniques. It also comes under data visualization techniques in excel. However, to some people, it can be hard to understand the chart while comparing to the line and bar type chart.



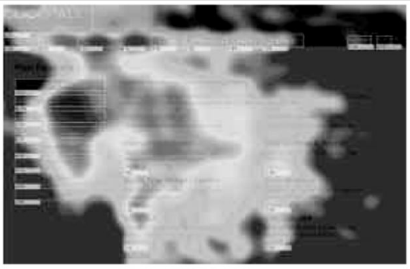
Line Charts: To make your data simple and more appealing you can simply use the line charts technique. Line chart basically displays the relationship between two patterns. Also, it is one of the most used techniques worldwide.



Combo charts & Bar Charts: Bars charts are also one of the most commonly used techniques when it comes to comparing two different patterns. The bar charts can display the data in a horizontal way or in a vertical way. It all depends on your needs.

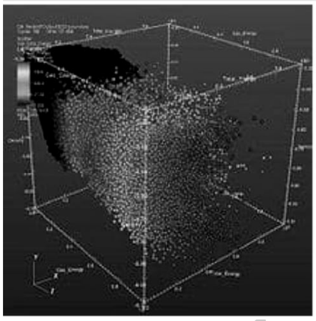


Area Chart: An **area chart** or **area** graph is similar to a line chart but provides graphically quantitative data. The areas can be filled with colours, hatch, pattern. This chart is generally used when comparing quantities which is depicted by area.



Heatmap: This type of chart is widely used by websites, mobile application makers, research institutes etc. These maps shows the concentration of activity/ entity over a particular area.

Network Diagrams: This is a powerful tool for finding out connections & correlations. It highlights and bridges the gaps. Shows how strongly one activity is connected to other.



Scattered 3 D plot: As the image shows it shows the distribution of entity in a 3 dimensional nature. It can be considered as showing location and concentration of gases in a box with different colours assigned to each gas.

Tree Maps: These are used to displaying large amounts of hierarchically **structured** (tree-**structured**) data. Generally size of each rectangle/ block refers to the quantity.

Q16. How data visualization is it being used?

Ans :

Data visualization can be helpful in many ways and just in case if you are wondering where it is being used. Then are some of the popular sectors:

- **Large data in a simple format** : Visualization it became easier for business owners to understand their large data in a simple format. The visualization method is also time saving so business does not have to spend much time to make a report or solve a query. They can easily do it in a less time and in a more appealing way.

- **Visual analytics:** It offers a story to the viewers. By using charts and graphs or images a person can easily exposure the whole concept. As well the viewers will be able to understand the whole thing in an easy way.
- **Complicated data** : It will look easy when it gets through the process of visualization. Complicated data report gets converted into a simple format. And it helps people to understand the concept in an easy way.
- **Visualization process :** It gets easier to the business owners to understand their product growth. Market competition in a better way. The visualization tools can be very helpful to monitor an email campaign. Or company's own initiative regarding something.

1.6 Data Queries

Q17. Explain the Data queries.

Ans :

- **A query is a request for data or information:** From a database table or combination of tables. This data may be generated as results returned by Structured QueryLanguage (SQL) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools.
- **A query is a request for information from a database**. : Query language: Many database systems require you to make requests for information in the form of a stylized query that must be written in a special query language.
- **A query is an inquiry into the database:** using the SELECT statement. A query is used to extract data from the database in a readable format according to the user'srequest. For instance, if you have an employee table, you might issue a SQL statement that returns the employee who is paid the most.
- **Queries can accomplish a few different tasks :** Primarily, queries are used to find specific data by filtering specific criteria. In a relational database, which contains records or rows of information, the SQL SELECT statement query allows the user to choose data and return it from the database to an application.

1.7 Statistical Methods for Summering Date

Q18. Explain the different statistical methods for summerizing data

Ans :

Statistical method is used for long run forecasting. In this method, statistical and mathematical techniques are used to forecast demand. This relies on past data.

- **Trend projection method:** These are generally based on analysis of past sales patterns. These methods dispense with the need for costly market research because the necessary information is often already available in company files. This method is used in case the sales data of the firm under consideration relate to different time periods, i.e., it is a time – series data. There are five main techniques of mechanical extrapolation.
- **Trend line by observation:** This method of forecasting trend is elementary, easy and quick. It involves merely the plotting of actual sales data on a chart and them estimating just by observation where the trend line lies. The line can be extended towards a future period and corresponding sales forecast is read form the graph.

- **Least squares methods:** This technique uses statistical formulae to find the trend line which best fits the available data. The trend line is the estimating equation, which can be used for forecasting demand by extrapolating the line for future and reading the corresponding values of sales on the graph.
- **Time series analysis:** Where the surveys or market tests are costly and time – consuming, statistical and mathematical analysis of past sales data offers another methods to prepare the forecasts, that is, time series analysis.
- **Moving average method:** This method considers that the average of past events determine the future events. In other words, this method provides consistent results when the past events are consistent and unaffected by wide changes.
- **Exponential smoothing:** This is a more popular technique used for short run forecasts. This method is an improvement over moving averages method, unlike in moving averages method, all time periods here are given varying weight, that is , value of the given variable in the recent times are given higher weight and the values of the given variable in the distant past are given relatively lower weights for further processing.
- **Barometric Technique:** Simple trend projections are not capable of forecasting turning paints. Under Barometric method, present events are used to predict the directions of change in future. This is done with the help of economics and statistical indicators. Those are (1) Construction Contracts awarded for building materials (2) Personal income (3) Agricultural Income. (4) Employment (5) Gross national income (6) Industrial Production (7) Bank Deposits etc.
- **Simultaneous equation method:** In this method, all variable are simultaneously considered, with the conviction that every variable influence theothervariable economic environment. Hence, the set of equations equal the number of dependent variable which is also called endogenous variables.
- **Correlation and regression methods:** correlation and regression methods are statistical techniques. Correlation describes the degree of association between two variable such as sales and advertisement expenditure. When the two variable tend to change together, then they are said to be correlate.

1.8 Exploring Data Using Pivot Tables

Q19. How to explore data in pivot tables?

Ans :

Excel PivotTable helps in exploring and extracting important data from an Excel table or a range of data. There are many ways for doing this and you can choose the ones which are suitable for your data. Further, while exploring data, you can view various combinations immediately while changing your choices to pick the data values.

Following tasks can be performed with a PivotTable

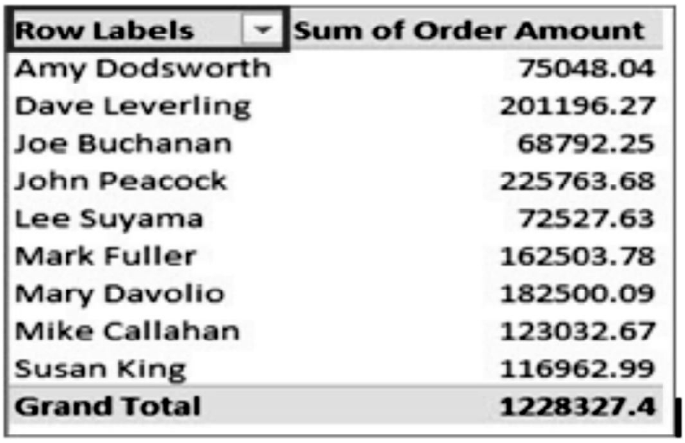
- Sort the data.
- Filter the data.
- Nest the PivotTable fields.
- Expand and Collapse the fields.
- Group and ungroup field values.

1. SORT THE DATA

Sorting data is helpful when you have large amounts of data in a PivotTable or PivotChart. You can sort in alphabetical order, from highest to lowest values, or from lowest to highest values. Sorting is one way of organizing your data so it's easier to find specific items that need more scrutiny.

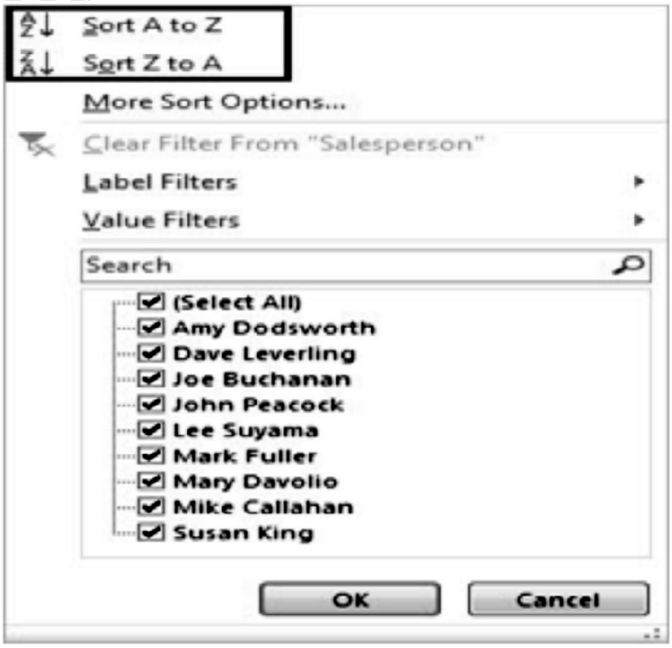
Follow these steps to sort in Excel Desktop:

1. In a PivotTable, click the small arrow next to **Row Labels** and **Column Labels** cells.
2. Click a field in the row or column you want to sort.
3. Click the arrow ▾ on **Row Labels** or **Column Labels**, and then click the sort option you want.



Row Labels ▾	Sum of Order Amount
Amy Dodsworth	75048.04
Dave Leverling	201196.27
Joe Buchanan	68792.25
John Peacock	225763.68
Lee Suyama	72527.63
Mark Fuller	162503.78
Mary Davolio	182500.09
Mike Callahan	123032.67
Susan King	116962.99
Grand Total	**1228327.4**

4. To sort data in ascending or descending order, click **Sort A to Z** or **Sort Z to A**.

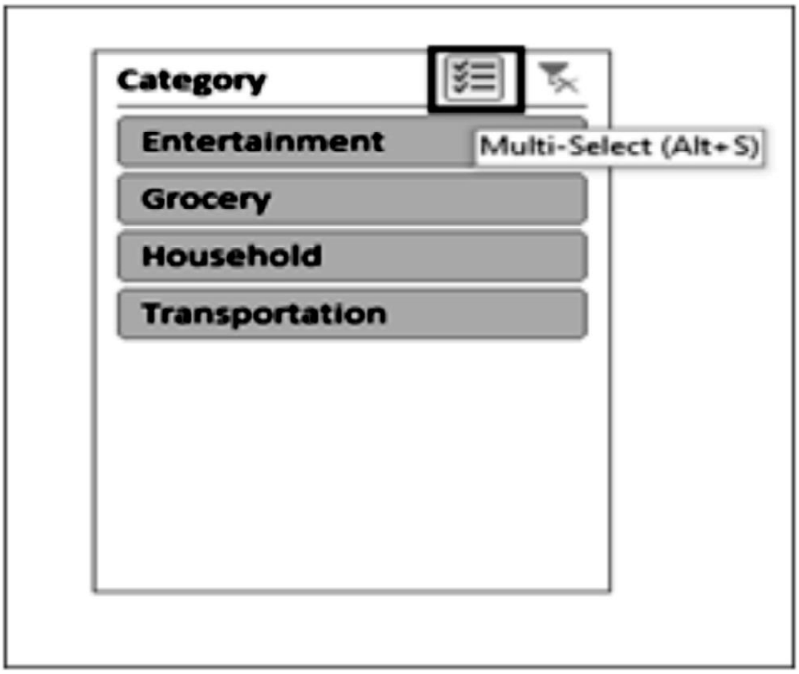


☑ (Select All)
☑ Amy Dodsworth
☑ Dave Leverling
☑ Joe Buchanan
☑ John Peacock
☑ Lee Suyama
☑ Mark Fuller
☑ Mary Davolio
☑ Mike Callahan
☑ Susan King

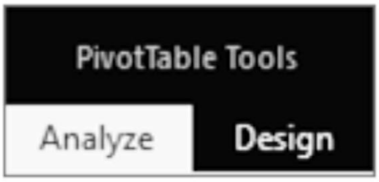
Text entries will sort in alphabetical order, numbers will sort from smallest to largest (or vice versa), and dates or times will sort from oldest to newest (or vice versa).

2. Filter the data

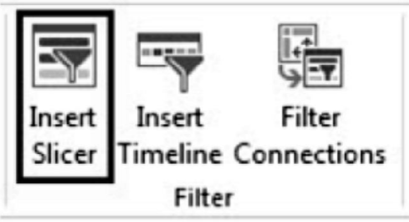
To focus on a smaller portion of a large amount of your PivotTable data for in-depth analysis, you can filter the data. There are several ways to do that. Start by inserting one or more slicers for a quick and effective way to filter your data. Slicers have buttons you can click to filter the data, and they stay visible with your data so you always know what fields are shown or hidden in the filtered PivotTable.



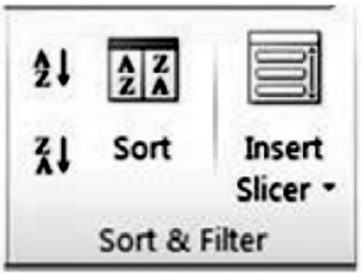
- Click anywhere in the PivotTable to show the **PivotTable Tools** on the ribbon.



- If you are using Excel 2016 or 2013, click **Analyze** > **Insert Slicer**.



- If you are using Excel 2010 or 2007, click **Options** > **Insert Slicer** > **Insert Slicer**.



- In the **Insert Slicers** dialog box, check the boxes of the fields you want to create slicers for.
- Click **OK**.

  A slicer appears for each field you checked in the **Insert Slicers** dialog box.
- In each slicer, click the items you want to show in the PivotTable.

  To choose more than one item, hold down Ctrl, and then pick the items you want to show.

3. Nest the PivotTable fields -

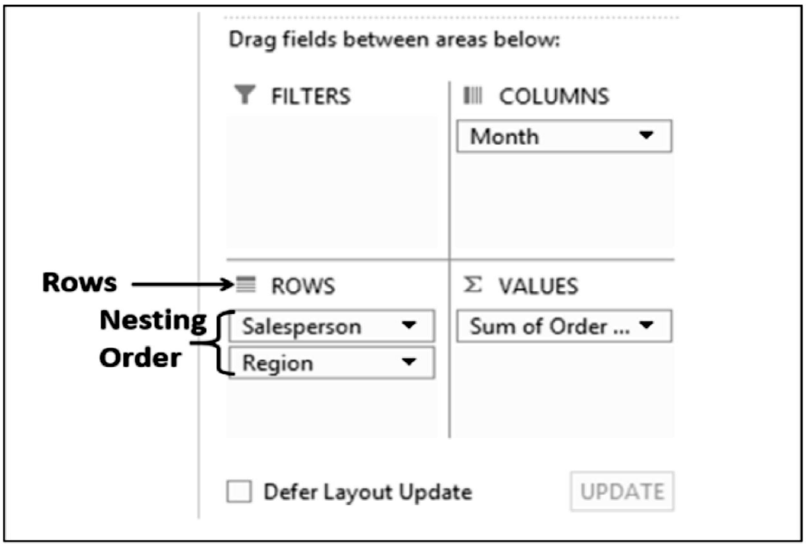
If you have more than one field in any of the PivotTable areas, then the PivotTable layout depends on the order you place the fields in that area. This is called the Nesting Order.

If you know how your data is structured, you can place the fields in the required order. If you are not sure about the structure of the data, you can change the order of the fields that instantly changes the layout of the PivotTable.

In this chapter, you will understand the nesting order of the fields and how you can change the nesting order.

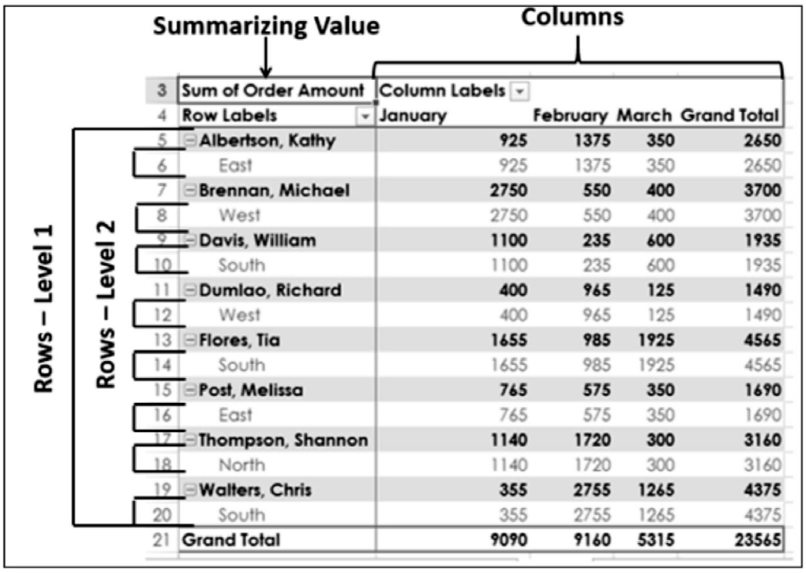
4. Nesting Order of the Fields

Consider the sales data example, where you have placed the fields in the following order -



As you can see, in the rows area there are two fields – salesperson and region in that order. This order of the fields is called nesting order i.e. Salesperson first and Region next.

In the PivotTable, the values in the rows will be displayed based on this order, as given below.



3	Sum of Order Amount	Column Labels			
4	Row Labels	January	February	March	Grand Total
5	⊟ **Albertson, Kathy**	**925**	**1375**	**350**	**2650**
6	East	925	1375	350	2650
7	⊟ **Brennan, Michael**	**2750**	**550**	**400**	**3700**
8	West	2750	550	400	3700
9	⊟ **Davis, William**	**1100**	**235**	**600**	**1935**
10	South	1100	235	600	1935
11	⊟ **Dumlao, Richard**	**400**	**965**	**125**	**1490**
12	West	400	965	125	1490
13	⊟ **Flores, Tia**	**1655**	**985**	**1925**	**4565**
14	South	1655	985	1925	4565
15	⊟ **Post, Melissa**	**765**	**575**	**350**	**1690**
16	East	765	575	350	1690
17	⊟ **Thompson, Shannon**	**1140**	**1720**	**300**	**3160**
18	North	1140	1720	300	3160
19	⊟ **Walters, Chris**	**355**	**2755**	**1265**	**4375**
20	South	355	2755	1265	4375
21	**Grand Total**	**9090**	**9160**	**5315**	**23565**

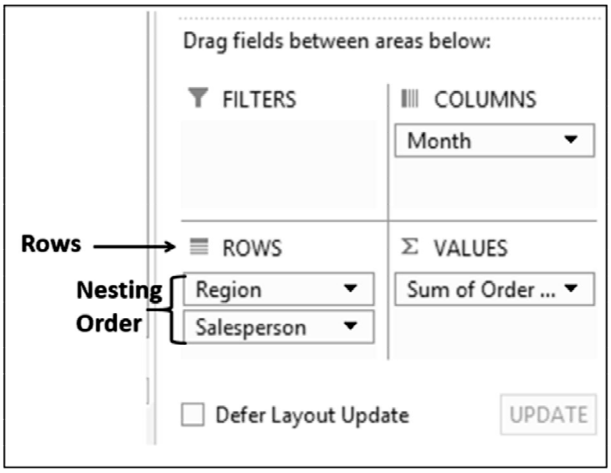
As you can observe, the values of the second field in the nesting order are embedded under each of the values of the first field.

In your data, each salesperson is associated with only one region, whereas most of the regions are associated with more than one salesperson. Hence, there is a possibility that if you reverse the nesting order, your PivotTable will look more meaningful.

5. Changing the Nesting Order

To change the nesting order of the fields in an area, just click the field and drag it to the position you want.

Click on the field Salesperson in the ROWS area, and drag it to below the field Region. Thus, you have changed the nesting order to – Region first and Salesperson next, as follows -



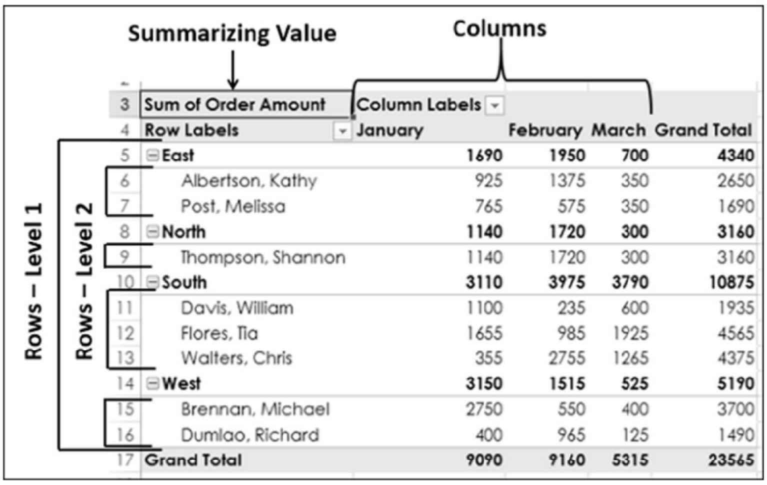
COLUMNS: Month

ROWS: Region, Salesperson

VALUES: Sum of Order ...

□ Defer Layout Update

The resulting PivotTable will be as given below H



3	Sum of Order Amount	Column Labels			
4	Row Labels	January	February	March	Grand Total
5	**⊟East**	**1690**	**1950**	**700**	**4340**
6	Albertson, Kathy	925	1375	350	2650
7	Post, Melissa	765	575	350	1690
8	**⊟North**	**1140**	**1720**	**300**	**3160**
9	Thompson, Shannon	1140	1720	300	3160
10	**⊟South**	**3110**	**3975**	**3790**	**10875**
11	Davis, William	1100	235	600	1935
12	Flores, Tia	1655	985	1925	4565
13	Walters, Chris	355	2755	1265	4375
14	**⊟West**	**3150**	**1515**	**525**	**5190**
15	Brennan, Michael	2750	550	400	3700
16	Dumlao, Richard	400	965	125	1490
17	**Grand Total**	**9090**	**9160**	**5315**	**23565**

You can clearly observe that the Layout with the nesting order – Region and then Salesperson yields a better and compact report than the one with the nesting order – Salesperson and then Region.

6. Expand and Collapse the fields

In case a Salesperson represents more than one area and you need to summarize the sales by Salesperson, then the previous Layout would have been a better option.

In a PivotTable or PivotChart, you can expand or collapse to any level of data detail, and even for all levels of detail in one operation. You can also expand or collapse to a level of detail beyond the next level. For example, starting at a country/region level, you can expand to a city level which expands both the state/province and city level. This can be a time-saving operation when you work with many levels of detail. In addition, you can expand or collapse all members for each field in an Online Analytical Processing (OLAP) data source.

You can also see the details that are used to aggregate the value in a value field.

7. Expand or collapse levels in a Pivot Table

In a PivotTable, do the following:

- Click the expand or collapse button next to the item that you want to expand or collapse.

  Note: If you don't see the expand or collapse buttons, see the Show or hide the expand and collapse buttons in a PivotTable section in this topic.
- Double-click the item that you want to expand or collapse.
- **Right-click** (**Ctrl+Click** on a Mac) the item, click **Expand/Collapse**, and then do one of the following:
- To see the details for the current item, click **Expand**.
- To hide the details for the current item, click **Collapse**.
- To hide the details for all items in a field, click **Collapse Entire Field**.
- To see the details for all items in a field, click **Expand Entire Field**.
- To see a level of detail beyond the next level, click **Expand To "<Field name>"**.
- To hide to a level of detail beyond the next level, click **Collapse To "<Field name>"**.

8. **Group and ungroup field values**

Grouping Data

In a pivot table, you can group dates, number and text fields. For example, group order dates by year and month, or group test scores in bands of 10. You can manually select text items in a pivot table field, and group the selected items. This lets you quickly see subtotals for a specific set of items in your pivot table.

Example: To group the items in a Date field by week

1. Right-click on one of the dates in the pivot table.
2. In the popup menu, click Group.
3. In the Grouping dialog box, select Days from the 'By' list.
4. For 'Number of days', select 7.
5. The week range is determined by the date in the 'Starting at' box, so adjust this if necessary.
6. Click OK.

9. **Ungrouping the Data**

To remove grouping, right-click any item in the grouped data, and click **Ungroup**.

If you ungroup numeric or date and time fields, all grouping for that field is removed. If you ungroup a group of selected items, only the selected items are ungrouped. The group field won't be removed from the Field List until all groups for the field are ungrouped.

For example : suppose you have four cities in the City field: Boston, New York, Los Angeles, and Seattle. You group them so that New York and Boston are in one group you name Atlantic, and Los Angeles and Seattle are in a group you name Pacific. A new field, City 2, appears in the Fields area and is placed in the Rows area of the Fields List.

As shown here, the City2 field is based on the City field, and is placed in the Rows area to group the selected cities.

Short Question and Answers

1. Data

A representation of facts, concepts or instructions in a formalised manner suitable for communication, interpretation, or processing by humans or by automatic means.

Three aspects of data can be identified.

- Record objective facts which will be understood in exactly the same way by everyone.
- Record absolutely any type of concept, with no guarantees as to its accuracy or validity.

2. Data Analytics

Ans :

Definition

Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements. Data analytics is also known as data analysis.

- **Exploratory data analysis:** At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to find patterns and relationships in data, and
- **Confirmatory data analysis (CDA),** which applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work. CDA is akin to the work of a judge or jury during a court trial - a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis.

3. Big Data

Ans :

Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.

Big data is a term that is used to describe data that is high volume, high velocity, and/ or high variety; requires new technolo- gies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

SAS describes Big Data as "a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis." What's important to keep in mind about Big Data is that the amount of data is not as important to an organization as the analytics that accompany it. When companies analyze Big Data, they are using Business Analytics to get the insights required for making better business decisions and strategic moves.

4. Dimensions of Big Data

Ans :

i) **Data Volume:** Data Volume can be measured by quality of transactions, events and amount of history. Big Data isn't just a description of raw volume. The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That's where Big Data analytics become necessary.

ii) **Data Variety :** It is the assortment of data. Traditionally data, especially operational data, is structured as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.). Wide variety of data: Internet data(Social media ,Social Network-Twitter, Face book),

iii) **DataVelocity** : It is the measure of how fast the data is coming in. Remember our Facebook example. 250 billion images may seem like a lot. But if you want your mind blown, consider this: Facebook users upload more than 900 million photos a day. So that 250 billion number from last year will seem like a drop in the bucket in a few months. Facebook has to handle a tsunami of photographs every day. It has to ingest it all, process it, file it, and somehow, later, be able to retrieve it.

iv) **Variability** : The increase in the range of values typical of a large data set - and value, which addresses the need for valuation of enterprise data.

5. Various Challenges in Business Analytics

Ans :

Penn State University's John Jordan described the challenges with Business Analytics: there is "a greater potential for privacy invasion, greater financial exposure in fast-moving markets, greater potential for mistaking noise for true insight, and a greater risk of spending lots of money and time chasing poorly defined problems or opportunities." Other challenges with developing and implementing Business Analytics include…

- **Executive Ownership:** Business Analytics requires buy-in from senior leadership and a clear corporate strategy for integrating predictive models.
- **IT Involvement:** Technology infrastructure and tools must be able to handle the data and Business Analytics processes.
- **Available Production Data vs. Cleansed Modeling Data :** Watch for technology infrastructure that restrict available data for historical modeling, and know the difference between historical data for model development and real-time data in production.
- **Project Management Office (PMO)** : The correct project management structure must be in place in order to implement predictive models and adopt an agile approach.

6. Data visualization

Ans :

It is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

- Today's data visualization tools go beyond the standard charts and graphs used in Microsoft Excel spreadsheets, displaying data in more sophisticated ways such as infographics, dials and gauges, geographic maps, sparklines, heat maps, and detailed bar, pie and fever charts. The images may include interactive capabilities, enabling users to manipulate them or drill into the data for querying and analysis. Indicators designed to alert users when data has been updated or predefined conditions occur can also be included.

7. Importance of data visualization

Ans :

- **Modern Business Intelligence**

  Data visualization has become the de facto standard for modern business intelligence (BI). The success of the two leading vendors in the BI space, Tableau and Qlik — both of which heavily emphasize visualization - has moved other vendors toward a more visual approach in their software. Virtually all BI software has strong data visualization functionality.

- **Democratizing Data**

  Data visualization tools have been important in democratizing data and analytics and making data-driven insights available to workers throughout an organization. They are typically easier to operate than traditional statistical analysis software or earlier versions of BI software. This has d to a rise in lines of business implementing data visualization tools on their own, without support from IT.

- **Data Visualization Software**

  It plays an important role in big data and advanced analyticsprojects. As businesses accumulated massive troves of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Visualization tools were a natural fit.

- **Advanced Analytics**

  Visualization is central for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

8. Data queries.

Ans :

- **A query is a request for data or information:** From a database table or combination of tables. This data may be generated as results returned by Structured QueryLanguage (SQL) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools.
- **A query is a request for information from a database**. : Query language: Many database systems require you to make requests for information in the form of a stylized query that must be written in a special query language.
- **A query is an inquiry into the database:** using the SELECT statement. A query is used to extract data from the database in a readable format according to the user'srequest. For instance, if you have an employee table, you might issue a SQL statement that returns the employee who is paid the most.

9. Data in pivot tables.

Ans :

Excel PivotTable helps in exploring and extracting important data from an Excel table or a range of data. There are many ways for doing this and you can choose the ones which are suitable for your data. Further, while exploring data, you can view various combinations immediately while changing your choices to pick the data values.

Following tasks can be performed with a PivotTable

- Sort the data.
- Filter the data.
- Nest the PivotTable fields.
- Expand and Collapse the fields.
- Group and ungroup field values.

10. Ungrouping the Data

Ans :

To remove grouping, right-click any item in the grouped data, and click **Ungroup**.

If you ungroup numeric or date and time fields, all grouping for that field is removed. If you ungroup a group of selected items, only the selected items are ungrouped. The group field won't be removed from the Field List until all groups for the field are ungrouped.

Example

Suppose you have four cities in the City field: Boston, New York, Los Angeles, and Seattle. You group them so that New York and Boston are in one group you name Atlantic, and Los Angeles and Seattle are in a group you name Pacific. A new field, City 2, appears in the Fields area and is placed in the Rows area of the Fields List.

As shown here, the City2 field is based on the City field, and is placed in the Rows area to group the selected cities.

UNIT II

Descriptive Statistical Measures: Population and samples, Measures of location, Measures of Dispersion, Measures of variability, measures of Association. Probability distribution and Data Modeling – Discrete Probability distribution, Continuous Probability distribution, Random sampling from Probability Distribution, Data Modeling and Distribution fitting.

2.1 Population and Samples

Q1. Explain the term Population and Sample in detail

Ans :

Population

The field of inferential statistics enables you to make educated guesses about the numerical characteristics of large groups. The logic of sampling gives you a way to test conclusions about such groups using only a small portion of its members.

A population is a group of phenomena that have something in common. The term often refers to a group of people, as in the following examples:

- All registered voters in Crawford County
- All members of the International Machinists Union
- All Americans who played golf at least once in the past year

But populations can refer to things as well as people:

- All widgets produced last Tuesday by the Acme Widget Company
- All daily maximum temperatures in July for major Metero cities
- All basal ganglia cells from a particular rhesus monkey

Often, researchers want to know things about populations but do not have data for every person or thing in the population. If a company's customer service division wanted to learn whether its customers were satisfied, it would not be practical (or perhaps even possible) to contact every individual who purchased a product. Instead, the company might select a sample of the population.

Sample

- Often, researchers want to know things about populations but do not have data for every person or thing in the population.
- If a company's customer service division wanted to learn whether its customers were satisfied, it would not be practical (or perhaps even possible) to contact every individual who purchased a product. Instead, the company might select a sample of the population.
- A sample is a smaller group of members of a population selected to represent the population. In order to use statistics to learn things about the population, the sample must be random.
- A random sample is one in which every member of a population has an equal chance of being selected.
- The most commonly used sample is a simple random sample. It requires that every possible sample of the selected size has an equal chance of being used.
- The field of inferential statistics enables you to make educated guesses about the numerical characteristics of large groups.

Sampling

The logic of sampling gives you a way to test conclusions about such groups using only a small portion of its members.

A population is a group of phenomena that have something in common. The term often refers to a group of people.

Examples for Sampling

1. All registered voters in Crawford County
2. All members of the International Machinists Union
3. All Americans who played golf at least once in the past year
4. But populations can refer to things as well as people:
5. All widgets produced last Tuesday by the Acme Widget Company
6. All daily maximum temperatures in July for major U.S. cities
7. All basal ganglia cells from a particular rhesus monkey

Random Sample

A sample is a smaller group of members of a population selected to represent the population. In order to use statistics to learn things about the population, the sample must be random. A random sample is one in which every member of a population has an equal chance of being selected. The most commonly used sample is a simple random sample. It requires that every possible sample of the selected size has an equal chance of being used.

Descriptive statistical measures

The mean or the standard deviation, or in our example the number of red, green and blue balls, are called parameters if they are calculated from the population. If you select for example 50 balls out of this box - a subset of the population - this set of 50 balls is called a sample.

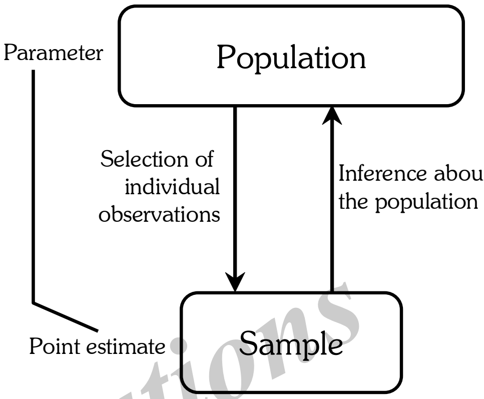
A descriptive measure referring to a sample - here the number of red balls in the box estimated from the sample of 50 balls - is called an estimate (or statistic). Parameters are depicted by Greek letters, estimates are depicted by Latin letters.

Random Variable

A Population is the set of all possible states of a random variable. The size of the population may be either infinite or finite.

A Sample is a subset of the population; its size is always finite.

He population, which is the basis of a statistical survey, has always to be defined in an exact way (which is not always easy) in order make sure that the results are comparable. The best way to go is to define not only the factual prerequisites ("what has to be analysed") but also the spatio-temporal general framework.


Parameter
Population
Selection of individual observations
Inference about the population
Point estimate
Sample

Examples

1. In our example of 500 balls, the size of the population is finite. In many cases - especially with real measurements - the population size is infinite. For example, if the variable of interest is the concentration of oxygen in air, measured with some analytical device. The population is the (infinite) set of all (possible) measurements (= results derived from the analytical instrument).
2. Another example is the number of cars driving on a particular section of a highway in the morning between 7am and 10am. The population of this variable is the number of cars using this highway during the defined time interval each day - as long as the highway exists.

2.2 Measures of Location

Q2. Explain about Measures of Location.

Ans :

A fundamental task in many statistical analyses is to estimate a location parameter for the distribution; i.e., to find a typical or central value that best describes the data.

The first step is to define what we mean by a typical value. For univariate data, there are three common definitions:

1. Mean

The mean is the sum of the data points divided by the number of data points. That is,

$$\bar{Y} = \sum_{i=1}^{N} Y_i / N$$

The mean is that value that is most commonly referred to as the average. We will use the term average as a synonym for the mean and the term typical value to refer generically to measures of location.

2. Median

The median is the value of the point which has half the data smaller than that point and half the data larger than that point. That is, if $X_1, X_2, \ldots ,X_N$ is a random sample sorted from smallest value to largest value, then the median is defined as:

$$\tilde{Y} = Y_{(N+1)/2} \text{ if N is odd}$$

$$\tilde{Y} = (Y_{N/2} + Y_{(N/2)+1})/2 \text{ if N is } \textit{even}$$

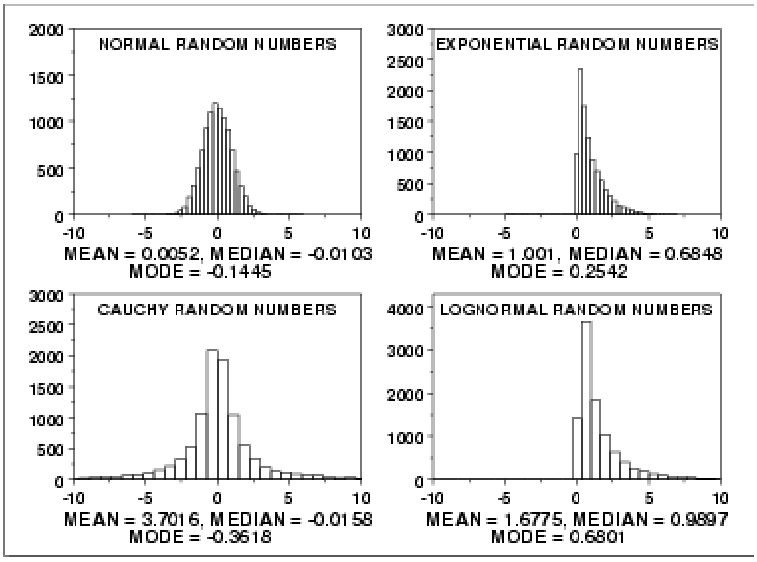
3. Mode

The mode is the value of the random sample that occurs with the greatest frequency. It is not necessarily unique. The mode is typically used in a qualitative fashion. For example, there may be a single dominant hump in the data perhaps two or more smaller humps in the data. This is usually evident from a histogram of the data.

When taking samples from continuous populations, we need to be somewhat careful in how we define the mode. That is, any specific value may not occur more than once if the data are continuous. What may be a more meaningful, if less exact measure, is the midpoint of the class interval of the histogram with the highest peak.

A natural question is why we have more than one measure of the typical value. The following example helps to explain why these alternative definitions are useful and necessary.

This plot shows histograms for 10,000 random numbers generated from a normal, an exponential, a Cauchy, and a lognormal distribution.



2.3 Measure of Dispersion

Q3. What is measures of dispersion.

Ans :

Dispersion in Statistics

It is a way of describing how spread out a set of data is. When a data set has a large value, the values in the set are widely scattered; when it is small the items in the set are tightly clustered. Very basically, this set of data has a small value:

1, 2, 2, 3, 3, 4

…and this set has a wider one:

0, 1, 20, 30, 40, 100

The spread of a data set can be described by a range of descriptive statistics including variance, standard deviation, and interquartile range. Spread can also be shown in graphs: dot plots, boxplots, and stem and leaf plots have a greater distance with samples that have a larger dispersion and vice versa.

Measures of Dispersion

- **Coefficient of Dispersion**

  A "catch-all" term for a variety of formulas, including distance between quartiles.

- **Standard deviation**

  Probably the most common measure. It tells you how spread out numbers are from the mean,

- **Index of Dispersion**

  A measure of dispersion commonly used with nominal variables.

- **Interquartile range (IQR)**

  Describes where the bulk of the data lies (the "middle fifty" percent).

- **Interdecile Range**

  The difference between the first decile (10%) and the last decile (90%).

- **Range**

  The difference between the smallest and largest number in a set of data.

- **Mean difference or difference in means**

  Measures the absolute difference between the mean value in two different groups in clinical trials.

- **Median absolute deviation (MAD)**

  The median of the absolute deviations from a data set's median.

- **Quartiles**

  Numbers that split the data into four quarters (first, second, third, and fourth quartiles).

2.4 Measure of Variability

Q4. Explain the various ways of Measure of variability?

Ans :

There are many ways to describe variability including :

(i) Range

(ii) Interquartile Range (IQR)

(iii) Variance

(iv) Standard Deviation

(i) Range

R = Maximum – Minimum

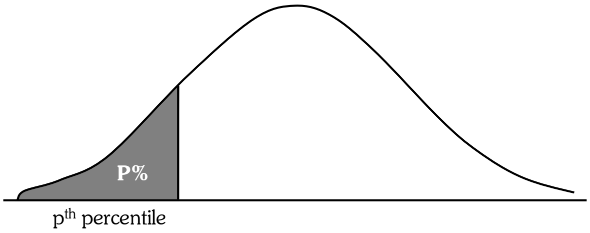
(a) Easy to calculate

(b) Very much affected by extreme values (ranges is not a resistant measure of variability)

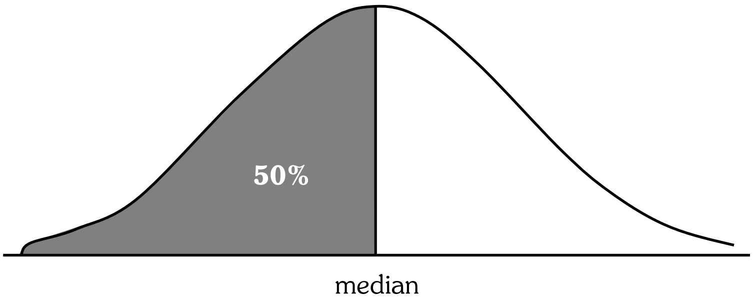
(ii) Interquartile Range (IQR)

In order to talk about interquartile range, we need to first talk about percentiles.

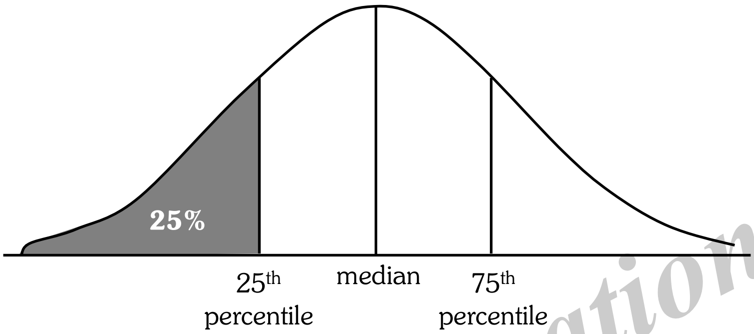
The pth percentile of the data set is a measurement such that after the data are ordered from smallest to largest, at most, p% of the data are at or below this value and at most, (100 – p)% at or above it.



Thus, the median is the 50th percentile. Fifty percent or the data values fall at or below the median.



Also, Q_1 = lower quartile = the 25th percentile and Q_3 = upper quartile = the 75th percentile.



Interquartile Range

It is the difference between upper and lower quartiles and denoted as IQR.

IQR = $Q_3 - Q_1$ = upper quartile - lower quartile = 75th percentile - 25th percentile.

Details about how to compute IQR will be given in Lesson 2.3.

Note: IQR is not affected by extreme values. It is thus a resistant measure of variability.

(iii) Variance

Two vending machines A and B drop candies when a quarter is inserted. The number of pieces of candy one gets is random. The following data are recorded for six trials at each vending machine:

Pieces of candy from vending machine A:

1, 2, 3, 3, 5, 4

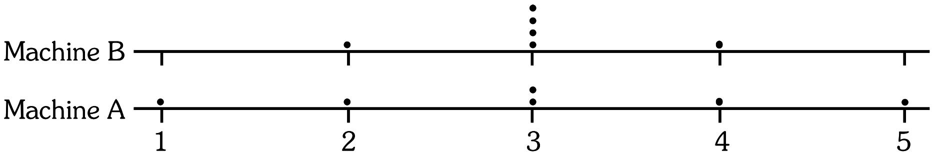
Mean = 3, Median = 3, Mode = 3

Pieces of candy from vending machine B:

2, 3, 3, 3, 3, 4

Mean = 3, Median = 3, Mode = 3

Dotplots for the pieces of candy from vending machine A and vending machine B:



They have the same center, but what about their spreads? One way to compare their spreads is to compute their standard deviations. In the following section, we are going to talk about how to compute the sample variance and the sample standard deviation for a data set.

Variance is the average squared distance from the mean.

Population variance is defined as:

$$\alpha^2 = \Sigma i = 1N\ (yi - \mu)\ /\ 2N$$

In this formula μ is the population mean and the summation is over all possible values of the population. N is the population size.

The sample variance that is computed from the sample and used to estimate α^2 is:

$$s2 = \Sigma i = 1n\ (yi - \bar{y}\ )2n - 1$$

Why do we divide by n – 1 instead of by n? Since μ is unknown and estimated by $\bar{y}$, the y_i's tend to be closer to $\bar{y}$ than to μ. To compensate, we divide by a smaller number, n – 1.

Sample Variance

It is the common default calculations used by software. When asked to calculate the variance or standard deviation of a set of data, assume - unless otherwise instructed - this is sample data and therefore calculating the sample variance and sample standard deviation.

Examples

Let's find S2 for the data set from vending machine A: 1, 2, 3, 3, 4, 5

$$\bar{y}\ = 1 + 2 + 3 + 3 + 4 + 56 = 3$$

$$s2 = (y1 - \bar{y}\ )2\ +\ +(yn - \bar{y}\ )\ 2n - 1$$

$$= (1 - 3)2 + (2 - 3)2 + (3 - 3)2$$

$$+(3 - 3)2 + (4 - 3)2 + (5 - 3)\ 26$$

$$-1 = 2$$

Calculate S^2 for the data set from vending machine B yourself and check that it is smaller than the S^2 for data set A. Work out your answer first, then click the graphic to compare answers.

(iv) Standard Deviation

The population standard deviation is notated by σ and found by $\sigma = \sigma^2 - \sqrt{\ }$ has the same unit as y_i's. This is a desirable property since one may think about the spread in terms of the original unit.

σ is estimated by the sample standard deviation *s* :

$$s = s2 - \sqrt{\ }$$

For the data set *A*,

$$s = 2 - \check{S} = 1.414 \text{ pieces of candy.}$$

2.5 Measures of Association

Q5. What is Measures of Association?

Ans :

Association

Association is concerned with how each variable is related to the other variable(s). In this case, the first measure that we will consider is the covariance between two variables j and k.

Population Covariance

The population covariance is a measure of the association between pairs of variables in a population. Here, the population covariance between variables j and k is

$$\sigma_{jk} = E\{(X_{ij} - \mu_j)\ (X_{ik} - \mu_k)\} \text{ for } i = 1, ..., n$$

Note that the product of the residuals $(X_{ij} - \mu_j)$ and $(X_{ik} - \mu_k)$ for variables j and k, respectively, is a function of the random variables X_{ij} and X_{ik}. Therefore, $(X_{ij} - \mu_j)(X_{ik} - \mu_k)$ is itself random, and has a population mean. The population covariance is defined to be the population mean of this product of residuals. We see that if either both variables are greater than their respective means, or if they are both less than their respective means, then the product of the residuals will be positive. Thus, if the value of variable j tends to be greater than its mean when the value of variable k is larger than its mean, and if the value of variable j tends to be less than its mean when the value of variable k is smaller than its mean, then the covariance will be positive.

Positive Population Covariances

It mean that the two variables are positively associated; variable j tends to increase with increasing values of variable k.

Negative Covariances Association

It can also occur. If one variable tends to be greater than its mean when the other variable is less than its mean, the product of the residuals will be negative, and you will obtain a negative population covariance. Variable j will tend to decrease with increasing values of variable k.

The population covariance σ_{jk} between variables j and k can be estimated by the sample covariance. This can be calculated using the formula below:

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{x}_j)(X_{ik}-\overline{x}_k) = \frac{\sum_{i=1}^{n} X_{ij}X_{ik} - (\sum_{i=1}^{n} X_{ij})(\sum_{i=1}^{n} X_{ik})/n}{n-1}$$

Just like in the formula for variance we have two expressions that make up this formula. The first half of the formula is most suitable for understanding the interpretation of the sample covariance, and the second half of the formula is used for calculation.

Looking at the first half of the expression, the product inside the sum is the residual differences between variable j and its mean times the residual differences between variable k and its mean. We can see that if either both variables tend to be greater than their respective means or less than their respective means, then the product of the residuals will tend to be positive leading to a positive sample covariance.

Conversely if one variable takes values that are greater than its mean when the opposite variable takes a value less than its mean, then the product will take a negative value. In the end, when you add up this product over all of the observations, you will end up with a negative covariance.

So, in effect, a positive covariance would indicate a positive association between the variables j and k. And a negative association is when the covariance is negative.

For computational purposes we will use the second half of the formula. For each subject, the product of the two variables is obtained, and then the products are summed to obtain the first term in the numerator. The second term in the numerator is obtained by taking the product of the sums of variable over the n subjects, then dividing the results by the sample size n. The difference between the first and second terms is then divided by n – 1 to obtain the covariance value.

Sample Covariance

Again, sample covariance is a function of the random data, and hence, is random itself. As before, the population mean of the sample covariance s_{jk} is equal the population covariance σ_{jk}; i.e.,

$$E(s_{jk}) = \sigma_{jk}$$

That is, the sample covariance s_{jk} is unbiased for the population covariance σ_{jk}.

The sample covariance is a measure of the association between a pair of variables:

- $s_{jk} = 0$ implies that the two variables are uncorrelated. (Note that this does not necessarily imply independence, we'll get back to this later.)
- $s_{jk} > 0$ implies that the two variables are positively correlated; i.e., values of variable j tend to increase with increasing values of variable k. The larger the covariance, the stronger the positive association between the two variables.
- $s_{jk} < 0$ implies that the two variables are negatively correlated; i.e., values of variable j tend to decrease with increasing values of variable k. The smaller the covariance, the stronger the negative association between the two variables.

Recall, that we had collected all of the population means of the p variables into a mean vector. Likewise, the population variances and covariances can be collected into the population variance-covariance matrix: This is also known by the name of population dispersion matrix.

$$\sum = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Note that the population variances appear along the diagonal of this matrix, and the covariance appear in the off-diagonal elements. So, the covariance between variables j and k will appear in row j and column k of this matrix.

The population variance-covariance matrix may be estimated by the sample variance-covariance matrix. The population variances and covariances in the above population variance-covariance matrix are replaced by the corresponding sample variances and covariances to obtain the sample variance-covariance matrix:

$$S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

Note that the sample variances appear along diagonal of this matrix and the covariances appear in the off-diagonal elements. So the covariance between variables j and k will appear in the jk-th element of this matrix.

Notes:

- S (the sample variance-covariance matrix) is symmetric; i.e., $s_{jk} = s_{kj}$.
- S is unbiased for the population variance covariance matrix Σ ; i.e.,

$$E(S) = \begin{pmatrix} E(s_1^2) & E(s_{12}) & \cdots & E(s_{1p}) \\ E(s_{21}) & E(s_2^2) & \cdots & E(s_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ E(s_{p1}) & E(s_{p2}) & \cdots & E(s_p^2) \end{pmatrix} = \Sigma$$

Matrix

It is a function of our random data, this means that the elements of this matrix are also going to be random, and the matrix on the whole is random as well. The statement Σ is unbiased means that the mean of each element of that matrix is equal to the corresponding elements of the population.

In matrix notation, the sample variance-covariance matrix may be computed used the following expressions:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})\,(X_i - \bar{x})'$$

$$= \frac{\sum_{i=1}^{n} X_i X_i' - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} X_i\right)' / n}{n-1}$$

Just as we have seen in the previous formulas, the first half of the formula is used in interpretation, and the second half of the formula is what is used for calculation purposes.

Looking at the second term you can see that the first term in the numerator involves taking the data vector for each subject and multiplying by its transpose. The resulting matrices are then added over the n subjects. To obtain the second term in the numerator, first compute the sum of the data vectors over the n subjects, then take the resulting vector and multiply by its transpose; then divide the resulting matrix by the number of subjects n. Take the difference between the two terms in the numerator and divide by n - 1.

Correlation

This suggests an alternative measure of association. The population correlation is defined to be equal to the population covariance divided by the product of the population standard deviations:

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$$

The population correlation may be estimated by substituting into the formula the sample covariances and standard deviations:

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n} X_{ij} X_{ik} - \left(\sum_{i=1}^{n} X_{ij}\right)\left(\sum_{i=1}^{n} X_{ik}\right)\Big/ n}{\sqrt{\left\{\sum_{i=1}^{n} X_{ij}^2 - \left(\sum_{i=1}^{n} X_{ij}\right)^2 \Big/ n\right\} \left\{\sum_{i=1}^{n} X_{ik}^2 - \left(\sum_{i=1}^{n} X_{ik}\right)^2 \Big/ n\right\}}}$$

It is very important to note that the population as well as the sample correlation must lie between -1 and 1.

$$-1 \leq \rho_{jk} \leq 1$$

$$-1 \leq r_{jk} \leq 1$$

Therefore:

$\rho_{jk} = 0$ indicates, as you might expect, that the two variables are uncorrelated .

ρ_{jk} close to +1 will indicate a strong positive dependence

ρ_{jk} close to –1 indicates a strong negative dependence

Sample correlation coefficients also have similar interpretation.

Correlation Matrix

For a collection of p variables, the correlation matrix is a p × p matrix that displays the correlations between pairs of variables. For instance, the value in the jth row and kth column gives the correlation between variables x_j and x_k . The correlation matrix is symmetric so that the value in the kth row and jth column is also the correlation between variables x_j and x_k . The diagonal elements of the correlation matrix are all identically equal to 1.

The sample correlation matrix is denoted as R.

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

2.6 PROBABILITY DISTRIBUTION

Q6. What is probability distribution? Explain the two classes of probability distribution?

Ans :

A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events. For instance, if the random variable X is used to denote the outcome of a coin toss (“the experiment”), then the probability distribution of X would take the value 0.5 for X = heads, and 0.5 for X = tails (assuming the coin is fair). Examples of random phenomena can include the results of an experiment or survey.

A probability distribution is defined in terms of an underlying sample space, which is the set of all possible outcomes of the random phenomenon being observed. The sample space may be the set of real numbers or a higher-dimensional vector space, or it may be a list of non-numerical values; for example, the sample space of a coin flip would be {heads, tails}.

Probability distributions are generally divided into two classes.

A discrete probability distribution (applicable to the scenarios where the set of possible outcomes is discrete, such as a coin toss or a roll of dice) can be encoded by a discrete list of the probabilities of the outcomes, known as a probability mass function.

A continuous probability distribution (applicable to the scenarios where the set of possible outcomes can take on values in a continuous range (e.g. real numbers), such as the temperature on a given day) is typically described by probability density functions (with the probability of any individual outcome actually being 0). The normal distribution is a commonly encountered continuous probability distribution. More complex experiments, such as those involving stochastic processes defined in continuous time, may demand the use of more general probability measures.

A probability distribution whose sample space is the set of real numbers is called univariate, while a distribution whose sample space is a vector space is called multivariate. A univariate distribution gives the probabilities of a single random variable taking on various alternative values; a multivariate distribution (a joint probability distribution) gives the probabilities of a random vector – a list of two or more random variables – taking on various combinations of values. Important and commonly encountered univariate probability distributions include the binomial distribution, the hypergeometric distribution, and the normal distribution. The multivariate normal distribution is a commonly encountered multivariate distribution.

2.7 Data Modeling

Q7. What is Data Modeling and Explain how data models deliver benefit?

Ans :

Data modeling is a process used to define and analyze data requirements needed to support the business processes within the scope of corresponding information systems in organizations. Therefore, the process of data modeling involves professional data modelers working closely with business stakeholders, as well as potential users of the information system.

There are three different types of data models produced while progressing from requirements to the actual database to be used for the information system.

- **Conceptual Data Modeling:** Identifies the highest-level relationships between different entities.
- **Enterprise Data Modeling:** Similar to conceptual data modeling, but addresses the unique requirements of a specific business.
- **Logical Data Modeling:** Illustrates the specific entities, attributes and relationships involved in a business function. Serves as the basis for the creation of the physical data model.
- **Physical Data Modeling:** Represents an application and database-specific implementation of a logical data model.

Data modeling techniques and methodologies are used to model data in a standard, consistent, predictable manner in order to manage it as a resource. The use of data modeling standards is strongly recommended for all projects requiring a standard means of defining and analyzing data within an organization, e.g., using data modeling:

- To assist business analysts, programmers, testers, manual writers, IT package selectors, engineers, managers, related organizations and clients to understand and use an agreed semi-formal model the concepts of the organization and how they relate to one another
- To manage data as a resource
- For the integration of information systems
- For designing databases/data warehouses (aka data repositories)

Data modeling may be performed during various types of projects and in multiple phases of projects. Data models are progressive; there is no such thing as the final data model for a business or application. Instead a data model should be considered a living document that will change in response to a changing business. The data models should ideally be stored in a repository so that they can be retrieved, expanded, and edited over time.

Whitten et al. (2004) determined two types of data modeling:

- **Strategic Data Modeling**

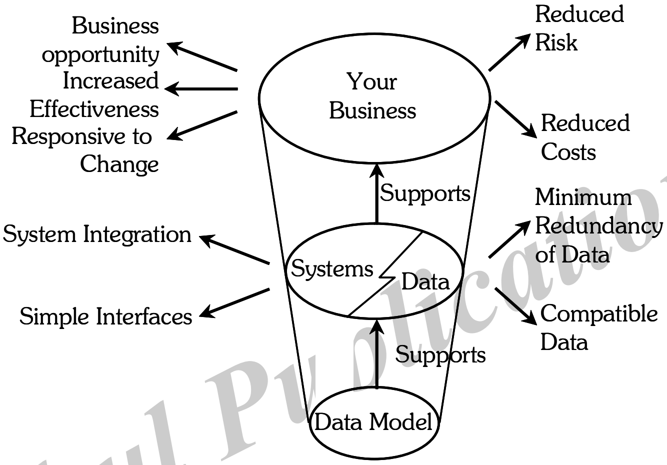
  This is part of the creation of an information systems strategy, which defines an overall vision and architecture for information systems. Information engineering is a methodology that embraces this approach.

- **Data Modeling During Systems Analysis**

  In systems analysis logical data models are created as part of the development of new databases.

  Data modeling is also used as a technique for detailing business requirements for specific databases. It is sometimes called database modeling because a data model is eventually implemented in a database.

Data Model



Data Models Deliver Benefit

Data models provide a framework for data to be used within information systems by providing specific definition and format. If a data model is used consistently across systems then compatibility of data can be achieved. If the same data structures are used to store and access data then different applications can share data seamlessly. The results of this are indicated in the diagram. However, systems and interfaces are often expensive to build, operate, and maintain. They may also constrain the business rather than support it. This may occur when the quality of the data models implemented in systems and interfaces is poor.

Data Models is Relatively Quick and Efficient

- Business rules, specific to how things are done in a particular place, are often fixed in the structure of a data model. This means that small changes in the way business is conducted lead to large changes in computer systems and interfaces. So, business rules need to be implemented in a flexible way that does not result in complicated dependencies, rather the data model should be flexible enough so that changes in the business can be implemented within the data model in a relatively quick and efficient way.

To Minimize Misinterpretation and Duplication

- Entity types are often not identified, or are identified incorrectly. This can lead to replication of data, data structure and functionality, together with the attendant costs of that duplication in development and maintenance. Therefore, data definitions should be made as explicit and easy to understand as possible to minimize misinterpretation and duplication.

To Interfaces within Different Systems

- Data models for different systems are arbitrarily different. The result of this is that complex interfaces are required between systems that share data. These interfaces can account for between 25-70% of the cost of current systems. Required interfaces should be considered inherently while designing a data model, as a data model on its own would not be usable without interfaces within different systems.

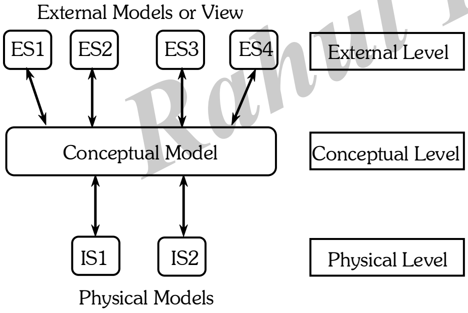
To Meet the Business Needs

- Data cannot be shared electronically with customers and suppliers, because the structure and meaning of data has not been standardised. To obtain optimal value from an implemented data model, it is very important to define standards that will ensure that data models will both meet business needs and be consistent.

Q8. Explain the different schemes of data modeling?

Ans:

There are three schemas of data modeling Conceptual, logical and physical schemas:



The ANSI/SPARC three level architecture. This shows that a data model can be an external model (or view), a conceptual model, or a physical model. This is not the only way to look at data models, but it is a useful way, particularly when comparing models.

In 1975 ANSI described three kinds of data-model instance:

Conceptual Schema

Describes the semantics of a domain (the scope of the model). For example, it may be a model of the interest area of an organization or of an industry. This consists of entity classes, representing kinds of things of significance in the domain, and relationships assertions about associations between pairs of entity classes. A conceptual schema specifies the kinds of facts or propositions that can be expressed using the model. In that sense, it defines the allowed expressions in an artificial "language" with a scope that is limited by the scope of the model. Simply described, a conceptual schema is the first step in organizing the data requirements.

Logical Schema

It describes the structure of some domain of information. This consists of descriptions of (for example) tables, columns, object-oriented classes, and XML tags. The logical schema and conceptual schema are sometimes implemented as one and the same.

Physical Schema

It describes the physical means used to store data. This is concerned with partitions, CPUs, tablespaces, and the like.

According to ANSI, this approach allows the three perspectives to be relatively independent of each other. Storage technology can change without affecting either the logical or the conceptual schema. The table/column structure can change without (necessarily) affecting the conceptual schema. In each case, of course, the structures must remain consistent across all schemas of the same data model.

Q9. Explain the different types of data modeling?

Ans :

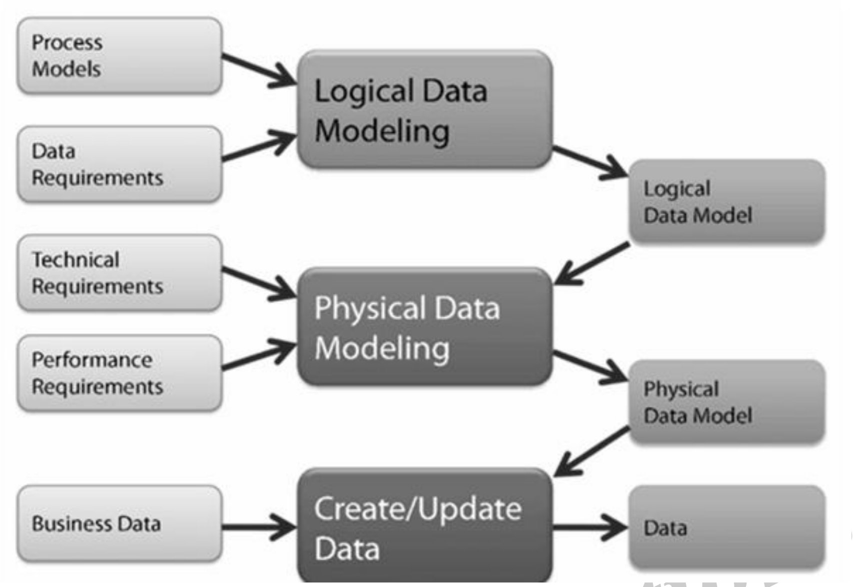
Types of Data Models

There are mainly three different types of data models:

1. **Conceptual:** This Data Model defines **WHAT** the system contains. This model is typically created by Business stakeholders and Data Architects. The purpose is to organize, scope and define business concepts and rules.

2. **Logical:** Defines **HOW** the system should be implemented regardless of the DBMS. This model is typically created by Data Architects and Business Analysts. The purpose is to developed technical map of rules and data structures.

3. **Physical**: This Data Model describes **HOW** the system will be implemented using a specific DBMS system. This model is typically created by DBA and developers. The purpose is actual implementation of the database.



Conceptual Model

The main aim of this model is to establish the entities, their attributes, and their relationships. In this Data modeling level, there is hardly any detail available of the actual Database structure.

The 3 basic tenants of Data Model are

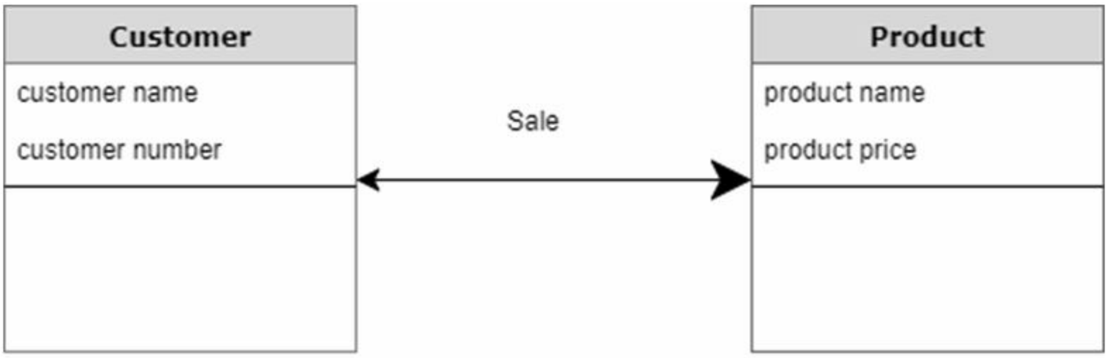
Entity: A real-world thing

Attribute: Characteristics or properties of an entity

Relationship: Dependency or association between two entities

For example

- Customer and Product are two entities. Customer number and name are attributes of the Customer entity
- Product name and price are attributes of product entity
- Sale is the relationship between the customer and product



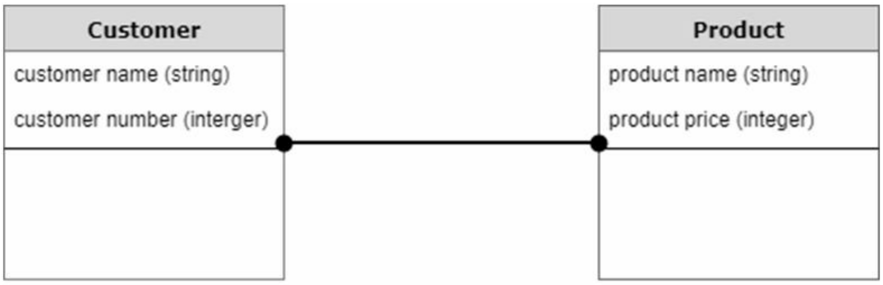
Characteristics of a conceptual data model

- Offers Organisation-wide coverage of the business concepts.
- This type of Data Models are designed and developed for a business audience.
- The conceptual model is developed independently of hardware specifications like data storage capacity, location or software specifications like DBMS vendor and technology. The focus is to represent data as a user will see it in the "real world."

Conceptual data models known as Domain models create a common vocabulary for all stakeholders by establishing basic concepts and scope.

Logical Data Model

Logical data models add further information to the conceptual model elements. It defines the structure of the data elements and set the relationships between them.



The advantage of the Logical data model is to provide a foundation to form the base for the Physical model. However, the modeling structure remains generic.

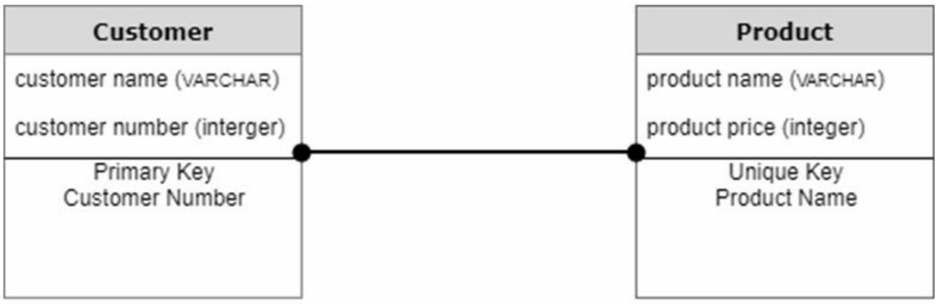
At this Data Modeling level, no primary or secondary key is defined. At this Data modeling level, you need to verify and adjust the connector details that were set earlier for relationships.

Characteristics of a Logical Data Model

- Describes data needs for a single project but could integrate with other logical data models based on the scope of the project.
- Designed and developed independently from the DBMS.
- Data attributes will have datatypes with exact precisions and length.
- Normalization processes to the model is applied typically till 3NF.

Physical Data Model

A Physical Data Model describes the database specific implementation of the data model. It offers an abstraction of the database and helps generate schema. This is because of the richness of meta-data offered by a Physical Data Model.



This type of Data model also helps to visualize database structure. It helps to model database columns keys, constraints, indexes, triggers, and other RDBMS features.

Characteristics of a Physical Data Model

- The physical data model describes data need for a single project or application though it maybe integrated with other physical data models based on project scope.
- Data Model contains relationships between tables that which addresses cardinality and nullability of the relationships.
- Developed for a specific version of a DBMS, location, data storage or technology to be used in the project.
- Columns should have exact datatypes, lengths assigned and default values.
- Primary and Foreign keys, views, indexes, access profiles, and authorizations, etc. are defined.

Q10. Explain the advantages and disadvantages of Data Model?

Ans :

Advantages and Disadvantages of Data Model:

Advantages of Data Model

- The main goal of a designing data model is to make certain that data objects offered by the functional team are represented accurately.
- The data model should be detailed enough to be used for building the physical database.
- The information in the data model can be used for defining the relationship between tables, primary and foreign keys, and stored procedures.
- Data Model helps business to communicate the within and across organizations.
- Data model helps to documents data mappings in ETL process
- Help to recognize correct sources of data to populate the model

Disadvantages of Data Model

- To developer Data model one should know physical data stored characteristics.
- This is a navigational system produces complex application development, management. Thus, it requires a knowledge of the biographical truth.
- Even smaller change made in structure require modification in the entire application.
- There is no set data manipulation language in DBMS.

Conclusion

- Data modeling is the process of developing data model for the data to be stored in a Database.
- Data Models ensure consistency in naming conventions, default values, semantics, security while ensuring quality of the data.
- Data Model structure helps to define the relational tables, primary and foreign keys and stored procedures.
- There are three types of conceptual, logical, and physical.
- The main aim of conceptual model is to establish the entities, their attributes, and their relationships.
- Logical data model defines the structure of the data elements and set the relationships between them.
- A Physical Data Model describes the database specific implementation of the data model.
- The main goal of a designing data model is to make certain that data objects offered by the functional team are represented accurately.

The biggest drawback is that even smaller change made in structure require modification in the entire application.

2.7.1 Discrete Probability Distribution

Q11. What is discrete probability distribution and explain what are the most common applications used.

Ans :

If a random variable is a discrete variable, its probability distribution is called a discrete probability distribution.

An example will make this clear. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable.

The probability distribution for this statistical experiment appears below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

Because the random variable is discrete, the above table is an example of a discrete probability distribution.

The most common applications of discrete probability distribution are

- Binomial distribution,
- Poisson distribution,
- Geometric distribution and
- Bernoulli distribution

1. Binomial Distribution

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes – no question, and each with its own boolean-valued outcome: a random variable containing a single bit of information: success/yes/true/one (with probability p) or failure/no/false/zero (with probability q = 1 " p). A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., n = 1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution remains a good approximation, and is widely used.

2. Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day. If receiving any particular piece of mail does not affect the arrival times of future pieces of mail, i.e., if pieces of mail from a wide range of sources arrive independently of one another, then a reasonable assumption is that the number of pieces of mail received in a day obeys a Poisson distribution. Other examples that may follow a Poisson include the number of phone calls received by a call center per hour and the number of decay events per second from a radioactive source.

3. Geometric Distribution

The geometric distribution represents the number of failures before you get a success in a series of Bernoulli trials. This discrete probability distribution is represented by the probability density function:

$$f(x) = (1 - p)^{x-1} p$$

For example, you ask people outside a polling station who they voted for until you find someone that voted for the independent candidate in a local election. The geometric distribution would represent the number of people who you had to poll before you found someone who voted independent. You would need to get a certain number of failures before you got your first success.

If you had to ask 3 people, then X=3; if you had to ask 4 people, then X=4 and so on. In other words, there would be X-1 failures before you get your success.

The geometric distribution represents the number of failures before you get a success in a series of Bernoulli trials. This discrete probability distribution is represented by the probability density function:

$$f(x) = (1 - p)^{x-1} p$$

For example, you ask people outside a polling station who they voted for until you find someone that voted for the independent candidate in a local election. The geometric distribution would represent the number of people who you had to poll before you found someone who voted independent. You would need to get a certain number of failures before you got your first success.

If you had to ask 3 people, then X=3; if you had to ask 4 people, then X=4 and so on. In other words, there would be X-1 failures before you get your success.

If X=n, it means you succeeded on the *n*th try and failed for n-1 tries. The probability of failing on your first try is 1-p. For example, if p = 0.2 then your probability of success is .2 and your probability of failure is 1 – 0.2 = 0.8. Independence (i.e. that the outcome of one trial does not affect the next) means that you can multiply the probabilities together. So the probability of failing on your second try is (1-p)(1-p) and your probability of failing on the nth-1 tries is $(1-p)^{n-1}$. If you succeeded on your 4th try, n = 4, n – 1 = 3, so the probability of failing up to that point is (1-p)(1-p)(1-p) = $(1-p)^3$.

Example

Sample question

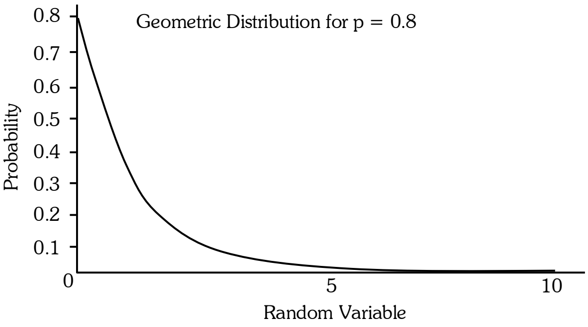
If your probability of success is 0.2, what is the probability you meet an independent voter on your third try?

Inserting 0.2 as p and with X = 3, the probability density function becomes:

$$f(x) = (1 - p)^{x-1} * p$$

$$P(X = 3) = (1 - 0.2)^{3-1} (0.2)$$

$$P(X = 3) = (0.8)^2 * 0.2 = 0.128.$$

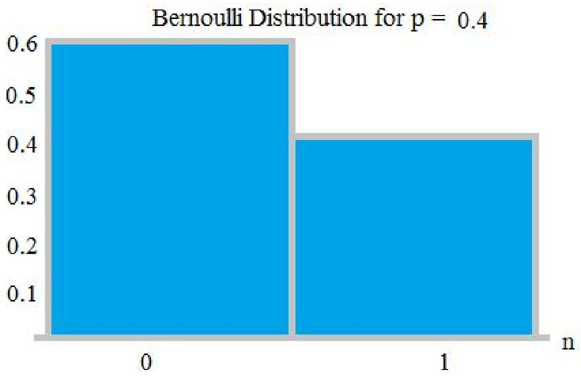


Theoretically, there are an infinite number of geometric distributions. The value of any specific distribution depends on the value of the probability p.

4. Bernoulli Distribution

A Bernoulli distribution is a discrete probability distribution for a Bernoulli trial - a random experiment that has only two outcomes (usually called a “Success” or a “Failure”). For example, the probability of getting a heads (a “success”) while flipping a coin is 0.5. The probability of “failure” is 1 – P (1 minus the probability of success, which also equals 0.5 for a coin toss). It is a special case of the binomial distribution for n = 1. In other words, it is a binomial distribution with a single trial (e.g. a single coin toss).

The probability of a failure is labeled on the x-axis as 0 and success is labeled as 1. In the following Bernoulli distribution, the probability of success (1) is 0.4, and the probability of failure (0) is 0.6:



The probability density function (pdf) for this distribution is $p^x (1-p)^{1-x}$, which can also be written as:

$$P(n) = \begin{cases} 1-p & \text{for } n=0 \\ p & \text{for } n=1 \end{cases}$$

The expected value for a random variable, X, from a Bernoulli distribution is:

$E[X] = p.$

For example, if p = .04, then E[X] = 0.4.

The variance of a Bernoulli random variable is:

$Var[X] = p(1 - p).$

Bernoulli Trial

A Bernoulli trial is one of the simplest experiments you can conduct in probability and statistics. It's an experiment where you can have one of two possible outcomes. For example, "Yes" and "No" or "Heads" and "Tails." A few more examples:

- **Coin tosses**: Record how many coins land heads up and how many land tails up.
- **Births**: How many boys are born and how many girls are born each day.
- **Rolling Dice**: The probability of a roll of two die resulting in a double six.

Coin tossing as a game of probability and chance has been around since Roman times.

Bernoulli trials are usually phrased in terms of **success** and **failure**. Success doesn't mean success in the usual way - it just refers to an outcome you want to keep track of. For example, you might want to find out how many boys are born each day, so you call a boy birth a "success" and a girl birth a "failure." In the dice rolling example, a double six die roll would be your "success" and everything else rolled would be considered a "failure."

2.7.2 Continuous Probability Distribution

Q12. What is continuous probability distribution explain with its example.

Ans :

Any variable can have two types of values. Either the values can be fix numbers which are also known as discrete values or a specified range that is known as continuous values. Based on these types of values a data set is defined as continuous or discrete. In continuous data type, the values can be lying anywhere within the range that is specified.

For Example

The time required to drive back from office to home is always continuous.

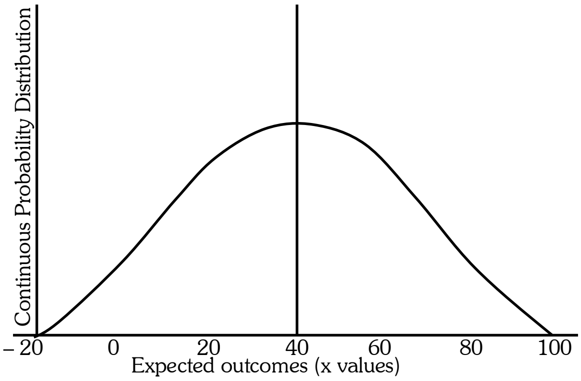
So the probability distribution over a random variable "Y" where "Y" takes continuous values is termed as continuous probability distribution.

There are three basic differences between a continuous and a discrete probability distribution:

i) The probability that a continuous variable will take a specific value is equal to zero.

ii) Because of this, we can never express continuous probability distribution in a tabular form.

iii) Thus we require an equation or a formula to describe such kind of distribution. Such equation is termed as probability density function.



Formula

We will discuss about probability distribution function here to understand the concept of continuous probability distribution:

A probability density function given over a rage $a \le x \le b$ satisfies the following:

$$f(x) \ge 0$$

for all values of 'x' lying between a and b.

The total area covered under the curve of the function lying in the range a and b is equal to 1.

So the probability of this continuous variable can now be found out by integrating this density function with respect to 'x' over the interval [a, b].

$$P(a \le x \le b) = \int bd\ f(x)dx$$

Common examples of continuous probability distributions are:

1. Uniform Distribution
2. Normal Distribution
3. Chi Squared Distribution etc.

1. Uniform Distribution

The continuous uniform distribution or rectangular distribution is a family of symmetric probability distributions such that for each member of the family, all intervals of the same length on the distribution's support are equally probable. The support is defined by the two parameters, a and b, which are its minimum and maximum values. The distribution is often abbreviated U(a,b). It is the maximum entropy probability distribution for a random variate X under no constraint other than that it is contained in the distribution's support.

2. Normal Distribution

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the "bell curve," although the tonal qualities of such a bell would be less than pleasing. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss.Abraham de Moivre first discovered the normal distribution.

The Normal distributions can differ in their means and in their standard deviations. The below Figure shows three normal distributions. The top (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in middle (the middle distribution) has a mean of 0 and a standard deviation of 1, and the distribution in bottom (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

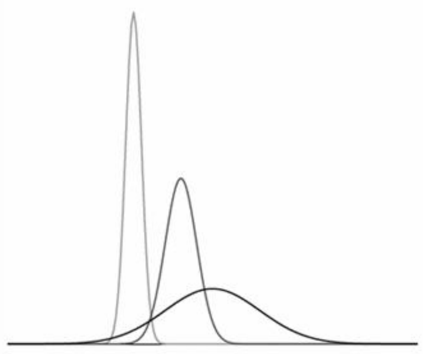


Figure 1: Normal distributions differing in mean and standard deviation

The density of the normal distribution (the height for a given value on the x axis) is shown below. The parameters μ and σ are the mean and standard deviation, respectively, and define the normal distribution. The symbol *e* is the base of the natural logarithm and π is the constant pi.

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Since this is a non-mathematical treatment of statistics, do not worry if this expression confuses you. We will not be referring back to it in later sections.

Seven features of normal distributions are listed below. These features are illustrated in more detail in the remaining sections of this chapter.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).

6. 68% of the area of a normal distribution is within one standard deviation of the mean.

7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

3. Chi Squared Distribution

The chi-squared distribution (also chi-square or χ^2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. The chi-square distribution is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing or in construction of confidence intervals. When it is being distinguished from the more general noncentral chi-squared distribution, this distribution is sometimes called the central chi-squared distribution.

The chi-squared distribution is used in the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.

2.7.3 Random Sampling from Probability Distribution

Q13. What is random sampling from probability distribution?

Ans :

Random Sampling

Random sampling is a procedure for sampling from a population in which (a) the selection of a sample unit is based on chance and (b) every element of the population has a known, non-zero probability of being selected.

Random sampling helps produce representative samples by eliminating voluntary response bias and guarding against under coverage bias.

Other names for random sampling include representative and proportionate sampling because all groups should be proportionately represented. Consider what might happen if a telephone directory were used as a source for randomly selecting survey participants. Some people have no phone, others have multiple phones and corresponding listings. Still others have unlisted phone numbers. In affluent areas unlisted phone numbers may approach half the population! Now-a-days many are giving up lands lines and use cell phone exclusively. Cell phone directories are controversial at best. Pollsters commonly use computers to generate and dial phone numbers in an attempt to circumvent these problems. However, many people consider such use of the telephone as an invasion of their privacy and refusals or hang-ups may well significantly influence the outcome. Some of us have learned to recognize these computer dialers and quickly hang up. Such are the pitfalls which must be carefully considered in designing an experiment, study, or survey.

In this technique, each member of the population has an equal chance of being selected as subject. The entire process of sampling is done in a single step with each subject selected independently of the other members of the population.

There are many methods to proceed with simple random sampling.

A random sample can be selected by using:

1. Lottery Method.
2. Random Numbers

1. Lottery Method

One of the most primitive and mechanical would be the lottery method. In this method each member gets a unique number. Each number is placed and mixed thoroughly and one of the blindfolded researcher picks the number from the bowl. Individuals bearing the number picked by the researcher are the subjects for the study. This method is advisable to be used for a population with a small number of members.

2. Random Numbers

Random number tables, more recently known as random number generators, tell researchers to select subjects at an interval generated randomly. To use we need to first identify N units in the population with numbers from 1 to N.

Randomly select any page of the random number table and pick the numbers in any row, column or diagonal at random. Population units corresponding to the numbers selected constitute the random sample.

Random sampling from probability distribution

When simulating any system with randomness, sampling from a probability distribution is necessary. Usually, you'll just need to sample from a normal or uniform distribution and thus can use a built-in random number generator. However, for the time when a built-in function does not exist for your distribution, here's a simple algorithm

2.7.4 Data Modeling

Q14. What is Data Modelling?

Ans :

Data modeling is the process of creating a data model for the data to be stored in a Database. This data model is a conceptual representation of

- Data objects
- The associations between different data objects
- The rules.

Data modeling helps in the visual representation of data and enforces business rules, regulatory compliances, and government policies on the data. Data Models ensure consistency in naming conventions, default values, semantics, security while ensuring quality of the data.

Data model emphasizes on what data is needed and how it should be organized instead of what operations need to be performed on the data. Data Model is like architect's building plan which helps to build a conceptual model and set the relationship between data items.

Data Models techniques

The two types of Data Models techniques are

1. Entity Relationship (E-R) Model
2. UML (Unified Modelling Language)

Uses of Data Model

The primary goal of using data model are:

- Ensures that all data objects required by the database are accurately represented. Omission of data will lead to creation of faulty reports and produce incorrect results.
- A data model helps design the database at the conceptual, physical and logical levels.
- Data Model structure helps to define the relational tables, primary and foreign keys and stored procedures.
- It provides a clear picture of the base data and can be used by database developers to create a physical database.
- It is also helpful to identify missing and redundant data.
- Though the initial creation of data model is labor and time consuming, in the long run, it makes your IT infrastructure upgrade and maintenance cheaper and faster.

2.7.5 Distribution Fitting

Q15. Explain in detail about distribution fitting?

Ans :

Probability distribution fitting or simply distribution fitting is the fitting of a probability distribution to a series of data concerning the repeated measurement of a variable phenomenon.

The aim of distribution fitting is to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval.

There are many probability distributions (see list of probability distributions) of which some can be fitted more closely to the observed frequency of the data than others, depending on the characteristics of the phenomenon and of the distribution. The distribution giving a close fit is supposed to lead to good predictions.

In distribution fitting, therefore, one needs to select a distribution that suits the data well.

Techniques of fitting

The following techniques of distribution fitting exist:

1. **Parametric Methods**
   - Method of moments
   - Maximum spacing estimation
   - Method of l-moments
   - Maximum likelihood estimation

2. Regression Method

Parametric Methods, by which the parameters of the distribution are calculated from the data series.The parametric methods are:

i) **Method of Moments:** The method of moments is a method of estimation of population parameters. One starts with deriving equations that relate the population moments (i.e., the expected values of powers of the random variable under consideration) to the parameters of interest. Then a sample is drawn and the population moments are estimated from the sample. The equations are then solved for the parameters of interest, using the sample moments in place of the (unknown) population moments. This results in estimates of those parameters. The method of moments was introduced by Pafnuty Chebyshev in 1887.

ii) **Maximum spacing estimation (mse or msp), or maximum product of spacing estimation (mps):** It is a method for estimating the parameters of a univariate statistical model. The method requires maximization of the geometric mean of spacings in the data, which are the differences between the values of the cumulative distribution function at neighbouring data points.

The concept underlying the method is based on the probability integral transform, in that a set of independent random samples derived from any random variable should on average be uniformly distributed with respect to the cumulative distribution function of the random variable. The MPS method chooses the parameter values that make the observed data as uniform as possible, according to a specific quantitative measure of uniformity.

One of the most common methods for estimating the parameters of a distribution from data, the method of maximum likelihood (MLE), can break down in various cases, such as involving certain mixtures of continuous distributions. In these cases the method of maximum spacing estimation may be successful.

iii) **Method of L-Moments:** It is a sequence of statistics used to summarize the shape of a probability distribution. They are linear combinations of order statistics (L-statistics) analogous to conventional moments, and can be used to calculate quantities analogous to standard deviation, skewness and kurtosis, termed the L-scale, L-skewness and L-kurtosis respectively (the L-mean is identical to the conventional mean). Standardised L-moments are called L-moment **ratios** and are analogous to standardized moments. Just as for conventional moments, a theoretical distribution has a set of population L-moments. Sample L-moments can be defined for a sample from the population, and can be used as estimators of the population L-moments.

iv) **Maximum Likelihood Estimation (mle) Method** : is a method of estimating the parameters of a statistical model, given observations. MLE attempts to find the parameter values that maximize the likelihood function, given the observations. The resulting estimate is called a maximum likelihood estimate, which is also abbreviated as MLE.

The method of maximum likelihood is used with a wide range of statistical analyses. As an example, suppose that we are interested in the heights of adult female penguins, but are unable to measure the height of every penguin in a population (due to cost or time constraints). Assuming that the heights are normally distributed with some unknown mean and variance, the mean and variance can be estimated with MLE while only knowing the heights of some sample of the overall population. MLE would accomplish that by taking the mean and variance as parameters and finding particular parametric values that make the observed results the most probable given the normal model.

Regression Method, using a transformation of the cumulative distribution function so that a linear relation is found between the cumulative probability and the values of the data, which may also need to be transformed, depending on the selected probability distribution. In this method the cumulative probability needs to be estimated by the plotting position.

Short Question and Answers

1. Population

Ans :

The field of inferential statistics enables you to make educated guesses about the numerical characteristics of large groups. The logic of sampling gives you a way to test conclusions about such groups using only a small portion of its members.

A population is a group of phenomena that have something in common. The term often refers to a group of people, as in the following examples:

- All registered voters in Crawford County
- All members of the International Machinists Union
- All Americans who played golf at least once in the past year.

2. Sample

Ans :

- Often, researchers want to know things about populations but do not have data for every person or thing in the population.
- If a company's customer service division wanted to learn whether its customers were satisfied, it would not be practical (or perhaps even possible) to contact every individual who purchased a product. Instead, the company might select a sample of the population.
- A sample is a smaller group of members of a population selected to represent the population. In order to use statistics to learn things about the population, the sample must be random.
- A random sample is one in which every member of a population has an equal chance of being selected.
- The most commonly used sample is a simple random sample. It requires that every possible sample of the selected size has an equal chance of being used.
- The field of inferential statistics enables you to make educated guesses about the numerical characteristics of large groups.

3. Random Sample

Ans :

A sample is a smaller group of members of a population selected to represent the population. In order to use statistics to learn things about the population, the sample must be random. A random sample is one in which every member of a population has an equal chance of being selected. The most commonly used sample is a simple random sample. It requires that every possible sample of the selected size has an equal chance of being used.

4. Probability Distribution

Ans :

A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events. For instance, if the random variable X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 for X = heads, and 0.5 for X = tails (assuming the coin is fair). Examples of random phenomena can include the results of an experiment or survey.

5. Characteristics of a Conceptual Data Model

Ans :

- Offers Organisation-wide coverage of the business concepts.
- This type of Data Models are designed and developed for a business audience.
- The conceptual model is developed independently of hardware specifications like data storage capacity, location or software specifications like DBMS vendor and technology. The focus is to represent data as a user will see it in the "real world."

6. Probability Distribution

Ans :

If a random variable is a discrete variable, its probability distribution is called a discrete probability distribution.

An example will make this clear. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable.

7. Poisson Distribution

Ans :

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

8. Bernoulli Distribution

Ans :

A Bernoulli distribution is a discrete probability distribution for a Bernoulli trial - a random experiment that has only two outcomes (usually called a "Success" or a "Failure"). For example, the probability of getting a heads (a "success") while flipping a coin is 0.5. The probability of "failure" is 1 – P (1 minus the probability of success, which also equals 0.5 for a coin toss). It is a special case of the binomial distribution for n = 1. In other words, it is a binomial distribution with a single trial (e.g. a single coin toss).

9. Lottery Method

Ans :

One of the most primitive and mechanical would be the lottery method. In this method each member gets a unique number. Each number is placed and mixed thoroughly and one of the blindfolded researcher picks the number from the bowl. Individuals bearing the number picked by the researcher are the subjects for the study. This method is advisable to be used for a population with a small number of members.

10. Data Modelling

Ans :

Data modeling is the process of creating a data model for the data to be stored in a Database. This data model is a conceptual representation of

- Data objects
- The associations between different data objects
- The rules.

Data modeling helps in the visual representation of data and enforces business rules, regulatory compliances, and government policies on the data. Data Models ensure consistency in naming conventions, default values, semantics, security while ensuring quality of the data.

Data model emphasizes on what data is needed and how it should be organized instead of what operations need to be performed on the data. Data Model is like architect's building plan which helps to build a conceptual model and set the relationship between data items.

11. Distribution Fitting

Ans :

Probability distribution fitting or simply distribution fitting is the fitting of a probability distribution to a series of data concerning the repeated measurement of a variable phenomenon.

The aim of distribution fitting is to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval.

There are many probability distributions (see list of probability distributions) of which some can be fitted more closely to the observed frequency of the data than others, depending on the characteristics of the phenomenon and of the distribution. The distribution giving a close fit is supposed to lead to good predictions.

UNIT III

Predictive Analytics : Karl Pearson Correlation Techniques - Multiple Correlation-Spearman's rank correlation-Simple and Multiple regression-Regression by the method of least squares - Building good regression models - Regression with categorical independent variables - Linear Discriminant Analysis - One way and Two way ANOVA

3.1 Karl Pearsons Correlation

Q1. What is Karl Pearson's of correlation ?

Ans :

Correlation is the study of the linear relationship between two variables. When there is a relationship of quantitative measure between two set of variables, the appropriate statistical tool for measuring the relationship and expressing each in a precise way is known as correlation.

For example, there is a relationship between the heights and weights of persons, demand and prices of commodities etc.

Meaning and Definition of Correlation

Correlation analysis is the statistical tool we can used to describe the degree to which one variable is linearly related to another.

According to Croxton and Cowden, "The appropriate statistical tool for discovering and measuring the relationship of quantitative nature and expressing it in brief formula is known as correlation".

According to Tippet, "The effects of correlation are to reduce the range of uncertainty of our prediction".

The coefficient of correlation measures the degree of relationship between two set of figures. As the reliability of estimates depend upon the closeness of the relationship it is imperative that utmost care be taken while interpreting the value of coefficient of correlation, otherwise wrong conclusion can be drawn.

Significance of Measuring Correlation

1. Correlation is very useful to economists to study the relationship between variables, like price and quantity demanded. To businessmen, it helps to estimate costs, sales, price and other related variables.
2. In economic theory we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply and quantity demanded; convenience, amenities and service standards are related to customer retention; yield a crop related to quantity of fertilizer app . . type of soil, quality of seeds, rainfall and so on. Correlation analysis helps in measuring the degree at association and direction of such relationship.
3. The relation between variables can be verified and tested for significance, with the help of the correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.
4. The coefficient of correlation is a relative measure and we can compare the relationship between variables, which are expressed in different units.
5. Correlations are useful in the areas of healthcare such as determining the validity and reliability of clinical measures or in expressing how health problems are related to certain biological or environmental factors. For example, correlation coefficient can be used to determine the degree of inter-observer reliability for two doctors who are assessing a patient's disease.

6. Sampling error can also be calculated.

7. Correlation is the basis for the concept of regression and ratio of variation.

8. The decision making is heavily felicitated by reducing the range of uncertainty and hence empowering the predictions.

Correlation is a statistical tool for studying the relationship between two or more variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of relationship between the two variable. Two variables said to be correlated, if the change in one variable results in a corresponding change in the other.

Q2. Explain the different types of correlation.

Ans :

Types of Correlation

Broadly speaking, there are four types of correlation, namely,

A) Positive correlation,

B) Negative correlation,

C) Linear correlation and

D) Non-Linear Correlation.

A) Positive correlation

If the values of two variables deviate in the same direction i.e., if increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, the corresponding correlation is said to be positive or direct.

Examples

i) Sales revenue of a product and expenditure on Advertising.

ii) Amount of rain fall and yield of a crop (up to a point)

iii) Price of a commodity and quantity of supply of a commodity

iv) Height of the Parent and the height of the Child.

v) Number of patients admitted into a Hospital and Revenue of the Hospital.

vi) Number of workers and output of a factory.

i) Perfect Positive Correlation : If the variables X and Y are perfectly positively related to each other then, we get a graph as shown in fig.below.

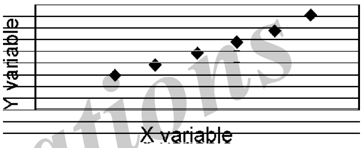


Fig. : Perfect Positive Correlation (r = +1)

ii) Very High Positive Correlation : If the variables X and Y are related to each other with a very high degree of positive relationship then we can notice a graph as in figure below.

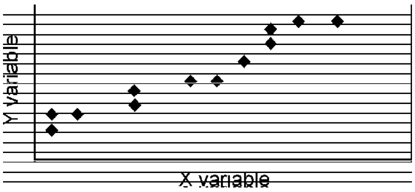


Fig. : Very High Positive Correlation (r = nearly + 1)

iii) Very Low Positive Correlation : If the variables X and Y are related to each other with a very low degree of positive relationship then we can notice a graph as in fig.below.

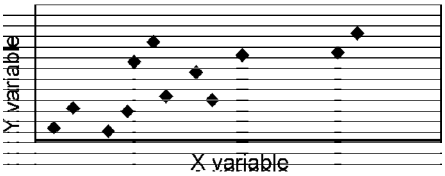


Fig.: Very Low Positive Correlation (r = near to +0)

B) Negative Correlation

Correlation is said to be negative or inverse if the variables deviate in the opposite direction i.e., if the increase (decrease) in the values of one variable results, on the average, in a corresponding decrease (increase) in the values of the other variable.

Examples

1. Price and demand of a commodity.
2. Sales of Woolen garments and the day temperature.

i) **Perfect Negative Correlation :** If the variables X and Y are perfectly negatively related to each other then, we get a graph as shown in fig.below.

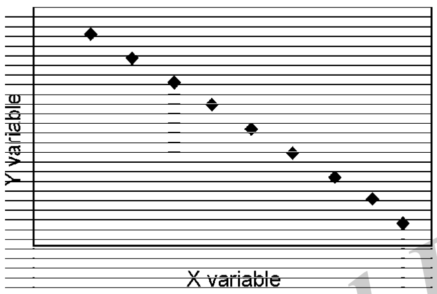


Fig. : Perfect Negative Correlation (r = –1)

ii) **Very High Negative Correlation :** If the variables X and Y are related to each other with a very high degree of negative relationship then we can notice a graph as in fig.below.

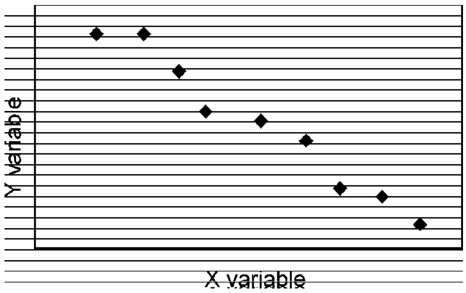


Fig. : Very High Negative Correlation (r = near to –1)

iii) **Very low Negative Correlation :** If the variables X and Y are related to each other with a very low degree of negative relationship then we can notice a graph as in fig.below.

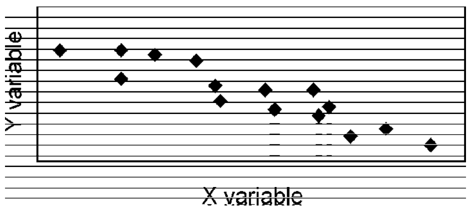


Fig.: Very Low Negative Correlation (r = near to 0 but negative)

iv) **No Correlation :** If the scatter diagram show the points which are highly spread over and show no trend or patterns we can say that there is no correlation between the variables.

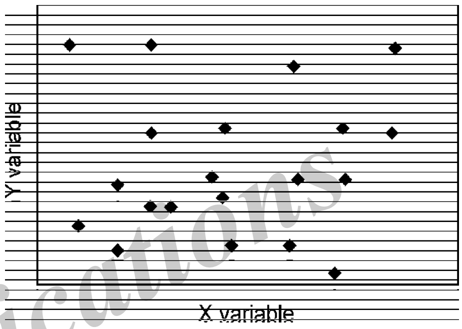


Fig. : No Correlation (r = 0)

C) Linear Correlation

Two variables are said to be linearly related if corresponding to a unit change in one variable there is a constant change in the other variable over the entire range of the values.

If two variables are related linearly, then we can express the relationship as

$$Y = a + b X$$

where '**a**' is called as the "intercept" (If X = 0, then Y = a) and 'b' is called as the "rate of change" or slope.

If we plot the values of X and the corresponding values of Y on a graph, then the graph would be a straight lines as shown in fig.below.

Example

X	1	2	3	4	5
Y	8	11	14	17	20

For a unit change in the value of x, a constant 3 units changes in the value of y can be noticed. The above can be expressed as : Y = 5 + 3x.

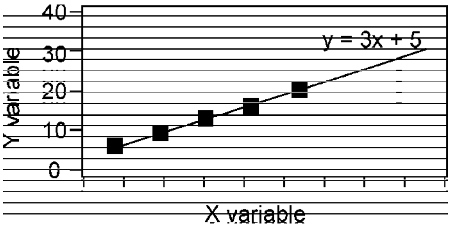


Fig.: Linear Correlation

D) Non Linear (Curvilinear) Correlation

If corresponding to a unit change in one variable, the other variable does not change in a constant rate, but change at varying rates, then the relationship between two variables is said to be nonlinear or curvilinear as shown in fig.below. In this case, if the data are plotted on the graph, we do not get a straight line curve.

Mathematically, the correlation is non-linear if the slope of the plotted curve is not constant. Data relating to Economics, Social Science and Business Management do exhibit often non-linear relationship. We confine ourselves to linear correlation only.

Example

X	–3	–2	–1	0	1	2	3
Y	9	4	1	0	1	4	9

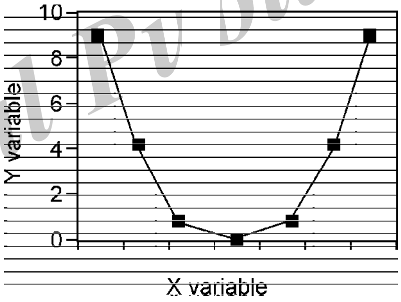


Fig.: Non Linear Correlation

Properties of Correlation

i) The value of correlation ‘r’ varies between [–1, +1]. This indicates that the r values does not exceed unity.

ii) Sign of the correlation sign of the Covariance.

iii) If r = –1 variables and perfectly negatively correlated.

iv) If r = + 1 variables are perfectly positively correlated.

If r = 0 variables are not correlated in a linear fashion. There may be non-linear relationship between variables.

Correlation coefficient is independent of change of scale and shifting of origin. In other words, Shifting the origin and change the scale do not have any effect on the value of correlation.

Q3. What are the methods of correlation ?

Ans :

Correlation can be determined by the following methods:

1. **Graphic Methods**

   (i) Scatter Diagram

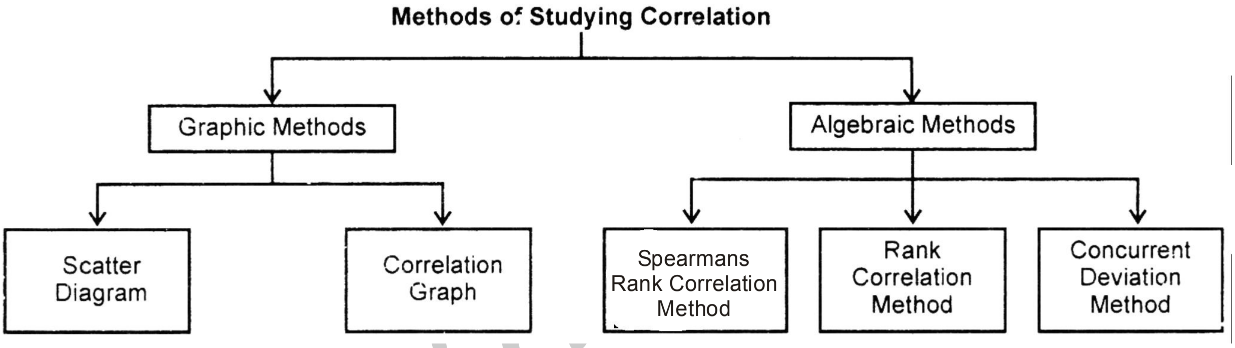
   (ii) Correlation Graph

2. **Algebraic Methods**

   (i) Karl Pearson's Coefficient of Correlation

   (ii) Spearman's Rank Correlation Method

   (iii Concurrent Deviation Method


Methods of Studying Correlation
Graphic Methods
Algebraic Methods
Scatter Diagram
Correlation Graph
Spearmans Rank Correlation Method
Rank Correlation Method
Concurrent Deviation Method

Q4. Explain the properties of Karl Pearson's Coefficient of Correlation.

Ans :

1. It is based on Arithmetic Mean and Standard Deviation.
2. It lies between -1
3. It measures both direction as well as degree of change. If r is less than 0, there is negative correlation, which means the direction of change of the two variables will be opposite. If r is more than 0, there is positive correlation, which means the direction of change of the two variables will be same. Higher the value of r, greater is the degree of correlation. Hence, Karl Pearson's coefficient of correlation is said to be the ideal measure of correlation.
4. It is independent of change in scale. In other words, if a constant amount is added/ subtracted from all values of a variable, the value of r does not change.
5. It is independent of change in origin. Thus, if a constant amount is multiplied with or divides all values of a variable, 'r' does not change.
6. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X.

7. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.

8. It takes into account all items of the variable(s).

9. It does not prove causation but is simply a measure of co-variation.

10. Correlation coefficient of two variables X and Y is the Geometric Mean of two regression coefficients, regression coefficient of X on Y and regression coefficient of Y on X. Symbolically,

$$r = \text{Square root of } (b_{xy} \times b_{yx})$$

11. Correlation coefficient can be calculated between two unrelated variables and such a number can be misleading. Such correlation is called accidental correlation, spurious correlation or non sense correlation.

Merits of Karl Pearson's Coefficient of Correlation

1. It takes into account all items of the variable(s).
2. It is a numerical measure and hence more objective.
3. It measures both direction as well as degree of change.
4. It facilitates comparisons between two series.
5. It is capable of further Algebraic treatment
6. It is more practical and hence popular and is more commonly used.

Demerits of Karl Pearson's Coefficient of Correlation

1. It is not easy to calculate as complex formulae are involved.
2. It is more time consuming compared to methods such as rank correlation
3. It assumes a linear relationship between the two variables which may not be correct
4. It is impacted by extreme values as it is based on mean and standard deviation.
5. It is not easy to interpret.

Q5. Explain the Computation of Karl Pearson's Coefficient of Correlation.

Ans :

i) Direct Method : When deviations are taken from actual mean

$$r = \frac{\sum xy}{N\sigma \times \sigma y}$$

However, this formula is transformed in the following form

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Where

$$x = X - \bar{X}, \text{ and } y = Y - \bar{Y}$$

Steps :

1. Find the means of the two series (X , Y)
2. Take the deviations of X series from the mean of X and denote these deviations as x.
3. Square these deviations and obtain the total. Denote it as Sx^2.
4. Take the deviations of Y series from the Mean of Y and denote these deviations as y.
5. Square these deviations, obtain the total and denote it as Sy^2.

6, Multiply the deviations of X and Y series, obtain the total and denote it Sxy.

7. Substitute the above values in the formula.

Short-Cut Method

When deviations are taken from assumed mean

When actual mean is in fraction, then the above formula becomes tedious. In such cases, assumed mean is used for calculating correlation. The formula is

$$r = \frac{\sum dxdy - \frac{\sum dx \cdot \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

Where

$\Sigma dxdy$ = Sum of the product of the deviations of X and Y series from their assumed means.

Σdx^2 = Sum of the squares of the deviations of X series from an assumed mean.

Σdy^2 = Sum of the squares of the deviations of Y series from an assumed mean.

Σdx = Sum of the deviations of X series from an assumed mean.

Σdy = Sum of the deviations of Y Series from an assumed mean.

N = No. of Pairs of observations.

The values of coefficient of correlation as obtained by above formulae will always lie between ±1. When there is perfect positive correlation its value is +1 and when there is perfect negative correlation, its value is-1. When r = 0 means that there is no relationship between the two variables. We normally get values which lie between +1 and -1.

Q6. Explain the Probable Error of the Coefficient of correlation and its interpretation ?

Ans :

The probable error of the coefficient of correlation helps in interpretation. The probable error of the coefficient of correlation is obtained as follows :

$$\text{P.E. of } r = 0.6745 \frac{1-r^2}{\sqrt{N}}$$

Where

r = Coefficient of correlation;

N = Number of pairs of observations.

If the probable error is added to and subtracted from the coefficient of correlation, it would give two such limits within which we can reasonably expect the value of coefficient of correlation to vary.

Symbolically P (rho) = r ± P. E.

Where 'P' denotes the correlation in the population. Suppose, the Coefficient of correlation for a pair of 10 observations is 0.8 and its P.E. is 0.05. the limits of the correlation in the population would be r ± P. E. i.e. 0.8 ± 0 – .05 or 0.75 – 0.85. If the value of r is less than the probable error then r is not at all significant, i.e. there is no evidence of correlation. If the value of *r* is more than six times the probable error, it is significant. Hence it can be said that *r* is significant, when

$$r > 6 \text{ P. E.} \quad \text{or} \quad \frac{r}{\text{P.E}} > 6$$

Example 1 :

Find the value of the correlation coefficient from the following table :

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Step 1: *Make a chart.* Use the given data, and add three more columns :

xy, x2, and y2.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99			
2	21	65			
3	25	79			
4	42	75			
5	57	87			
6	59	81			

Step 2 : Multiply x and y together to fill the xy column. For example, row 1 would be 43 × 99 = 4,257.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257		
2	21	65	1365		
3	25	79	1975		
4	42	75	3150		
5	57	87	4959		
6	59	81	4779		

Step 3 : Take the square of the numbers in the x column, and put the result in the x2 column.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	
2	21	65	1365	441	
3	25	79	1975	625	
4	42	75	3150	1764	
5	57	87	4959	3249	
6	59	81	4779	3481	

Step 4 : Take the square of the numbers in the y column, and put the result in the y2 column.

Subject	Age X	Glucose Level Y	XY	X2	Y2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561

Step 5 : Add up all of the numbers in the columns and put the result at the bottom of the column. The Greek letter sigma (Ó) is a short way of saying "sum of."

Subject	Age X	Glucose Level Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561

Step 6 : Use the following correlation coefficient formula.

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2]\,[n\sum y^2-(\sum y)^2]}}$$

The answer is : 2868 / 5413.27 = 0.529809

Click here if you want easy, step-by-step instructions for solving this formula.

From our table :

- Σx = 247
- Σy = 486
- Σxy = 20,485
- Σx2 = 11,409
- Σy2 = 40,022
- n is the sample size, in our case = 6

The correlation coefficient

$= 6(20{,}485) - (247 \times 486)/[\sqrt{[[6(11{,}409)} - (2472)] \times [6(40{,}022) - 4862]]]$

= 0.5298

The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation.

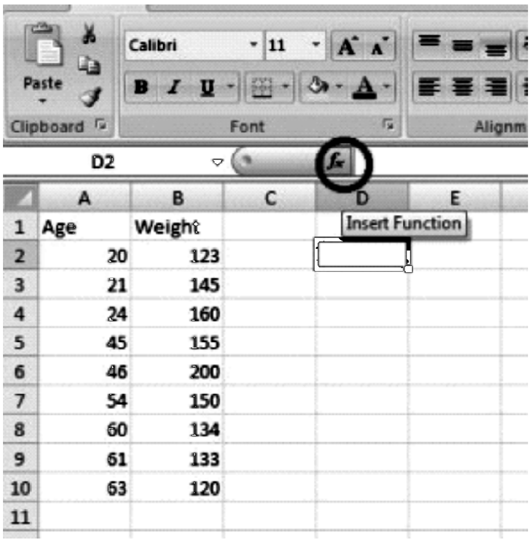
Example 2 :

How to Compute the Pearson Correlation Coefficient Excel 2007 ?

Step 1 : Type your data into two columns in Excel. For example, type your "x" data into column A and your "y" data into column B.

Step 2 : Select any empty cell.

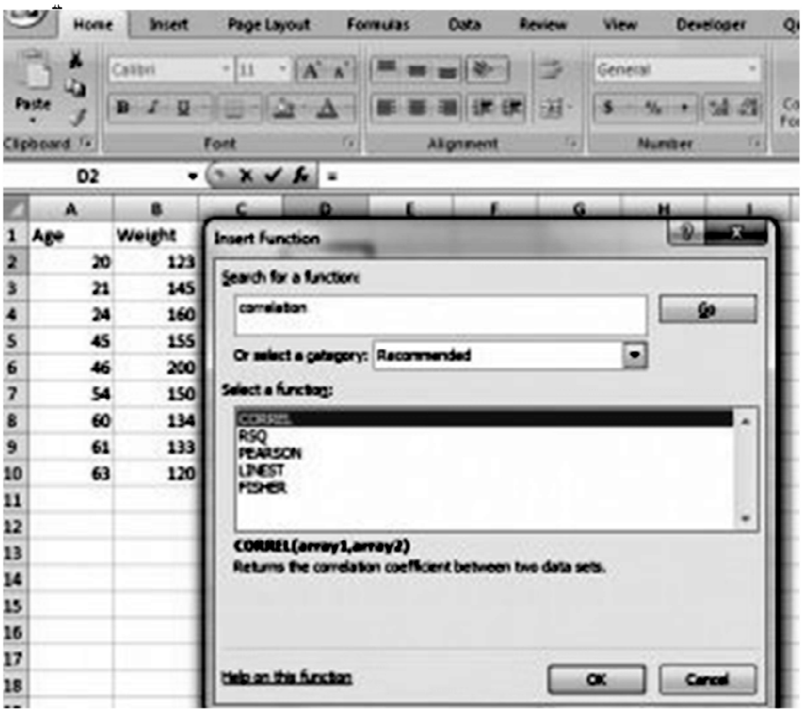
Step 3 : Click the function button on the ribbon.



	A	B	C	D	E
1	Age	Weight			
2	20	123			
3	21	145			
4	24	160			
5	45	155			
6	46	200			
7	54	150			
8	60	134			
9	61	133			
10	63	120			
11					

Step 4 : Type “correlation” into the ‘Search for a function’ box.

Step 5 : Click “Go.” CORREL will be highlighted.



D2 =

	A	B
1	Age	Weight
2	20	123
3	21	145
4	24	160
5	45	155
6	46	200
7	54	150
8	60	134
9	61	133
10	63	120

Insert Function

Search for a function:

correlation

Or select a category: Recommended

Select a function:

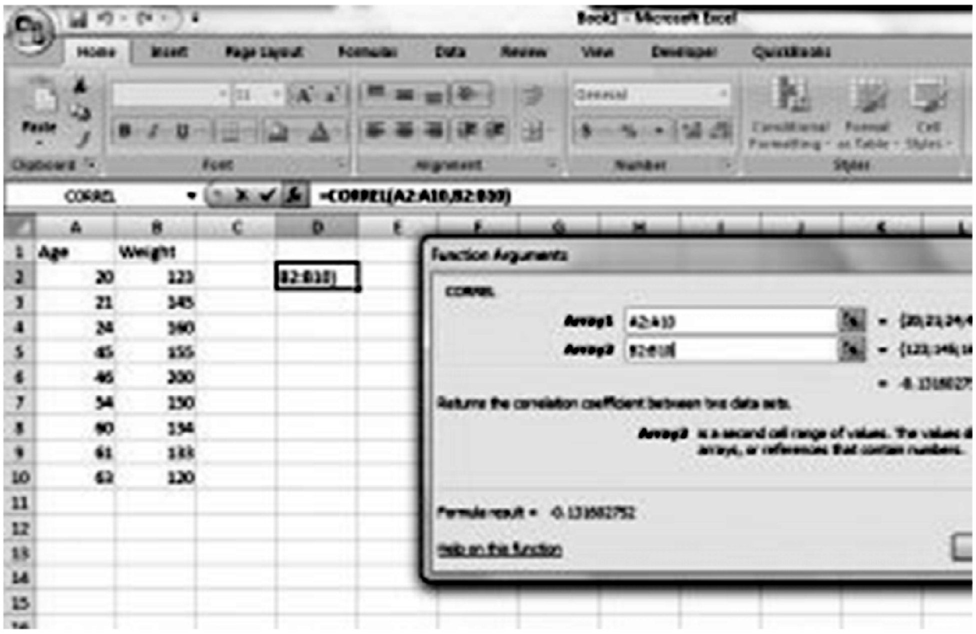
CORREL
RSQ
PEARSON
LINEST
FISHER

CORREL(array1,array2)

Returns the correlation coefficient between two data sets.

Step 6: Click “OK.”

Step 7: Type the location of your data into the ”Array 1" and ”Array 2" boxes. For this example, type “A2:A10” into the Array 1 box and then type “B2:B10” into the Array 2 box.



CORREL =CORREL(A2:A10,B2:B10)

	A	B	C	D
1	Age	Weight		
2	20	123		B2:B10)
3	21	145		
4	24	160		
5	45	155		
6	46	200		
7	54	150		
8	60	134		
9	61	133		
10	63	120		

Function Arguments

CORREL

Array1 A2:A10 = {20;21;24;4

Array2 B2:B10 = {123;145;1

= -0.1316827

Returns the correlation coefficient between two data sets.

Array2 is a second cell range of values. The values s arrays, or references that contain numbers.

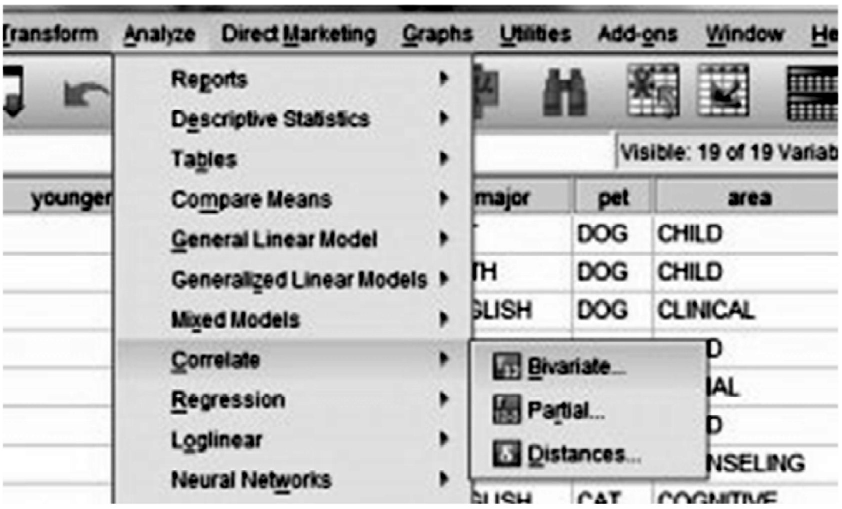
Formula result = -0.131682752

Step 8 : Click "OK." The result will appear in the cell you selected in Step 2. For this particular data set, the correlation coefficient(r) is -0.1316.

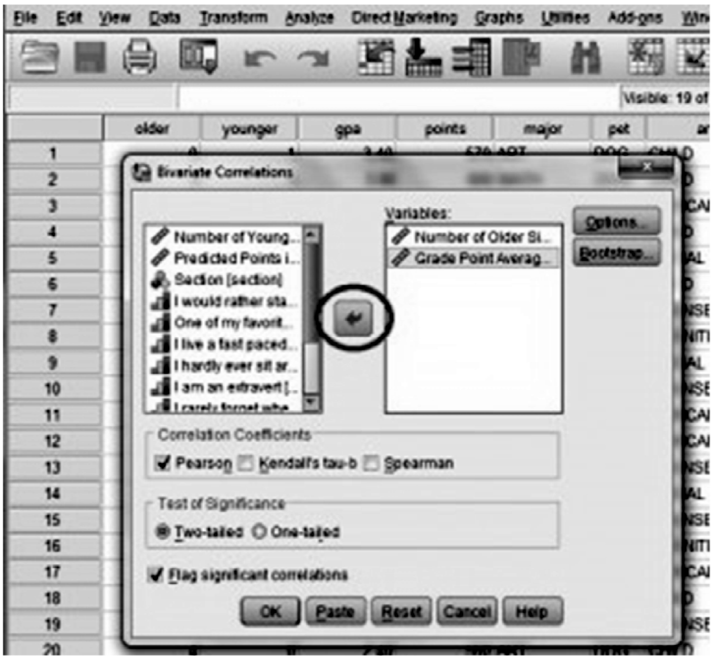
Example 2

How to calculate Correlation Coefficient in SPSS: Overview.

Step 1 : Click "Analyze," then click "Correlate," then click "Bivariate." The Bivariate Correlations window will appear.



Step 2: Click one of the variables in the left-hand window of the Bivariate Correlations pop-up window. Then click the center arrow to move the variable to the "Variables:" window. Repeat this for a second variable.



Variables:
Number of Older Si...
Grade Point Averag...

Correlation Coefficients
☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Step 3 : Click the "Pearson" check box if it isn't already checked. Then click either a "one-tailed" or "two-tailed" test radio button. If you aren't sure if your test is one-tailed or two-tailed, see: Is it a one-tailed test or two-tailed test?

Step 4 : Click "OK" and read the results. Each box in the output gives you a correlation between two variables.

3.2 Multiple Correlation

Q7. Explain the Multiple Correlation between more than two variable.

Ans :

The coefficient of multiple correlation is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables

We can also calculate the correlation between more than two variables.

Definition 1:

Given variables *x*, *y* and *z*, we define the multiple correlation coefficient.

$$R_{z,xy} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xz}r_{yz}r_{xy}}{1 - r_{xy}^2}}$$

where r_{xz}, r_{yz}, r_{xy} are as defined in Definition 2 of Basic Concepts of Correlation. Here *x* and *y* are viewed as the independent variables and *z* is the dependent variable.

We also define the **multiple coefficient of determination** to be the square of the multiple correlation coefficient.

The multiple correlation coefficient and multiple coefficient of determination are written simply as *R* and R^2 respectively. These definitions may also be expanded to more than two independent variables. With just one independent variable the multiple correlation coefficient is simply *r*.

Unfortunately *R* is not an unbiased estimate of the population multiple correlation coefficient, which is evident for small samples. A relatively unbiased version of *R* is given by *R* adjusted.

Definition 2

If *R* is $R_{z,xy}$ as defined above (or similarly for more variables) then the **adjusted** multiple coefficient of determination is

$$R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

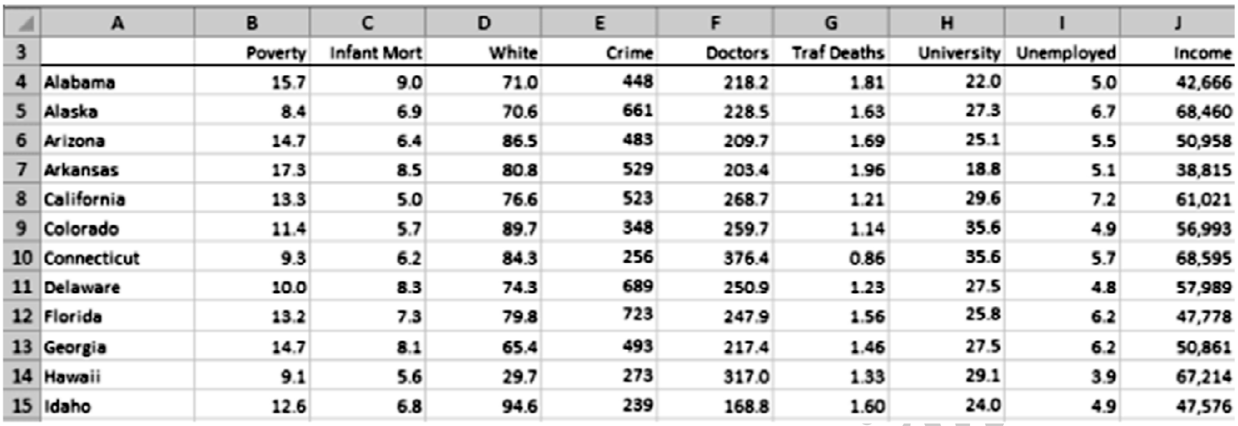
where *k* = the number of independent variables and *n* = the number of data elements in the sample for *z* (which should be the same as the samples for *x* and *y*).

Excel Data Analysis Tools

In addition to the various correlation functions described elsewhere, Excel provides the **Covariance** and **Correlation** data analysis tools. The **Covariance** tool calculates the pairwise population covariances for all the variables in the data set. Similarly the **Correlation** tool calculates the various correlation coefficients as described in the following example.

Example

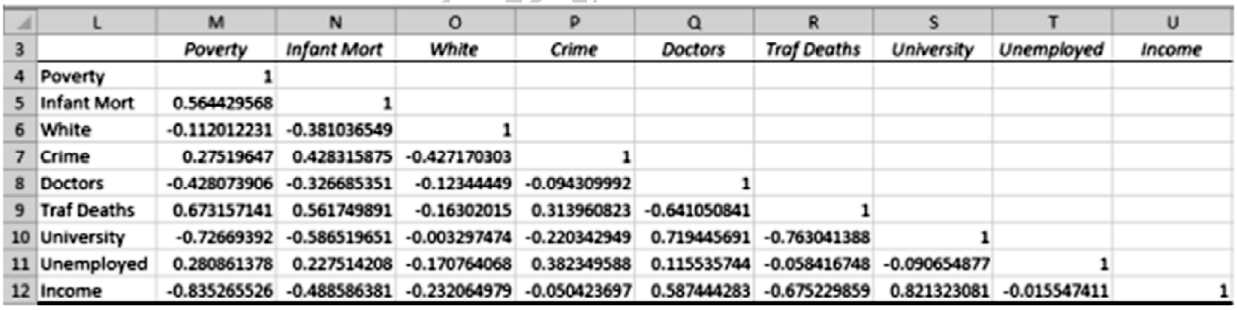
We expand the data in Example 2 of Correlation Testing via the t Test to include a number of other statistics. The data for the first few states are as described in the Figure 1 :



	A	B	C	D	E	F	G	H	I	J
3		Poverty	Infant Mort	White	Crime	Doctors	Traf Deaths	University	Unemployed	Income
4	Alabama	15.7	9.0	71.0	448	218.2	1.81	22.0	5.0	42,666
5	Alaska	8.4	6.9	70.6	661	228.5	1.63	27.3	6.7	68,460
6	Arizona	14.7	6.4	86.5	483	209.7	1.69	25.1	5.5	50,958
7	Arkansas	17.3	8.5	80.8	529	203.4	1.96	18.8	5.1	38,815
8	California	13.3	5.0	76.6	523	268.7	1.21	29.6	7.2	61,021
9	Colorado	11.4	5.7	89.7	348	259.7	1.14	35.6	4.9	56,993
10	Connecticut	9.3	6.2	84.3	256	376.4	0.86	35.6	5.7	68,595
11	Delaware	10.0	8.3	74.3	689	250.9	1.23	27.5	4.8	57,989
12	Florida	13.2	7.3	79.8	723	247.9	1.56	25.8	6.2	47,778
13	Georgia	14.7	8.1	65.4	493	217.4	1.46	27.5	6.2	50,861
14	Hawaii	9.1	5.6	29.7	273	317.0	1.33	29.1	3.9	67,214
15	Idaho	12.6	6.8	94.6	239	168.8	1.60	24.0	4.9	47,576

Fig. 1 – Data for Example 1

Using Excel's **Correlation** data analysis tool we can compute the pairwise correlation coefficients for the various variables in the table in Figure 1. The results are shown in Figure 2.



	L	M	N	O	P	Q	R	S	T	U
3		*Poverty*	*Infant Mort*	*White*	*Crime*	*Doctors*	*Traf Deaths*	*University*	*Unemployed*	*Income*
4	Poverty	1								
5	Infant Mort	0.564429568	1							
6	White	-0.112012231	-0.381036549	1						
7	Crime	0.27519647	0.428315875	-0.427170303	1					
8	Doctors	-0.428073906	-0.326685351	-0.12344449	-0.094309992	1				
9	Traf Deaths	0.673157141	0.561749891	-0.16302015	0.313960823	-0.641050841	1			
10	University	-0.72669392	-0.586519651	-0.003297474	-0.220342949	0.719445691	-0.763041388	1		
11	Unemployed	0.280861378	0.227514208	-0.170764068	0.382349588	0.115535744	-0.058416748	-0.090654877	1	
12	Income	-0.835265526	-0.488586381	-0.232064979	-0.050423697	0.587444283	-0.675229859	0.821323081	-0.015547411	1

Fig 2 – Correlation coefficients for data in Example 1

We can also single out the first three variables, poverty, infant mortality and white (i.e. the percentage of the population that is white) and calculate the multiple correlation coefficients, assuming poverty is the dependent variable, as defined in Definition 1 and 2. We use the data in Figure 2 to obtain the values r_{PW}, r_{P1} and r_{W1}.

$$R_{P, IW} = \sqrt{\frac{r_{IP}^2 + r_{WP}^2 - 2r_{IP}r_{WP}r_{IW}}{1 - r_{IW}^2}} = \sqrt{\frac{.564^2 + (-.112)^2 - 2(.564)(-.112)(-.381)}{1-(-.381)^2}} = 0.575$$

$$R_{P,IW}^2 = 0.575^2 = 0.331$$

$$\text{Adjusted} \quad R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

$$= 1 - \frac{(1-.331)(50-1)}{50-2-1}$$

$$= .3025$$

Definition 3 :

Given *x*, *y* and *z* as in Definition 1, the **partial correlation** of *x* and *z*holding *y* constant is defined as follows :

$$r_{zx,y} = \frac{r_{zx} - r_{zy}r_{xy}}{\sqrt{1-r_{zy}^2}\sqrt{1-r_{xy}^2}}$$

In the **semi-partial correlation**, the correlation between *x* and *y* is eliminated, but not the correlation between *x* and *z* and *y* and *z* :

$$r_{z(x,y)} = \frac{r_{zx} - r_{zy}r_{xy}}{\sqrt{1-r_{xy}^2}}$$

Observation : Suppose we look at the relationship between GPA (grade point average) and Salary 5 years after graduation and discover there is a high correlation between these two variables.

3.3 Spearman's Rank Correlation

Q8. Explain about spearman's rank correlation.

Ans :

'Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This problem arises while dealing with qualitative characteristics such as honesty, beauty, character, morality etc. which cannot be measured quantitatively, but can be arranged serially. In such cases, Karl Pearson's Coefficient of Correlation may not serve the purpose.

Another method was developed by Edward Spearman to study correlation between such attributes. In this method, the change in a variable with respect to a change in another variable is not measured by means of absolute change as it is difficult to quantify the absolute measure. However, if the movement of the two variables is similar, they should be getting similar, if not identical, ranks. Thus, if the difference in ranks is minimal, then there is a case of positive correlation. If the difference in ranks is huge, then it indicates negative correlation.

Properties of Spearman's Rank Correlation

1. It is based on subjective ranking of variables.
2. It lies between -1

3. It measures both direction as well as degree of change. If r is less than 0, there is negative correlation, which means the direction of change of the two variables will be opposite. If r is more than 0, there is positive correlation, which means the direction of change of the two variables will be same. Higher the value of r, greater is the degree of correlation.

4. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X.

5. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.

6. It is not impacted by extreme values as only ranking matters.

Merits of Spearman's Rank Correlation

1. It is easy to understand and calculate
2. It is not impacted by extreme values.
3. It is a numerical measure and provides objectivity to subjective ranking.
4. It is the only method of finding correlation with respect to qualitative factors such as honesty, beauty, etc.
5. It measures both direction as well as degree of change.
6. It facilitates comparisons between two series.
7. It can be applied to irregular data also.
8. It is ideal when the number of observations is very small.

Demerits of Spearman's Rank Correlation

1. It cannot be applied to grouped data
2. It lacks the precision of Karl Pearson's Coefficient of Correlation.
3. All the information concerning the variable is not used.
4. The computation becomes complicated as the number of observations increase.

Computation: Spearman's correlation is computed by using the following formula:

Where

r_s = rank coefficient of correlation.

D = difference of rank between paired items in two series.

N = No. of Pairs of observation.

Spearman's rank correlation coefficient lies between +1 and -1. Under this method two types of problems can be given. (1) Where ranks are given (2) Where ranks are not given.

1. **When ranks are given :**

Steps :

1) Take the difference of the two ranks and denote the difference as D.

2) Square these differences and obtain the total. Denote it as $\cdot D^2$ and apply the formula.

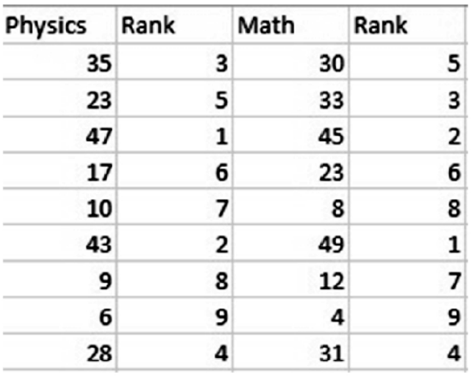
Example :

The scores for nine students in physics and math are as follows:

Physics :	35, 23, 47, 17, 10, 43, 9, 6, 28
Mathematics :	30, 33, 45, 23, 8, 49, 12, 4, 31

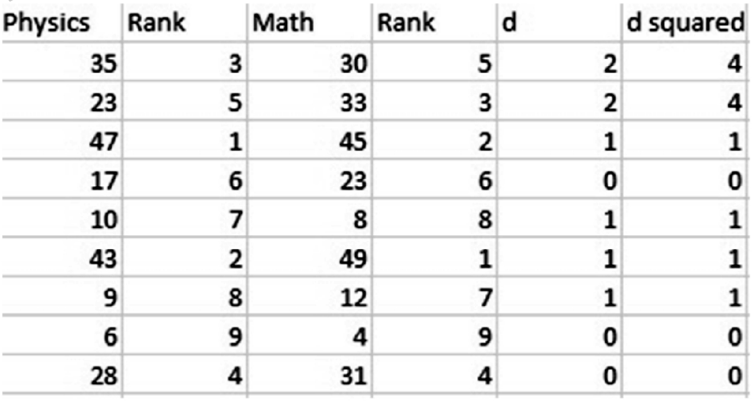
Compute the student's ranks in the two subjects and compute the Spearman rank correlation.

Step 1: Find the ranks for each individual subject. I used the Excel rank function to find the ranks. If you want to rank by hand, order the scores from greatest to smallest; assign the rank 1 to the highest score, 2 to the next highest and so on :



Physics	Rank	Math	Rank
35	3	30	5
23	5	33	3
47	1	45	2
17	6	23	6
10	7	8	8
43	2	49	1
9	8	12	7
6	9	4	9
28	4	31	4

Step 2: Add a third column, d, to your data. The d is the difference between ranks. For example, the first student's physics rank is 3 and math rank is 5, so the difference is 3 points. In a fourth column, square your d values.



Physics	Rank	Math	Rank	d	d squared
35	3	30	5	2	4
23	5	33	3	2	4
47	1	45	2	1	1
17	6	23	6	0	0
10	7	8	8	1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0

Step 3 : Find d –squared values

Step 4 : Sum (add up) all of your d-squared values.

4 + 4 + 1 + 0 + 1 + 1 + 1 + 0 + 0 = 12. You'll need this for the formula (the Σ d2 is just "the sum of d-squared values").

Step 5 : Insert the values into the formula. These ranks are not tied, so use the first formula :

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

= 1 – (6*12)/(9(81–1))

= 1 – 72/720

= 1 – 0.1

= 0.9

The Spearman Rank Correlation for this set of data is 0.9.

3.4 Regression Analysis

Q9. Explain the concept of regression analysis.

Ans :

Regression takes its name from studies made by Sir Francis Galton. He compared the heights of persons to the heights of their parents. His major conclusion was that the off-springs of unusually tall persons tend to be shorter than their parents while children of usually short parents tend to be shorter than their parents.

In a sense, the successive generations of off-springs from tall persons "regress" downward toward the height of the population. While the reverse is true originally for short families. But the distribution of heights for the total population, continues to have the same variability from generation to generation.

Regression Variables

i) **Independent Variable (Regressor or Predictor or Explanatory).** The variable which influences the values of the other variable or which is used for prediction of the value of the other variable is called independent variable.

ii) **Dependent Variable (Regressed or Explained Variable).** The variable whose value is influenced or is to be predicted is called dependent variable.

Regression test generates lines of regression of the two variables which helps in estimating the values. Lines of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x. Similarly, line of regression of x on y is the line which gives the best estimate for the value of x for any specified value of y.

Types of Regression

a) **Simple Regression.** The regression analysis confined to the study of only two variables at a time is termed as simple regression.

b) **Multiple Regression.** The regression analysis for studying more than two variables at a time is termed as multiple regression.

c) **Linear Regression.** If the regression curve is a straight line, the regression is termed as linear regression. The equation of such a curve is the equation of a straight line i.e., first degree equation in variables x and y.

d) **Nonlinear Regression.** If the curve of the regression is not a straight line, the regression is termed as curved or non-linear regression. The regression equation will be a functional relation between variables x and y involving terms in x and y of degree more than one.

Applications / Utility of Regression Test

Regression lines or equations are useful in the predictions of values of one variable for a specified value of the other variable.

Example

i) For pharmaceutical firms which are interested in studying the effect of new drugs in patients, regression test helps in such predictions.

ii) When price and demand are related, we can estimate or predict the future demand for a specified price.

iii) When crop yield depends on the amount of rainfall, then regression test can predict crop yield for a particular amount of rainfall.

iv) If advertising expenditure and sales are related, then regression analysis helps in estimating the advertising expenditure for a required amount of sales (or) sales expected for a particular advertising expenditure.

v) When capital employed and profits earned are related, the test can be used to predict profits for a specified amount of capital invested.

Limitations of Regression Analysis

Some of the limitations of regression analysis are as follows :

1. Regression analysis assumes that linear relationship exists among the related variables. But in the ‘area of social sciences, linear relationship may not exist among the related variables.
2. When regression analysis is used to evaluate the value of dependent variable based on independent variable, it is assumed that the static conditions of relationship exist between ‘them. These statistic conditions do not exist in social sciences, so, this assumption “minimizes the use’ of regression analysis in social science.
3. The value of dependent variable can be evaluated based on independent variable by using regression analysis but only upto some limits. If the circumstances go beyond the limits, then resists would be inaccurate.

Even though, there are many limitations of regression ‘technique, it is still regarded as a very useful statistical tool for estimating or predicting the value of dependent variable.

Q10. What is the comparison of correlation and regression analysis ?

Ans :

S.No.	Correlation Analysis	Regression Analysis
1.	Correlation is a measure of the 'degree and direction' of relationship between the variables.	Regression studies 'nature' of relationship between the variables.
2.	Correlation means the relationship between two or more variables which vary in sympathy so that the movements in one variable tend to be accompanied by the corresponding movements in others.	Regression means stepping back or returning to average value and is a mathematical measure expressing the average relationship between the variables.
3.	Correlation does not indicate the cause and effect relationship between the variables.	Regression clearly indicates the cause and effect relationship between the variables.
4.	Correlation cannot say which variable is the dependent variable and which is the independent variable.	In regression, the variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
5.	Correlation coefficient is a relative measure of the linear relationship.	Regression coefficients are absolute measures indicating the change in the value of one variable for a unit change in the value of the other variable.
6.	Correlation analysis cannot be used for predicting or estimating value.	Regression analysis is very helpful in predicting and estimating value of one variable given the value of another variable.
7.	Correlation coefficients are symmetric i.e., $r_{yx} = r_{xy}$.	Regression coefficients are asymmetric i.e., $b_{xy} \neq b_{yx}$.
8.	The range of r is +1 to – 1.	The range of b_{xy} and b_{yx} is not restricted.
9.	Correlation coefficient can be calculated from regression coefficients.	Regression coefficients cannot be directly compared from correlation coefficient.
10.	There may be non-sense correlation between variables due to chance.	There is no such thing in regression.

Q11. Explain about regression equation.

Ans :

Regression is mainly concerned with the estimation of unknown value of one variable from the known value of other variable of the given observations. For doing so, there must be a relation between two variables. This relationship is mathematically expressed in the form of equation known as "Regression Equation " or " Estimating Equation".

The regression equation which states and explains the linear relationship between two variables is known as 'Linear Regression Equation'. Basically, as there are two regression lines, there would be two regression equations i.e.,

1. Regression equation of K on X and
2. Regression equation of X on Y.

The regression equation of Y on X is considered for predicting the value of Y when a specific value of X is given. Whereas the regression equation of X on Y is used for predicting the unknown value of X when a specific value of Y is given.

Formation of Regression Equations

There are two ways of forming regression equations as follows,

a) Normal equation and

b) Regression coefficient.

Formation of Regression Equation through Normal Equation

Generally, the situations where perfect linear relationship exists between the two variables X and Y, usually there would be two regression lines and when there are two regression lines, there would be two regression equations as follows,

1. The regression equation of Y on X is denoted as Y. = a + bX.
2. The regression equation of X on Y is denoted as X = a + bY.

In the above equations 'a' and 'b' are two unknown constants which ascertains the positions of the regression line. Therefore, these constants are known as parameters of the regression lines.

The parameter 'a' ascertains the level of a fitted line, whereas 'b' ascertains the slope of the line. Y_C and X_C are the symbols stating and showing the values of Y and X calculated from the relationship for given X or Y.

Regression Equation of Y on X

$Y = a + bX$

By applying the least square principle, the values of 'a' and 'b' are determined in such a way $Y_C = a + bX$ is minimum.

The normal equation for determining the value of a and b are,

$\Sigma y = Na + b\Sigma x$ ...(1)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ ...(2)

Regression Equation of X on Y

$Y_c = a + by$

The normal equation for obtaining the values of a and b are,

$\Sigma x = Na + b\Sigma y$ ...(1)

$\Sigma xy = a\Sigma y + b\Sigma y^2$ ... (2)

After calculating the values of N, Σx, Σy, Σx^2, "Σy^2, substitute them in regression equation Y on X and X on Y for ascertaining the values of a and b. Lastly, by substituting the values of a and b in regression equation, the required best fitting straight line is obtained.

b) Regression Coefficients

To estimate values of population parameter β_0 and β_1, under certain assumptions, the fitted or estimated regression equation representing the straight line regression model is written as :

$$\hat{y} = a + bx$$

where

y = estimated average (mean) value of dependent variable y for a given value of independent variable *x*.

a or b_0 = y - intercept that represents average value of $\hat{y}$

b = slope of regression line that represents the expected change in the value of y for unit change in the value of *x*

To determine the value of $\hat{y}$ for a given value of *x*, this equation requires the determination of two unknown constants a (intercept) and *b* (also called regression coefficient). Once these constants are calculated, the regression line can be used to compute an estimated value of the dependent variable *y* for a given value of independent variable *x*.

The particular values of *a* and *b* define a specific linear relationship between *x* and *y* based on sample data. The coefficient '*a*' represents the level of fitted line (i.e., the distance of the line above or below the origin) when *x* equals zero, whereas coefficient '*b*' represents the slope of the line (a measure of the change in the estimated value of *y* for a one-unit change in).

The regression coefficient 'b' is also denoted as :

- b_{yx} (regression coefficient of y on x) in the regression line, y = a + bx
- b_{xy} (regression coefficient of x on y) in the regression line, x = c + dy.

3.4.1 Simple Regression

Q12. Discuss briefly about simple regression.

Ans :

Simple regression represents the relationship between two variables where one of them is independent variables 'X' and other variable is dependent variable 'Y'.

The relationship between two variables can be of three types. They are,

i) Linear Relationship

The graph of linear relationship between two variables looks as follows,

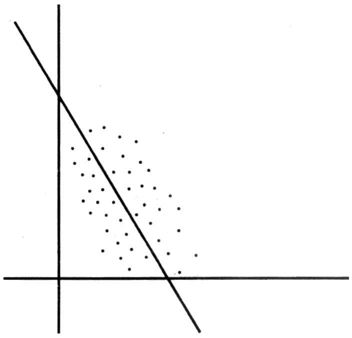


Figure: Linear Relationship

ii) Non-Linear Relationship

The graph of non-linear relationship between two variables looks as follows,

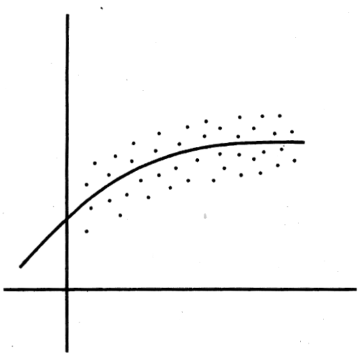


Figure: Non-linear Relationship

No Relationship

The graph of no relationship between two variables looks as follows,

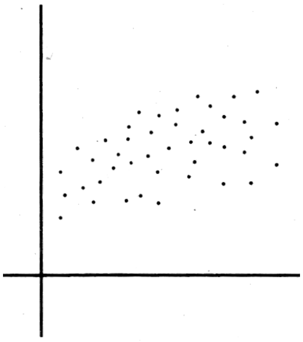


Figure: No Relationship

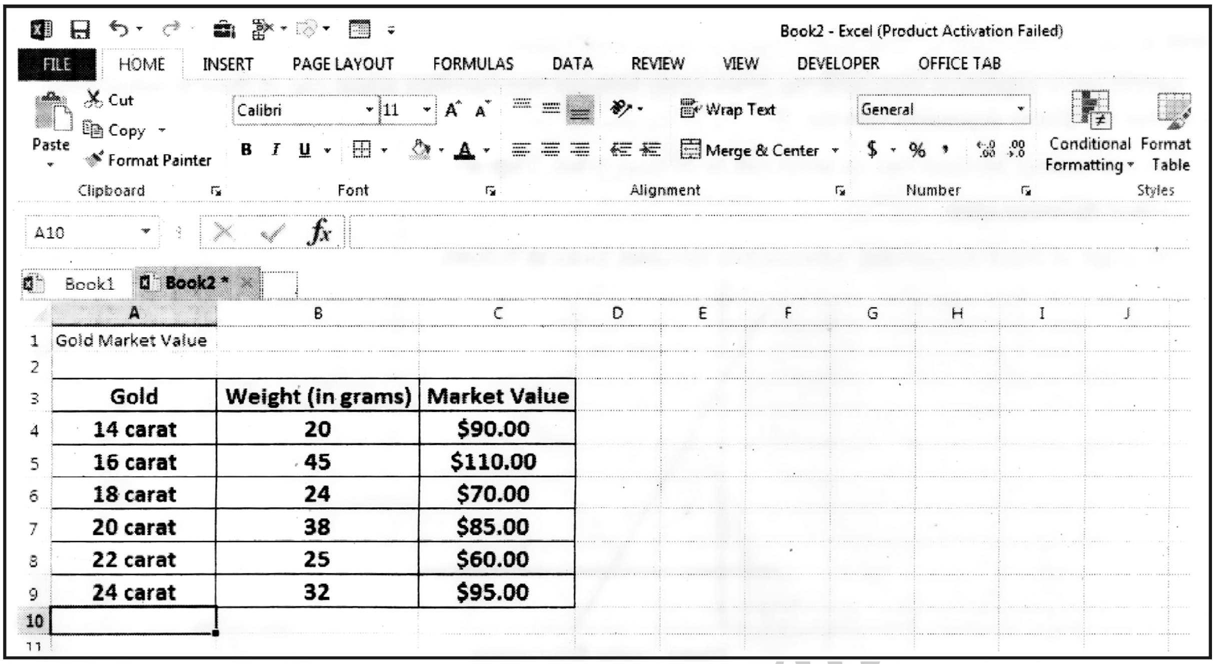
Q13. Explain the concept of simple linear regression with excel.

Ans :

In Microsoft Excel, the information regarding statistical properties of regression analysis are provided by the software tools of regression analysis. The regression tool can be used not only for simple regression but, also for multiple regression.

The steps to be followed for generating regression analysis output are as follows,

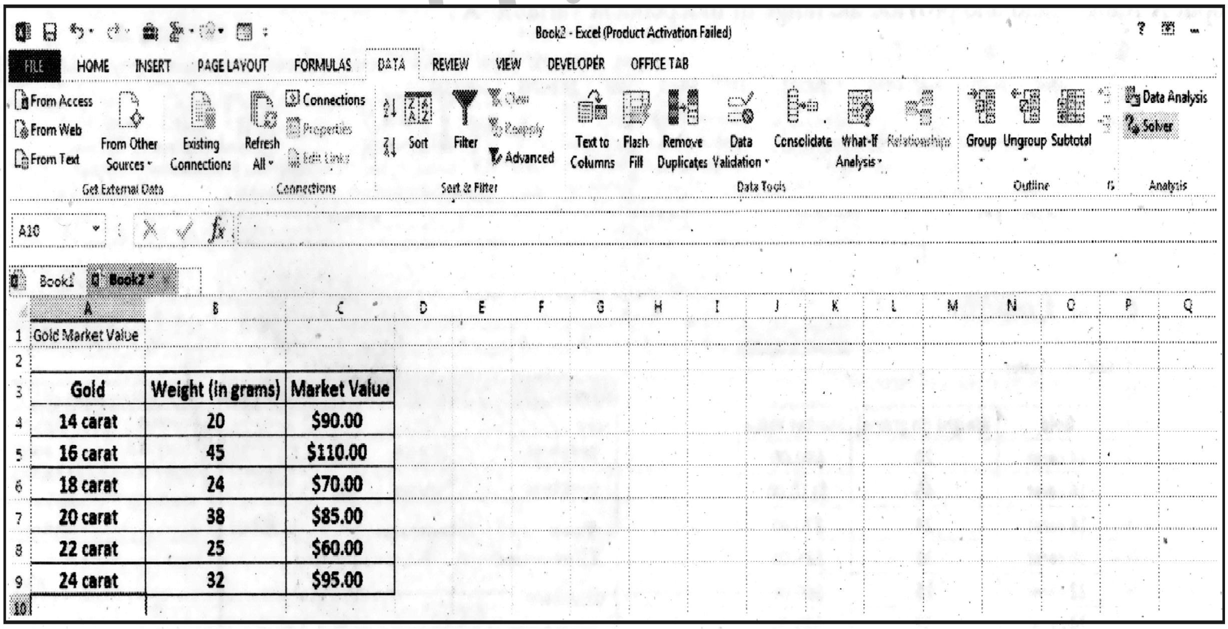
1. Select the data wherein user want to apply regression.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$90.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

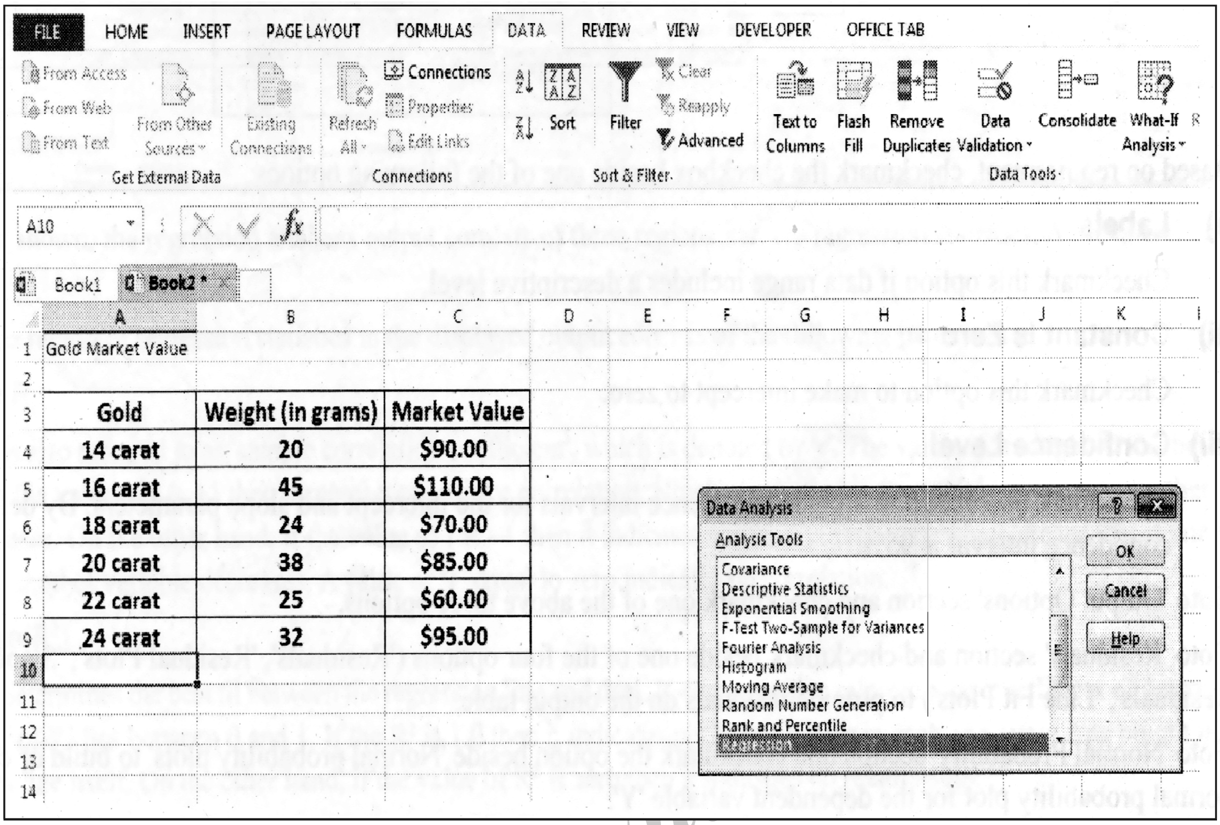
2. Click on 'Data' tab present on excel ribbon and the click 'Data Analysis' command present under' Analysis' group.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$90.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

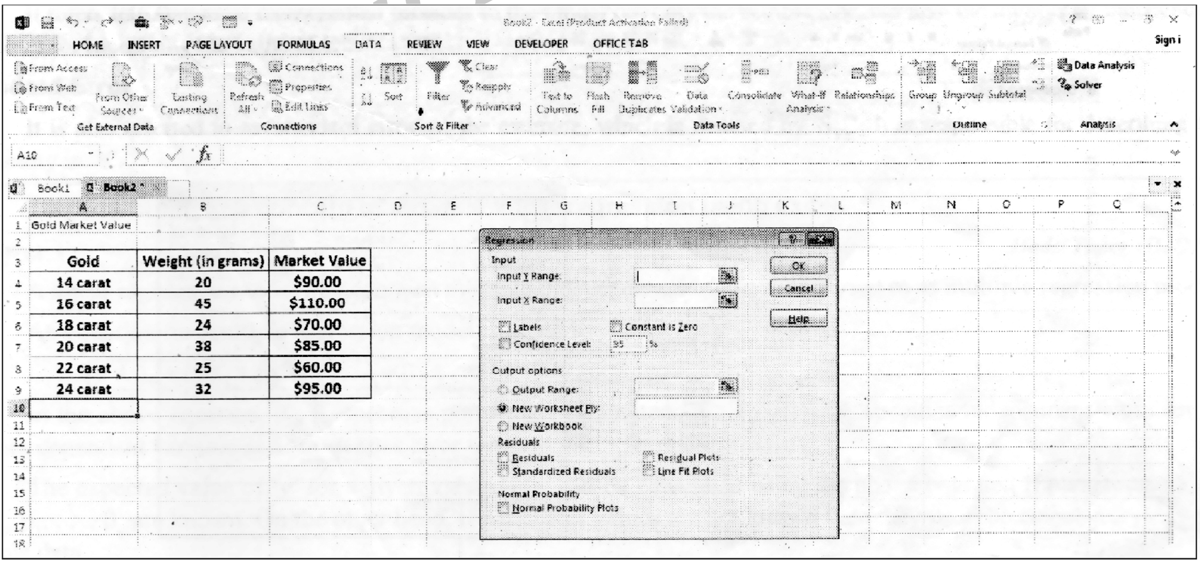
3. As a result, 'Data Analysis' dialog box will be displayed. Goto 'Analysis Tools' section and select 'Regression' option from the menu list and then click "OK" button.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$90.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

4. As a result, 'Regression' window appears on screen.

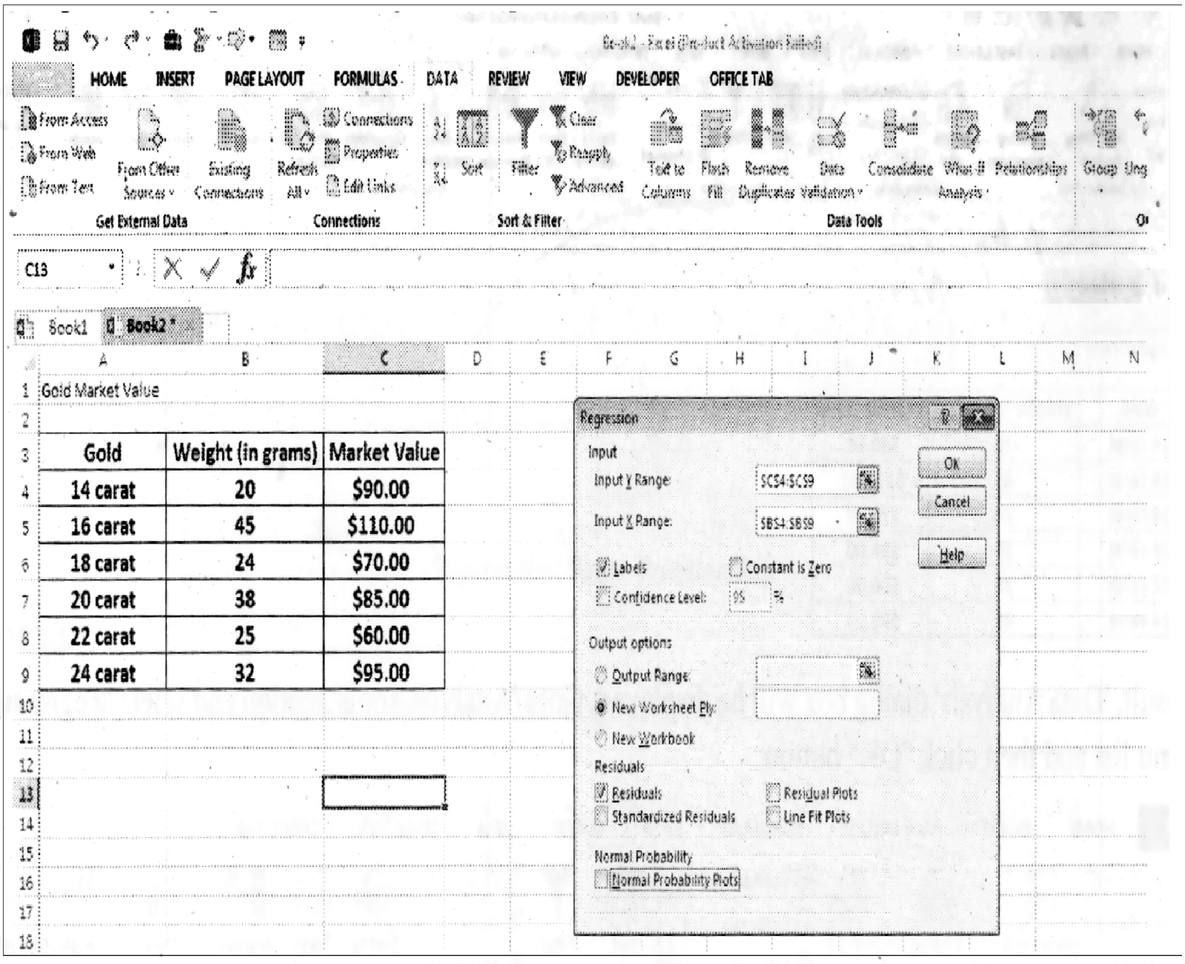


Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$90.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

Input
Input Y Range:
Input X Range:
□ Labels □ Constant is Zero
□ Confidence Level: 95 %
Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook
Residuals
□ Residuals □ Residual Plots
□ Standardized Residuals □ Line Fit Plots
Normal Probability
□ Normal Probability Plots

5. In the 'Regression' dialog box, goto 'Input Y Range' field and provide the range of dependent variable 'Y'. Similarly, Goto 'Input X Range' field and provide the range of independent variable 'X'.



Gold Market Value

Gold	Weight (in grams)	Market Value
14 carat	20	$90.00
16 carat	45	$110.00
18 carat	24	$70.00
20 carat	38	$85.00
22 carat	25	$60.00
24 carat	32	$95.00

Regression

Input

Input Y Range: $C$4:$C$9

Input X Range: $B$4:$B$9

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

6. Based on requirement, checkmark the checkbox beside one of the following options.

   i) Labels : Checkmark this option if data range includes a descriptive level.

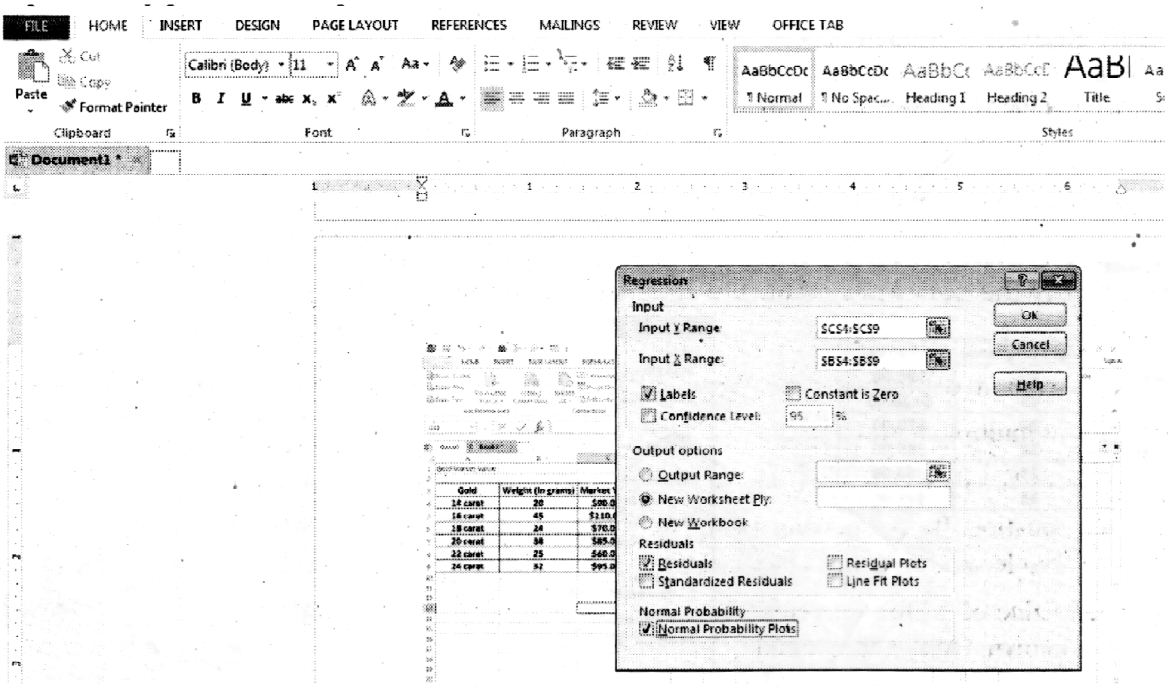
   ii) Constant is Zero : Checkmark this option to make intercept to zero.

   iii) Confidence Level : Checkmark this option to include confidence intervals for the intercept and slope parameters. By default, the confidence interval is 95%.

7. Goto 'Output Options' section and checkmark one of the above three options.

8. Goto 'Residuals' section and checkmark beside one of the four options ('Residuals', 'Residual Plots', 'Standardized Residuals', 'Line Fit Plots') to provide residuals on the output table.

9. Goto 'Normal Probability' section and checkmark the option beside 'Normal probability plots' to build or construct normal probability plot for the dependent variable 'Y'.



Regression

Input

Input Y Range: $C$4:$C$9

Input X Range: $B$4:$B$9

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

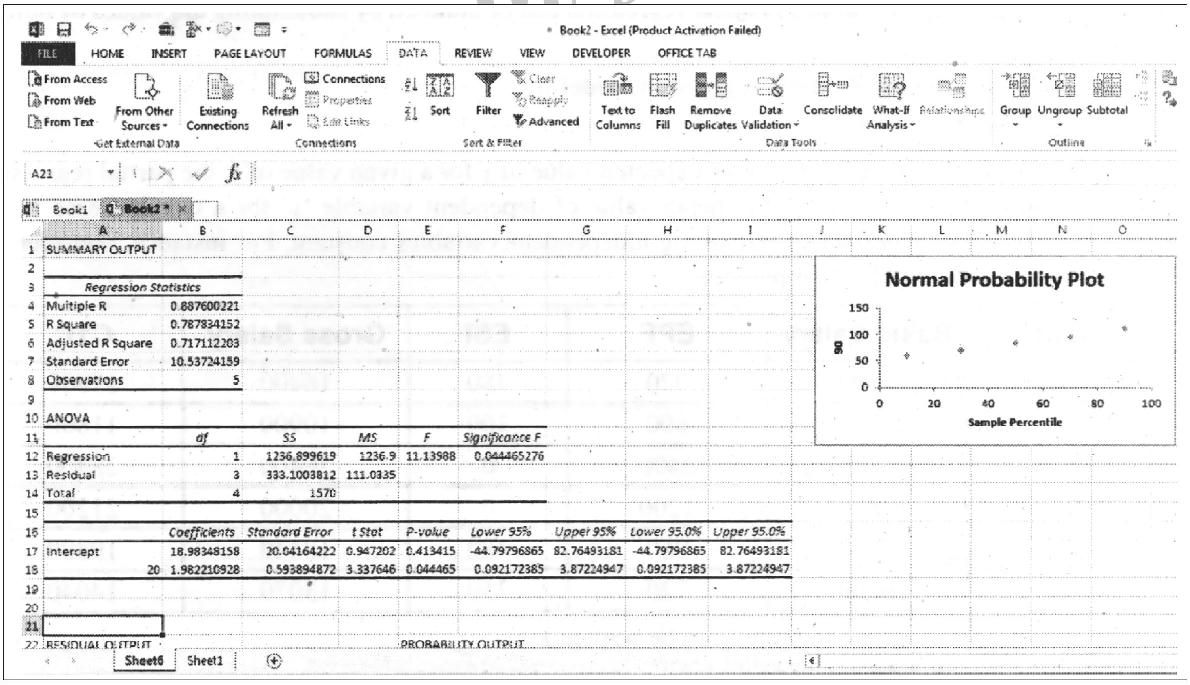
☑ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☑ Normal Probability Plots

10. Click on "OK" button. As a result, the regression analysis output will be displayed on the screen.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.887600221
R Square	0.787834152
Adjusted R Square	0.717112203
Standard Error	10.53724159
Observations	5

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1236.899619	1236.9	11.13988	0.044465276
Residual	3	333.1003812	111.0335		
Total	4	1570			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	18.98348158	20.04164222	0.947202	0.413415	-44.79796865	82.76493181	-44.79796865	82.76493181
20	1.982210928	0.593894872	3.337646	0.044465	0.092172385	3.87224947	0.092172385	3.87224947

RESIDUAL OUTPUT PROBABILITY OUTPUT

As shown, the regression analysis output consists of three regions namely regression statistics, Annova and unlabelled section.

The region of regression statistics in the displayed output consists of the following parameters,

Multiple R

It is also referred to as 'sample correlation coefficient', which is denoted by 'r'. The value of'multiple R' lies between - 1 and +1. If the value of r is +1 then it represents positive correlation, which means that if one variable increases another variable also increases. On the other hand, if the value of r is -1 then it indicates negative correlation, which means that if one viarable decreases another variable decreases. A value of'r' equal to zero indicates no correlation.

R-Square(R^2)

It determines the best fit between the regression line and data. R-sqaure is also referred to as 'coefficient of determination'. The value of R^2 lies between 0 and 1. If the R^2 is 1.0 then it indicates perfect fit where in each and every data point falls on the regression line itself. On the other hand, if the value of R^2 is zero then it indicates no relationship.

Adjusted R Square

It refers to a statistical measure that includes in the model not only the sample size, but also the number of independent variables for modifying the value of R^2.

Standard Error

It is also referred to as 'standard error' of the estimate, which is denoted by 'S_{YX}'. It is responsible for describing the variability in 'Y'.

3.4.2 Multiple Regression

Q14. Discuss briefly the concept of multiple regression using excel.

Ans :

Multiple Regression

The regression analysis for studying more than two variables at a time is termed as multiple regression.

A linear regression model that contains multiple independent variables is referred to as multiple regression model. An equation for multiple regression model is given by,

$$\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots.. + \beta_p X_p + \in}$$

In the above equation, β_0, β_1 specifies population parameters, X_1, X_2, ... X_p specifies independent-variables, Y defines dependent variable and '$\in$' defines error term.

The expected value of 'y' for a given value of V can be calculated using the above equation if parameter values of β_0, β_1, . . ., β_q are known. On the other hand, if parameter values are not known then they must be calculated using the sample data.

The estimated regression equation for multiple linear regression can be attained by substituting the values of sample statistics b_0, b_1, ... , b_p in β_0, β_1, ... , β_p.

The estimated regression equation in multiple regression model is,

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

In the above equation, y refers to point estimator of expected value of y for a given value of x, the partial regression coefficients b_0, b_1, ... ,b_p indicates the change in the mean value of dependent variable 'y' for a unit increase in the independent variables, while holding the values of remaining independent variables constant. For instance, consider the following excel file containing salary details of employees.

Employee	Dept	Basic Salary	EPF	ESI	Gross Salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0	20000	21200
Praveen	ECE	8500	960	480	18440	19880
Anusha	EE	6000	720	350	13070	14040

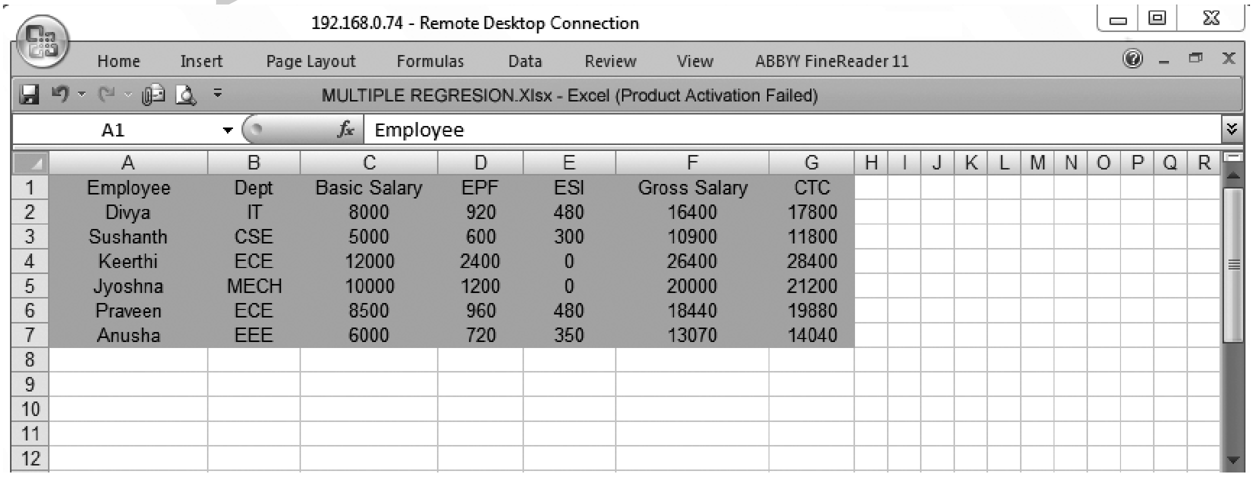
In the above table, the multiple regression model can be written as,

$$\text{CTC} = b_0 + b_x \text{ Basic Salary} + b_2 \text{ EPF} + b_3 \text{ ESI} + b_4 \text{ Gross Salary}$$

Therefore, b, indicates the change in the mean value of CTC for a unit increase in the associated independent variable 'EPF' while holding all remaining independent variables 'Basic Salary', 'EPF', 'ESI' and 'Gross Salary Constant like simple linear regression, multiple linear regression also follows the least squares technique for estimating both intercept and slope coefficients.

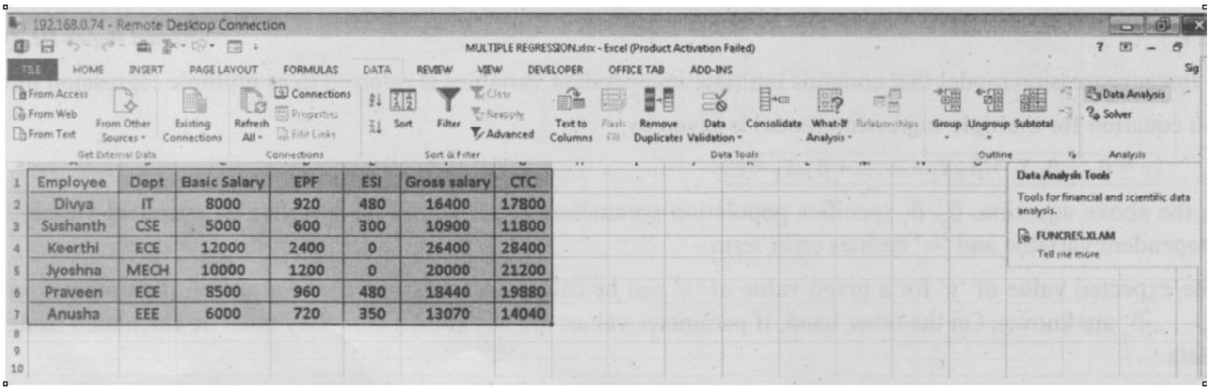
The steps to be followed for generating regression analysis output in case of multiple linear regression are given below,

1. Select the data wherein user want to apply regression.


192.168.0.74 - Remote Desktop Connection
Home
Insert
Page Layout
Formulas
Data
Review
View
ABBYY FineReader 11
MULTIPLE REGRESION.Xlsx - Excel (Product Activation Failed)
A1
Employee

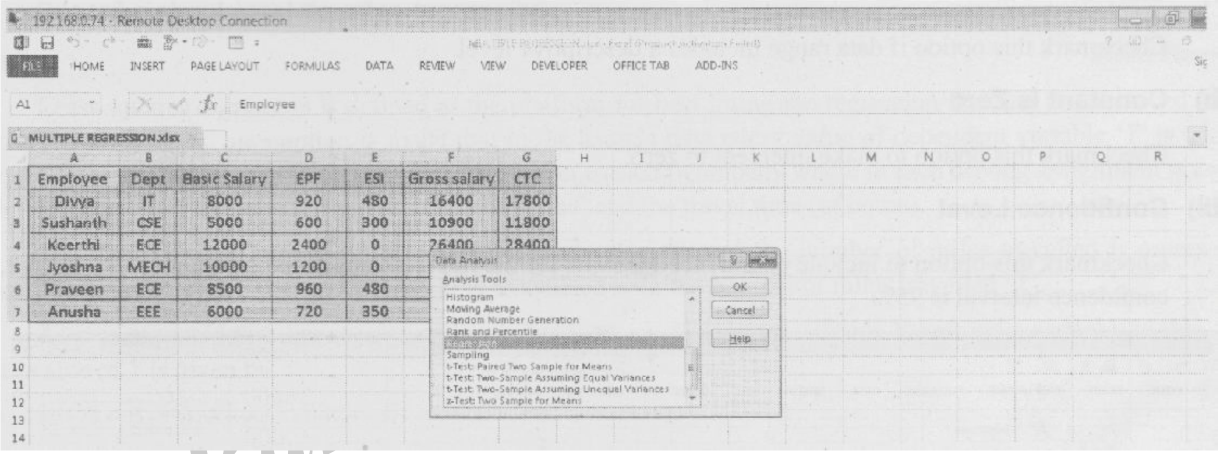
	A	B	C	D	E	F	G
1	Employee	Dept	Basic Salary	EPF	ESI	Gross Salary	CTC
2	Divya	IT	8000	920	480	16400	17800
3	Sushanth	CSE	5000	600	300	10900	11800
4	Keerthi	ECE	12000	2400	0	26400	28400
5	Jyoshna	MECH	10000	1200	0	20000	21200
6	Praveen	ECE	8500	960	480	18440	19880
7	Anusha	EEE	6000	720	350	13070	14040

2. Click on 'Data' tab present on excel ribbon and the click 'Data Analysis' command present under 'Analysis' group.



Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0	20000	21200
Praveen	ECE	8500	960	480	18440	19880
Anusha	EEE	6000	720	350	13070	14040

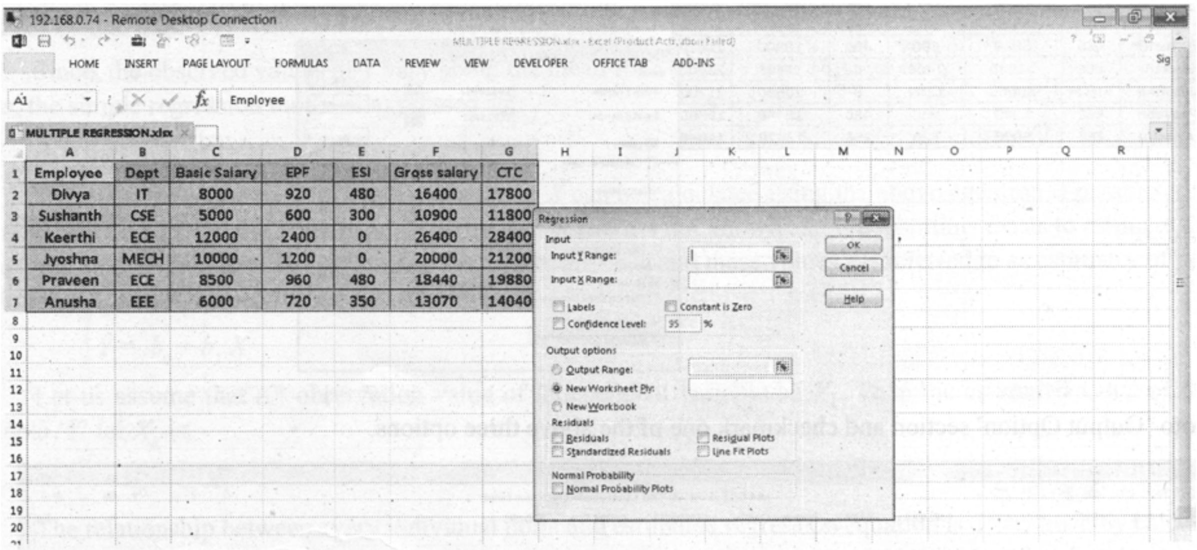
3. As a result, 'Data Analysis' dialog box will be displayed. Goto 'Analysis Tools' section and select 'Regression' option from the menu list and then click "OK" button.


192.168.0.74 - Remote Desktop Connection
MULTIPLE REGRESSION.xlsx - Excel (Product Activation Failed)
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER OFFICE TAB ADD-INS
A1
Employee
MULTIPLE REGRESSION.xlsx
Data Analysis
OK
Cancel
Help

Analysis Tools

Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means

4. As a result, 'Regression' window appears on screen.


192.168.0.74 - Remote Desktop Connection
MULTIPLE REGRESSION.xlsx - Excel (Product Activation Failed)
HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER OFFICE TAB ADD-INS
A1
Employee
MULTIPLE REGRESSION.xlsx
Regression
OK
Cancel
Help

Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

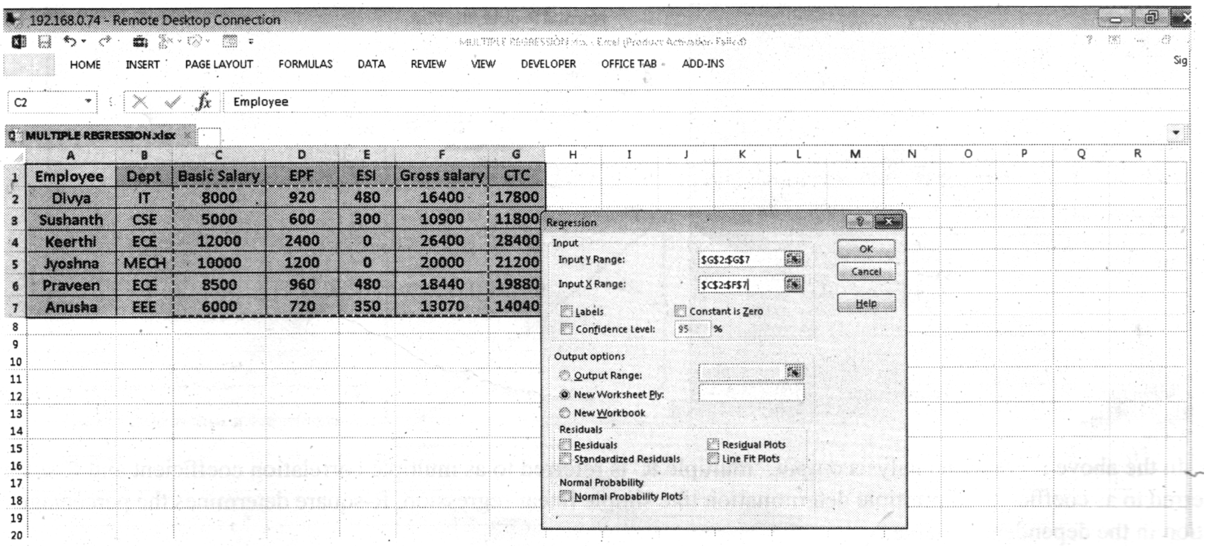
□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

5. In the 'Regression', dialog box, Goto 'Input Y Range' field and provide the range of dependent variable Y. Similarly, Goto 'Input X Range' field and provide the entire range of independent variable ˡJC.


192.168.0.74 - Remote Desktop Connection

C2 Employee

MULTIPLE REGRESSION.xlsx

Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0	20000	21200
Praveen	ECE	8500	960	480	18440	19880
Anusha	EEE	6000	720	350	13070	14040

Regression

Input

Input Y Range: $G$2:$G$7

Input X Range: $C$2:$F$7

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

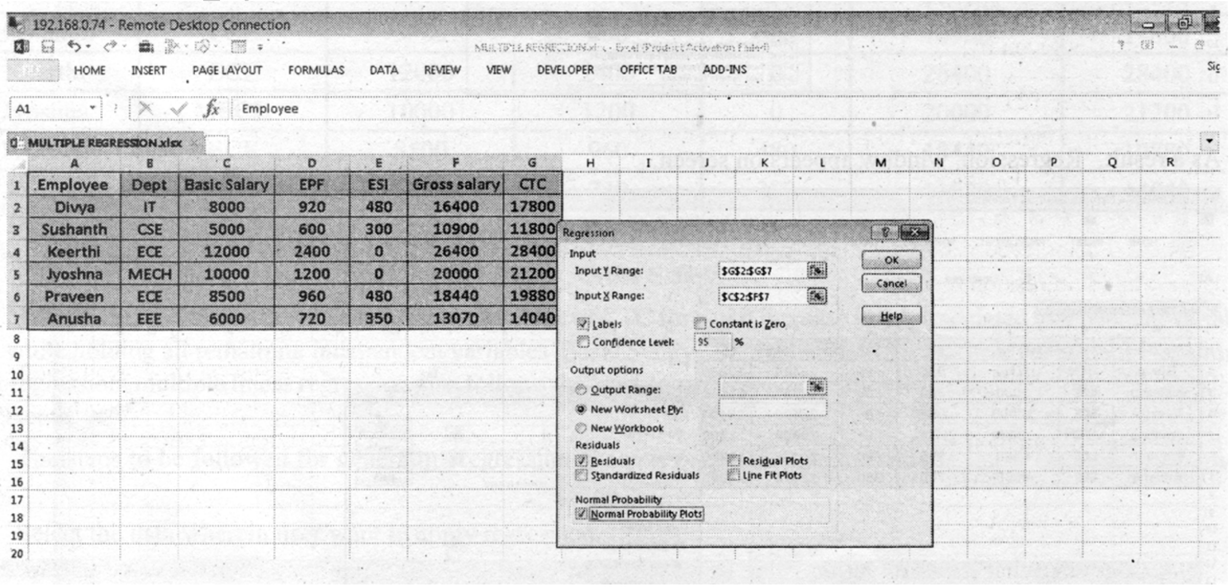
□ Normal Probability Plots

6. Based on requirement, checkmark the checkbox beside one of the following options.

(i) Labels: Checkmark this option if data range includes a descriptive level.

(ii) Constant is Zero: Checkmark this option to make intercept to zero.

(iii) Confidence Level: Checkmark this option to include confidence intervals for the intercept and slope parameters. By default, the confidence interval is 95%.


192.168.0.74 - Remote Desktop Connection

A1 Employee

MULTIPLE REGRESSION.xlsx

Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0	20000	21200
Praveen	ECE	8500	960	480	18440	19880
Anusha	EEE	6000	720	350	13070	14040

Regression

Input

Input Y Range: $G$2:$G$7

Input X Range: $C$2:$F$7

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

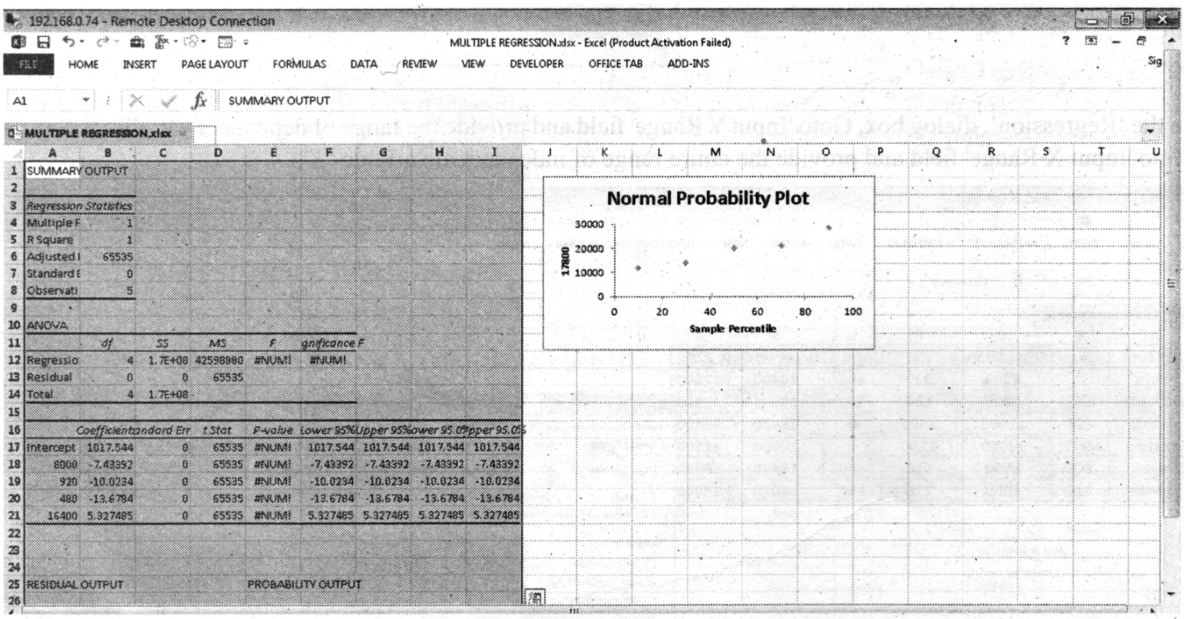
☑ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

☑ Normal Probability Plots

7. Goto 'Output Option' section and checkmark one of the above three options.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple F	1							
5	R Square	1							
6	Adjusted I	65535							
7	Standard E	0							
8	Observati	5							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressio	4	1.7E+08	42598880	#NUM!	#NUM!			
13	Residual	0	0	65535					
14	Total	4	1.7E+08						
15									
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	1017.544	0	65535	#NUM!	1017.544	1017.544	1017.544	1017.544
18	8000	-7.43392	0	65535	#NUM!	-7.43392	-7.43392	-7.43392	-7.43392
19	920	-10.0234	0	65535	#NUM!	-10.0234	-10.0234	-10.0234	-10.0234
20	480	-13.6784	0	65535	#NUM!	-13.6784	-13.6784	-13.6784	-13.6784
21	16400	5.327485	0	65535	#NUM!	5.327485	5.327485	5.327485	5.327485
22									
23									
24									
25	RESIDUAL OUTPUT				PROBABILITY OUTPUT				
26									

In the above regression analysis output, 'multiple R' is referred to as multiple correlation coefficient and R square is referred to as coefficient of multiple determination like simple linear regression, R-square determines the percentage of variation in the dependent variable.

3.5 Regression by the Method of Least Squares

Q15. What is regression by the method of lease square?

Ans :

The least squares method is a form of mathematical regression analysis that finds the line of best fit for a dataset, providing a visual demonstration of the relationship between the data points.

Least Squares Method

The least squares method is a form of mathematical regression analysis that finds the line of best fit for a dataset, providing a visual demonstration of the relationship between the data points. Each point of data is representative of the relationship between a known independent variable and an unknown dependent variable.

The least squares method provides the overall rationale for the placement of the line of best fit among the data points being studied. The most common application of the least squares method, referred to as linear or ordinary, aims to create a straight line that minimizes the sum of the squares of the errors generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value and the value anticipated based on the model.

This method of regression analysis begins with a set of data points to be graphed. An analyst using the least squares method will seek a line of best fit that explains the potential relationship between an independent variable and a dependent variable. In regression analysis, dependent variables are designated on the vertical Y axis and independent variables are designated on the horizontal X-axis. These designations will form the equation for the line of best fit, which is determined from the least squares method.

Least-Squares Regression

Let us suppose that the given data (x_i, y_i), i = 1..n is inexact and has substantial error, right from their source where they are obtained. Experimental data is usually scattered and is a good example for inexact data. Polynomial interpolation is inappropriate in such cases. To understand this let us look at the following graphical representation of some scattered data:

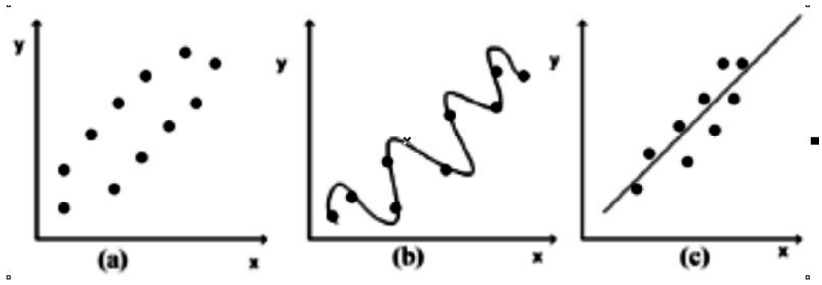


Fig.:1(a) Scattered Data, 1(b) A polynomial fit oscillating beyond the range of the data; 1(c) An approximate fit for data.

Now a look at the data in figure 1(a) tells us that the data has increasing trend i.e. higher values of y are associated with higher values of x. As in figure 1(b) if we fit an eight order interpolation polynomial, it passes through the data exactly but oscillates due to the scattered nature of data and also goes well beyond the range suggested by data. Hence a more appropriate way is to find a function as shown in figure 1(c), that fits the shape or general trend of the data. One of the standard techniques for finding such a fit is Least-Square Regression.

Q16. Explain the Least Square Method.

Ans :

The principle of least squares is one of the popular methods for finding a curve fitting a given data. Say (x_1, y_1), (x_2, y_2), ... (x_n, y_n) be n observations from an experiment. We are interested in finding a curve

$$y = f(x) \qquad \text{... (1)}$$

Closely fitting the given data of size 'n'. Now at $x = x_1$ while the observed value of y is y_1, the expected value of y from the curve (1) is $f(x_1)$. Let us define the residual by

$$e_1 = y_t - f(x_t) \qquad \text{... (2)}$$

Likewise, the residuals at all other points x_2, .., x_n are given by

$$e_2 = y_2 - f(x_2) \qquad \text{... (3)}$$

$$e_n = y_n - f(x_n)$$

Some of the residuals e_i's may be positive and some may be negative. We would like to find the curve fitting the given data such that the residual at any x_i is as small as possible. Now since some of the residuals are positive and others are negative and as we would like to give equal importance to all the residuals it is desirable to consider sum of the squares of these residuals, say E and thereby find the curve that minimizes E. Thus, we consider

$$E = \sum_{i=1}^{n} e_i^2 \qquad \text{... (4)}$$

and find the best representative curve (1) that minimizes (4).

3.5.1 Least Square Fit of a Straight Line

Q17. Explain about Least Square Fit of a Straight Line.

Ans :

Suppose that we are given a data set $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ of n observations from an experiment. Say that we are interested in fitting a straight line,

$$y = ax + b$$

to the given data. Find the 'n' residuals e_i by:

$$e_i = y_i - (ax_i + b), i = 1, 2....n \qquad ... (2)$$

Now consider the sum of the squares of $e_i's$ i.e

$$E = \sum_{i=1}^{n} e_i^2$$

x	y	xy	x^2
0	1	0	0
1	5	5	1
2	10	20	4
3	22	66	9
4	38	152	16
$\sum_i x_i = 10$	$\sum_i y_i = 76$	$\sum_j x_j y_i = 243$	$\sum_j x_i^2 = 30$

Therefore the normal equations are given by:

$$30a + 10b = 243 \qquad ... (3)$$

$$10a + 5b = 76 \qquad ... (4)$$

On solving (3) and (4) we get

$$a = 9.1, b = -3 \qquad ... (5)$$

Hence the required fit for the given data is

$$y = 9.1\,x - 3 \qquad ... (6)$$

3.6 Building Good Regressional Models

Q18. Describe briefly the systematic approach for building good regression models.

Ans :

A regression model containing exclusively significant independent variables is referred to as good regression model. In this model, the significance of variable cannot be predicted by adding or deleting independent variables from a model. As the significance, variables varies from one model to another, it is necessary to employ more systematic approach.

In regression model, whenever an indepen-dent variable is added the value of R^2 becomes greater than or equal to the value of R^2 in the original model. The value of R^2 do not serve better in building good regression models. The best approach followed for building good regression models is using the parameter 'Adjusted R-square'. An adjusted R^2 not only generates impact on the number of independent variables but, also on the sample size. The value of adjusted R^2 increases or decreases depending on the addition or deletion of independent variables from a model. Moreover, an increase in adjusted R^2 specifies improvement in the model.

The process of building good regression models involves two approaches. They are,

(i) Using P-values

(ii) Using t-statistics.

The steps to be followed for building good regression models using P-values are given below,

1. Develop regression model using independent variables.
2. Determine the significance of independent variables by analysing the P-values.
3. Determine the independent variable containing P-value greater than the specified level of significance.
4. Delete the independent variable identified in step (3) and calculate the value of adjusted R^2.
5. Repeat the above process till the model contains only significant variables.

This approach using P-values determines a significant model having highest adjusted R^2.

The second approach of building good regression models is using f-statistics. It operates similarly as first approach except the fact that it uses r-values in place of P-values. If the value of t is less than 1 then there is decrease in standard error. In this case, if the variable is deleted then adjusted R^2 increases. On the other hand, if value of t is greater than 1, then there is increase in standard error and decrease in the value of adjusted R^2.

The two approaches using P-values and r-statistics for determining the significant variables requires great amount of evaluation. Additionally, the number of independent variables increases, the number of models also increases. For instance, a collection of eight independent variables requires construction of a total of 256 (= 28) models.

Due to this reason, it becomes difficult for the user to eliminate or remove variables that are insignificant.

3.7 Regression with Categorical Independent Variables

Q19. What is regression with categorical independent variables.

Ans :

Categorical variables commonly occur in research settings.

- Another term sometimes used to describe for categorical variables is that of qualitative variables.
- A strict definition of a qualitative or categorical variable is that of a variable that has a finite number of levels.
- Continuous (or quantitative) variables, alternatively, have infinitely many levels.
- Often this is assumed more than practiced.
- Quantitative variables often have countably many levels.
- Level of precision of an instrument can limit the number of levels of a quantitative variable.
- Categorical variables can occur in many different research designs:
- Experimental research.
- Quasi-experimental research.
- Non experimental/Observational research.

- Such variables can be used with regression for:
- Prediction.
- Explanation.
- Because of nature of categorical variables, emphasis of regression is not on linear trends but on differences between means (of Y) at each level of the category.
- Not all categorical variables are ordered (like cereal box type, gender, etc...).
- When considering differences in the mean of the dependent variable, the type of analysis being conducted by a regression is commonly called an Analysis of variance (ANOVA).

3.8 Linear Discriminant Analysis

Q20. Explain linear discriminant analysis.

Ans :

Linear discriminant analysis is typically used to identify the characteristics which can accurately discriminate between the respondents who fall in one category from those who fall in another category.

For example, LDA can be used to study successful salesmen and unsuccessful salesmen in order to determine the characteristics which are possessed by successful salesman but not possessed by unsuccessful salesman.

Once the characteristics of successful salesman have been identified, the information can be used to recruit individuals with characteristics similar to those possessed by successful salesmen.

LDA can also be used to study owners and non-owners of videotape recorders, or to study beer drinkers who prefer different brands of beer. In each of these situations, a researcher can use LDA as an attempt to determine the characteristics which are possessed by one category of respondent but not possessed by the other categories of respondents. Such information can be useful to a manufacturer of videotape recorders or to a brewer of a certain brand of beer.

By knowing how the respondents in their target market are different from the respondents not in their target market, the companies involved will have a better definition of their target market and this knowledge can help them greatly to design more effective marketing programs.

LDA is applied to a large scatter diagram of data, which represents the characteristics of individual salesmen (Example : education, experience, and so on). Some of the data points in the scatter diagram belong to salesmen who fall into one category (Example : successful), while the rest of the data points belong to salesmen who fall into another category (Example : unsuccessful).

LDA attempts to find a straight line which, when placed in the scatter diagram, accurately discriminates or separates one category from the other. In this example, all or most of the respondents one side of the line will be successful salesmen, who posses certain characteristics, and all or most of the respondents on the other side of the line will be unsuccessful salesmen who possess different characteristics.

3.9 One Way and Two Way Anova

Q21. What is ANOVA ? Explain the assumptions ?

Ans :

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis. Basically, you're testing groups to see if there's a difference between them. Examples of when you might want to test different groups:

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if one college outperforms the other.

- A procedure for comparing more than two groups – independent variable: smoking status
  - non-smoking
  - one pack a day
  - two packs a day – dependent variable: number of coughs per day
  - k = number of conditions
- Statistical technique specially designed to test whether the means of more than 2 quantitative populations are equal. Developed by Sir Ronald A. Fisher in 1920's.
- Analysis of Variance (ANOVA) is a hypothesis-testing technique used to test the equality of two or more population (or treatment) means by examining the variances of samples that are taken.
- ANOVA allows one to determine whether the differences between the samples are simply due to random error (sampling errors) or whether there are systematic treatment effects that causes the mean in one group to differ from the mean in another.
- Most of the time ANOVA is used to compare the equality of three or more means, however
- when the means from two samples are compared using ANOVA it is equivalent to using a t-test to compare the means of independent samples.
- ANOVA is based on comparing the variance (or variation) between the data samples to variation within each particular sample. If the between variation is much larger than the within variation, the means of different samples will not be equal. If the between and within variations are approximately the same size, then there will be no significant difference between sample means.

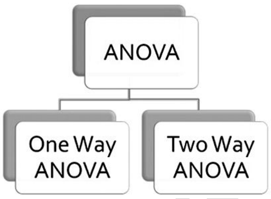
Assumptions of ANOVA

(i) All populations involved follow a normal distribution.

(ii) All populations have the same variance (or standard deviation).

(iii) The samples are randomly selected and independent of one another.

Since ANOVA assumes the populations involved follow a normal distribution, ANOVA falls into a category of hypothesis tests known as parametric tests. If the populations involved did not follow a normal distribution, an ANOVA test could not be used to examine the equality of the sample means. Instead, one would have to use a non-parametric test (or distribution-free test), which is a more general form of hypothesis testing that does not rely on distributional assumptions.



Q22. Explain One Way ANOVA with example?

Ans :

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups (although you tend to only see it used when there are a minimum of three, rather than two groups).

One Way Anova:

One way ANOVA or single factor ANOVA:

- Determines means of ≥ 3 independent groups significantly different from one another.
- Only 1 independent variable (factor/grouping variable) with ≥ 3 levels
- Grouping variable- nominal
- Outcome variable- interval or ratio
- One-Way Analysis of Variance (ANOVA) Example Problem

Consider this example:

Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use $\alpha = 5\%$.

Table : ANOVA

Compact Cars	Midsize Cars	Full-size Cars
643	469	484
655	427	456
702	525	402

$\bar{X}$	666.67	473.67	447.33
S	31.18	49.17	41.68

1. State the null and alternative hypotheses

The null hypothesis for an ANOVA always assumes the population means are equal. Hence, we may write the null hypothesis as:

H_0: $\mu_1 = \mu_2 = \mu_3$ - The mean head pressure is statistically equal across the three types of cars.

Since the null hypothesis assumes all the means are equal, we could reject the null hypothesis if only mean is not equal. Thus, the alternative hypothesis is:

H_a: At least one mean pressure is not statistically equal.

2. Calculate the appropriate test statistic

The test statistic in ANOVA is the ratio of the between and within variation in the data. It follows an F distribution.

Total Sum of Squares – the total variation in the data. It is the sum of the between and within variation.

Total Sum of Squares (SST) = $\sum_{i=1}^{r} \sum_{j=1}^{c} (X_{ij} - \bar{\bar{X}})^2$, where r is the number of rows in the table, c is the number of columns, $\bar{\bar{X}}$ is the grand mean, and X_{ij} is the ith observation in the jth column.

Using the data in Table ANOVA.1 we may find the grand mean :

$$\bar{\bar{X}} = \frac{\Sigma X_{ij}}{N} = \frac{(643+655+702+469+427+525+484+456+402)}{9} = 529.22$$

$$\text{SST} = (643 - 529.22)^2 + (655 - 529.22)^2 + (702 - 529.22)^2 + (469 - 529.22)^2$$
$$+ \ldots + (402 - 529.22)^2 = 96303.55$$

Between Sum of Squares (or Treatment Sum of Squares) – variation in the data between the different samples (or treatments).

Treatment Sum of Squares (SSTR) = $\Sigma\ r_j\ (\bar{X}_j - \bar{\bar{X}})^2$, where r_j is the number of rows in the jth treatment and $\bar{X}_j$ is the mean of the jth treatment.

Using the data in Table ANOVA.1,

$$\text{SSTR} = [3 * (666.67 - 529.22)^2] + [3 *(473.67 - 529.22)^2] + [3 - (447.33 - 529.22)^2]$$

$$= 86049.55$$

Within variation (or Error Sum of Squares) – Variation in the data from each individual treatment.

Error Sum of Squares (SSE) = $\sum\sum (X_{ij} - \bar{X}_j)^2$

From Table ANOVA.1,

$$\text{SSE} = [(643 - 666.67)^2 + (655 - 666.67)^2 + (702 - 666.67)^2]$$
$$+ [(469 - 473.67)^2 + (427 - 473.67)^2 + (525 - 473.67)^2]$$
$$+ [(484 - 447.33)^2 + (456 - 447.33)^2 + (402 - 447.33)^2] = 10254.$$

Note that SST = SSTR + SSE (96303.55 = 86049.55 + 10254).

Hence, you only need to compute any two of three sources of variation to conduct an ANOVA.

Especially for the first few problems you work out, you should calculate all three for practice.

The next step in an ANOVA is to compute the "average" sources of variation in the data using SST, SSTR, and SSE.

Total Mean Squares (MST) = $\frac{\text{SST}}{\text{N}-1}$ → "Average total variation in the data" (N is the total number of observations)

$$\text{MST} = \frac{96303.55}{(9-1)} = 12037.94$$

Mean Square Treatment (MSTR) = $\frac{\text{SSTR}}{\text{c}-1}$ → "average between variation" (c is the number of columns in the data table)

$$\text{MSTR} = \frac{86049.55}{(3-1)} = 43024.78$$

Mean Square Error (MSE) = $\frac{\text{SSE}}{\text{N}-\text{c}}$ → "average within variation"

$$\text{MSE} = \frac{10254}{(9-3)} = 1709$$

Note: MST $\neq$ MSTR + MSE

The test statistic may now be calculated. For a one-way ANOVA the test statistic is equal to the ratio of MSTR and MSE. This is the ratio of the "average between variation" to the "average within variation." In addition, this ratio is known to follow an F distribution. Hence,

$$\text{F} = \frac{\text{MSTR}}{\text{MSE}} = \frac{43024.78}{1709} = 25.17$$

The intuition here is relatively straightforward. If the average between variation rises relative to the average within variation, the F statistic will rise and so will our chance of rejecting the null hypothesis.

3. Obtain the Critical Value

To find the critical value from an F distribution you must know the numerator (MSTR) and denominator (MSE) degrees of freedom, along with the significance level.

F^{CV} has df1 and df2 degrees of freedom, where df1 is the numerator degrees of freedom equal to c–1 and df2 is the denominator degrees of freedom equal to N–c.

In our example, df1 = 3 – 1 = 2 and df2 = 9 – 3 = 6. Hence we need to find $F^{CV}_{2.6}$ corresponding to $\alpha = 5\%$. Using the F tables in your text we determine that $F^{CV}_{2.6} = 5.14$.

4. Decision Rule

You reject the null hypothesis if: F (observed value) > F^{CV} (critical value). In our example 25.17 > 5.14, so we reject the null hypothesis.

5. Interpretation

Since we rejected the null hypothesis, we are 95% confident $(1 - \alpha)$ that the mean head pressure is not statistically equal for compact, midsize, and full size cars. However, since only one mean must be different to reject the null, we do not yet know which mean(s) is/are different. In short, an ANOVA test will test us that at least one mean is different, but an additional test must be conducted to determine which mean(s) is/are different.

Determining which Mean(s) Is / Are Different

If you fail to reject the null hypothesis in an ANOVA then you are done. You know, with some level of confidence, that the treatment means are statistically equal. However, if you reject the null then you must conduct a separate test to determine which mean(s) is/are different.

There are several techniques for testing the differences between means, but the most common test is the Least Significant Difference Test.

Least Significant Difference (LSD) for a balanced sample: $\sqrt{\frac{2 * MSE * F_{1,\,N\text{-}c}}{r}}$, where MSE is the mean square error and r is the number of rows in each treatment.

In the example above, LSD = $\sqrt{\frac{(2)\,(1709)\,(5.99)}{3}} = 82.61$

Thus, if the absolute value of the difference between any two treatment means is greater than 82.61, we may conclude that they are not statistically equal.

Compact cars vs. Midsize cars:

|666.67 – 473.67| = 193. Since 193 > 82.61 → mean head pressure is statistically different between compact and midsize cars.

Midsize cars vs. Full-size cars:

|473.67 – 447.33| = 26.34. Since 26.34 < 82.61 → mean head pressure is statistically equal between midsize and full-size cars.

Compact vs. Full-size:

Work this on your own.

One-way ANOVA in Excel

You may conduct a one-way ANOVA using Excel.

(Preliminary step) First, make sure that the "Analysis ToolPak" is installed.

Under "Tools" is the option "Data Analysis" present?

If yes – ToolPak is installed.

If no – select "Add-ins."

Check the boxes entitled "Analysis ToolPak" and "Analysis ToolPak – VBA" and click "OK". This will install the "Data Analysis ToolPak."

1. Under "Tools" select "Data Analysis"

   In the window that appears select "ANOVA: One factor" and click "OK."

2. Using your mouse highlight the cells containing the data.
3. Select "Columns" if each treatment is its own column or "Row" if each treatment is its own row.
4. Set your level of significance. (The default is 5% or 0.05.)
5. Click "OK" and the ANOVA output will appear on a new worksheet.

ANOVA Results from Excel:

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	3	2000	666.6667	972.3333
Column 2	3	1421	473.6667	2417.333
Column 3	3	1342	447.3333	1737.333

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	86049.55556	2	43024.78	25.17541	0.001207	5.143249
Within Groups	10254	6	1709			
Total	96303.55556	8				

The results under the heading "SUMMARY" simply provides you with summary statistics for each of your samples. The results of the ANOVA test are provided under the heading "ANOVA." Comparing these figures with the example above, it should be simple to determine the meaning of the Excel output.

Q23. Explain the Two Way ANOVA with an example.

Ans :

The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable. For example, you could use a two-way ANOVA to understand whether there is an interaction between gender and educational level on test anxiety amongst university students, where gender (males/females) and education level (undergraduate/postgraduate) are your independent variables, and test anxiety is your dependent variable. Alternately, you may want to determine whether there is an interaction between physical activity level and gender on blood cholesterol concentration in children, where physical activity (low/moderate/high) and gender (male/female) are your independent variables, and cholesterol concentration is your dependent variable.

Assumptions

➢ Two independent variables should each consist of two or more categorical, independent groups.

➢ Dependent variable should be measured at the continuous level (i.e., they are interval or ratio variables).

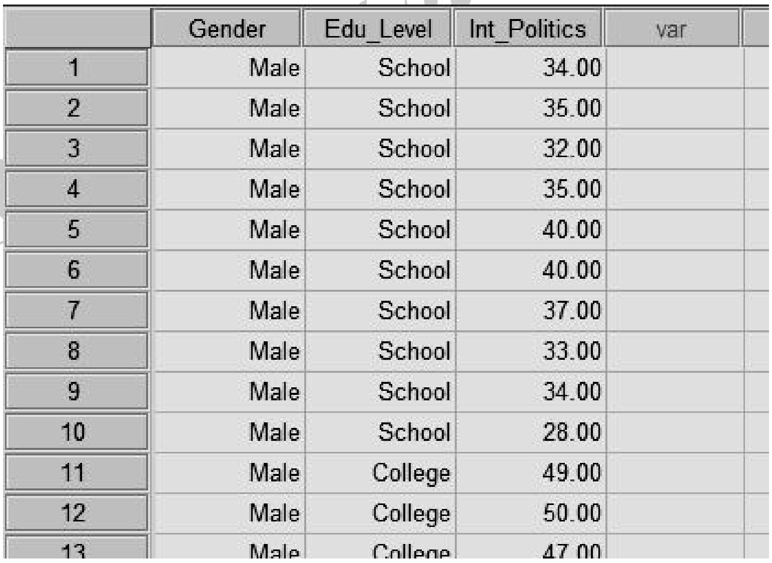
➢ It should have independence of observations, which means that there is no relationship between the observations in each group or between the groups themselves.

- There should be no significant outliers.
- Dependent variable should be approximately normally distributed for each combination of the groups of the two independent variables.
- There needs to be homogeneity of variances for each combination of the groups of the two independent variables.

Example

A researcher was interested in whether an individual's interest in politics was influenced by their level of education and gender. They recruited a random sample of participants to their study and asked them about their interest in politics, which they scored from 0 to 100, with higher scores indicating a greater interest in politics. The researcher then divided the participants by gender (Male / Female) and then again by level of education (School / College / University). Therefore, the dependent variable was "interest in politics", and the two independent variables were "gender" and "education".

In SPSS Statistics, we separated the individuals into their appropriate groups by using two columns representing the two independent variables, and labelled them Gender and Edu_Level. For Gender, we coded "males" as **1** and "females" as **2**, and for Edu_Level, we coded "school" as **1**, "college" as **2** and "university" as **3**. The participants' interest in politics – the dependent variable – was entered under the variable name, Int_Politics. The setup for this example can be seen below :

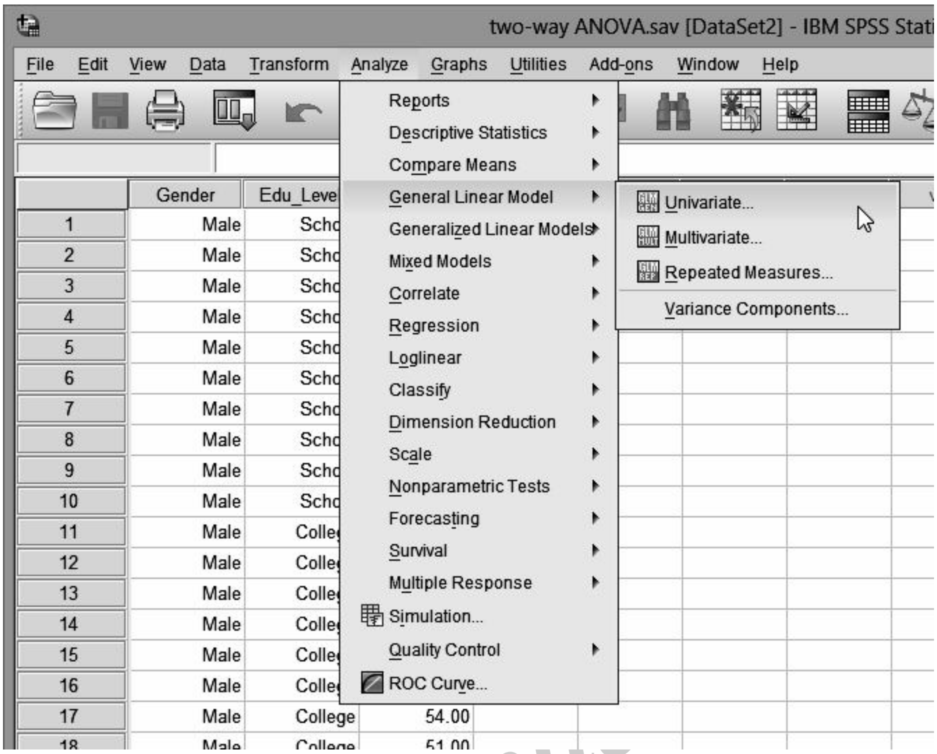


	Gender	Edu_Level	Int_Politics	var
1	Male	School	34.00	
2	Male	School	35.00	
3	Male	School	32.00	
4	Male	School	35.00	
5	Male	School	40.00	
6	Male	School	40.00	
7	Male	School	37.00	
8	Male	School	33.00	
9	Male	School	34.00	
10	Male	School	28.00	
11	Male	College	49.00	
12	Male	College	50.00	
13	Male	College	47.00	

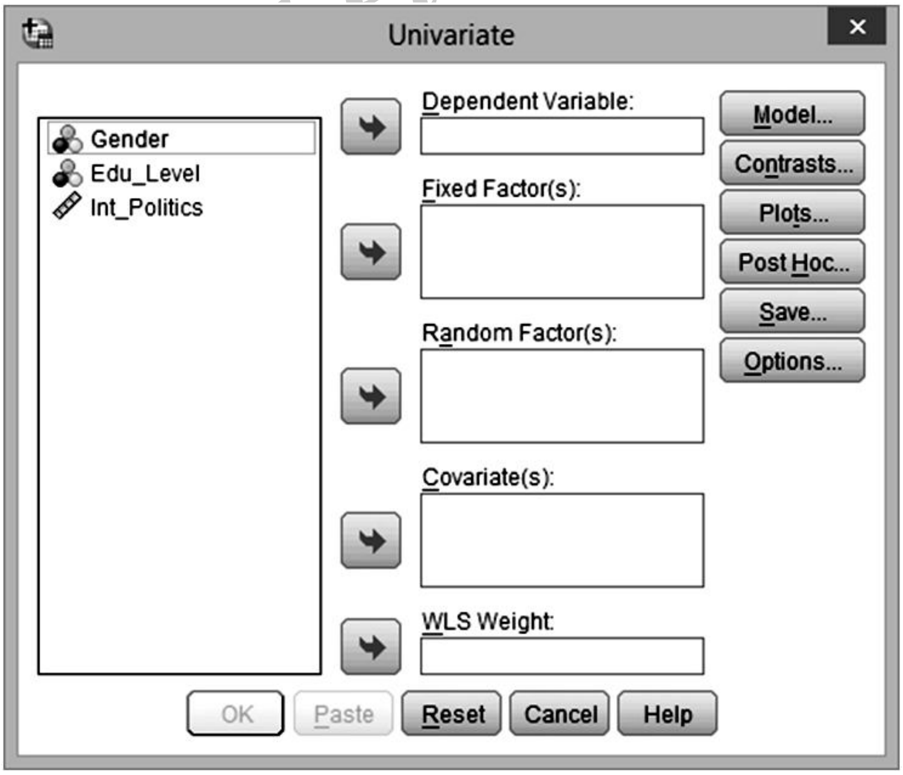
Test Procedure in SPSS Statistics

The 14 steps below show you how to analyse your data using a two-way ANOVA in SPSS Statistics when the six assumptions in the previous section, Assumptions, have not been violated. At the end of these 14 steps, we show you how to interpret the results from this test. If you are looking for help to make sure your data meets assumptions #4, #5 and #6, which are required when using a two-way ANOVA and can be tested using SPSS Statistics, you can learn more in our enhanced guides here.

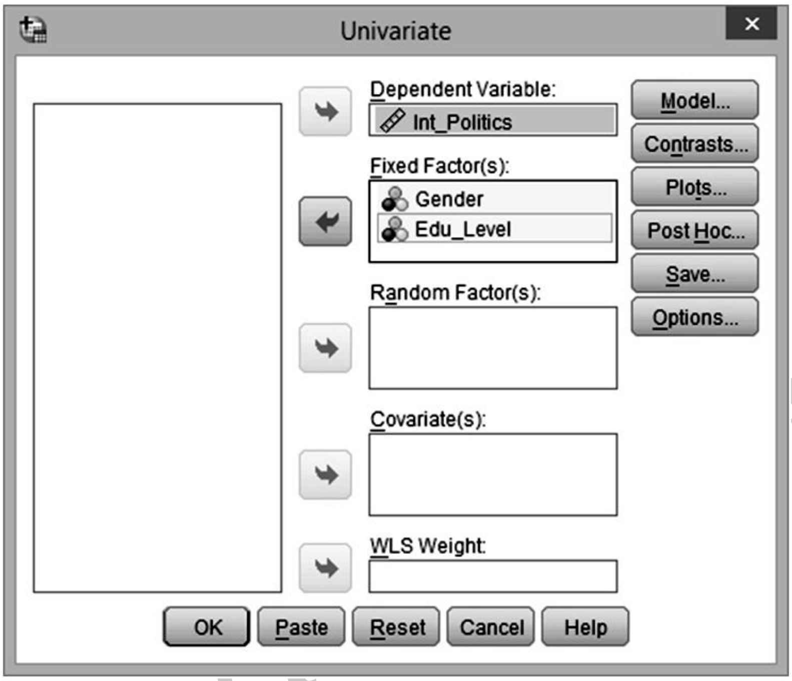
Step 1 : Click Analyze > General Linear Model > Univariate... on the top menu, as shown below:



Step 2 : You will be presented with the Univariate dialogue box, as shown below:



Step 3 : Transfer the dependent variable, Int_Politics, into the Dependent Variable: box, and transfer both independent variables, Gender and Edu_Level, into the Fixed Factor(s): box. You can do this by drag-and-dropping the variables into the respective boxes or by using the ➡ button. If you are using older versions of SPSS Statistics you will need to use the latter method. You will end up with a screen similar to that shown below:



Univariate

Dependent Variable: Int_Politics

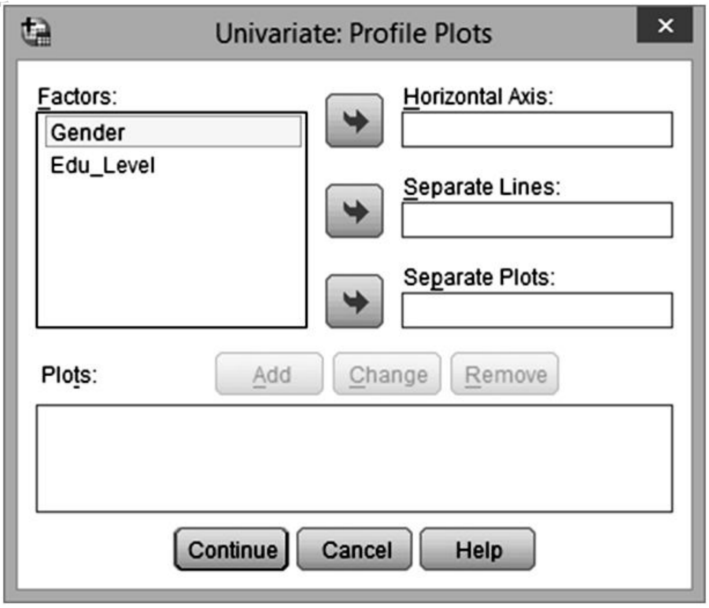
Fixed Factor(s): Gender, Edu_Level

Random Factor(s):

Covariate(s):

WLS Weight:

Step 4 : Click on the Plots... **button. You will be presented with the Univariate:** Profile Plots dialogue box, as shown below:



Univariate: Profile Plots

Factors: Gender, Edu_Level

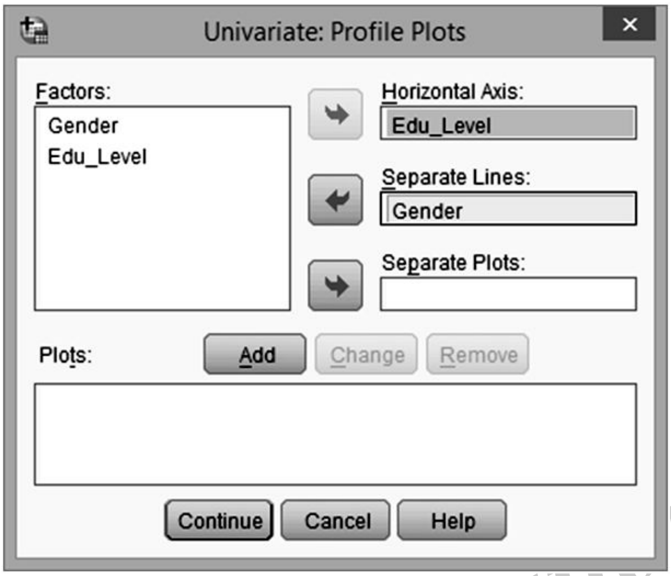
Horizontal Axis:

Separate Lines:

Separate Plots:

Plots:

Step 5: Transfer the independent variable, Edu_Level, from the Factors: box into the Horizontal Axis: box, and transfer the other independent variable, Gender, into the Separate Lines: box. You will be presented with the following screen:



Univariate: Profile Plots

Factors: Gender, Edu_Level

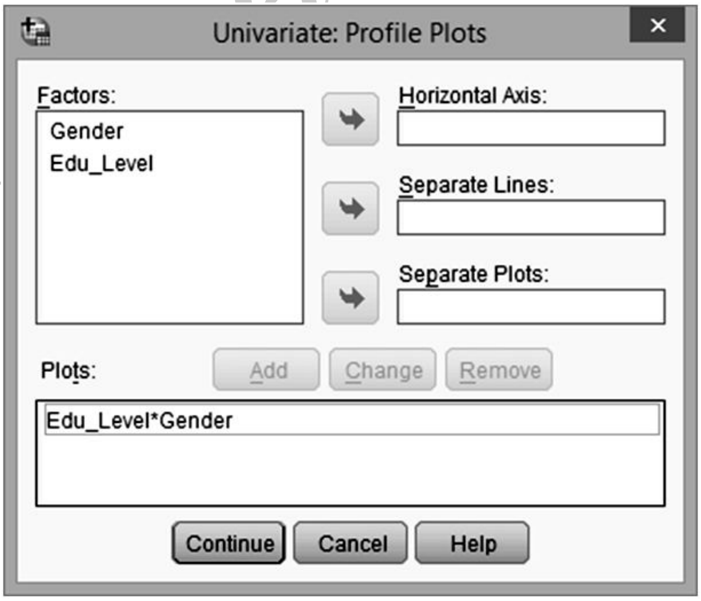
Horizontal Axis: Edu_Level

Separate Lines: Gender

Separate Plots:

Plots:

Step 6: Click the Add button. You will see that "Edu_Level*Gender" has been added to the Plots: box, as shown below:



Univariate: Profile Plots

Factors: Gender, Edu_Level

Horizontal Axis:

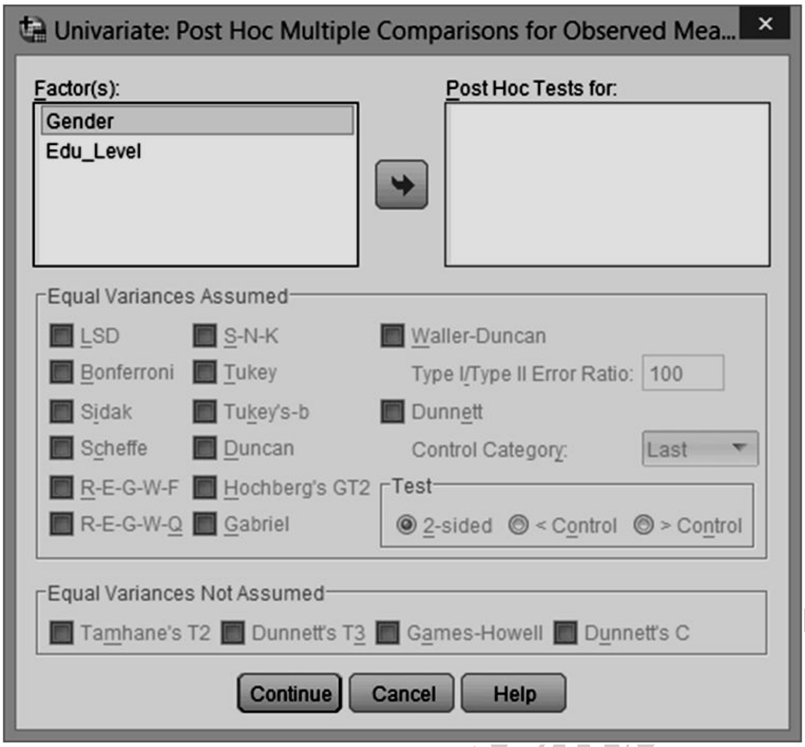
Separate Lines:

Separate Plots:

Plots: Edu_Level*Gender

Step 7: Click the Continue button. This will return you to the Univariate dialogue box.

Step 8: Click the Post Hoc... button. You will be presented with the Univariate: Post Hoc Multiple Comparisons for Observed Means dialogue box, as shown below:



Univariate: Post Hoc Multiple Comparisons for Observed Mea...

Factor(s): Gender, Edu_Level

Post Hoc Tests for:

Equal Variances Assumed

□ LSD □ S-N-K □ Waller-Duncan

□ Bonferroni □ Tukey Type I/Type II Error Ratio: 100

□ Sidak □ Tukey's-b □ Dunnett

□ Scheffe □ Duncan Control Category: Last

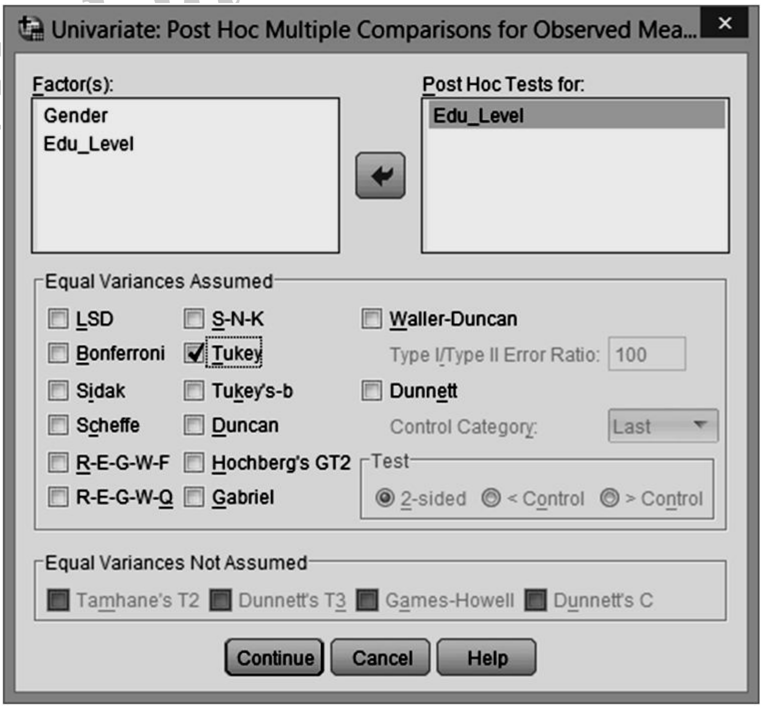
□ R-E-G-W-F □ Hochberg's GT2 Test

□ R-E-G-W-Q □ Gabriel ● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

□ Tamhane's T2 □ Dunnett's T3 □ Games-Howell □ Dunnett's C

Step 9: Transfer Edu_Level from the Factor(s): box to the Post Hoc Tests for: box. This will make the –Equal Variances Assumed– area become active (lose the "grey sheen") and present you with some choices for which post hoc test to use. For this example, we are going to select Tukey, which is a good, all-round post hoc test.



Univariate: Post Hoc Multiple Comparisons for Observed Mea...

Factor(s): Gender, Edu_Level

Post Hoc Tests for: Edu_Level

Equal Variances Assumed

□ LSD □ S-N-K □ Waller-Duncan

□ Bonferroni ☑ Tukey Type I/Type II Error Ratio: 100

□ Sidak □ Tukey's-b □ Dunnett

□ Scheffe □ Duncan Control Category: Last

□ R-E-G-W-F □ Hochberg's GT2 Test

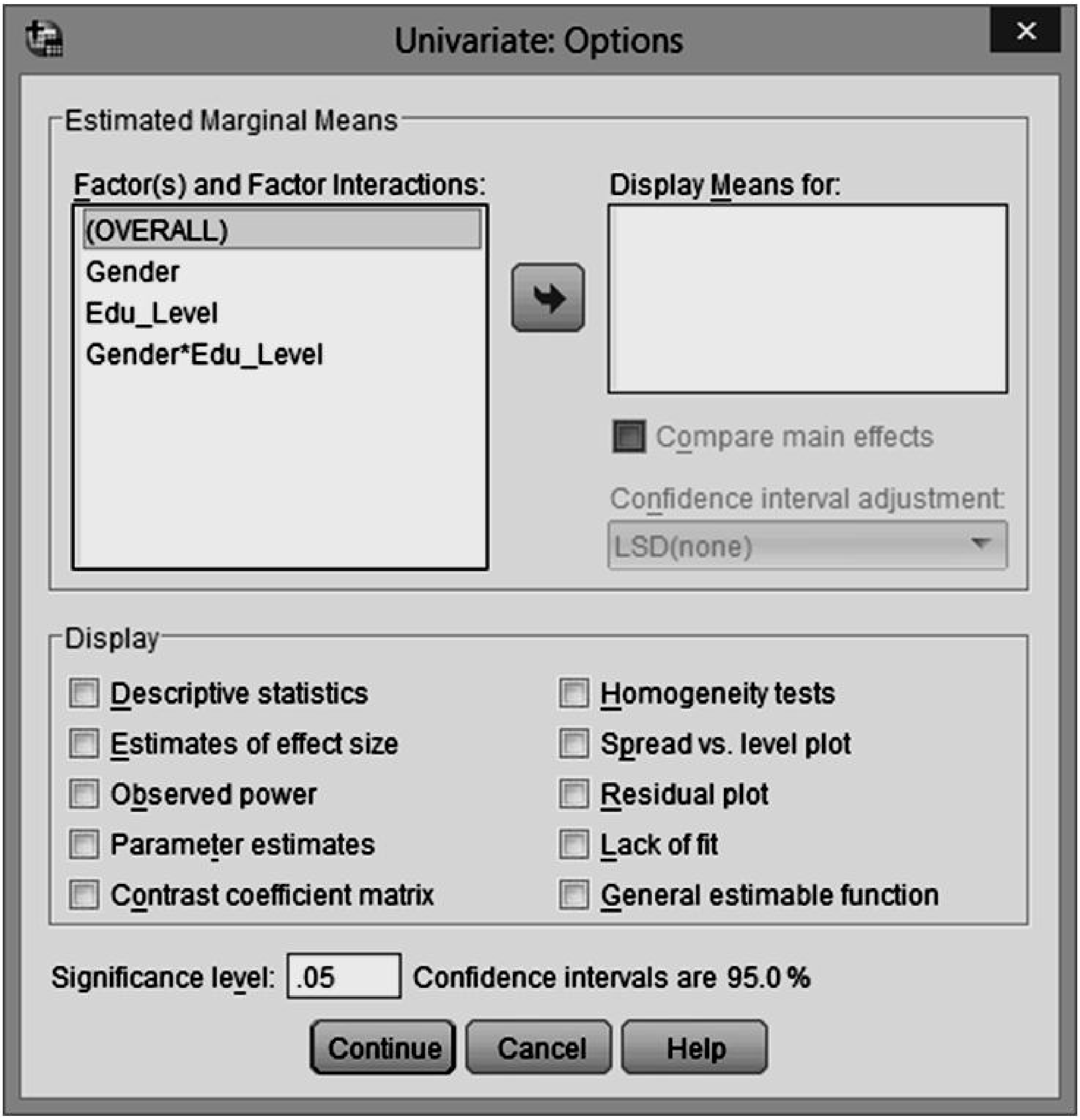
□ R-E-G-W-Q □ Gabriel ● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

□ Tamhane's T2 □ Dunnett's T3 □ Games-Howell □ Dunnett's C

Step 10: Click the Continue button to return to the Univariate dialogue box.

Step 11: Click the Options... button. This will present you with the Univariate: Options dialogue box, as shown below:



Univariate: Options

Estimated Marginal Means

Factor(s) and Factor Interactions:

(OVERALL)
Gender
Edu_Level
Gender*Edu_Level

Display Means for:

■ Compare main effects

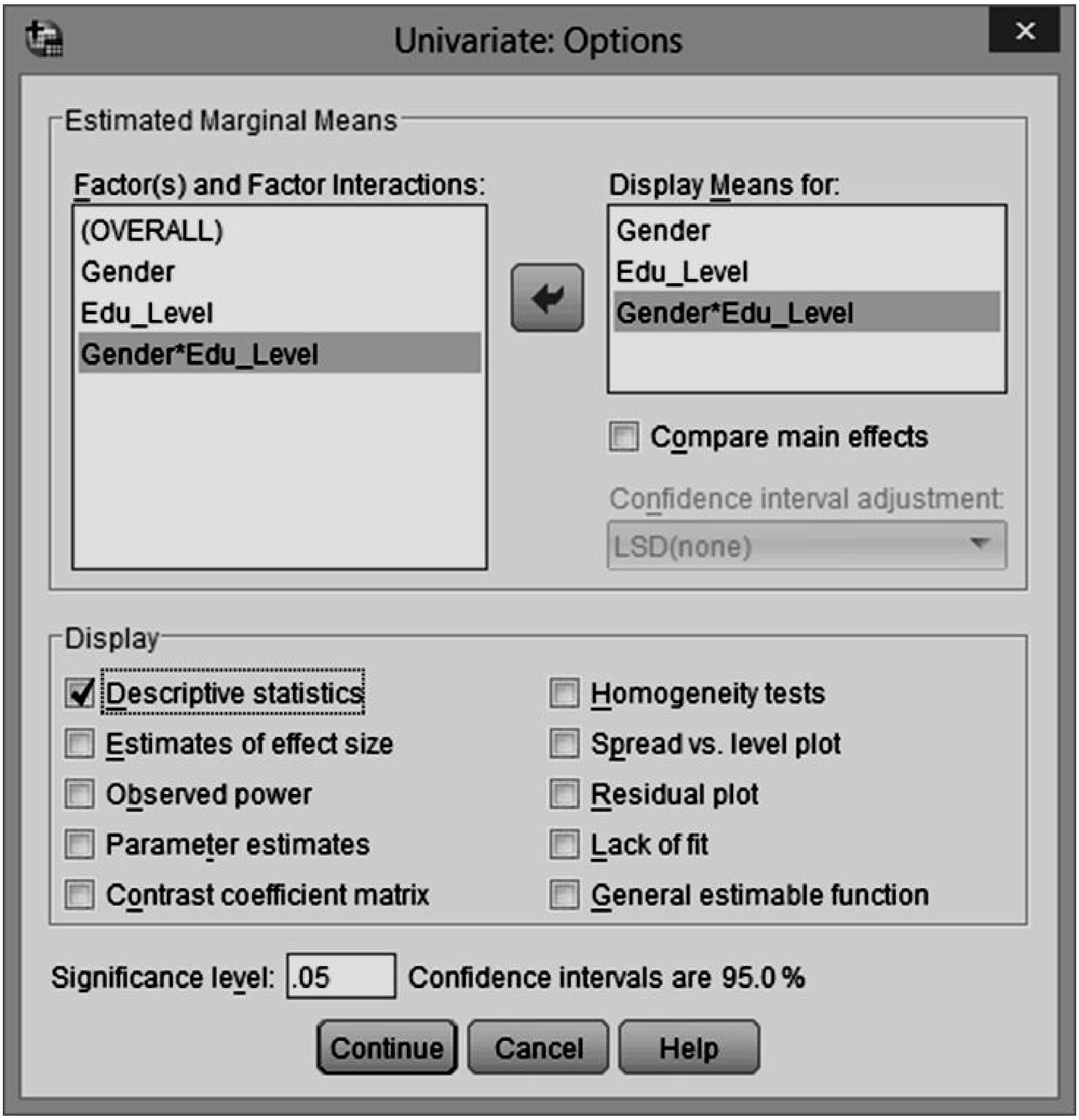
Confidence interval adjustment: LSD(none)

Display

□ Descriptive statistics
□ Estimates of effect size
□ Observed power
□ Parameter estimates
□ Contrast coefficient matrix
□ Homogeneity tests
□ Spread vs. level plot
□ Residual plot
□ Lack of fit
□ General estimable function

Significance level: .05 Confidence intervals are 95.0 %

Step 12: Transfer Gender, Edu_Level and Gender*Edu_Level from the Factor(s) and Factor Interactions: box into the Display Means for: box. In the –Display– area, tick the Descriptive Statistics option. You will presented with the following screen:



Univariate: Options

Estimated Marginal Means

Factor(s) and Factor Interactions:
(OVERALL)
Gender
Edu_Level
Gender*Edu_Level

Display Means for:
Gender
Edu_Level
Gender*Edu_Level

□ Compare main effects

Confidence interval adjustment: LSD(none)

Display

☑ Descriptive statistics
□ Estimates of effect size
□ Observed power
□ Parameter estimates
□ Contrast coefficient matrix
□ Homogeneity tests
□ Spread vs. level plot
□ Residual plot
□ Lack of fit
□ General estimable function

Significance level: .05 Confidence intervals are 95.0 %

Step 13: Click the Continue button to return to the Univariate dialogue box.

Step 14: Click the OK button to generate the output.

Short Question and Answers

Q1. Karl Pearson's of correlation

Ans :

Correlation analysis is the statistical tool we can used to describe the degree to which one variable is linearly related to another.

According to Croxton and Cowden, "The appropriate statistical tool for discovering and measuring the relationship of quantitative nature and expressing it in brief formula is known as correlation".

According to Tippet, "The effects of correlation are to reduce the range of uncertainty of our prediction".

The coefficient of correlation measures the degree of relationship between two set of figures. As the reliability of estimates depend upon the closeness of the relationship it is imperative that utmost care be taken while interpreting the value of coefficient of correlation, otherwise wrong conclusion can be drawn.

Q2. Properties of Karl Pearson's Coefficient of Correlation.

Ans :

1. It is based on Arithmetic Mean and Standard Deviation.
2. It lies between -1
3. It measures both direction as well as degree of change. If r is less than 0, there is negative correlation, which means the direction of change of the two variables will be opposite. If r is more than 0, there is positive correlation, which means the direction of change of the two variables will be same. Higher the value of r, greater is the degree of correlation. Hence, Karl Pearson's coefficient of correlation is said to be the ideal measure of correlation.
4. It is independent of change in scale. In other words, if a constant amount is added/ subtracted from all values of a variable, the value of r does not change.
5. It is independent of change in origin. Thus, if a constant amount is multiplied with or divides all values of a variable, 'r' does not change.
6. It is independent of direction. In other words, Correlation of X and Y is same as Correlation of Y and X.
7. It is a pure number without any units. In other words, it is independent of the unit of measurement of the 2 variables.
8. It takes into account all items of the variable(s).

Q3. Spearman's rank correlation.

Ans :

'Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This problem arises while dealing with qualitative characteristics such as honesty, beauty, character, morality etc. which cannot be measured quantitatively, but can be arranged serially. In such cases, Karl Pearson's Coefficient of Correlation may not serve the purpose.

Another method was developed by Edward Spearman to study correlation between such attributes. In this method, the change in a variable with respect to a change in another variable is not measured by means of absolute change as it is difficult to quantify the absolute measure. However, if the movement of the two variables is similar, they should be getting similar, if not identical, ranks. Thus, if the difference in ranks is minimal, then there is a case of positive correlation. If the difference in ranks is huge, then it indicates negative correlation.

Q4. Merits of Spearman's Rank Correlation

Ans :

1. It is easy to understand and calculate
2. It is not impacted by extreme values.

3. It is a numerical measure and provides objectivity to subjective ranking.
4. It is the only method of finding correlation with respect to qualitative factors such as honesty, beauty, etc.
5. It measures both direction as well as degree of change.
6. It facilitates comparisons between two series.
7. It can be applied to irregular data also.
8. It is ideal when the number of observations is very small.

Q5. Regression Analysis.

Ans :

Regression takes its name from studies made by Sir Francis Galton. He compared the heights of persons to the heights of their parents. His major conclusion was that the off-springs of unusually tall persons tend to be shorter than their parents while children of usually short parents tend to be shorter than their parents.

In a sense, the successive generations of off-springs from tall persons "regress" downward toward the height of the population. While the reverse is true originally for short families. But the distribution of heights for the total population, continues to have the same variability from generation to generation.

Regression Variables

i) **Independent Variable (Regressor or Predictor or Explanatory).** The variable which influences the values of the other variable or which is used for prediction of the value of the other variable is called independent variable.

ii) **Dependent Variable (Regressed or Explained Variable).** The variable whose value is influenced or is to be predicted is called dependent variable.

Regression test generates lines of regression of the two variables which helps in estimating the values. Lines of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x. Similarly, line of regression of x on y is the line which gives the best estimate for the value of x for any specified value of y.

Q6. Limitations of Regression Analysis

Ans :

Some of the limitations of regression analysis are as follows :

1. Regression analysis assumes that linear relationship exists among the related variables. But in the 'area of social sciences, linear relationship may not exist among the related variables.
2. When regression analysis is used to evaluate the value of dependent variable based on independent variable, it is assumed that the static conditions of relationship exist between 'them. These statistic conditions do not exist in social sciences, so, this assumption "minimizes the use' of regression analysis in social science.
3. The value of dependent variable can be evaluated based on independent variable by using regression analysis but only upto some limits. If the circumstances go beyond the limits, then resists would be inaccurate.

Q7. Least Squares Method

Ans :

The least squares method is a form of mathematical regression analysis that finds the line of best fit for a dataset, providing a visual demonstration of the relationship between the data points. Each point of data is representative of the relationship between a known independent variable and an unknown dependent variable.

The least squares method provides the overall rationale for the placement of the line of best fit among the data points being studied. The most common application of the least squares method, referred to as linear or ordinary, aims to create a straight line that minimizes the sum of the squares of the errors generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value and the value anticipated based on the model.

This method of regression analysis begins with a set of data points to be graphed. An analyst using the least squares method will seek a line of best fit that explains the potential relationship between an independent variable and a dependent variable. In regression analysis, dependent variables are designated on the vertical Y axis and independent variables are designated on the horizontal X-axis. These designations will form the equation for the line of best fit, which is determined from the least squares method.

Q8. Building good regression models.

Ans :

A regression model containing exclusively significant independent variables is referred to as good regression model. In this model, the significance of variable cannot be predicted by adding or deleting independent variables from a model. As the significance, variables varies from one model to another, it is necessary to employ more systematic approach.

In regression model, whenever an independent variable is added the value of R^2 becomes greater than or equal to the value of R^2 in the original model. The value of R^2 do not serve better in building good regression models. The best approach followed for building good regression models is using the parameter 'Adjusted R-square'. An adjusted R^2 not only generates impact on the number of independent variables but, also on the sample size. The value of adjusted R^2 increases or decreases depending on the addition or deletion of independent variables from a model. Moreover, an increase in adjusted R^2 specifies improvement in the model.

The process of building good regression models involves two approaches. They are,

(i) Using P-values

(ii) Using t-statistics.

The steps to be followed for building good regression models using P-values are given below,

1. Develop regression model using independent variables.
2. Determine the significance of independent variables by analysing the P-values.
3. Determine the independent variable containing P-value greater than the specified level of significance.
4. Delete the independent variable identified in step (3) and calculate the value of adjusted R^2.
5. Repeat the above process till the model contains only significant variables.

Q9. Linear Discriminant Analysis

Ans :

Linear discriminant analysis is typically used to identify the characteristics which can accurately discriminate between the respondents who fall in one category from those who fall in another category.

For example, LDA can be used to study successful salesmen and unsuccessful salesmen in order to determine the characteristics which are possessed by successful salesman but not possessed by unsuccessful salesman.

Once the characteristics of successful salesman have been identified, the information can be used to recruit individuals with characteristics similar to those possessed by successful salesmen.

LDA can also be used to study owners and non-owners of videotape recorders, or to study beer drinkers who prefer different brands of beer. In each of these situations, a researcher can use LDA as an attempt to determine the characteristics which are possessed by one category of respondent but not possessed by the other categories of respondents. Such information can be useful to a manufacturer of videotape recorders or to a brewer of a certain brand of beer.

By knowing how the respondents in their target market are different from the respondents not in their target market, the companies involved will have a better definition of their target market and this knowledge can help them greatly to design more effective marketing programs.

Q10. Assumptions of ANOVA

Ans :

(i) All populations involved follow a normal distribution.

(ii) All populations have the same variance (or standard deviation).

(iii) The samples are randomly selected and independent of one another.

Since ANOVA assumes the populations involved follow a normal distribution, ANOVA falls into a category of hypothesis tests known as parametric tests. If the populations involved did not follow a normal distribution, an ANOVA test could not be used to examine the equality of the sample means. Instead, one would have to use a non-parametric test (or distribution-free test), which is a more general form of hypothesis testing that does not rely on distributional assumptions.

UNIT IV

Data Mining: Scope of Data Mining, Data Exploration and Reduction, Unsupervised learning – cluster analysis, Association rules, Supervised learning-Partition Data, Classification Accuracy, prediction Accuracy, k-nearest neighbors, Classification and regression trees, Logistics Regression.

4.1 Scope of Data Mining

Q1. What is data mining and explain the steps involved in this process ?

Ans :

Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue.

Data mining is also known as data discovery and knowledge discovery.

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection and warehousing as well as computer processing.

Example

Grocery stores are well-known users of data mining techniques. Many supermarkets offer free loyalty cards to customers that give them access to reduced prices not available to non-members. The cards make it easy for stores to track who is buying what, when they are buying it and at what price. The stores can then use this data, after analyzing it, for multiple purposes, such as offering customers coupons targeted to their buying habits and deciding when to put items on sale or when to sell them at full price. Data mining can be a cause for concern when only selected information, which is not representative of the overall sample group, is used to prove a certain hypothesis.

Steps Involved in this Process

1. **Problem Definition**

   A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition.

   In the problem definition phase, data mining tools are not yet required.

2. **Data Exploration**

   Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital.

   In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

3. **Data Preparation**

   Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value.

In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

4. **Modeling**

Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model.

In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.

The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

Evaluation

Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

- Does the model achieve the business objective?
- Have all business issues been considered?

At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

5. **Deployment**

Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

The Intelligent Miner products assist you to follow this process. You can apply the functions of the Intelligent Miner products independently, iteratively, or in combination.

The following figure shows the phases of the Cross Industry Standard Process for data mining (CRISP DM) process model.

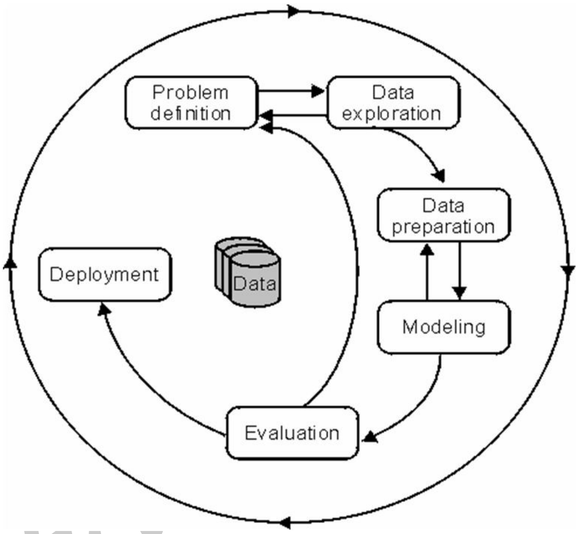


Fig.1: The CRISP DM process model

IM Modeling helps you to select the input data, explore the data, transform the data, and mine the data. With IM Visualization you can display the data mining results to analyze and interpret them. With IM Scoring, you can apply the model that you have created with IM Modeling.

Q2. Explain the scope and techniques of data mining?

Ans :

Scope of Data Mining

1. Data mining process the work in such a manner that it allows business to more proactive to grow substantially.
2. It optimizes large database within the short time and works business intelligence which is more vital to organizational growth.
3. It represents the data in some logical order or may be in pattern to identify the sequential way of processing of data.
4. It includes tree-shaped structure to understand the hierarchy of data and representation of the set of information described in the database.

5. Brings a genetic way of classification of different sets of data items to view the data in the quick glance.

The most commonly used techniques in data mining are:

- **Artificial neural networks:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms:** Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbour method:** A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ^{3} 1). Sometimes called the k-nearest neighbour technique.
- **Rule induction:** The extraction of useful if-then rules from data based on statistical significance.

4.2 Data Exploration and Reduction

Q3. What is data exploration ? Explain the Steps of Data Exploration and Preparation

Ans :

Data exploration is an informative search used by data consumers to form true analysis from the information gathered. Often, data is gathered in a non-rigid or controlled manner in large bulks. For true analysis, this unorganized bulk of data needs to be narrowed down. This is where data exploration is used to analyze the data and information from the data to form further analysis.

Data often converges in a central warehouse called a data warehouse. This data can come from various sources using various formats. Relevant data is needed for tasks such as statistical reporting, trend spotting and pattern spotting. Data exploration is the process of gathering such relevant data.

Steps of Data Exploration and Preparation

Remember the quality of your inputs decide the quality of your output. So, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here. With my personal estimate, data exploration, cleaning and preparation can take up to 70% of your total project time.

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

Finally, we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

1. Variable Identification

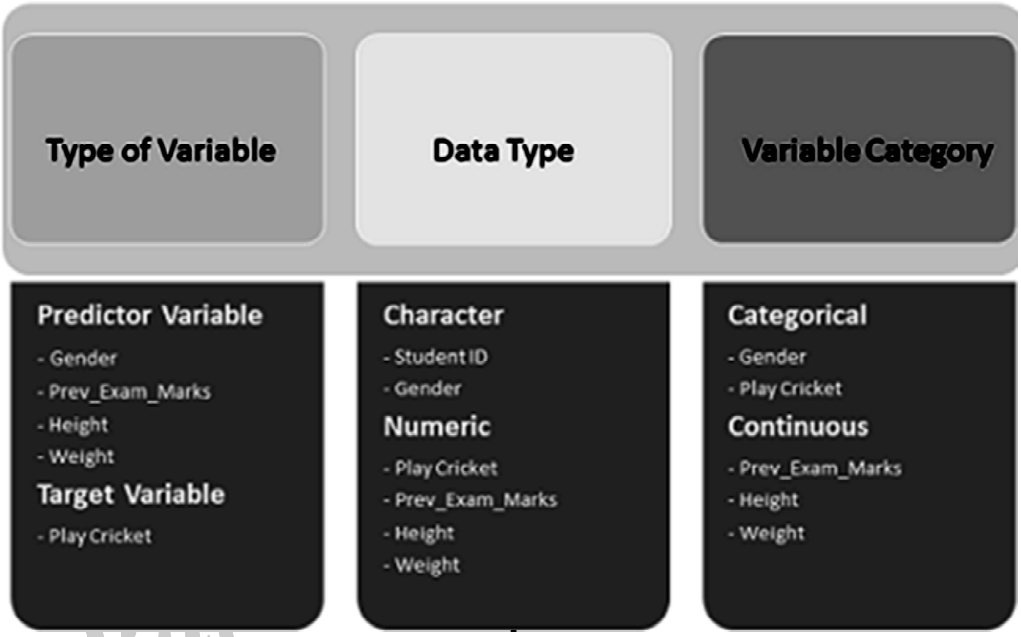
First, identify **Predictor** (Input) and **Target** (output) variables. Next, identify the data type and category of the variables.

Let's understand this step more clearly by taking an example.

Example

Suppose, we want to predict, whether the students will play cricket or not (refer below data set). Here you need to identify predictor variables, target variable, data type of variables and category of variables. Below, the variables have been defined in different category:

Student ID	Gender	Prev Exam Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



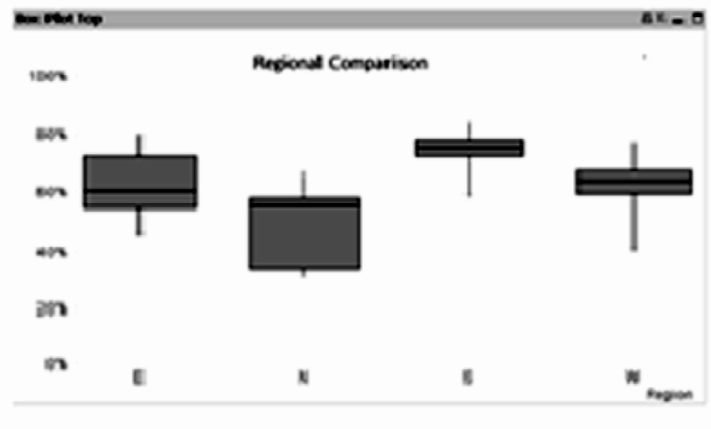
Type of Variable	Data Type	Variable Category
Predictor Variable - Gender - Prev_Exam_Marks - Height - Weight **Target Variable** - Play Cricket	**Character** - Student ID - Gender **Numeric** - Play Cricket - Prev_Exam_Marks - Height - Weight	**Categorical** - Gender - Play Cricket **Continuous** - Prev_Exam_Marks - Height - Weight

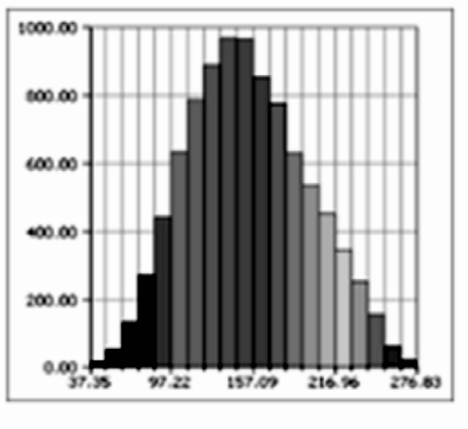
2. Univariate Analysis

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. Let's look at these methods and statistical measures for categorical and continuous variables individually:

(i) **Continuous Variables:** In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods as shown below:

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	





Note: Univariate analysis is also used to highlight missing and outlier values. In the upcoming part of this series, we will look at methods to handle missing and outlier values. To know more about these methods, you can refer course descriptive statistics from Udacity.

(ii) Categorical Variables: For categorical variables, we'll use frequency table to understand distribution of each category. We can also read as percentage of values under each category. It can be be measured using two metrics, Count and Count% against each category. Bar chart can be used as visualization.

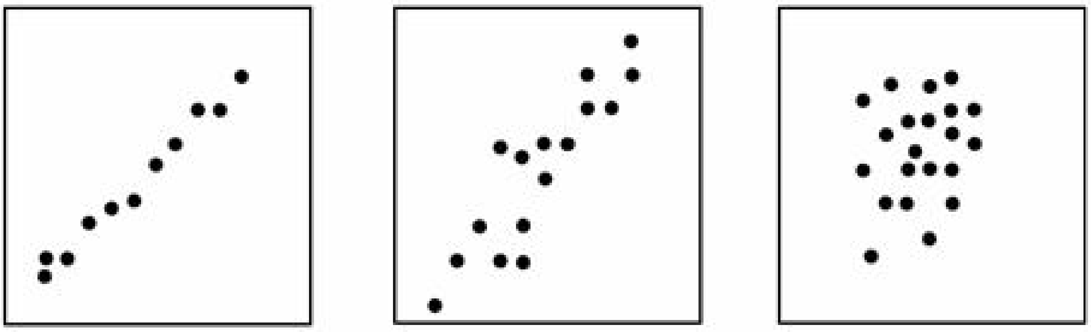
Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Let's understand the possible combinations in detail:

Continuous and Continuous

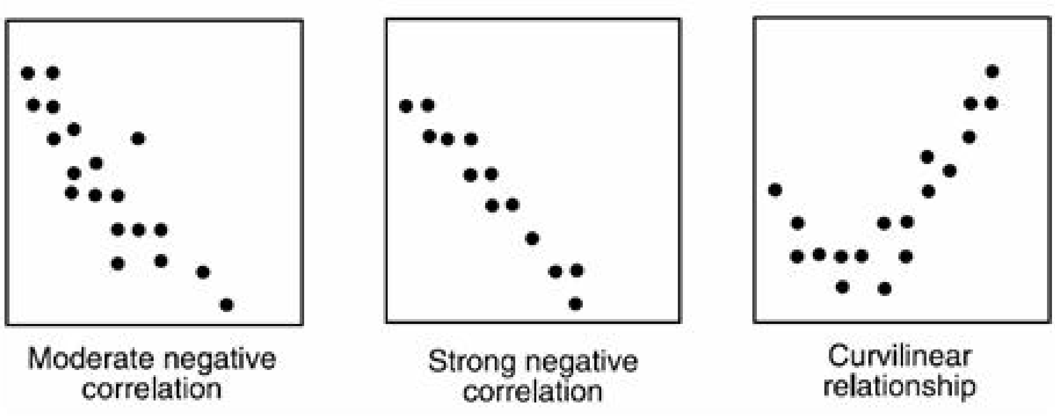
While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.



Strong positive correlation

Moderate positive correlation

No correlation



Scatter plot shows the relationship between two variable but does not indicates the strength of relationship amongst them. To find the strength of the relationship, we use Correlation. Correlation varies between -1 and +1.

- –1: perfect negative linear correlation
- +1:perfect positive linear correlation and
- 0: No correlation

Correlation can be derived using following formula:

Correlation = Covariance(X,Y) / SQRT(Var(X)* Var(Y))

Various tools have function or functionality to identify correlation between variables. In Excel, function CORREL() is used to return the correlation between two variables and SAS uses procedure PROC CORR to identify the correlation. These function returns Pearson Correlation value to identify the relationship between two variables:

X	65	72	78	65	72	70	65	68
Y	72	69	79	69	84	75	60	73

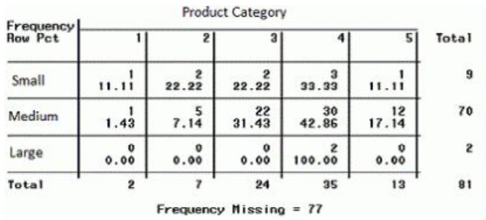
Metrics	**Formula**	**Value**
Co-Variance (X,Y)	= COVAR(E6:L6,E7:L7)	18.77
Variance (X)	= VAR.P(E6:L6)	18.48
Variance (Y)	= VAR.P(E7:L7)	45.23
Correlation	= G10/SQRT(G11*G12)	0.65

In above example, we have good positive relationship(0.65) between two variables X and Y.

Categorical and Categorical

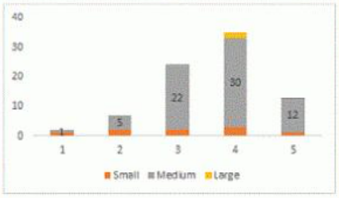
To find the relationship between two categorical variables, we can use following methods:

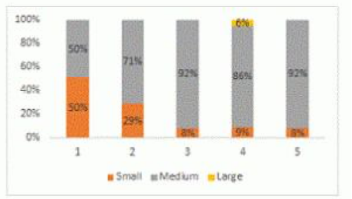
- **Two-way table:** We can start analyzing the relationship by creating a two-way table of count and count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.
- **Stacked Column Chart:** This method is more of a visual form of Two-way table.



Frequency Row Pct	Product Category 1	2	3	4	5	Total
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77





Chi-Square Test

This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

Probability of 0: It indicates that both categorical variable are dependent

Probability of 1: It shows that both variables are independent.

Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence. The chi-square test statistic for a test of independence of two categorical variables is found by:

$X^2 = \Sigma(O - E)^2 / E$ whereO represents the observed frequency. E is the expected frequency under the null hypothesis and computed by:

$$E = \frac{\text{row total} \times \text{column total}}{\text{Sample size}}$$

From previous two-way table, the expected count for product category 1 to be of small size is 0.22. It is derived by taking the row total for Size (9) times the column total for Product category (2) then dividing by the sample size (81). This is procedure is conducted for each cell. Statistical Measures used to analyze the power of relationship are:

- Cramer's V for Nominal Categorical Variable
- Mantel-Haenszed Chi-Square for ordinal categorical variable.

Different data science language and tools have specific methods to perform chi-square test. In SAS, we can use Chisq as an option with Procfreq to perform this test.

Categorical and Continuous

While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance we can perform Z-test, T-test or ANOVA.

Z-Test/ T-Test

Either test assess whether mean of two groups are statistically different from each other or not. If the probability of Z is small then the difference of two averages is more significant. The T-test is very similar to Z-test but it is used when number of observation for both categories is less than 30.

$$Z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$S^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_3^2}{N_1 + N_2 - 2}$$

Where

- $\bar{X}_1, \bar{X}_{2\text{: Averages}}$
- $S_1^2, S^2_{2\text{:Variances}}$
- $N_1, N_{2\text{: Counts}}$
- t : has t distribution with $N_1 + N_2 - 2$ degree of freedom

ANOVA

It assesses whether the average of more than two groups is statistically different.

Example

Suppose, we want to test the effect of five different exercises. For this, we recruit 20 men and assign one type of exercise to 4 men (5 groups). Their weights are recorded after a few weeks. We need to find out whether the effect of these exercises on them is significantly different or not. This can be done by comparing the weights of the 5 groups of 4 men each.

Till here, we have understood the first three stages of Data Exploration, Variable Identification, Uni-Variate and Bi-Variate analysis. We also looked at various statistical and visual methods to identify the relationship between variables.

Now, we will look at the methods of Missing values Treatment. More importantly, we will also look at why missing values occur in our data and why treating them is necessary.

Q4. What is data reduction and explain data reduction techniques?

Ans :

Data reduction is the process of minimizing the amount of data that needs to be stored in a data storage environment. Data reduction can increase storage efficiency and reduce costs.

Bottom of Form

Data reduction can be achieved using several different types of technologies. The best-known data reduction technique is data deduplication, which eliminates redundant data on storage systems. The deduplication process typically occurs at the storage block level. The system analyzes the storage to see if duplicate blocks exist, and gets rid of any redundant blocks. The remaining block is shared by any file that requires a copy of the block. If an application attempts to modify this block, the block is copied prior to modification so that other files that depend on the block can continue to use the unmodified version, thereby avoiding file corruption.

Some storage arrays track which blocks are the most heavily shared. Those blocks that are shared by the largest number of files may be moved to a memory- or flash storage-based cache so they can be read as efficiently as possible.

Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts.

When information is derived from instrument readings there may also be a transformation from analog to digital form. When the data are already in digital form the 'reduction' of the data typically involves some editing, scaling, coding, sorting, collating, and producing tabular summaries. When the observations are discrete but the underlying phenomenon is continuous then smoothing and interpolation are often needed. Often the data reduction is undertaken in the presence of reading or measurement errors. Some idea of the nature of these errors is needed before the most likely value may be determined.

- Data Reduction techniques are usually categorized into three main families:

i) Dimensionality Reduction

Dimensionality Reduction ensures the reduction of thenumber of attributesorrandom variables in the data set. Dimensionality reduction is a series of techniques in machine learning and statistics to reduce the number of random variables to consider. It involves feature selection and feature extraction. Dimensionality reduction makes analyzing data much easier and faster for machine learning algorithms without extraneous variables to process, making machine learning algorithms faster and simpler in turn.

ii) Sample Numerosity Reduction

Replacesthe original data by an alternative smaller data representation This is a technique of choosing smaller forms or data representation to reduce the volume of data.

These techniques may be parametric or nonparametric.

a) Parametric

For parametric methods, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. (Outliers may also be stored.)

Example: Log-linear models, which estimate discrete multidimensional probability distributions.

b) Nonparametric

Nonparametric methods are used for storing reduced representations of the data include histograms, clustering, and sampling.

Regression and Log-Linear Models

- Regression and log-linear models can be used to approximate the given data.
- In (simple) linear regression, the data are modeled to fit a straight line.
- Multiple linear regression is an extension of (simple) linear regression, which allows a response variable y to be modeled as a linear function of two or more predictor variables.
- Log-linear models approximate discrete multidimensional probability distributions.
- Log-linear models can be used to estimate the probability of each point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations.
- This allows a higher-dimensional data space to be constructed from lower dimensional spaces.
- Log-linear models are therefore also useful for dimensionality reduction and data smoothing
- Regression and log-linear models can both be used on sparse data, although their application may be limited.
- While both methods can handle skewed data, regression does exceptionally well. Regression can be computationally intensive when applied to high dimensional data, whereas log-linear models show good scalability for up to 10 or so dimensions.

iii) CardinalityReduction

Transformations applied to obtain a reduced representation of the original data.

The term cardinality refers to the uniqueness of data values contained in a particular column (attribute) of a databasetable. The lower the cardinality, the more duplicated elements in a column. Thus, a column with the lowest possible cardinality would have the same value for every row. SQL databases use cardinality to help determine the optimal query plan for a given query.

4.3 Unsupervised Learning - Clustral Analysis

Q5. What is unsupervised learning and explain about most common cluster learning method?

Ans :

Unsupervised learning is the training of an artificial intelligence (AI) algorithm using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

In unsupervised learning, an AI system may group unsorted information according to similarities and differences even though there are no categories provided. AI systems capable of unsupervised learning are often associated with generative learning models, although they may also use a retrieval-based approach (which is most often associated with supervised learning). Chatbots, self-driving cars, facial recognition programs, expert systems and robots are among the systems that may use either supervised or unsupervised learning approaches.

In unsupervised learning, an AI system is presented with unlabeled, uncategorised data and the system's algorithms act on the data without prior training. The output is dependent upon the coded algorithms. Subjecting a system to unsupervised learning is one way of testing AI.

Unsupervised learning algorithms can perform more complex processing tasks than supervised learning systems. However, unsupervised learning can be more unpredictable than the alternate model. While an unsupervised learning AI system might, for example, figure out on its own how to sort cats from dogs, it might also add unforeseen and undesired categories to deal with unusual breeds, creating clutter instead of order.

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.

The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The clusters are modeled using a measure of similarity which is defined upon metrics such as Euclidean or probabilistic distance.

Common clustering algorithms include:

- **Hierarchical clustering**: builds a multilevel hierarchy of clusters by creating a cluster tree
- **k-Means clustering**: partitions data into k distinct clusters based on distance to the centroid of a cluster
- **Gaussian mixture models**: models clusters as a mixture of multivariate normal density components
- **Self-organizing maps**: uses neural networks that learn the topology and distribution of the data
- **Hidden Markov models**: uses observed data to recover the sequence of states.

Unsupervised learning methods are used in bioinformatics for sequence analysis and genetic clustering; in data mining for sequence and pattern mining; in medical imaging for image segmentation; and in computer vision for object recognition.

Q6. Explain about cluster analysis with its methods?

Ans :

Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

Cluster Analysis has been used in marketing for various purposes. Segmentation of consumers in cluster analysis is used on the basis of benefits sought from the purchase of the product. It can be used to identify homogeneous groups of buyers.

Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

The variables on which the cluster analysis is to be done should be selected by keeping past research in mind. It should also be selected by theory, the hypotheses being tested, and the judgment of the researcher. An appropriate measure of distance or similarity should be selected; the most commonly used measure is the Euclidean distance or its square.

1. **A hierarchical procedure in cluster analysis:** It is characterized by the development of a tree like structure. A hierarchical procedure can be agglomerative or divisive. Agglomerative methods in cluster analysis consist of linkage methods, variance methods, and centroid methods. Linkage methods in cluster analysis are comprised of single linkage, complete linkage, and average linkage.

2. **The non-hierarchical methods in cluster analysis** : It is frequently referred to as K means clustering. The two-step procedure

can automatically determine the optimal number of clusters by comparing the values of model choice criteria across different clustering solutions. The choice of clustering procedure and the choice of distance measure are interrelated. The relative sizes of clusters in cluster analysis should be meaningful. The clusters should be interpreted in terms of cluster centroids.

There are certain concepts and statistics associated with cluster analysis:

- Agglomeration schedule in cluster analysis gives information on the objects or cases being combined at each stage of the hierarchical clustering process.
- Cluster Centroid is the mean value of a variable for all the cases or objects in a particular cluster.
- A dendrogram is a graphical device for displaying cluster results.
- Distances between cluster centers in cluster analysis indicate how separated the individual pairs of clusters are. The clusters that are widely separated are distinct and therefore desirable.
- Similarity/distance coefficient matrix in cluster analysis is a lower triangle matrix containing pairwise distances between objects or cases.

4.3.1 Association Rules

Q7. Explain the concept Association rules in data mining?

Ans :

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk."

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

Association rule mining is the data mining process of finding the rules that may govern associations and causal objects between sets of items.

So in a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together. For example, peanut butter and jelly are often bought together because a lot of people like to make PB&J sandwiches.

Also surprisingly, diapers and beer are bought together because, as it turns out, that dads are often tasked to do the shopping while the moms are left with the baby.

The main applications of association rule mining:

- **Basket data analysis:** Is to analyze the association of purchased items in a single basket or single purchase as per the examples given above.
- **Cross marketing:** Is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and manufacturers have cross marketing campaigns with oil and gas companies for obvious reasons.
- **Catalog design:** The selection of items in a business' catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related.

Q8. Explain developing assocition rules using X-Lminer?

Ans :

Developing Association Rules Using XLMiner Step 1

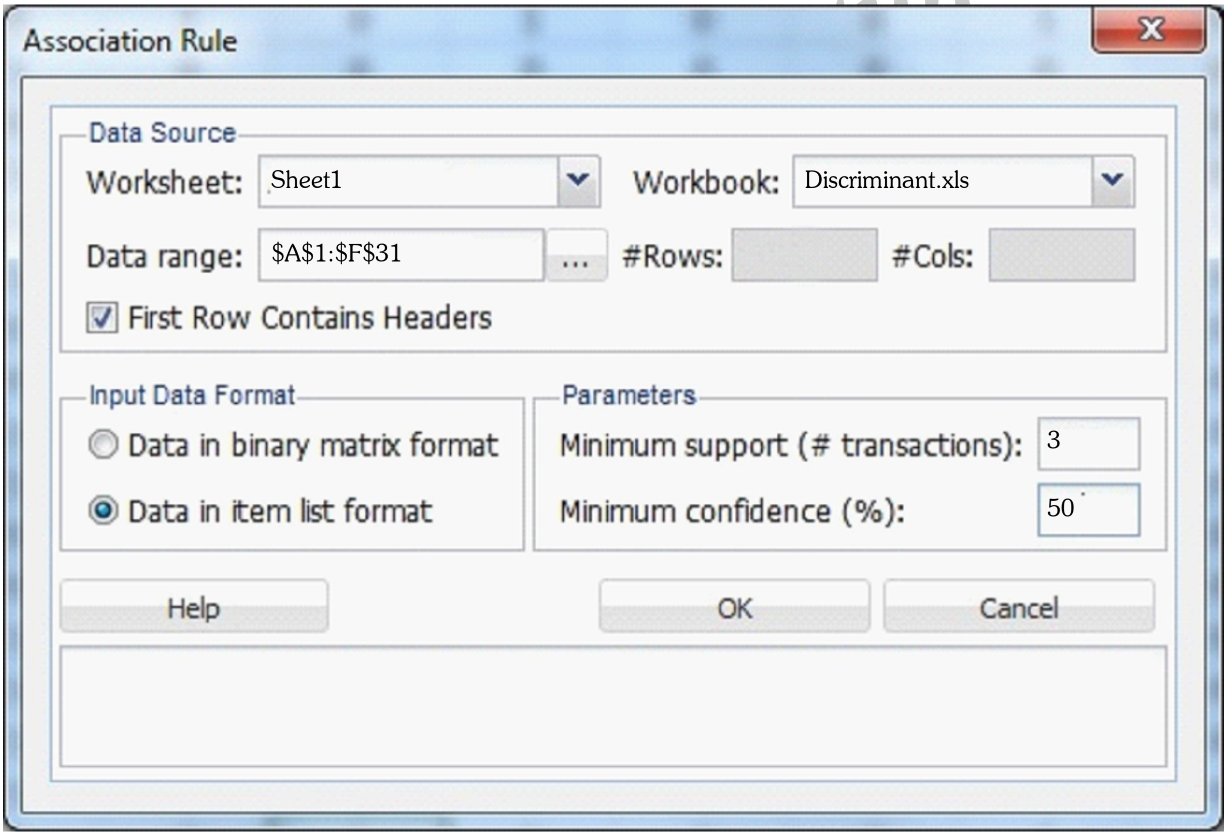
Select a cell from the data.

Step 2

Click on XLMiner, select DataMining followed by Association and Association Rules.

- ➢ In 'Data Source' area of Association Rules Dialog box.
- ➢ Enter the values into the fields Worksheet, Workbook and Data range.
- ➢ Select the checkbox of First Row Contains Headers
- ➢ In 'Input Data Format' area select Radio button of Data in item list format.
- ➢ In 'Parameters' area enter the values into Minimum support and Minimum confidence.
- ➢ Click on OK.

The worksheet shows the results in selected format along with the associated rules.



Association Rule

Data Source

Worksheet: Sheet1

Workbook: Discriminant.xls

Data range: $A$1:$F$31

#Rows:

#Cols:

☑ First Row Contains Headers

Input Data Format

○ Data in binary matrix format

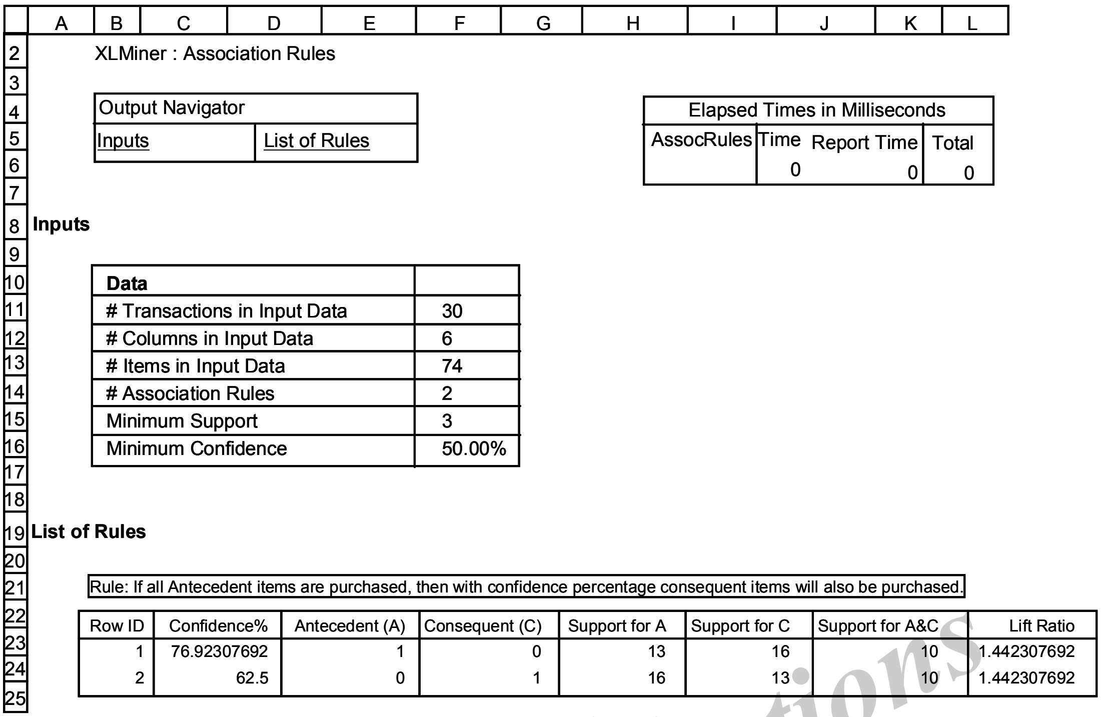
● Data in item list format

Parameters

Minimum support (# transactions): 3

Minimum confidence (%): 50

Figure: Association Rules Dialog



XLMiner : Association Rules

Output Navigator	
Inputs	List of Rules

Elapsed Times in Milliseconds		
AssocRules Time	Report Time	Total
0	0	0

Inputs

Data	
# Transactions in Input Data	30
# Columns in Input Data	6
# Items in Input Data	74
# Association Rules	2
Minimum Support	3
Minimum Confidence	50.00%

List of Rules

Rule: If all Antecedent items are purchased, then with confidence percentage consequent items will also be purchased.

Row ID	Confidence%	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A&C	Lift Ratio
1	76.92307692	1	0	13	16	10	1.442307692
2	62.5	0	1	16	13	10	1.442307692

Figure: Assocaiton Rules for PC Purchse Data

4.4 Supervised Learning

Q9. What is supervised learning explain the steps involves to solve a given problem of supervised learning?

Ans :

Supervised learning, in the context of artificial intelligence (AI) and machine learning, is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing.

Supervised machine learning systems provide the learning algorithms with known quantities to support future judgments. Chatbots, self-driving cars, facial recognition programs, expert systems and robots are among the systems that may use either supervised or unsupervised learning. Supervised learning systems are mostly associated with retrieval-based AI but they may also be capable of using a generative learning model.

Training data for supervised learning includes a set of examples with paired input subjects and desired output (which is also referred to as the supervisory signal). In supervised learning for image processing, for example, an AI system might be provided with labelled pictures of vehicles in categories such as cars and trucks. After a sufficient amount of observation, the system should be able to distinguish between and categorize unlabeled images, at which time training can be said to be complete.

Supervised learning models have some advantages over the unsupervised approach, but they also have limitations. The systems are more likely to make judgments that humans can relate to, for example, because humans have provided the basis for decisions. However, in the case of a retrieval-based method, supervised learning systems have trouble dealing with new information. If a system with categories for cars and trucks is presented with a bicycle, for example, it would have to be incorrectly lumped in one category or the other. If the AI system was generative, however, it may not know what the bicycle is but would be able to recognize it as belonging to a separate category.

In order to solve a given problem of supervised learning, one has to perform the following steps:

1. **Determine the type of training examples**: Before doing anything else, the user should decide what kind of data is to be used as a training set. In case of handwriting analysis, for example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.

2. **Gather a training set**: The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.

3. **Determine the input feature representation of the learned function**: The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.

4. **Determine the structure of the learned function and corresponding learning algorithm**: For example, the engineer may choose to use support vector machines or decision trees.

5. **Complete the design**: Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a *validation* set) of the training set, or via cross-validation.

6. **Evaluate the accuracy of the learned function**: After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

4.4.1 Partition Data

Q10. Briefly explain about data partitioning with its Techniques.

Ans :

Introduction to Partitioning

A big problem can be solved easily when it is chopped into several smaller sub-problems. That is what the partitioning technique does. It divides the big database containing data metrics and indexes into smaller and handy slices of data called as partitions. The partitioned tables are directly used by the SQL queries without any alteration. Once the database is partitioned, the data definition language can easily work on the smaller partitioned slices, instead of handling the giant database altogether. This is how partitioning cuts down the problems in managing the large database tables.

The partitioning key consists of a single or supplementary columns with the intention of determining the partition wherever the rows will be stored. Spark modifies the partitions by using these partition keys.

All the smaller partitioned slices share the same logical features, but they do carry different physical features.

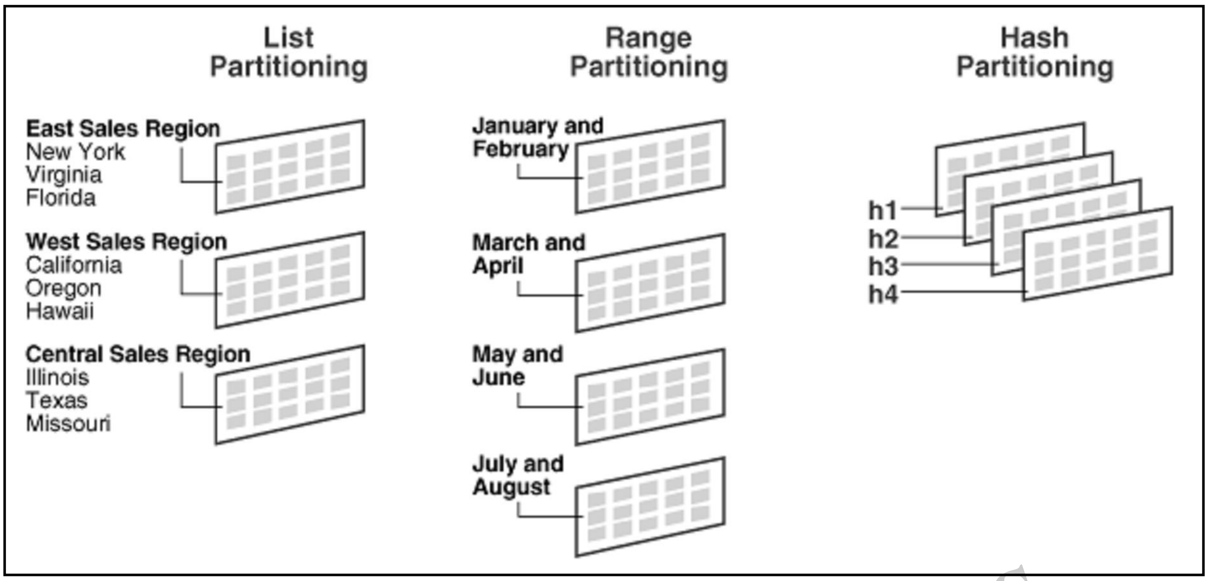
Partitioning Key Extensions

The key extensions assist in signifying the keys used for the partitioning processes. These extensions are

- **Reference Partitioning** : Reference partitioning facilitates the division of two databases associated with one another by referential limitations. By activating the primary as well as the foreign keys, it produces a new partition key from another active relationship.

- **Virtual Column-Based Partitioning :**The partition of a database is possible even when the partition keys are physically unavailable. This is possible by the Virtual Column-Based Partitioning method which creates logical partition keys using the columns of the data table.

Partitioning Techniques



Spark provides three information allocation processes and they are:

- Hash
- Range
- List

Using these information allocation processes, the database tables are partitioned using two methods and they are:

1. Single-Level Partitioning

Any data table is addressed by identifying one of the above data distribution methodologies, using one or more columns as the partitioning key. The techniques are:

- Hash Partitioning
- Range Partitioning
- List Partitioning
- **Hash Partitioning:** Oracle has got a hash algorithm for recognizing the partition tables. This algorithm uniformly divides the rows into various partitions in order to make all the partitions of identical dimensions. The process carried on by using this hash algorithm to divide the database tables into smaller divisions is termed as the hash partitioning.

  Hash partitioning is the perfect means for sharing out data consistently among different devices. This method of partitio-ning is an user-friendly partitioning system, particularly when the information to be detached has no apparent partitioning key.
- **Range Partitioning:** Range partitio-ning divides the information into a number of partitions depending on ranges of values of the particular partitioning keys for every partition of data. It is a popular partitioning scheme which is normally used with dates. For example, representing the days of the May month, it will have a table with the column name as May and rows with dates from 1st of May to 31st of May.

All the partitions smaller than a particular partition comes before the VALUES LESS THAN clause, while all the partitions higher than a particular partition comes after the VALUES LESS THAN clause of the particular partition. For representing the highest range partition, the MAXVALUE clause is used.

List Partitioning

List partitioning allows to openly organize the rows, which are divided into partitions by spelling out a roll of distinct standards for the partitioning key in an account for every division. Using this scheme of partitioning, even dissimilar and shuffled information tables can be managed in a comfortable approach.

In order to avoid the errors during the partition of rows in the giant database, the addition of the probable terms into the table formed by the list partitioning method can be avoided by using the default partition process.

SUBMIT

2. Composite Partitioning

The composite partitioning method includes a minimum of two partitioning procedures on the data. Initially, the database table will be divided by using one partition procedure and then the output partition slices are again partitioned further by using another partitioning procedure.

4.4.2 Classificaiton Accuracy

Q11. What is classification accuracy?

Ans :

Accuracy is one metric for evaluating classifi-cation models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Accuracy = Number of correct predictions Total number of predictions

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

Accuracy = TP + TNTP + TN + FP + FN

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Let's try calculating accuracy for the following model that classified 100 tumors as malignant (the positive class) or benign (the negative class):

True Positive (TP)	**False Positive (FP)**
Reality: Malignant	Reality: Benign
ML model predicted: Malignant	ML model predicted: Malignant
Number of TP results: 1	Number of FP results: 1
False Negative (FN):	True Negative (TN):
Reality: Malignant	Reality: Benign
ML model predicted: Benign	ML model predicted: Benign
Number of FN results: 8	Number of TN results: 90

Accuracy = TP + TNTP + TN + FP + FN = 1 + 901 + 90 + 1 + 8 = 0.91

Accuracy comes out to 0.91, or 91% (91 correct predictions out of 100 total examples). That means our tumor classifier is doing a great job of identifying malignancies, right?

Actually, let's do a closer analysis of positives and negatives to gain more insight into our model's performance.

Of the 100 tumor examples, 91 are benign (90 TNs and 1 FP) and 9 are malignant (1 TP and 8 FNs).

Of the 91 benign tumors, the model correctly identifies 90 as benign. That's good. However, of the 9 malignant tumors, the model only correctly identifies 1 as malignant a terrible outcome, as 8 out of 9 malignancies go undiagnosed!

While 91% accuracy may seem good at first glance, another tumor-classifier model that always predicts benign would achieve the exact same accuracy (91/100 correct predictions) on our examples. In other words, our model is no better than one that has zero predictive ability to distinguish malignant tumors from benign tumors.

Accuracy alone doesn't tell the full story when you're working with a class-imbalanced data set, like this one, where there is a significant disparity between the number of positive and negative labels.

4.4.3 Predictive Accuracy

Q12. What is predictive accuracy?

Ans :

The accuracy paradox for predictive analytics states that predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy. It may be better to avoid the accuracy metric in favor of other metrics such as precision and recall.

Accuracy is often the starting point for analyzing the quality of a predictive model, as well as an obvious criterion for prediction. Accuracy measures the ratio of correct predictions to the total number of cases evaluated. It may seem obvious that the ratio of correct predictions to cases should be a key metric. A predictive model may have high accuracy, but be useless.

In an example predictive model for an insurance fraud application, all cases that are predicted as high-risk by the model will be investigated. To evaluate the performance of the model, the insurance company has created a sample data set of 10,000 claims. All 10,000 cases in the validation sample have been carefully checked and it is known which cases are fraudulent. A table of confusion assists analyzing the quality of the model. The definition of accuracy, the table of confusion for model M_1^{Fraud}, and the calculation of accuracy for model M_1^{Fraud} is shown below.

where

A(M) = (TN+TP) / (TN+FP+FN+TP)

TN is the number of true negative cases

FP is the number of false positive cases

FN is the number of false negative cases

TP is the number of true positive cases

Formula 1: Definition of Accuracy

	Predicted Negative	**Predicted Positive**
Negative Cases	9,700	150
Positive Cases	50	100

Table 1: Table of Confusion for Fraud Model M_1^{Fraud}

Formula 2: Accuracy for model M_1^{Fraud}

With an accuracy of 98.0% model M_1^{Fraud} appears to perform fairly well. The paradox lies in the fact that accuracy can be easily improved to 98.5% by always predicting "no fraud". The table of confusion and the accuracy for this trivial "always predict negative" model M_2^{Fraud} and the accuracy of this model are shown below.

	Predicted Negative	**Predicted Positive**
Negative Cases	9,850	0
Positive Cases	150	0

Table 2: Table of Confusion for Fraud Model M_2^{Fraud}.

Formula 3: Accuracy for model M_2^{Fraud}

Model M_2^{Fraud} reduces the rate of inaccurate predictions from 2% to 1.5%. This is an apparent improvement of 25%. The new model M_2^{Fraud} shows fewer incorrect predictions and markedly improved accuracy, as compared to the original model M_1^{Fraud}, but is obviously useless.

The alternative model M_2^{Fraud} does not offer any value to the company for preventing fraud. The less accurate model is more useful than the more accurate model.

Caution is advised when using accuracy in the evaluation of predictive models; it is appropriate only if the cost of a false positive (false alarm) is equal to the cost of a false negative (missed prediction). Otherwise, a more appropriate loss function should be determined.

4.4.4 K-Nearest Neighbors

Q13. What is *k*-nearest neighbors and explain about KNN classifier

Ans :

The k-nearest neighbors (k-NN) is a non - parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the *k*-NN algorithm is that it is sensitive to the local structure of the data.The algorithm is not to be confused with *k*-means, another popular machine learning technique.

The KNN classifier is also a **non parametric** and **instance-based** learning algorithm.

- **Non-parametric** means it makes no explicit assumptions about the functional form of h, avoiding the dangers of mismodeling the underlying distribution of the data. For example, suppose our data is highly non-Gaussian but the learning model we choose assumes a Gaussian form. In that case, our algorithm would make extremely poor predictions.
- **Instance-based** learning means that our algorithm doesn't explicitly learn a model. Instead, it chooses to memorize the training instances which are subsequently used as "knowledge" for the prediction phase. Concretely, this means that only when a query to our database is made (i.e. when we ask it to predict a label given an input), will the algorithm use the training instances to spit out an answer.

4.4.5 Classification and Regresion Trees

Q14. What is Classification and regression trees?

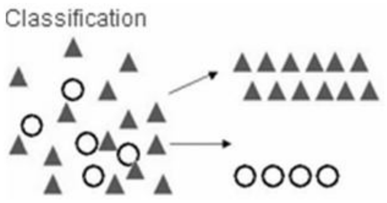
Ans :

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

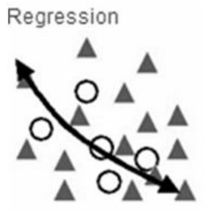
In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making). This page deals with decision trees in data mining.

He CART or Classification & Regression Trees methodology was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone as an umbrella term to refer to the following types of decision trees:

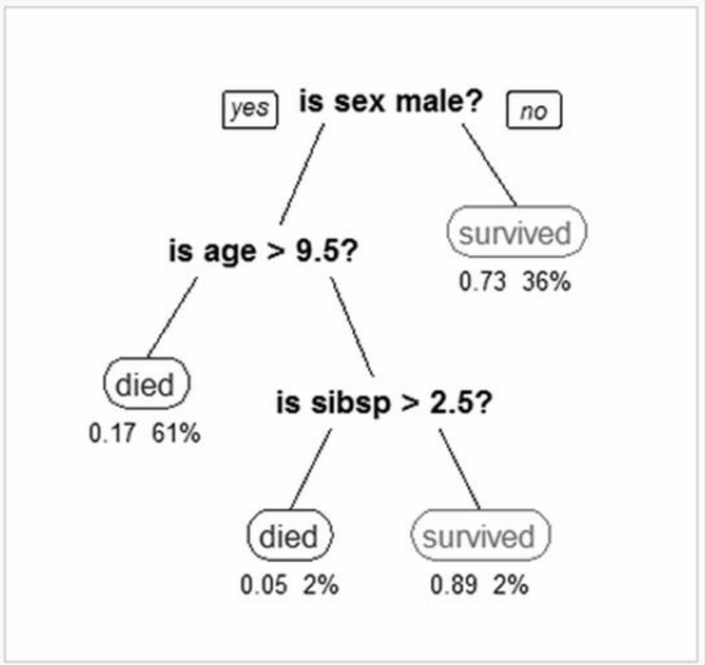
- **Classification Trees**: where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.



- **Regression Trees**: where the target variable is continuous and tree is used to predict it's value.



The CART algorithm is structured as a sequ-ence of questions, the answers to which determine what the next question, if any should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions. A simple example of a decision tree is as follows:



A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

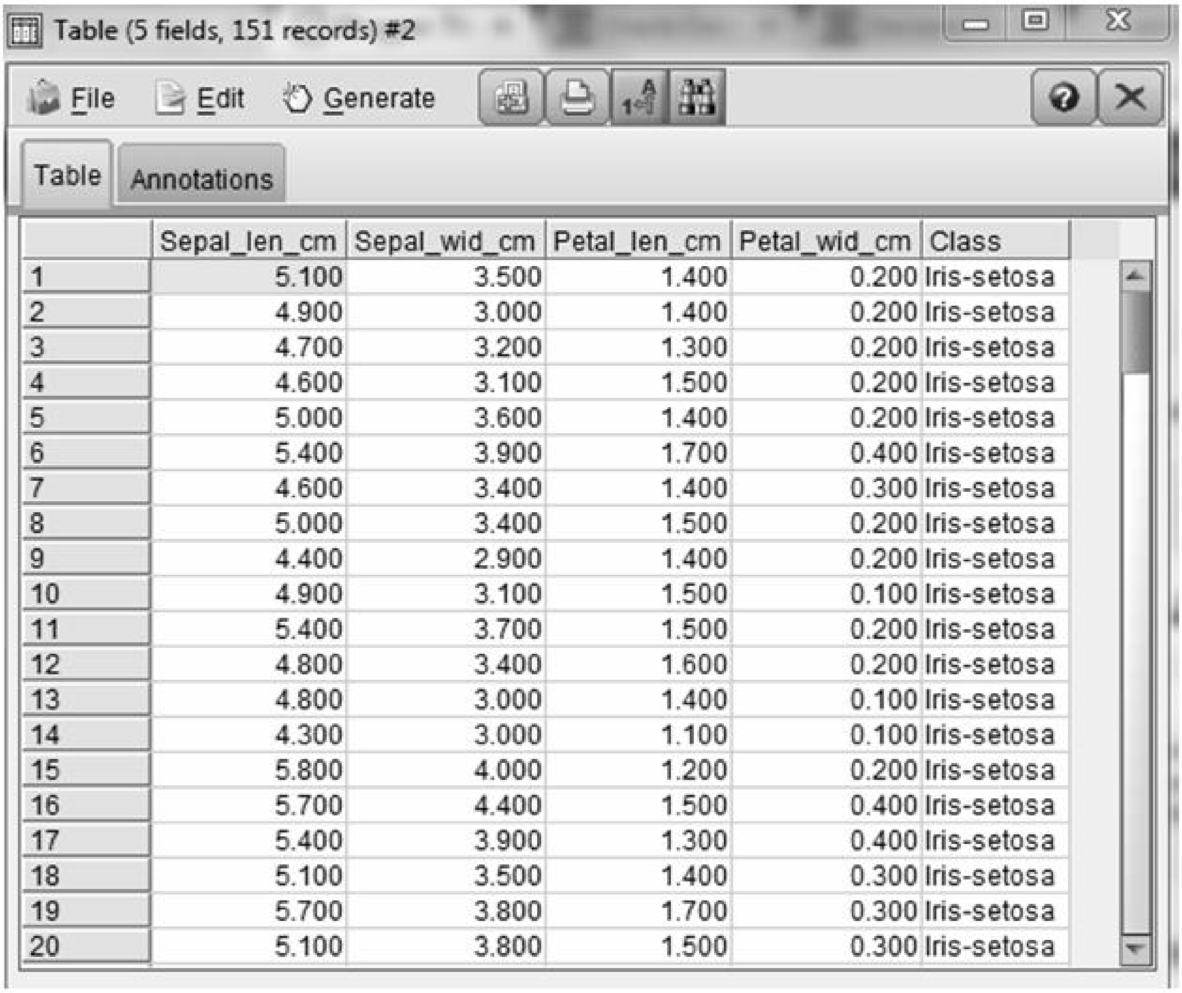
Q15. What are the main elements of CART ?

Ans :

The main elements of CART (and any decision tree algorithm) are:

1. Rules for splitting data at a node based on the value of one variable;
2. Stopping rules for deciding when a branch is terminal and can be split no more; and
3. Finally, a prediction for the target variable in each terminal node.

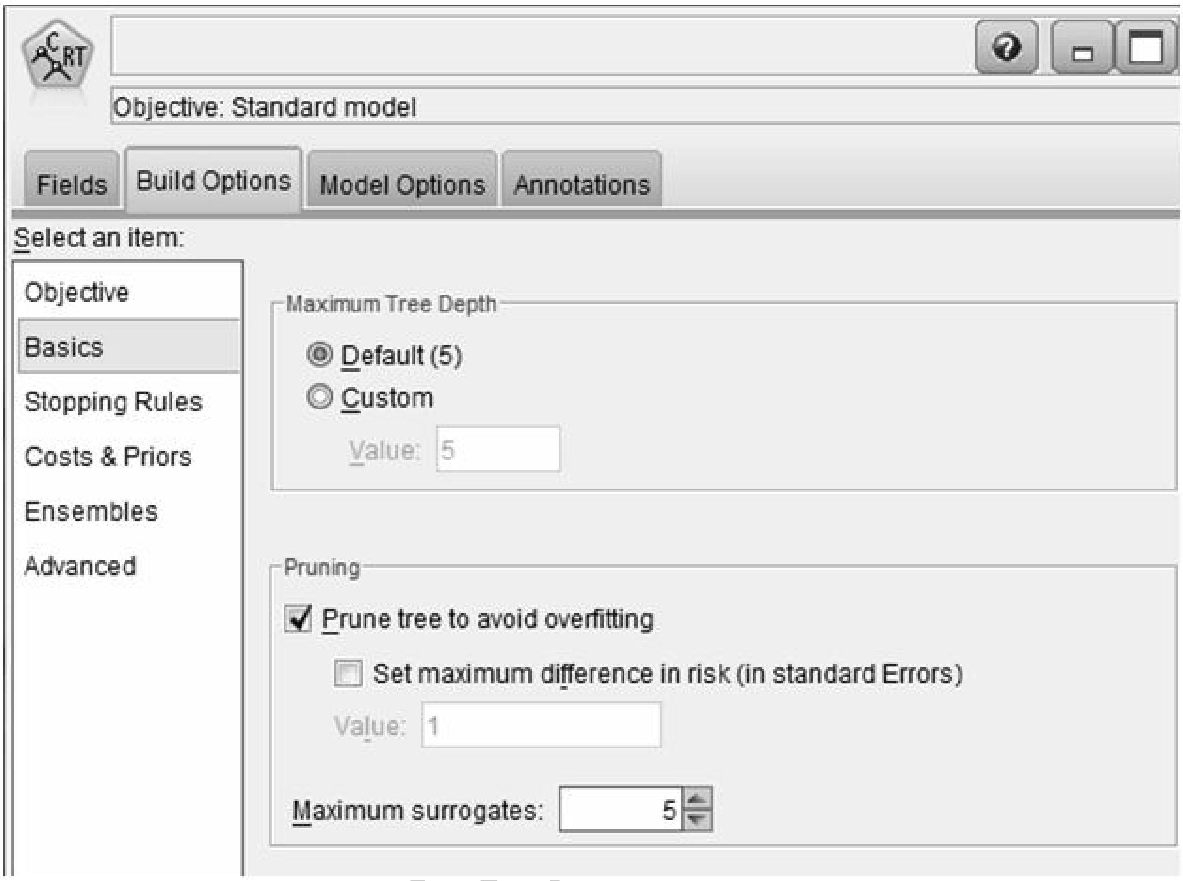
In order to understand this better, let us consider the Iris dataset. The dataset consists of 5 variables and 151 records as shown below:



	Sepal_len_cm	Sepal_wid_cm	Petal_len_cm	Petal_wid_cm	Class
1	5.100	3.500	1.400	0.200	Iris-setosa
2	4.900	3.000	1.400	0.200	Iris-setosa
3	4.700	3.200	1.300	0.200	Iris-setosa
4	4.600	3.100	1.500	0.200	Iris-setosa
5	5.000	3.600	1.400	0.200	Iris-setosa
6	5.400	3.900	1.700	0.400	Iris-setosa
7	4.600	3.400	1.400	0.300	Iris-setosa
8	5.000	3.400	1.500	0.200	Iris-setosa
9	4.400	2.900	1.400	0.200	Iris-setosa
10	4.900	3.100	1.500	0.100	Iris-setosa
11	5.400	3.700	1.500	0.200	Iris-setosa
12	4.800	3.400	1.600	0.200	Iris-setosa
13	4.800	3.000	1.400	0.100	Iris-setosa
14	4.300	3.000	1.100	0.100	Iris-setosa
15	5.800	4.000	1.200	0.200	Iris-setosa
16	5.700	4.400	1.500	0.400	Iris-setosa
17	5.400	3.900	1.300	0.400	Iris-setosa
18	5.100	3.500	1.400	0.300	Iris-setosa
19	5.700	3.800	1.700	0.300	Iris-setosa
20	5.100	3.800	1.500	0.300	Iris-setosa

In this data set, "Class" is the target variable while the other four variables are independent variables. In other words, the "Class" is dependent on the values of the other four variables. We will use IBM SPSS Modeler v15 to build our tree. To do this, we attach the CART node to the data set. Next, we choose our options in building out our tree as follows:

On this screen, we pick the maximum tree depth, which is the most number of "levels" we want in the decision tree. We also choose the option of "Pruning" the tree which is used to avoid over-fitting. More about pruning in a different blog post.



Objective: Standard model

Select an item: Basics

Maximum Tree Depth

● Default (5)

○ Custom

Value: 5

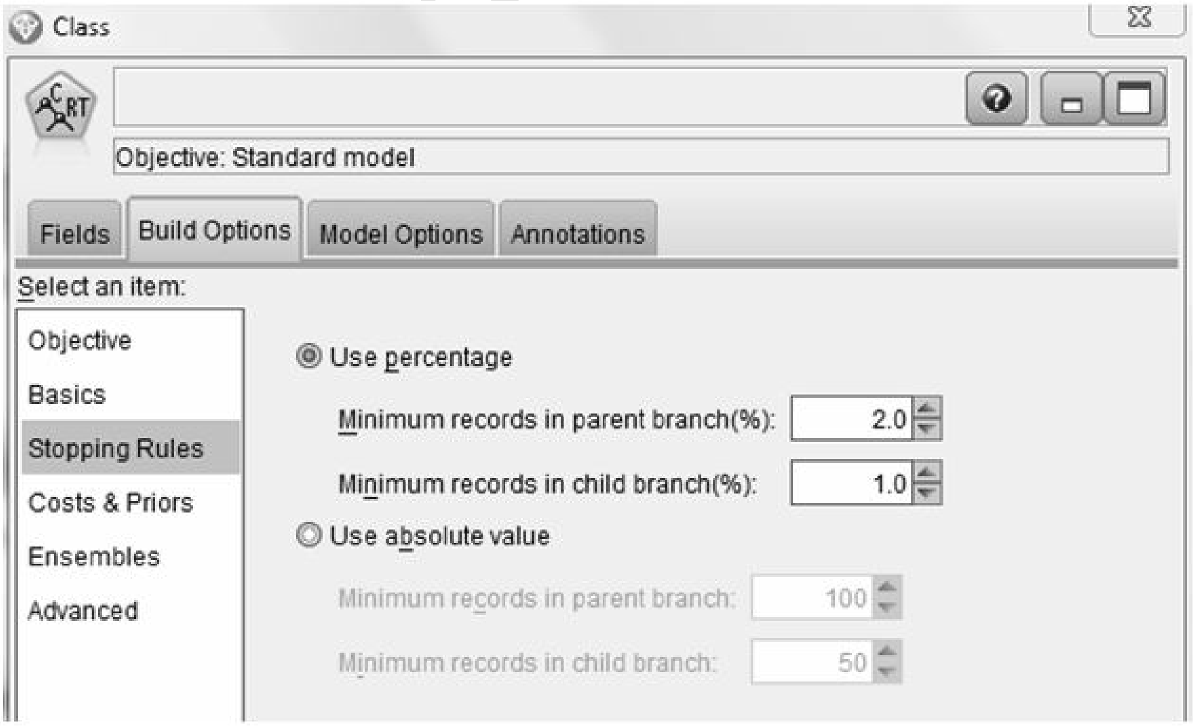
Pruning

☑ Prune tree to avoid overfitting

□ Set maximum difference in risk (in standard Errors)

Value: 1

Maximum surrogates: 5



Objective: Standard model

Select an item: Stopping Rules

● Use percentage

Minimum records in parent branch(%): 2.0

Minimum records in child branch(%): 1.0

○ Use absolute value

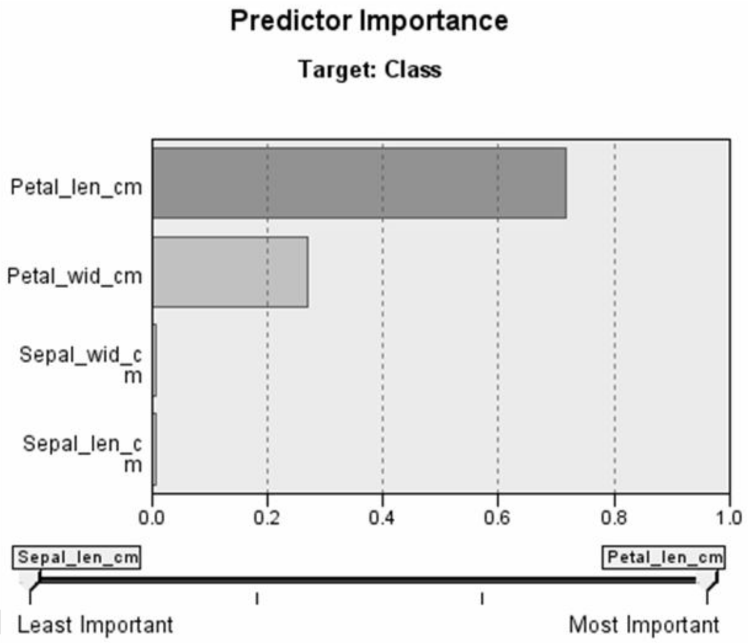
Minimum records in parent branch: 100

Minimum records in child branch: 50

On this screen, we choose stopping rules, which determine when further splitting of a node stops or when further splitting is not possible. In addition to maximum tree depth discussed above, stopping rules typically include reaching a certain minimum number of cases in a node, reaching a maximum number of nodes in the tree, etc. Conditions under which further splitting is impossible include when

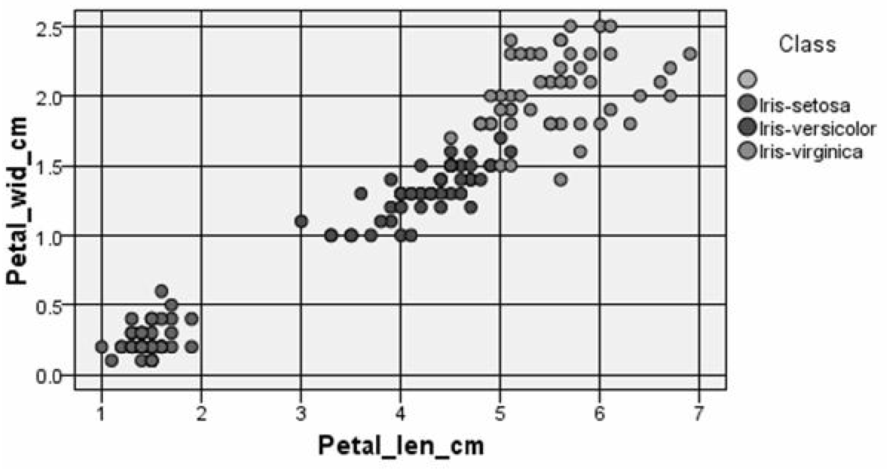
- Only one case is left in a node;
- All other cases are duplicates of each other; and
- The node is pure (all target values agree).

Next we run the CART node and examine the results. We first look at Predictor Importance, which represents the most important variables used in splitting the tree:

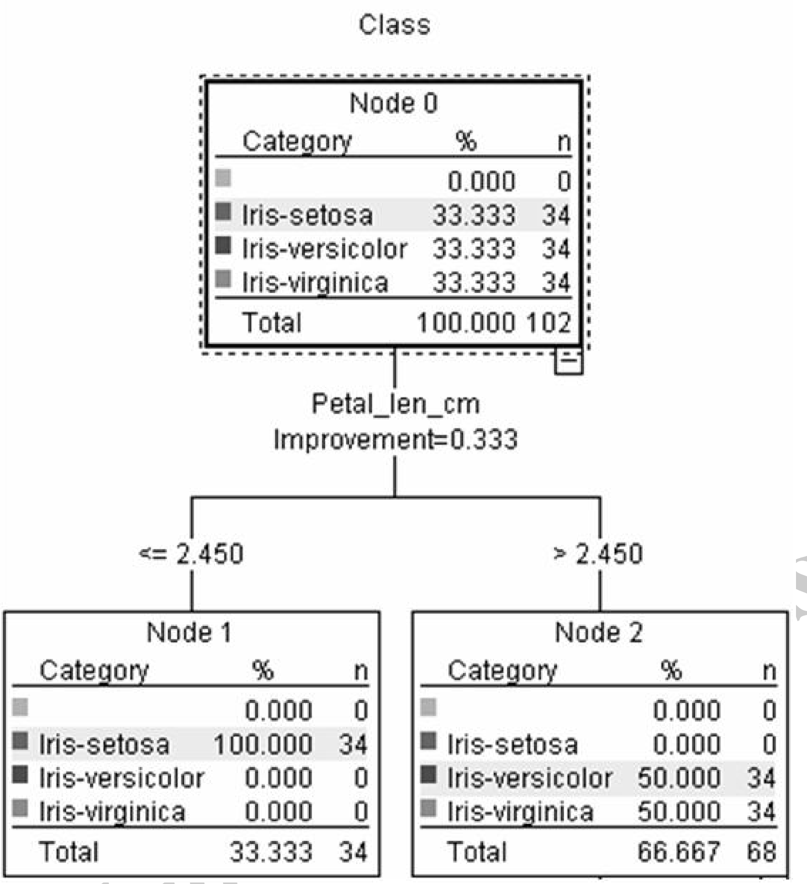


From the chart above, we note that the most important predictor (by a long distance) is the length of the Petal followed by the width of the Petal.

A scatter plot of the data by plotting Petal length by Petal width also reflects the predictor importance:



This should also be reflected in the decision tree generated by the CART. Let us examine this next:



Node 0

Category	%	n
■	0.000	0
■ Iris-setosa	33.333	34
■ Iris-versicolor	33.333	34
■ Iris-virginica	33.333	34
Total	100.000	102

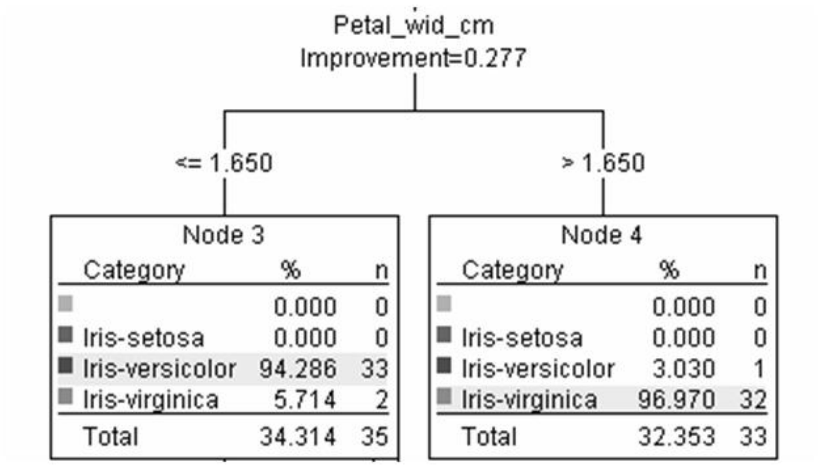
Node 1

Category	%	n
■	0.000	0
■ Iris-setosa	100.000	34
■ Iris-versicolor	0.000	0
■ Iris-virginica	0.000	0
Total	33.333	34

Node 2

Category	%	n
■	0.000	0
■ Iris-setosa	0.000	0
■ Iris-versicolor	50.000	34
■ Iris-virginica	50.000	34
Total	66.667	68

As can be seen, the first node is split based on our most important predictor, the length of the Petal. The question posed is "Is the length of the petal greater than 2.45 cms?". If not, then the class in which the Iris falls is "setosa". If yes, then the class could be either "versicolor" or "virginica". Since we have completely classified "setosa" in Node 1, that becomes a terminal node and no additional questions are posed there. However, we still need to Node 2 still needs to be broken down to separate out "versicolor" and "virginica". Therefore, the next question needs to be posed which is based on our second most important predictor, the width of the Petal.



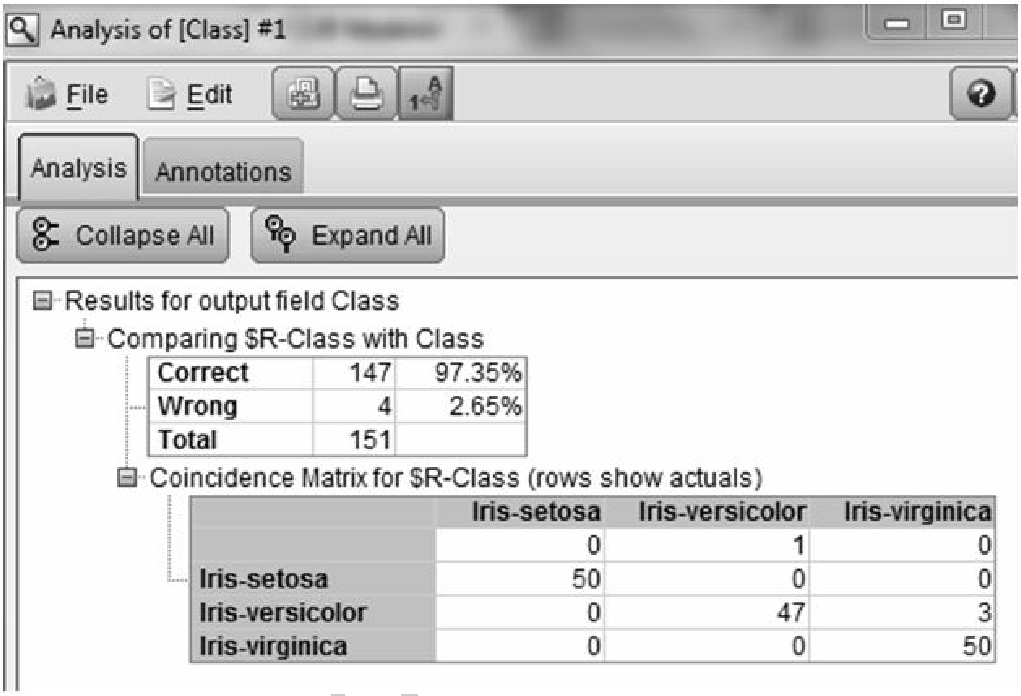
Node 3

Category	%	n
■	0.000	0
■ Iris-setosa	0.000	0
■ Iris-versicolor	94.286	33
■ Iris-virginica	5.714	2
Total	34.314	35

Node 4

Category	%	n
■	0.000	0
■ Iris-setosa	0.000	0
■ Iris-versicolor	3.030	1
■ Iris-virginica	96.970	32
Total	32.353	33

As expected, in this case, the question relates to the width of the Petal. From the nodes, we can see that by asking the second question, the decision tree has almost completely split the data separately into "versicolor" and "virginica". We can continue splitting them further until there is no overlap between classes in each node; however, for the purposes of this post, we will stop our decision tree here. We attach an Analysis node to see the overall accuracy of our predictions:



Results for output field Class

Comparing $R-Class with Class

Correct	147	97.35%
Wrong	4	2.65%
Total	151	

Coincidence Matrix for $R-Class (rows show actuals)

	Iris-setosa	Iris-versicolor	Iris-virginica
	0	1	0
Iris-setosa	50	0	0
Iris-versicolor	0	47	3
Iris-virginica	0	0	50

From the analysis, we can see that the CART algorithm has classified "setosa" and "virginica" accurately in all cases and accurately classified "versicolor" in 47 of the 50 cases giving us an overall accuracy of 97.35%.

Q16. Explain the features and advantages of CART?

Ans :

Some useful features and advantages of CART.

- CART is nonparametric and therefore does not rely on data belonging to a particular type of distribution.
- CART is not significantly impacted by outliers in the input variables.
- You can relax stopping rules to "overgrow" decision trees and then prune back the tree to the optimal size. This approach minimizes the probability that important structure in the data set will be overlooked by stopping too soon.
- CART incorporates both testing with a test data set and cross-validation to assess the goodness of fit more accurately.
- CART can use the same variables more than once in different parts of the tree. This capability can uncover complex interdependencies between sets of variables.
- CART can be used in conjunction with other prediction methods to select the input set of variables.

4.4.6 Logistic Regression

Q17. What is logistic regression and explain how to enter data in it ?

Ans :

Logistic Regressionis the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of indepen-dent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

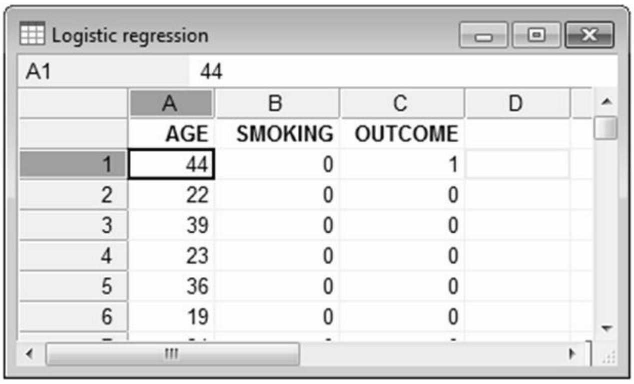
$$\text{odds} = \frac{p}{1-p} = \frac{\text{Pr obability of presence of characteristic}}{\text{probability of absence of characteristic}}$$

and

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

How to enter data

In the following example there are two predictor variables: AGE and SMOKING. The dependent variable, or response variable is OUTCOME. The dependent variable OUTCOME is coded 0 (negative) and 1 (positive).

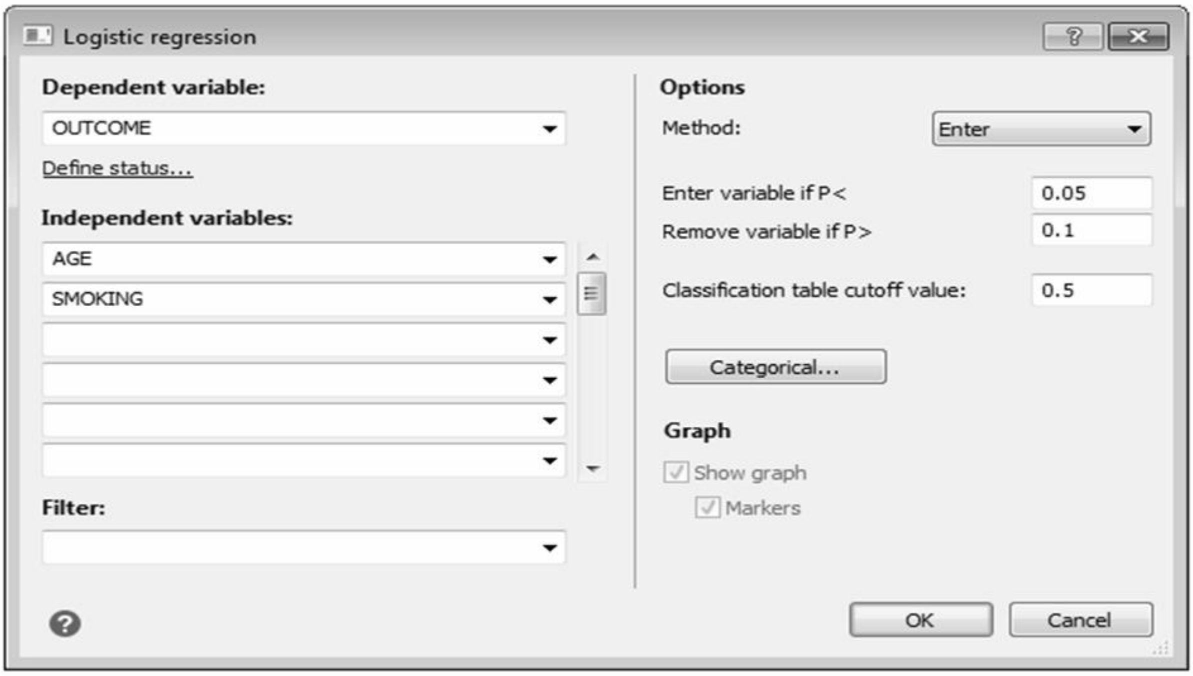


Logistic regression

A1 44

	A	B	C	D
	AGE	SMOKING	OUTCOME	
1	44	0	1	
2	22	0	0	
3	39	0	0	
4	23	0	0	
5	36	0	0	
6	19	0	0	

Required input



Logistic regression

Dependent variable: OUTCOME

Define status...

Independent variables: AGE, SMOKING

Filter:

Options

Method: Enter

Enter variable if P< 0.05

Remove variable if P> 0.1

Classification table cutoff value: 0.5

Categorical...

Graph

☑ Show graph

☑ Markers

OK Cancel

Dependent variable

The variable whose values you want to predict. The dependent variable must be binary or dichotomous, and should only contain data coded as 0 or 1. If your data are coded differently, you can use the Define status tool to recode your data.

Independent variables

Select the different variables that you expect to influence the dependent variable.

Filter

(Optionally) enter a data filter in order to include only a selected subgroup of cases in the analysis.

Options

- **Method:** Select the way independent variables are entered into the model.
  - **Enter:** enter all variables in the model in one single step, without checking
  - **Forward:** enter significant variables sequentially
  - **Backward:** first enter all variables into the model and next remove the non-significant variables sequentially
  - **Stepwise:** enter significant variables sequentially; after entering a variable in the model, check and possibly remove variables that became non-significant.
- Enter variable if P<

  A variable is entered into the model if its associated significance level is less than this P-value.
- Remove variable if P>

  A variable is removed from the model if its associated significance level is greater than this P-value.

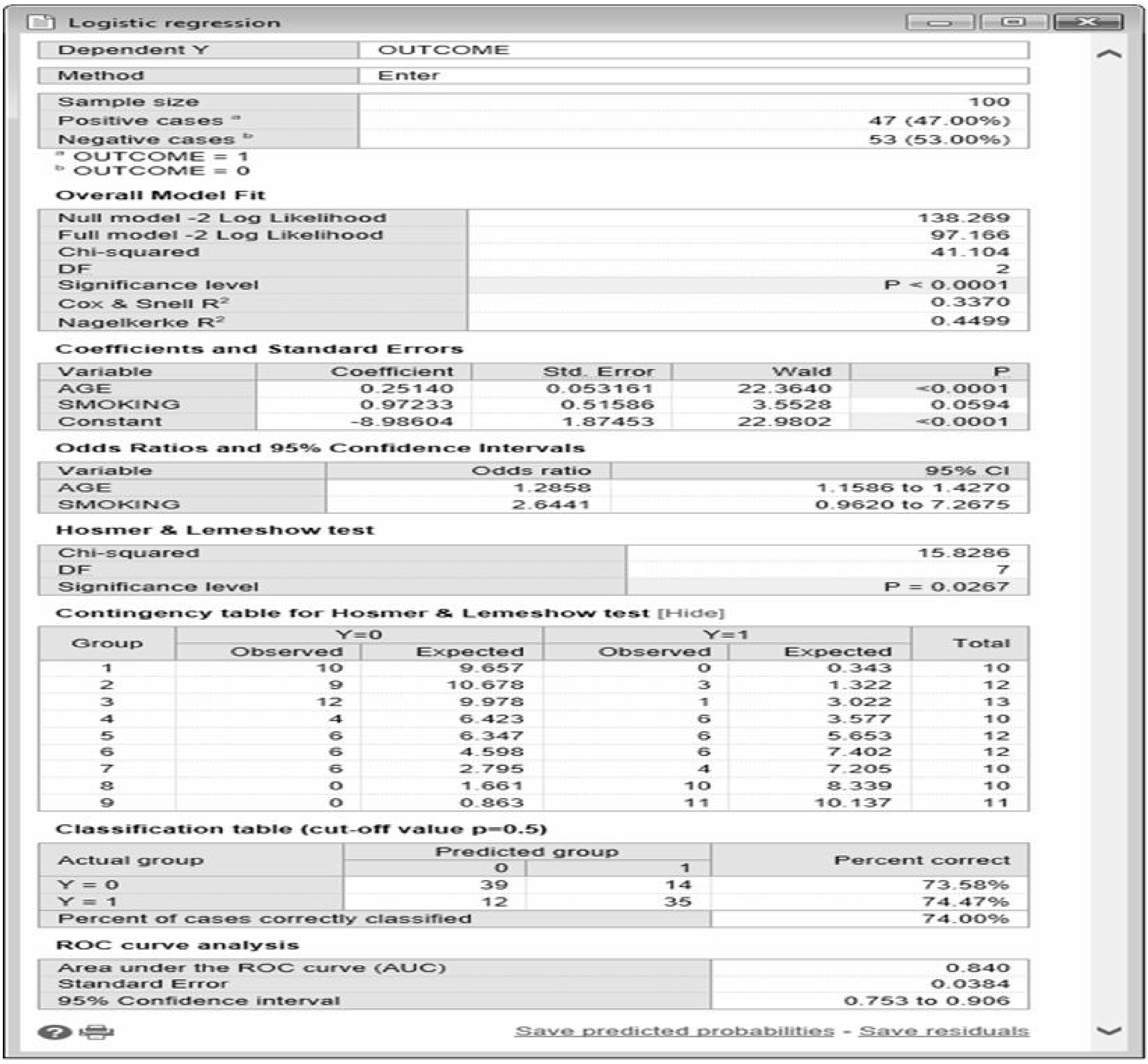
- **Classification table cutoff value:** a value between 0 and 1 which will be used as a cutoff value for a classification table. The classification table is a method to evaluate the logistic regression model. In this table the observed values for the dependent outcome and the predicted values (at the selected cut-off value) are cross-classified.
- **Categorical:** click this button to identify nominal categorical variables.

Graph

The option to plot a graph that shows the logistic regression curve is only available when there is just one single independent variable.

Results

After you click the OK button, the following results are displayed:



Logistic regression

Dependent Y	OUTCOME
Method	Enter

Sample size	100
Positive cases [a]	47 (47.00%)
Negative cases [b]	53 (53.00%)

[a] OUTCOME = 1
[b] OUTCOME = 0

Overall Model Fit

Null model -2 Log Likelihood	138.269
Full model -2 Log Likelihood	97.166
Chi-squared	41.104
DF	2
Significance level	P < 0.0001
Cox & Snell R^2	0.3370
Nagelkerke R^2	0.4499

Coefficients and Standard Errors

Variable	Coefficient	Std. Error	Wald	P
AGE	0.25140	0.053161	22.3640	<0.0001
SMOKING	0.97233	0.51586	3.5528	0.0594
Constant	-8.98604	1.87453	22.9802	<0.0001

Odds Ratios and 95% Confidence Intervals

Variable	Odds ratio	95% CI
AGE	1.2858	1.1586 to 1.4270
SMOKING	2.6441	0.9620 to 7.2675

Hosmer & Lemeshow test

Chi-squared	15.8286
DF	7
Significance level	P = 0.0267

Contingency table for Hosmer & Lemeshow test [Hide]

Group	Y=0		Y=1		Total
	Observed	Expected	Observed	Expected	
1	10	9.657	0	0.343	10
2	9	10.678	3	1.322	12
3	12	9.978	1	3.022	13
4	4	6.423	6	3.577	10
5	6	6.347	6	5.653	12
6	6	4.598	6	7.402	12
7	6	2.795	4	7.205	10
8	0	1.661	10	8.339	10
9	0	0.863	11	10.137	11

Classification table (cut-off value p=0.5)

Actual group	Predicted group		Percent correct
	0	1	
Y = 0	39	14	73.58%
Y = 1	12	35	74.47%
Percent of cases correctly classified			74.00%

ROC curve analysis

Area under the ROC curve (AUC)	0.840
Standard Error	0.0384
95% Confidence interval	0.753 to 0.906

Short Question and Answers

1. Data Mining

Ans :

Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue.

Data mining is also known as data discovery and knowledge discovery.

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection and warehousing as well as computer processing.

2. Scope of Data Mining

Ans :

1. Data mining process the work in such a manner that it allows business to more proactive to grow substantially.
2. It optimizes large database within the short time and works business intelligence which is more vital to organizational growth.
3. It represents the data in some logical order or may be in pattern to identify the sequential way of processing of data.
4. It includes tree-shaped structure to understand the hierarchy of data and representation of the set of information described in the database.
5. Brings a genetic way of classification of different sets of data items to view the data in the quick glance.

3. Steps of Data Exploration and Preparation

Ans :

Remember the quality of your inputs decide the quality of your output. So, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here. With my personal estimate, data exploration, cleaning and preparation can take up to 70% of your total project time.

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

Finally, we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

4. Data Reduction

Ans :

Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts.

When information is derived from instrument readings there may also be a transformation from analog to digital form. When the data are already in digital form the 'reduction' of the data typically involves some editing, scaling, coding, sorting, collating, and producing tabular summaries. When the observations are discrete but the underlying phenomenon is continuous then smoothing and interpolation are often needed. Often the data reduction is undertaken in the presence of reading or measurement errors. Some idea of the nature of these errors is needed before the most likely value may be determined.

5. Cluster Analysis

Ans :

Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

Cluster Analysis has been used in marketing for various purposes. Segmentation of consumers in cluster analysis is used on the basis of benefits sought from the purchase of the product. It can be used to identify homogeneous groups of buyers.

Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

6. Association Rules in Data Mining

Ans :

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk."

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

Association rule mining is the data mining process of finding the rules that may govern associations and causal objects between sets of items.

So in a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together. For example, peanut butter and jelly are often bought together because a lot of people like to make PB&J sandwiches.

7. Data partitioning

Ans :

A big problem can be solved easily when it is chopped into several smaller sub-problems. That is what the partitioning technique does. It divides the big database containing data metrics and indexes into smaller and handy slices of data called as partitions. The partitioned tables are directly used by the SQL queries without any alteration. Once the database is partitioned, the data definition language can easily work on the smaller partitioned slices, instead of handling the giant database altogether. This is how partitioning cuts down the problems in managing the large database tables.

The partitioning key consists of a single or supplementary columns with the intention of determining the partition wherever the rows will be stored. Spark modifies the partitions by using these partition keys.

All the smaller partitioned slices share the same logical features, but they do carry different physical features.

8. Predictive Accuracy

Ans :

The accuracy paradox for predictive analytics states that predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy. It may be better to avoid the accuracy metric in favor of other metrics such as precision and recall.

Accuracy is often the starting point for analyzing the quality of a predictive model, as well as an obvious criterion for prediction. Accuracy measures the ratio of correct predictions to the total number of cases evaluated. It may seem obvious that the ratio of correct predictions to cases should be a key metric. A predictive model may have high accuracy, but be useless.

9. k-nearest neighbors

Ans :

The k-nearest neighbors (k-NN) is a non - parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

10. Logistic Regression

Ans :

Logistic Regressionis the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

UNIT V

Simulation : Random Number Generation, Monte Carlo Simulation, What if Analysis, Verification and Validation, Advantages and Disadvantages of Simulation, Risk Analysis, Decision Tree Analysis

5.1 Simulation

Q1. What is simulation and explain different types of simulations?

Ans :

Simulation is an imitation of the operation of a real-world process or system. The act of simulating something first requires that a model be developed; this model represents the key characteristics, behaviors and functions of the selected physical or abstract system or process. The model represents the system itself, whereas the simulation represents the operation of the system over time.

Simulation is used in many contexts, such as simulation of technology for performance optimization, safety engineering, testing, training, education, and video games. Often, computer experiments are used to study simulation models. Simulation is also used with scientific modelling of natural systems or human systems to gain insight into their functioning, as in economics. Simulation can be used to show the eventual real effects of alternative conditions and courses of action. Simulation is also used when the real system cannot be engaged, because it may not be accessible, or it may be dangerous or unacceptable to engage, or it is being designed but not yet built, or it may simply not exist.

Key issues in simulation include acquisition of valid source information about the relevant selection of key characteristics and behaviours, the use of simplifying approximations and assumptions within the simulation, and fidelity and validity of the simulation outcomes. Procedures and protocols for model verification and validation are an ongoing field of academic study, refinement, research and development in simulations technology or practice, particularly in the field of computer simulation

Types of Simulation

Historically, simulations used in different fields developed largely independently, but 20th century studies of systems theory and cybernetics combined with spreading use of computers across all those fields have led to some unification and a more systematic view of the concept.

1. **Physical Simulation**

   It refers to simulation in which physical objects are substituted for the real thing (some circles use the term for computer simulations modelling selected laws of physics, but this article does not). These physical objects are often chosen because they are smaller or cheaper than the actual object or system.

2. **Interactive Simulation**

   It is a special kind of physical simulation, often referred to as a human in the loop simulation, in which physical simulations include human operators, such as in a flight simulator or a driving simulator.

3. **Continuous Simulation**

   It is a simulation where time evolves continuously based on numerical integration of Differential Equations.

4. **Discrete Event Simulation**

   It is a simulation where time evolves along events that represent critical moments, while the values of the variables are not relevant between two of them or result trivial to be computed in case of necessity.

5. **Stochastic Simulation**

It is a simulation where some variable or process is regulated by stochastic factors and estimated based on Monte Carlo techniques using pseudo-random numbers, so replicated runs from same boundary conditions are expected to produce different results within a specific confidence band.

6. **Deterministic Simulation**

It is a simulation where the variable are regulated by deterministic algorithms, so replicated runs from same boundary conditions produce always identical results.

7. **Hybrid Simulation**

It (sometime Combined Simulation) corresponds to a mix between Continuous and Discrete Event Simulation and results in integrating numerically the differential equations between two sequential events to reduce number of discontinuities.

8. **Stand Alone Simulation**

It is a Simulation running on a single workstation by itself.

9. **Distributed Simulation**

It is operating over distributed computers in order to guarantee access from/to different resources (e.g. multi users operating different systems, or distributed data sets); a classical example is Distributed Interactive Simulation (DIS).

10. **Parallel Simulation**

It is executed over multiple processor usually to distribute the computational workload as it is happening in High Performance Computing·

11. **Interoperable Simulation**

Where multiple models, simulators (often defined as Federates) interoperate locally on distributed over a network; a classical example is High Level Architecture.

Q2. Discuss the various steps involved in the process of simulation.

Ans :

The various steps involved in simulation process are as follows.

Step-1: Identify the Problem

The simulation process solves only those problems whose assumptions for analytical problems are not satisfied or there is no appropriate model of the system under consideration. For example, the arrival/service of pattern of the queuing system does not meet the criteria required to solve the problem by queuing theory.

Step-2: Identify the Decision Variables and Decide the Objective

After identifying the problem, the next step is to identify the decision variables and define the problem and list the objectives to be achieved from solution of the model. This not only facilitates the development of the model but also provides the basis for the evaluation of the simulation results. For example, in inventory situation, the demand, lead time and safety stock are considered as decision variables.

Step-3: Construction of an Appropriate Model

The third step in simulation process is the development of a suitable simulation model. For the development of the model, a clear understanding of the relationship among the system elements is required. This model may be a physical mathematical, mental conception or a combination of these. In general, many models involve physical scaled down model of an aeroplane or ship made up of wood or other material. As the physical models are expensive, the mathematical model showing the relationship between the system elements are preferred.

Step-4: Experimentation with the Model Constructed

This step involves comparing the model with the actual system under consideration as the model should represent the exact system in consideration. This step runs the model developed for study If the conditions are deterministic and constant, a single run is enough and if the conditions are stochastic in nature, then number of runs will be required to get the correct picture of the model performance. For the parameters subject to random variation, large amount of runs are required to get a reasonable degree of confidence that the results are truly indicative of the system behaviour.

Step-5: Evaluation of the Results

The last step in simulation process is examining the results of the problem as well as their reliability and accuracy. These interpretations depend on the extent to which the model portray the reality. If the simulation is complete, then select the best course of action or else make necessary changes in the decision variables and repeat the process from step 3. The closer the model is related to real system, the lesser will be the need for adjusting the results.

Q3. Write the practical applications of simulation.

Ans :

Simulation techniques is useful and applicable in reaching optimization in the following problems,

1. Queueing
2. Inventory
3. Capital budgeting and
4. Financial planning problems.

1. Queueing Problem

The optimal solution for queueing problems can be determined by reducing the effectiveness of queue length, average waiting time, service time, idle time and so on.

In case of uncertain events like bulk arrival, jockeying, balking, reneging etc., Simulation is the only method found to be suitable for optimization. It is also helpful to find approximate waiting time in case of uncertain working conditions. Depending on such information effective decisions can be made.

2. Inventory Problem

In many inventory problems, the analytical method will not be useful to solve and to find the optimal solution because of the complexities, for example, in case of storage problem, due to the existence of uncertain demand and supply distribution tends to be complex. In such situation, with the help of simulation model policy of reorder and order quantity can be designed. The main purpose of simulation method is to reduce the total cost of inventory thereby reducing carrying, holding, storage costs etc. By reducing such costs, optimization can be obtained.

3. Capital Budgeting

According to prof. David B. Hertz, simulation can also be used in capital budgeting i.e., to estimate the proposal yielding the best return for its investment. By adopting simulation model, different alternatives of financial evaluation can be determined and out of that, the best optimal solution can be selected which has less risk due to various uncertain factors like Selling Price (SP), market growth rate, market size etc., on financial parameters and yields maximum return on investment.

4. Financial Planning Problems

Due to varying situations and different stages of development, the companies are facing problems in financial decision making. If inappropriate decisions are made, then businesses will suffer huge losses. In order to avoid such losses, simulation methods are used. Hence the management needs to be careful while taking decisions during financial planning of the firm. With simulation model, the managers can simulate the overall behavior of the system under given situations if all inputs and decisions had actually occurred and must select those projects which yields maximum returns to the firm.

5.2 Random Number Generation

Q4. What is Random number generation and discuss the fundamental methods of it

Ans :

Random numbers are numbers that occur in a sequence such that two conditions are met: (1) the values are uniformly distributed over a defined interval or set, and (2) it is impossible to predict future values based on past or present ones. Random numbers are important in statistical analysis and probability theory.

The most common set from which random numbers are derived is the set of single-digit decimal numbers {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}. The task of generating random digits from this set is not trivial. A common scheme is the selection (by means of a mechanical escape hatch that lets one ball out at a time) of numbered ping-pong balls from a set of 10, one bearing each digit, as the balls are blown about in a container by forced-air jets. This method is popular in lotteries. After each number is selected, the ball with that number is returned to the set, the balls are allowed to blow around for a minute or two, and then another ball is allowed to escape.

Sometimes the digits in the decimal expansions of irrational numbers are used in an attempt to obtain random numbers. Most whole numbers have irrational square roots, so entering a string of six or eight digits into a calculator and then hitting the square root button can provide a sequence of digits that seems random. Other algorithms have been devised that supposedly generate random numbers. The problem with these methods is that they violate condition (2) in the definition of randomness. The existence of any number-generation algorithm produces future values based on past and/or current ones. Digits or numbers generated in this manner are called pseudorandom.

Statisticians, mathematicians, and scientists have long searched for the ideal source of random numbers. One of the best methods is the sampling of electromagnetic noise. This noise, generated by the chaotic movements of electrons, holes, or other charge carriers in materials and in space, is thought to be as close to "totally random" as any observable phenomenon.

Random number generation is the generation of a sequence of numbers or symbols that cannot be reasonably predicted better than by a random chance, usually through a hardware random-number generator (RNG).

The half dozen basic methods that all of these techniques are derived. The famous random variate generation/sampling techniques are derived from combinations of the following six fundamental methods:

1. Physical sources
2. Empirical resampling
3. Pseudo random generators
4. Simulation/Game-play
5. Rejection Sampling
6. Transform methods

The technical fights (such as: "is Gibbs sampling superior to, or even distinguishable from, Markov chain Monte Carlo?") are all in the details, history and citation conventions. Each field and particular method accretes its own traditions. We will quickly discuss the fundamental methods we listed. As we will see: complexity goes up as we move through the list (so at some point things are no longer fundamental but instead derived, allowing us to end the list).

1. Physical Sources

This is the most basic way (though not as practical in the computer age) to generate random variables. Observe the flip of a real coin, shuffle actual cards, mix numbered balls or count the number of ticks from an actual radioactive source. In all of these the randomness comes from physical principles (such chaotic dynamics for coin flips or quantum mechanics for radioactive decay).

These sources are "outside of computer science" so we will say the least about them.

2. Empirical Resampling

This is what used to be called "tables" (which were themselves often generated from physical processes). The observation is: that sometimes to run a simulation you need access to instances of random variables that are distributed in a very precise way- but you don't have a usable description of the desired distribution. You would think that in this case you could do nothing. But the principle of empirical resampling is that you can approximately generate new samples by taking samples (with repetition or replacement) from an old sample. This is the cornerstone of Bootstrap methods.

3. Pseudo Random Generators

In the computer age, to avoid need for external tables or expensive and slow peripherals we tend to use pseudo random generators. That is the output of deterministic iterative procedures as equivalent to true random sources. The science of pseudo randomness has evolved from cobbled together procedures passing ad-hoc tests (such as in Knuth Volume 2) to more formal pseudo randomness based on important properties (like provably being k-wise independent) or complexity (being computational indistinguishable from a truly random on a time or space bounded machine). Behind the canned routines of all of the basic "random generators" commonly available is a pseudo random source.

4. Simulation/Game-play

Another fundamental method is direct simulation or game play. If we wanted a random variable that was 1 with probability equal to the odds of being dealt a full house from a standard shuffled deck of 52 cards (and zero otherwise). We can generate such a variable by simulating shuffling a deck, drawing a hand and returning 1 if the hand draw is a full house (and returning 0 otherwise). Notice in this case we are combining many random variables to get a single result.

One of the most important simulation techniques is Markov chain Monte Carlo methods (related to Gibbs sampling, simulated annealing and many other variations). These method implement a complex procedure over a stream of random inputs to generate a more difficult to achieve sequence of random outputs.

5. Rejection Sampling

Rejection sampling is another way to convert one sequence of random variables into another. If we assume we can generate a random variable according to the distribution p(x) we can "rejection sample" to a new distribution using an "acceptance function" q(x) which returns a number in the interval [0,1]. Our procedure is to repeat the following: generate x with probability p(x), generate a random variable y with uniformly in the interval [0,1] if y q(x) accept x as our answer and quit (otherwise draw a new x and repeat).

When the distribution that rejection sampling draws with is such that if x and y had a ratio of being drawn of p(x)/p(y) then under the rejection procedure they have relative odds of (p(x)q(x))/(p(y)q(y)). An important special case is when q() is always 0 or 1, in this case we are drawing with relative odds proportional to p(x) from the subset of x with q(x)=1.

6. Transform Methods

A transform method is used when we have the ability to generate instances of a random variable according to one distribution and we would like instances according to another distribution.

One method is used when we have access to the inverse of the cumulative distribution function of the distribution we are trying to generate. In this case we can use this function to convert uniform variants from the interval [0,1] into our target distribution. The commutative distribution function is the function cdf() where cdf(x) is the probability a random variate generated according to our distribution is less than or equal to x. The inverse function functionicdf() where icdf(y) is such that cdf(icdf(y)) = y. For example the exponential distribution has an inverse cumulative distribution function icdf(y) = -ln(1-y)/lamda . So if y is generated uniformly in the interval [0,1] then icdf(y) is a random variable generated according to the exponential distribution with parameter lambda.

Q5. Explain the random number generation in excel.

Ans :

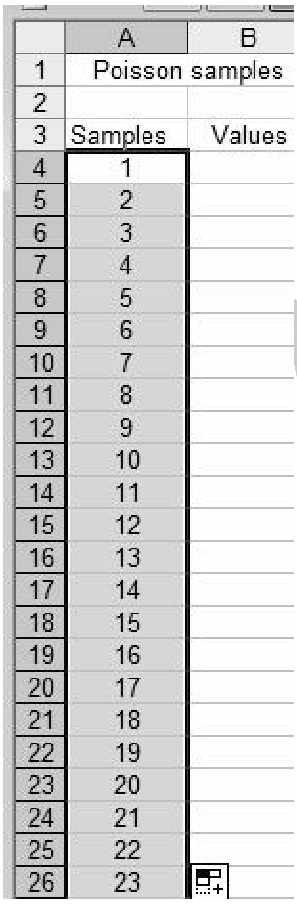
In excel, a random number can be generated easily with its random number generation tool. Even though RAND and 'RANDBETWEEN' function enables us to generate random numbers but the random number generation tool gives more advantage in creating random number population based on various distributions.

This option can be better understand with an example.

Let us generate 25 outcomes from a Poisson distribution with a mean of 12 and also displaying the same result in a histogram. The steps involved in generating random numbers in excel are as follows,

Step-1

Enter the following in the worksheet as shown below.



	A	B
1	Poisson samples	
2		
3	Samples	Values
4	1	
5	2	
6	3	
7	4	
8	5	
9	6	
10	7	
11	8	
12	9	
13	10	
14	11	
15	12	
16	13	
17	14	
18	15	
19	16	
20	17	
21	18	
22	19	
23	20	
24	21	
25	22	
26	23	

Step-2

Click on 'Data' tab and then click on 'Data Analysis'.

Step-3

Select 'Random Number Generation' by clicking on 'OK'. Random number generation dialog box will be displayed.

Step-4

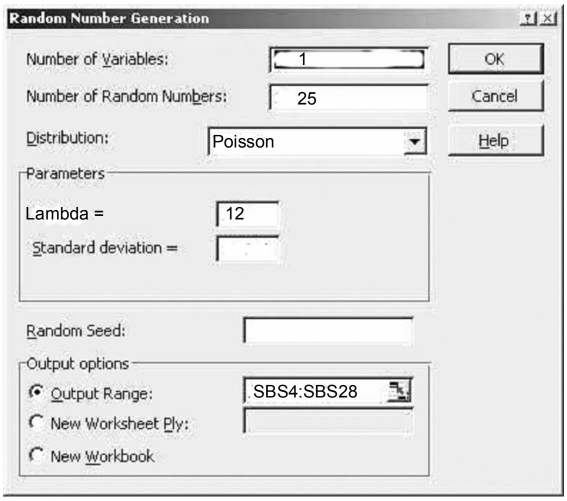
(i) Enter '1' in the 'Number of Variables' field (This option creates number of columns).

(ii) Enter '25' in the 'Number of Random Numbers' field (This option creates number of rows).

(iii) Select 'Poisson' from drop-down list of distribution.

(iv) Enter'12'in the Lamba = field.

(v) Select 'Output Range under Output options. 'Here, the random numbers will be generated under value'. Enter $B$4:$B$28 or you can select cell B4 to cell B25. After entering all details, the final dialog will appear as shown below.



Random Number Generation

Number of Variables: 1

Number of Random Numbers: 25

Distribution: Poisson

Parameters

Lambda = 12

Standard deviation =

Random Seed:

Output options

● Output Range: SBS4:SBS28

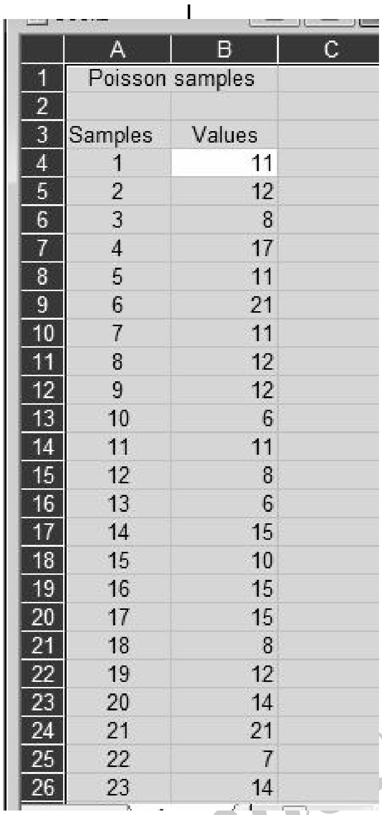
○ New Worksheet Ply:

○ New Workbook

Step-5

Press enter to generate random numbers.

As numbers are randomly generated, it will not be same as generated here which is shown below.



	A	B	C
1	Poisson samples		
2			
3	Samples	Values	
4	1	11	
5	2	12	
6	3	8	
7	4	17	
8	5	11	
9	6	21	
10	7	11	
11	8	12	
12	9	12	
13	10	6	
14	11	11	
15	12	8	
16	13	6	
17	14	15	
18	15	10	
19	16	15	
20	17	15	
21	18	8	
22	19	12	
23	20	14	
24	21	21	
25	22	7	
26	23	14	

Step-6

In order to create histogram for this results, select 'Data' tab. Click on 'Data analysis' select 'Histogram' from the list. The histogram dialog box will appear.

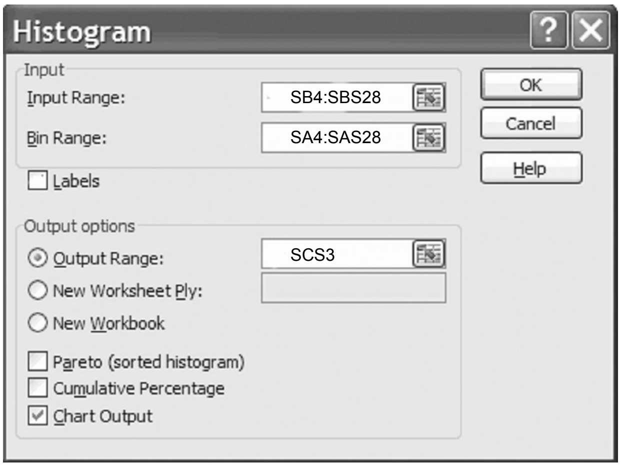
Step- 7

(i) Enter $B$4:$B$28 under 'Input Range' or you can select the cells by clicking on cell B4 and dragging till cell B28.

(ii) Enter $A$4:$A$28 under 'Bin Range' or you can select the cells by clicking on cell A4 and dragging till cell A28 (A bin range refers to a range of values which defines the limits for each column of the histogram. If bin range is omitted then excel creates ten equal intervals bins).

(iii) Enter $D$3 under 'output range' or you can select the cells by clicking on cell D3.

(iv) Tick mark 'chart output'. The final dialog box will appear after entering details is shown below,



Histogram

Input

Input Range: SB4:SBS28

Bin Range: SA4:SAS28

□ Labels

Output options

● Output Range: SCS3

○ New Worksheet Ply:

○ New Workbook

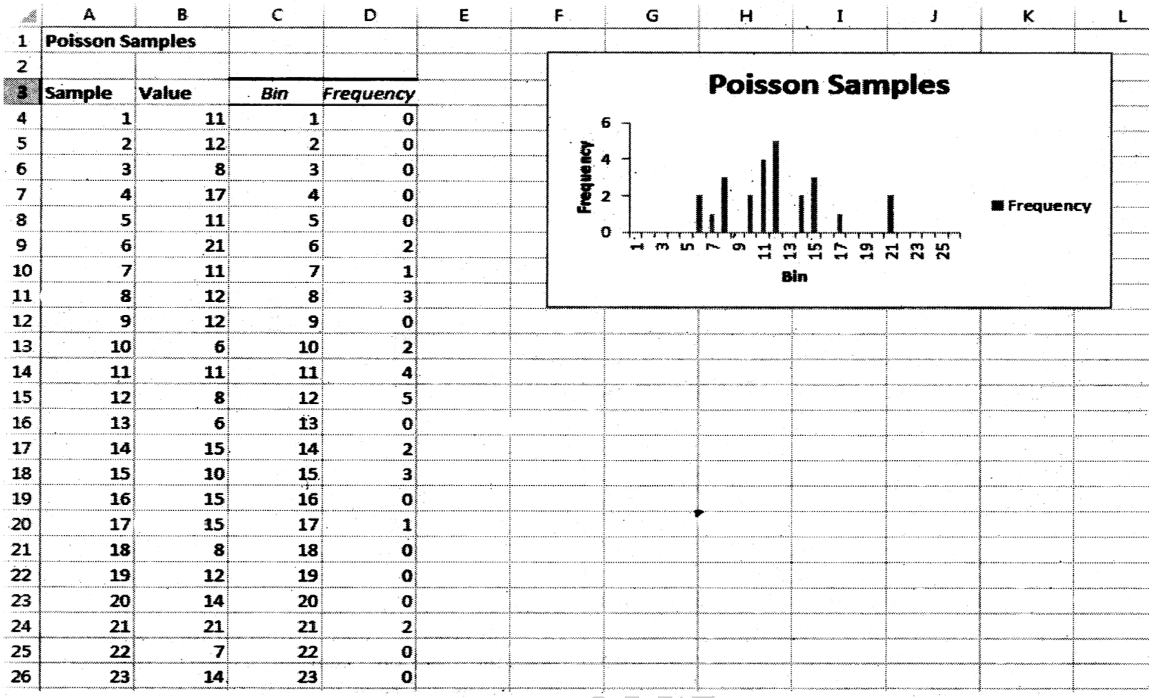
□ Pareto (sorted histogram)

□ Cumulative Percentage

☑ Chart Output

Step-8

Press Enter. The histogram for the generated random numbers is shown below.



Poisson Samples

Sample	Value	Bin	Frequency
1	11	1	0
2	12	2	0
3	8	3	0
4	17	4	0
5	11	5	0
6	21	6	2
7	11	7	1
8	12	8	3
9	12	9	0
10	6	10	2
11	11	11	4
12	8	12	5
13	6	13	0
14	15	14	2
15	10	15	3
16	15	16	0
17	15	17	1
18	8	18	0
19	12	19	0
20	14	20	0
21	21	21	2
22	7	22	0
23	14	23	0

5.3 Monte Carlo Simulation

Q6. Explain about Monte Carlo simulation and explain how it works.

Ans :

Monte Carlo simulation is a computerized mathematical technique that allows people to account for risk in quantitative analysis and decision making. The technique is used by professionals in such widely disparate fields as finance, project management, energy, manufacturing, engineering, research and development, insurance, oil & gas, transportation, and the environment.

Monte Carlo simulation furnishes the decision-maker with a range of possible outcomes and the probabilities they will occur for any choice of action. It shows the extreme possibilities - the outcomes of going for broke and for the most conservative decision - along with all possible consequences for middle-of-the-road decisions.

The technique was first used by scientists working on the atom bomb; it was named for Monte Carlo, the Monaco resort town renowned for its casinos. Since its introduction in World War II, Monte Carlo simulation has been used to model a variety of physical and conceptual systems.

Monte Carlo Simulation Works

Monte Carlo simulation performs risk analysis by building models of possible results by substituting a range of values - a probability distribution - for any factor that has inherent uncertainty. It then calculates results over and over, each time using a different set of random values from the probability functions. Depending upon the number of uncertainties and the ranges specified for them, a Monte Carlo simulation could involve thousands or tens of thousands of recalculations before it is complete. Monte Carlo simulation produces distributions of possible outcome values.

By using probability distributions, variables can have different probabilities of different outcomes occurring. Probability distributions are a much more realistic way of describing uncertainty in variables of a risk analysis.

Common probability distributions include:

1. **Normal**

   Or "bell curve." The user simply defines the mean or expected value and a standard deviation to describe the variation about the mean. Values in the middle near the mean are most likely to occur. It is symmetric and describes many natural phenomena such as people's heights. Examples of variables described by normal distributions include inflation rates and energy prices.

2. **Lognormal**

   Values are positively skewed, not symmetric like a normal distribution. It is used to represent values that don't go below zero but have unlimited positive potential. Examples of variables described by lognormal distributions include real estate property values, stock prices, and oil reserves.

3. **Uniform**

   All values have an equal chance of occurring, and the user simply defines the minimum and maximum. Examples of variables that could be uniformly distributed include manufacturing costs or future sales revenues for a new product.

4. **Triangular**

   The user defines the minimum, most likely, and maximum values. Values around the most likely are more likely to occur. Variables that could be described by a triangular distribution include past sales history per unit of time and inventory levels.

5. **PERT**

   The user defines the minimum, most likely, and maximum values, just like the triangular distribution. Values around the most likely are more likely to occur. However values between the most likely and extremes are more likely to occur than the triangular; that is, the extremes are not as emphasized. An example of the use of a PERT distribution is to describe the duration of a task in a project management model.

6. **Discrete**

   The user defines specific values that may occur and the likelihood of each. An example might be the results of a lawsuit: 20% chance of positive verdict, 30% change of negative verdict, 40% chance of settlement, and 10% chance of mistrial.

   During a Monte Carlo simulation, values are sampled at random from the input probability distributions. Each set of samples is called an iteration, and the resulting outcome from that sample is recorded. Monte Carlo simulation does this hundreds or thousands of times, and the result is a probability distribution of possible outcomes. In this way, Monte Carlo simulation provides a much more comprehensive view of what may happen. It tells you not only what could happen, but how likely it is to happen.

Q7. Explain how Monte Carlo simulation can be developed in Ms.Excel?

Ans :

We will develop a Monte Carlo simulation using Microsoft Excel and a game of dice. The Monte Carlo Simulation is a mathematical numerical method that uses random draws to perform calculations and complex problems.

Monte Carlo method was invented by Nicolas Metropolis in 1947 and seeks to solve complex problems using random and probabilistic methods. The term "Monte Carlo" originates from the administrative area of Monaco popularly known as a place where European elites gamble. We use the Monte Carlo method when the problem is too complex and difficult to do by direct calculation. A large number of iterations allows a simulation of the normal distribution.

The Monte Carlo simulation method computes the probabilities for integrals and solves partial differential equations, thereby introducing a statistical approach to risk in a probabilistic decision. Although many advanced statistical tools exist to create Monte Carlo simulations, it is easier to simulate the normal law and the uniform law using Microsoft Excel and bypass the mathematical underpinnings.

For the Monte Carlo simulation, we isolate a number of key variables that control and describe the outcome of the experiment and assign a probability distribution after a large number of random samples is performed. Let's take a game of dice as model.

Game of Dice

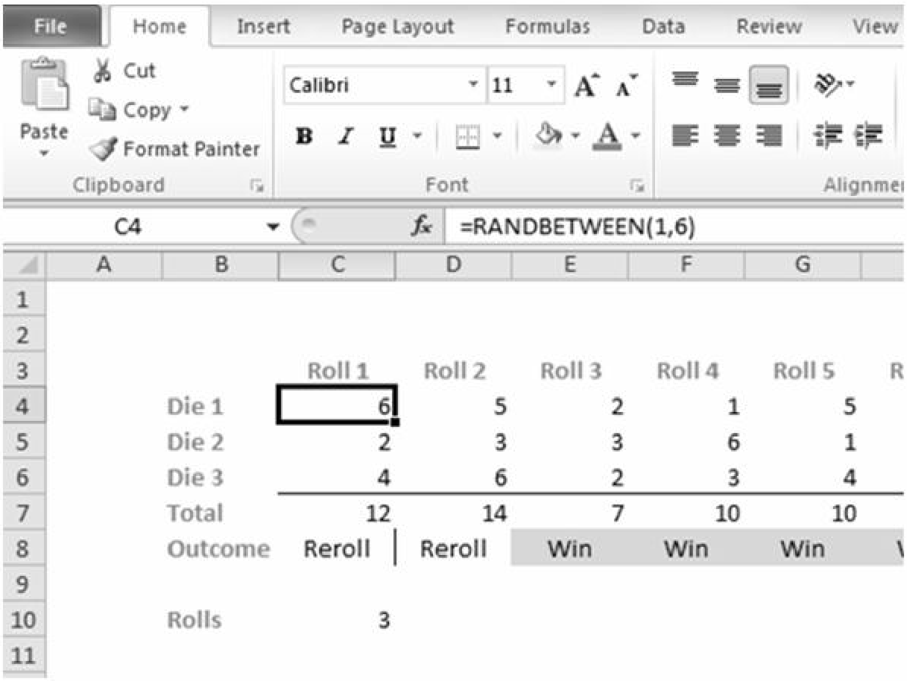
Here's how the dice game rolls:

- The player throws three dice that have 6 sides 3 times.
- If the total of the 3 throws is 7 or 11, the player wins.
- If the total of the 3 throws is: 3, 4, 5, 16, 17 or 18, the player loses.
- If the total is any other outcome, the player plays again and re-rolls the die.
- When the player throws the die again, the game continues in the same way, except that the player wins when the total is equal to the sum determined in the first round.

It is also recommended to use a data table to generate the results. Moreover, 5,000 results are needed to prepare the Monte Carlo simulation.

Step 1: Dice Rolling Events

First, we develop a range of data with the results of each of the 3 dice for 50 rolls. To do this, it is proposed to use the "RANDBETWEEN (1.6)" function. Thus, each time we click F9, we generate a new set of roll results. The "Outcome" cell is the sum total of the results from the 3 rolls.



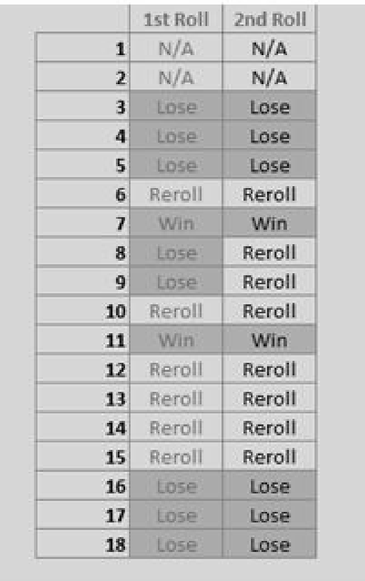
=RANDBETWEEN(1,6)

	A	B	C	D	E	F	G
1							
2							
3			Roll 1	Roll 2	Roll 3	Roll 4	Roll 5
4		Die 1	6	5	2	1	5
5		Die 2	2	3	3	6	1
6		Die 3	4	6	2	3	4
7		Total	12	14	7	10	10
8		Outcome	Reroll	Reroll	Win	Win	Win
9							
10		Rolls	3				
11							

Step 2 : Range of Outcomes

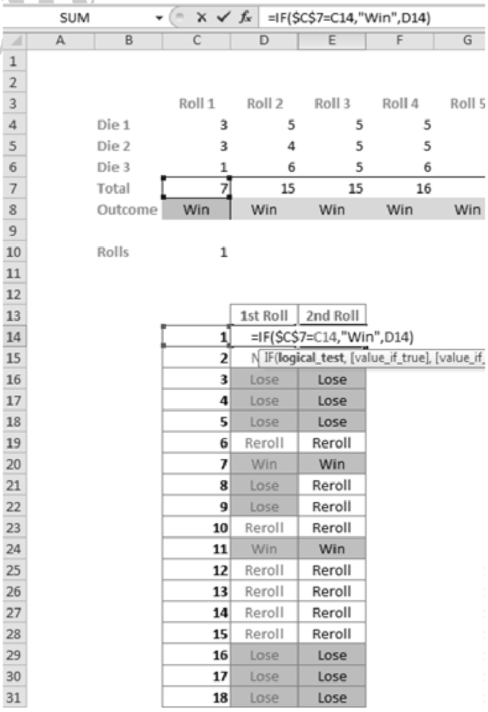
Then, we need to develop a range of data to identify the possible outcomes for the first round and subsequent rounds. There is provided below a 3-column data range. In the first column, we have the numbers 1 to 18. These figures represent the possible outcomes following rolling the dice 3 times: the maximum being 3*6=18. You will note that for cells 1 and 2, the findings are N/A since it is impossible to get a 1 or a 2 using 3 dice. The minimum is 3.

In the second column, the possible conclusions after the first round is included. As stated in the initial statement, either the player wins (Win) or loses (Lose) or he replays (Re-roll), depending on the result (the total of 3 dice rolls).



	1st Roll	2nd Roll
1	N/A	N/A
2	N/A	N/A
3	Lose	Lose
4	Lose	Lose
5	Lose	Lose
6	Reroll	Reroll
7	Win	Win
8	Lose	Reroll
9	Lose	Reroll
10	Reroll	Reroll
11	Win	Win
12	Reroll	Reroll
13	Reroll	Reroll
14	Reroll	Reroll
15	Reroll	Reroll
16	Lose	Lose
17	Lose	Lose
18	Lose	Lose

In the third column, the possible conclusions to subsequent rounds are registered. We can achieve these results using a function "If." This ensures that if the result obtained is equivalent to the result obtained in the first round, we win, otherwise we follow the initial rules of the original play to determine whether we re-roll the dice.

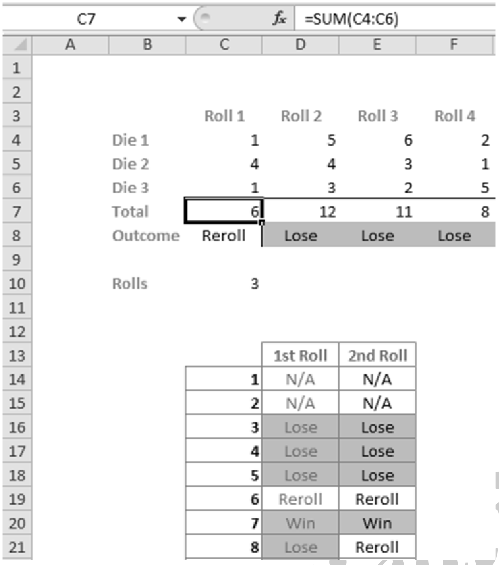


SUM =IF($C$7=C14,"Win",D14)

	A	B	C	D	E	F	G
1							
2							
3			Roll 1	Roll 2	Roll 3	Roll 4	Roll 5
4		Die 1	3	5	5	5	
5		Die 2	3	4	5	5	
6		Die 3	1	6	5	6	
7		Total	7	15	15	16	
8		Outcome	Win	Win	Win	Win	Win
9							
10		Rolls	1				
11							
12							
13				1st Roll	2nd Roll		
14			1	=IF($C$7=C14,"Win",D14)			
15			2	N IF(**logical_test**, [value_if_true], [value_if_			
16			3	Lose	Lose		
17			4	Lose	Lose		
18			5	Lose	Lose		
19			6	Reroll	Reroll		
20			7	Win	Win		
21			8	Lose	Reroll		
22			9	Lose	Reroll		
23			10	Reroll	Reroll		
24			11	Win	Win		
25			12	Reroll	Reroll		
26			13	Reroll	Reroll		
27			14	Reroll	Reroll		
28			15	Reroll	Reroll		
29			16	Lose	Lose		
30			17	Lose	Lose		
31			18	Lose	Lose		

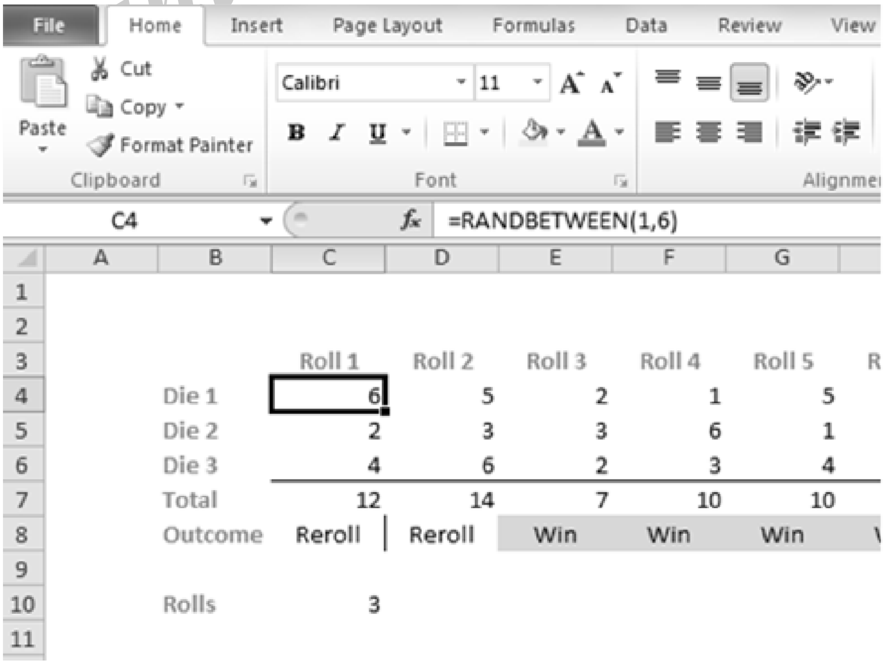
Step 3 : Conclusions

In this step, we identify the outcome of the 50 dice rolls. The first conclusion can be obtained with an index function. This function searches the possible results of the first round, the conclusion corresponding to the result obtained. For example, when obtaining 6, as is the case in the picture below, we play again.


C7 =SUM(C4:C6)

	A	B	C	D	E	F
1						
2						
3			Roll 1	Roll 2	Roll 3	Roll 4
4		Die 1	1	5	6	2
5		Die 2	4	4	3	1
6		Die 3	1	3	2	5
7		Total	6	12	11	8
8		Outcome	Reroll	Lose	Lose	Lose
9						
10		Rolls	3			
11						
12						
13				1st Roll	2nd Roll	
14			1	N/A	N/A	
15			2	N/A	N/A	
16			3	Lose	Lose	
17			4	Lose	Lose	
18			5	Lose	Lose	
19			6	Reroll	Reroll	
20			7	Win	Win	
21			8	Lose	Reroll	

One can get the findings of other dice rolls, using an "Or" function and an index function nested in an "If" function. This function tells Excel, "If the previous result is Win or Lose," stop rolling the dice because once we have won or lost we are done. Otherwise, we go to the column of the following possible conclusions and we identify the conclusion of the result.

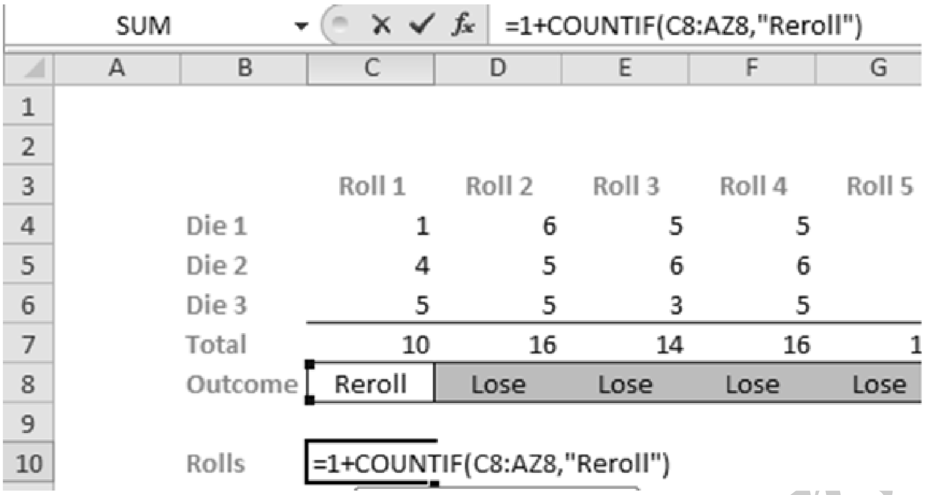


C4 =RANDBETWEEN(1,6)

	A	B	C	D	E	F	G
1							
2							
3			Roll 1	Roll 2	Roll 3	Roll 4	Roll 5
4		Die 1	6	5	2	1	5
5		Die 2	2	3	3	6	1
6		Die 3	4	6	2	3	4
7		Total	12	14	7	10	10
8		Outcome	Reroll	Reroll	Win	Win	Win
9							
10		Rolls	3				
11							

Step 4 :Number of Dice Rolls

Now, we determine the number of dice rolls required before losing or winning. To do this, we can use a “Count if” function, which requires Excel to count the results of “Re-roll” and add the number 1 to it. It adds one because we have one extra round, and we get a final result (win or lose).

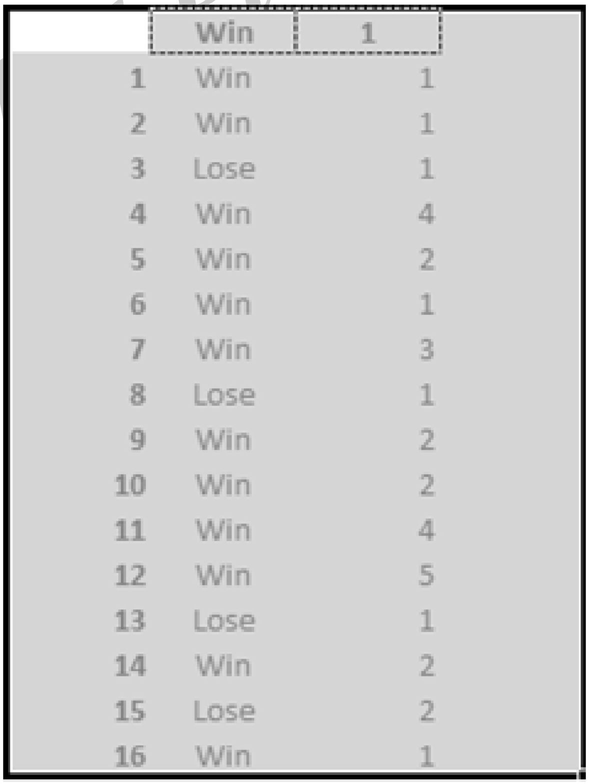


SUM =1+COUNTIF(C8:AZ8,"Reroll")

	A	B	C	D	E	F	G
1							
2							
3			Roll 1	Roll 2	Roll 3	Roll 4	Roll 5
4		Die 1	1	6	5	5	
5		Die 2	4	5	6	6	
6		Die 3	5	5	3	5	
7		Total	10	16	14	16	1
8		Outcome	Reroll	Lose	Lose	Lose	Lose
9							
10		Rolls	=1+COUNTIF(C8:AZ8,"Reroll")				

Step 5: Simulation

We develop a range to track the results of different simulations. To do this, we will create three columns. In the first column, one of the figures included is 5,000. In the second column we will look for the result after 50 dice rolls. In the third column, the title of the column, we will look for the number of dice rolls before obtaining the final status (win or lose).

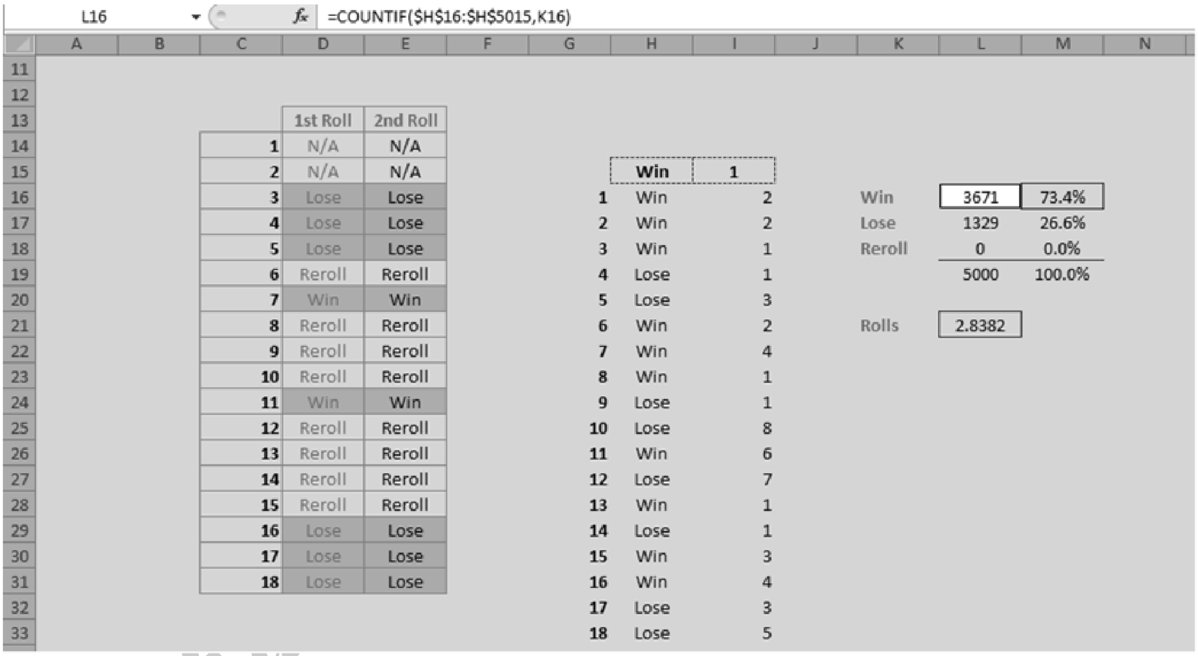


	Win	1
1	Win	1
2	Win	1
3	Lose	1
4	Win	4
5	Win	2
6	Win	1
7	Win	3
8	Lose	1
9	Win	2
10	Win	2
11	Win	4
12	Win	5
13	Lose	1
14	Win	2
15	Lose	2
16	Win	1

Then, we will create a sensitivity analysis table by using the feature data or Table Data table (this sensitivity will be inserted in the second table and third columns). In this sensitivity analysis, the numbers of events of 1 – 5,000 must be inserted into cell A1 of the file. In fact, one could choose any empty cell. The idea is simply to force a recalculation each time and thus get new dice rolls (results of new simulations) without damaging the formulas in place.

Step 6: Probability

We can finally calculate the probabilities of winning and losing. We do this using the "Countif" function. The formula counts the number of "win" and "lose" then divides by the total number of events, 5,000, to obtain the respective proportion of one and the other. We finally see below that the probability of getting a Win outcome is 73.2% and getting a Lose outcome is therefore 26.8%.



L16 =COUNTIF($H$16:$H$5015,K16)

	C	D	E	G	H	I	K	L	M
11									
12									
13		1st Roll	2nd Roll						
14	1	N/A	N/A						
15	2	N/A	N/A		Win	1			
16	3	Lose	Lose	1	Win	2	Win	3671	73.4%
17	4	Lose	Lose	2	Win	2	Lose	1329	26.6%
18	5	Lose	Lose	3	Win	1	Reroll	0	0.0%
19	6	Reroll	Reroll	4	Lose	1		5000	100.0%
20	7	Win	Win	5	Lose	3			
21	8	Reroll	Reroll	6	Win	2	Rolls	2.8382	
22	9	Reroll	Reroll	7	Win	4			
23	10	Reroll	Reroll	8	Win	1			
24	11	Win	Win	9	Lose	1			
25	12	Reroll	Reroll	10	Lose	8			
26	13	Reroll	Reroll	11	Win	6			
27	14	Reroll	Reroll	12	Lose	7			
28	15	Reroll	Reroll	13	Win	1			
29	16	Lose	Lose	14	Lose	1			
30	17	Lose	Lose	15	Win	3			
31	18	Lose	Lose	16	Win	4			
32				17	Lose	3			
33				18	Lose	5			

5.4 What-if Analysis

Q8. Explain about "What-if Analysis" in Excel?

Ans :

By using What-If Analysis tools in Excel, you can use several different sets of values in one or more formulas to explore all the various results.

For example, you can do What-If Analysis to build two budgets that each assumes a certain level of revenue. Or, you can specify a result that you want a formula to produce, and then determine what sets of values will produce that result. Excel provides several different tools to help you perform the type of analysis that fits your needs.

What-If Analysis in **Excel** allows you to try out different values (scenarios) for formulas. The following example helps you master what-if analysis quickly and easily.

What-if analysis is a useful way of being able to test out various scenarios in Excel. You can look at these things two different ways.

The first way is to change the input variables and see what impact that has on the output. The scenario manager and data tables work in this way and can be used to answer questions like, what would happen to our profits if the number of units we sell doubled, or what would happen to our profits if the cost price of each of our units sold increased by 10%.

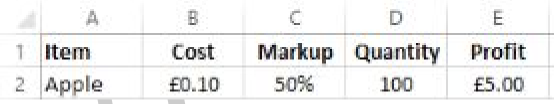
The second way is to say what outcome you would like to have and ask Excel to calculate what change in the inputs would be required to achieve this. The goal seek feature works this way and can answer questions such as how many units of a product need to be sold in order to reach a desired profit level.

Scenario Manager

The following example looks at scenarios showing the profit from selling 100 apples, with differing levels of mark-up.

The first step is to create some scenarios. The scenario manager can be found on the Data ribbon, under What-If Analysis. Click the Add button to start creating new scenarios. The next dialog box will ask for a name for the scenario and which cells are to be changed. A scenario can contain up to 32 changing cells, although in reality, most scenarios will use far fewer than this. Once the scenario name and cells to be changed have been selected, click the OK button and fill in the values for each cell to be used in the scenario. From here, click Add to continue adding more scenarios, or OK when finished.

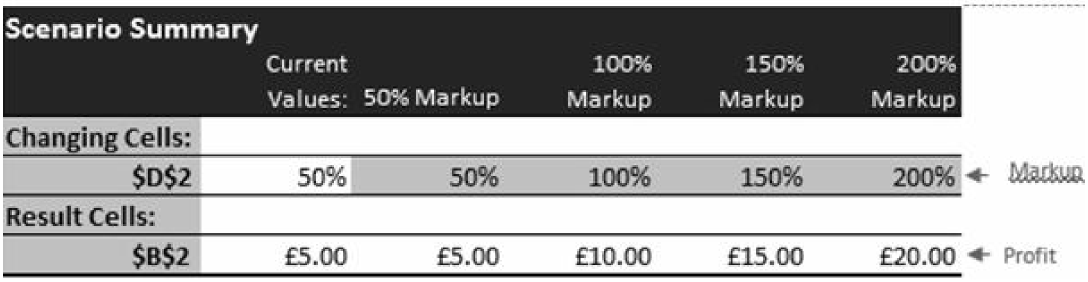
For this example, four scenarios have been created – for a 50%, 100%, 150% and 200% markup. There are two ways that a scenario can be applied – it can either be shown on the worksheet itself, or a summary can be created. The below table shows the results when a 50% markup is applied. This was achieved simply by selecting the 50% Markup scenario from the Scenario Manager and clicking the Show button. 50% has been entered into the Markup column, and the figure in the Profit column, which is calculated using a formula referring to this has updated accordingly.



	A	B	C	D	E
1	Item	Cost	Markup	Quantity	Profit
2	Apple	£0.10	50%	100	£5.00

To compare results from several scenarios simultaneously, the Summary option can be used. As with the Show option, the Summary report is accessed from the Scenario Manager dialog box. There are two options here; Scenario summary and Scenario Pivot Table Report. Although the option is there to display the results as a Pivot Table report, in most cases, the summary will display the results in a more user-friendly way.

The below shows the results from the scenario summary, demonstrating that the profit changes depending on the markup scenario used. Note that the Changing Cells and Result Cells are shown using their cell references. In this example, they have been left so that they can be referred against the table above, however in practice, it will make the report easier to read if the cells have been named.



Scenario Summary	Current Values:	50% Markup	100% Markup	150% Markup	200% Markup
Changing Cells:					
$D$2	50%	50%	100%	150%	200%
Result Cells:					
$B$2	£5.00	£5.00	£10.00	£15.00	£20.00

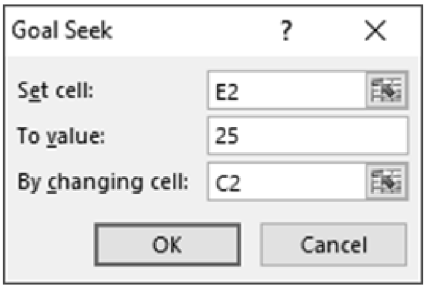
Note :

Current values column represents values of changing cells at time scenario summary report was created. Changing cells for each scenario are highlighted in gray.

Goal Seek

The scenario manager is good for known variables, however, sometimes it is desirable to work backwards. Using the same example of selling 100 apples, we might want to know what markup we would need to use in order to achieve a profit of £25. Goal Seek can be used to answer exactly these sorts of questions.

Goal Seek can be found on the Data ribbon, under What-If Analysis. It simply asks for three parameters. Set cell refers to the cell we want to contain the goal value. In this case it is E2, the cell showing the profit on apples. The value in the To value box should represent the goal, in this case 25, representing the desired £25 profit. Finally, the By changing cell box should show the cell reference for the cell to be manipulated in order to achieve the goal, in this case the markup.



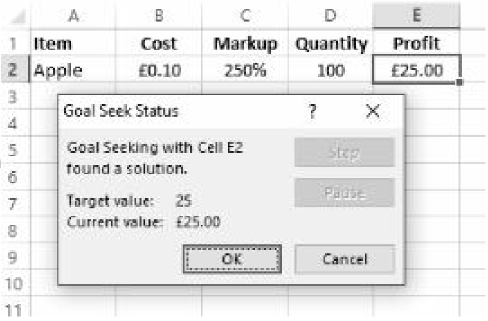
Goal Seek

Set cell: E2

To value: 25

By changing cell: C2

Upon pressing OK, Excel will look for a solution, displaying a dialog box like the one below when it has finished. Note that in the example below, cell C2 has been updated to show a 250% markup and cell E2 showing the profit has updated accordingly. Pressing OK will confirm these changes and commit them to the worksheet, while pressing cancel will see the cells revert back to their previous values.



	A	B	C	D	E
1	Item	Cost	Markup	Quantity	Profit
2	Apple	£0.10	250%	100	£25.00

Goal Seek Status

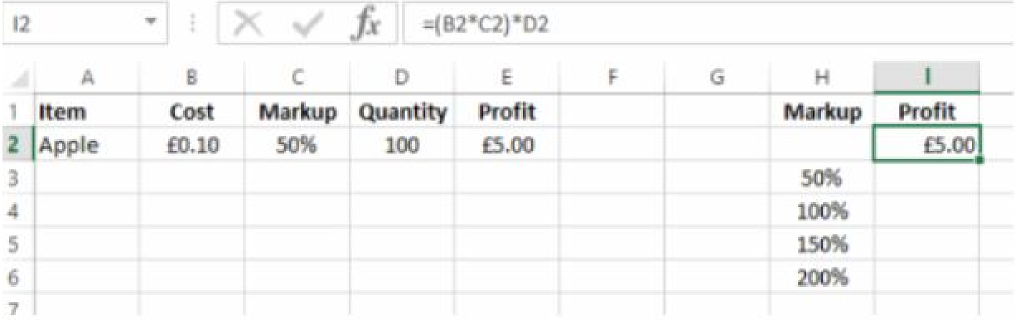
Goal Seeking with Cell E2 found a solution.

Target value: 25

Current value: £25.00

Data Table

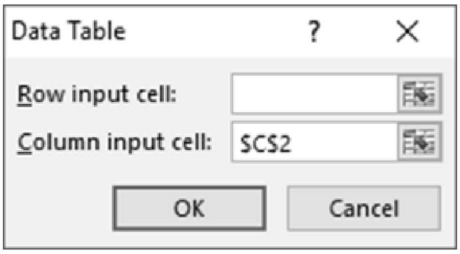
Scenario summaries give a table showing data from various scenarios, however, they do not update, if the data they are based upon changes. In the apple sales example, we might not need scenarios based on the cost price, as this is a non-controllable factor, yet it could still change in the future. The scenario summaries based upon it would not change if this was updated, whereas a data table would.



I2 =(B2*C2)*D2

	A	B	C	D	E	F	G	H	I
1	Item	Cost	Markup	Quantity	Profit			Markup	Profit
2	Apple	£0.10	50%	100	£5.00				£5.00
3								50%	
4								100%	
5								150%	
6								200%	
7									

To populate the data table, select the entire table (in the case of this example this would be cells H2 to I6) and navigate to What-If Analysis on the Data ribbon, and choose Data Table from the drop down menu. On the dialog box that appears, select the cell containing the variable as the column input cell. This tells Excel to use the value in the input column instead of this cell in the formula performing the calculation. Press OK, and Excel will populate the rest of the data table.

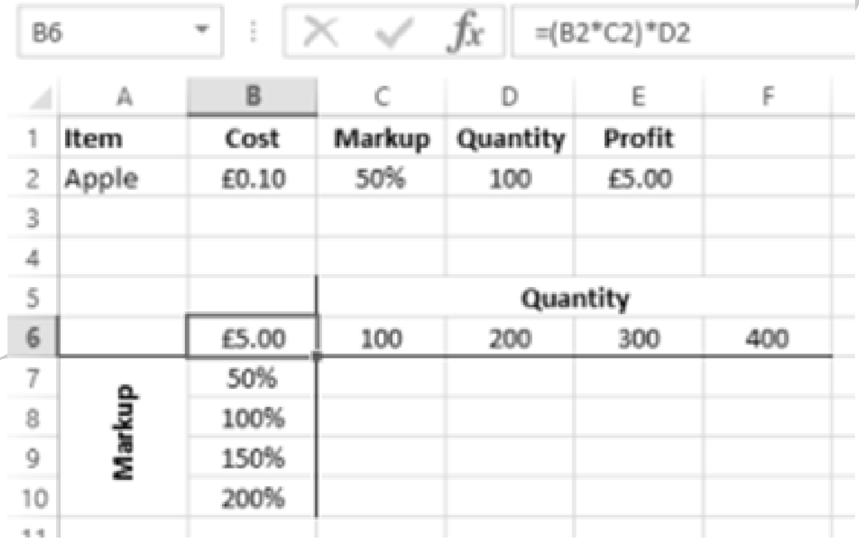


Data Table

Row input cell:

Column input cell: $C$2

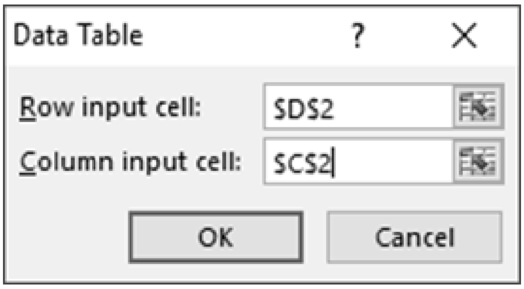
A two variable data table works in much the same way, however the layout is slightly different. Instead of consisting of two columns, there should be one column containing the values for the first variable, with a row containing the values for the second variable. The formula should go in the top left corner, where the two meet.



B6 =(B2*C2)*D2

	A	B	C	D	E	F
1	Item	Cost	Markup	Quantity	Profit	
2	Apple	£0.10	50%	100	£5.00	
3						
4						
5			Quantity			
6		£5.00	100	200	300	400
7	Markup	50%				
8		100%				
9		150%				
10		200%				

As with a single variable data table, highlight the entire table, so B6 to F10 in the example, and select Data Table from the What-If Analysis Drop down on the Data ribbon. This time, enter the cells containing both the row and column variables in the formula. Press OK, and Excel will populate the table.



Data Table

Row input cell: $D$2

Column input cell: $C$2

Q9. Discuss the steps involved in generating a scenarios report in excel 2013.

Ans :

In Microsoft excel 2013, user can generate two types of reports.

They are,

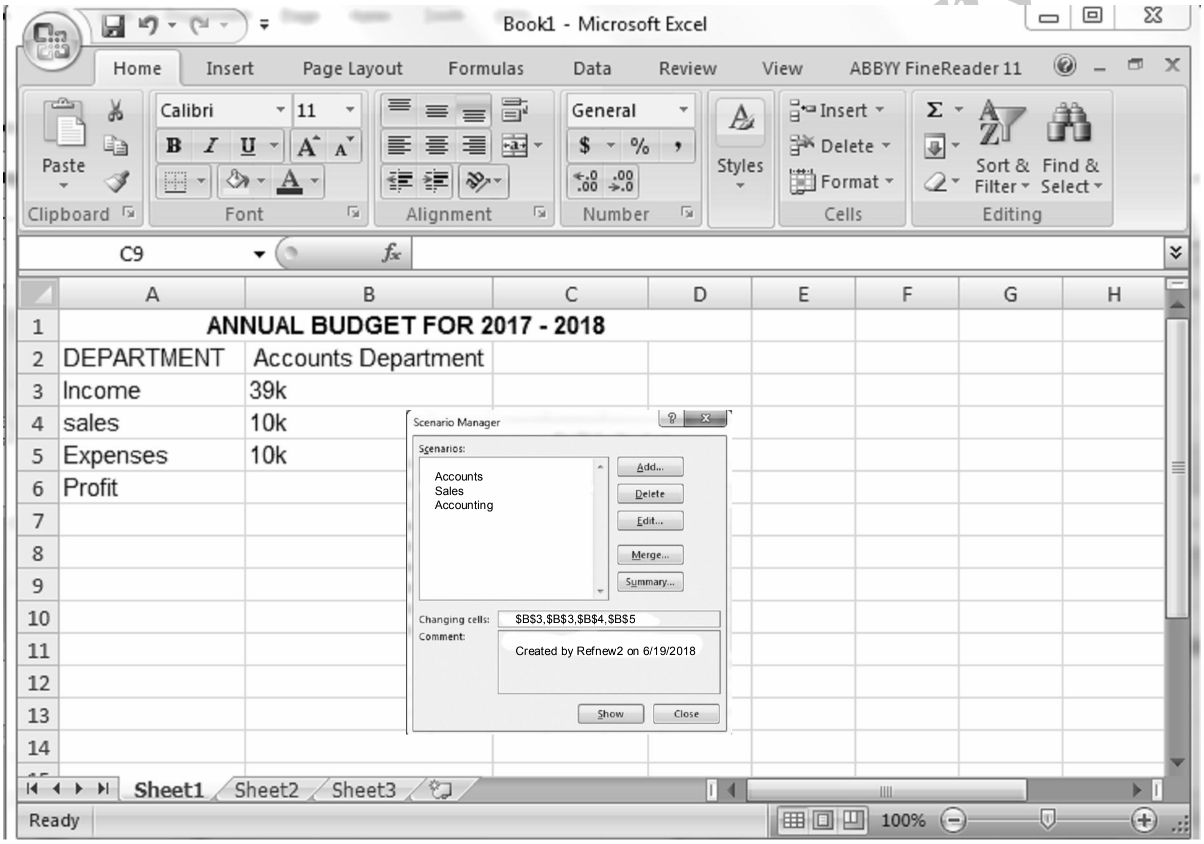
(i) Scenario summary report

(ii) Scenario pivot table report.

The former one follows the structure of worksheet outline and it is used in cases involving less complexity while the latter follows the structure of pivot table and it is used in cases where scenarios are defined with numerous result cells. The following steps are involved in generating scenario report,

1. Click on 'Data' tab and then click on 'What-if Analysis' command present under the 'Data Tools' group.

2. Select 'Scenario Manager' option from drop down menu. A 'Scenario Manager' dialog box appears on screen.



	A	B
1	ANNUAL BUDGET FOR 2017 - 2018	
2	DEPARTMENT	Accounts Department
3	Income	39k
4	sales	10k
5	Expenses	10k
6	Profit	

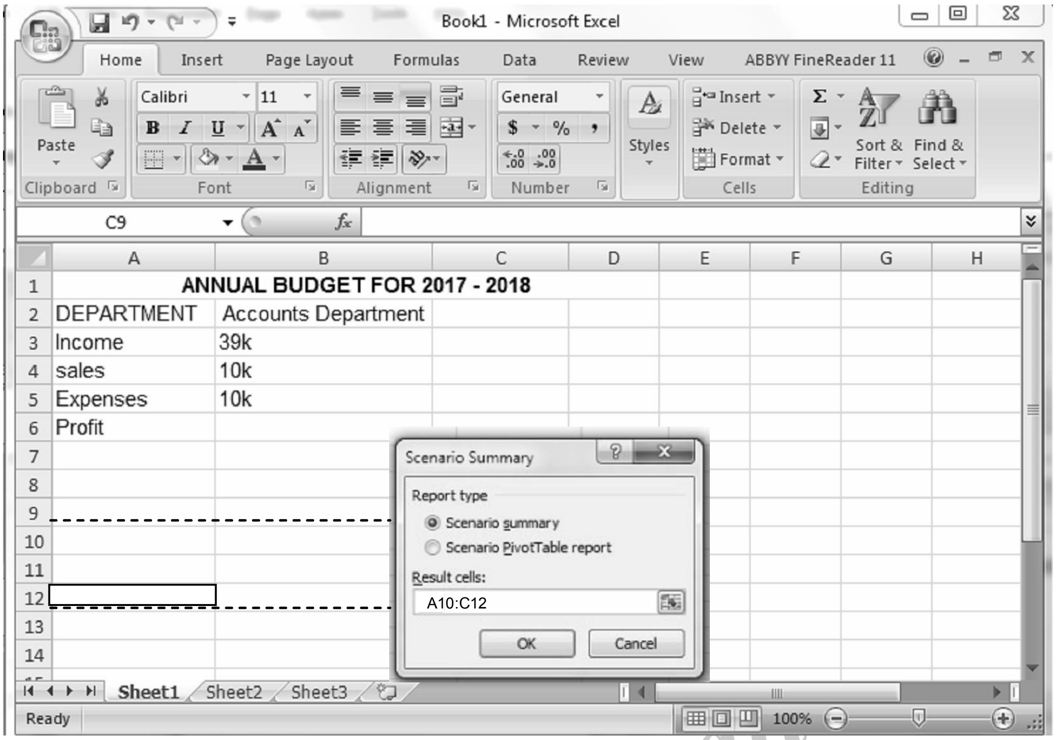
Scenario Manager

Scenarios:
Accounts
Sales
Accounting

Changing cells: $B$3,$B$3,$B$4,$B$5

Comment: Created by Refnew2 on 6/19/2018

3. Click on 'Summary' button in scenario manager dialog box. Which display's a 'Scenario summary' dialog box. Go to 'Report Type' group and check mark the option beside 'Scenario Summary' option to generate scenario summary report or check mark the option beside 'Scenario Pivot Table Report' to generate scenario pivot table report.



Book1 - Microsoft Excel

C9

	A	B
1	ANNUAL BUDGET FOR 2017 - 2018	
2	DEPARTMENT	Accounts Department
3	Income	39k
4	sales	10k
5	Expenses	10k
6	Profit	

Scenario Summary

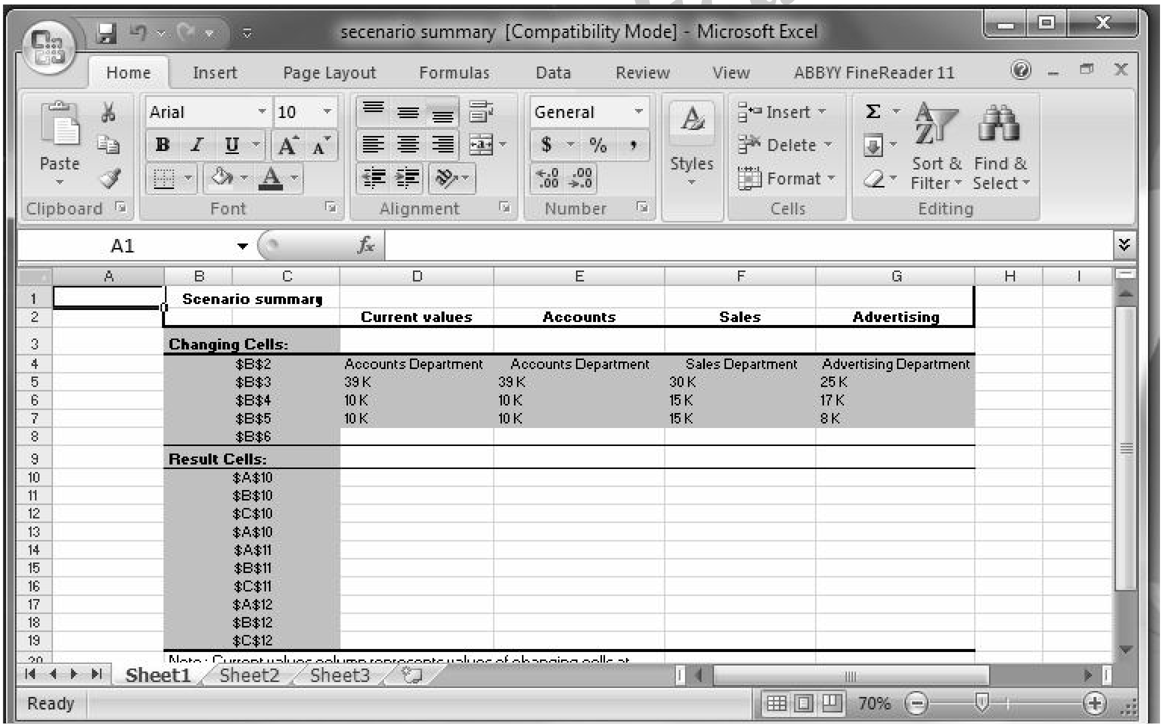
Report type

● Scenario summary

○ Scenario PivotTable report

Result cells: A10:C12

4. Go to results text box and enter the cell reference of the location of report to be generated.

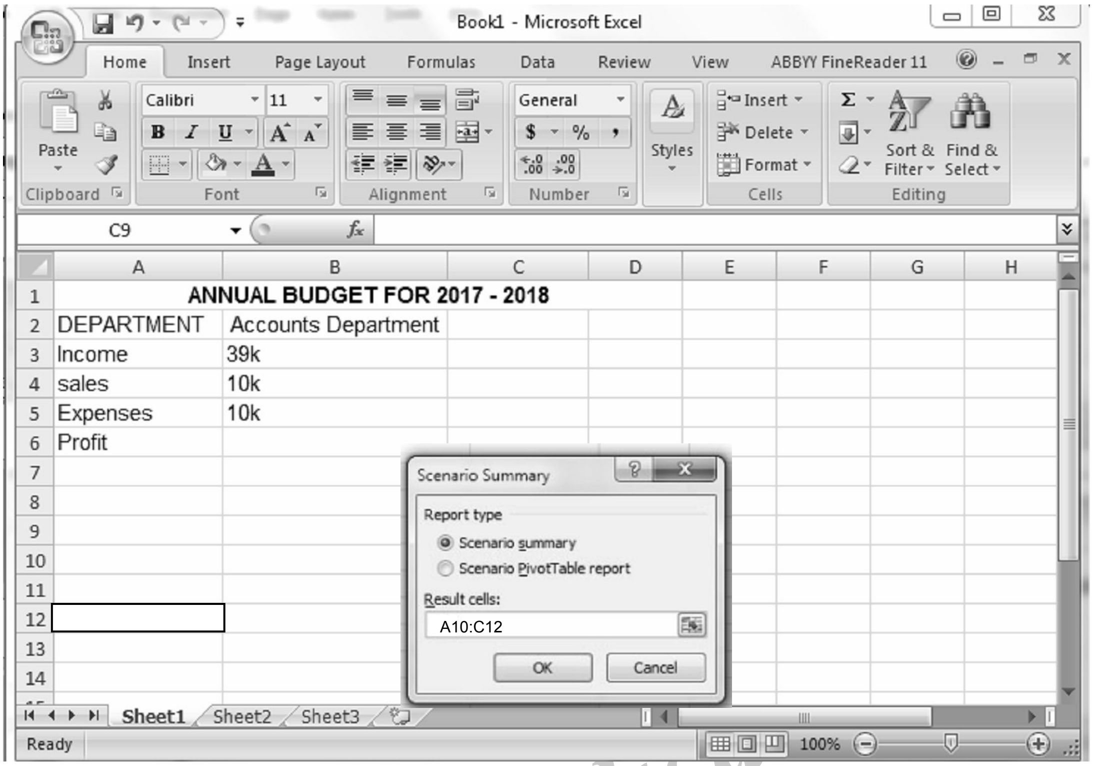


secenario summary [Compatibility Mode] - Microsoft Excel

A1

	B	C	D	E	F	G
1	Scenario summary					
2			Current values	Accounts	Sales	Advertising
3	Changing Cells:					
4		$B$2	Accounts Department	Accounts Department	Sales Department	Advertising Department
5		$B$3	39 K	39 K	30 K	25 K
6		$B$4	10 K	10 K	15 K	17 K
7		$B$5	10 K	10 K	15 K	8 K
8		$B$6				
9	Result Cells:					
10		$A$10				
11		$B$10				
12		$C$10				
13		$A$10				
14		$A$11				
15		$B$11				
16		$C$11				
17		$A$12				
18		$B$12				
19		$C$12				
20	Note : Current values column represents values of changing cells at					

5. Click on "OK" button, and click on the scenario summary option this creates a report type scenario summary in the new worksheet named 'scenario summary' and it appears besided the current worksheet. On the other hand, if user selects report of type scenario pivot table report then pivot table is generated in view worksheet named 'Summary Pivot Table' and it appears beside the current worksheet as follows.



C9

	A	B	C
1	ANNUAL BUDGET FOR 2017 - 2018		
2	DEPARTMENT	Accounts Department	
3	Income	39k	
4	sales	10k	
5	Expenses	10k	
6	Profit		

Scenario Summary

Report type

● Scenario summary

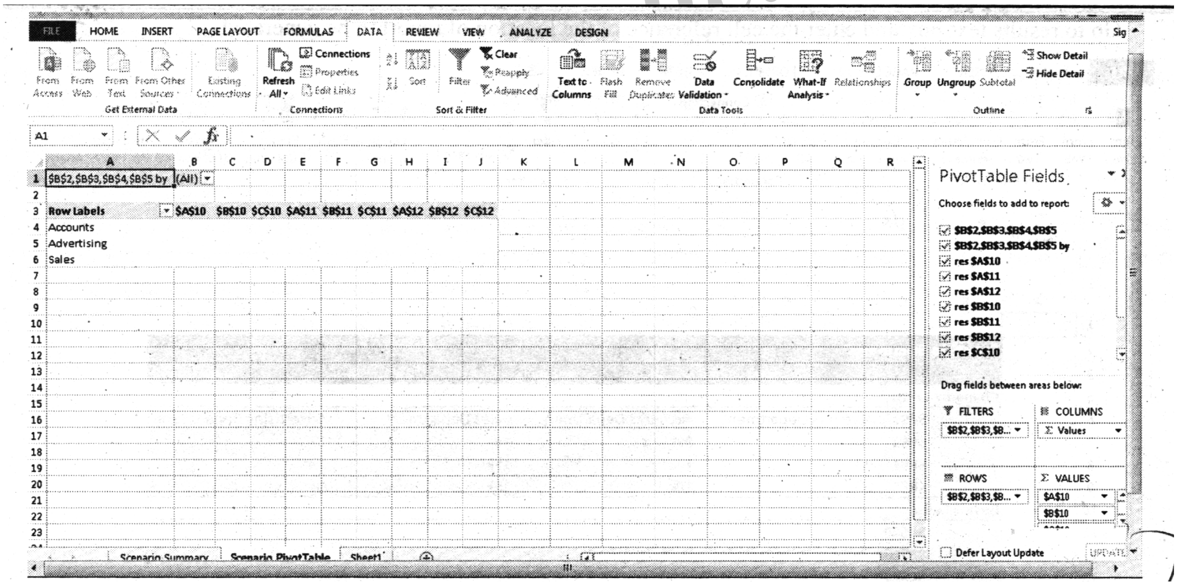
○ Scenario PivotTable report

Result cells:

A10:C12

OK Cancel

Sheet1 Sheet2 Sheet3



A1

	A	B	C	D	E	F	G	H	I	J
1	$B$2,$B$3,$B$4,$B$5 by	(All)								
2										
3	Row Labels	$A$10	$B$10	$C$10	$A$11	$B$11	$C$11	$A$12	$B$12	$C$12
4	Accounts									
5	Advertising									
6	Sales									

PivotTable Fields

Choose fields to add to report:

☑ $B$2,$B$3,$B$4,$B$5
☑ $B$2,$B$3,$B$4,$B$5 by
☑ res $A$10
☑ res $A$11
☑ res $A$12
☑ res $B$10
☑ res $B$11
☑ res $B$12
☑ res $C$10

Drag fields between areas below:

FILTERS: $B$2,$B$3,$B...
COLUMNS: Σ Values
ROWS: $B$2,$B$3,$B...
VALUES: $A$10, $B$10

5.5 Verification And Validation

Q10. What is verification explain with its methods ?

Ans :

Definition

The process of evaluating software to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase.

Verification is a static practice of verifying documents, design, code and program. It includes all the activities associated with producing high quality software: inspection, design analysis and specification analysis. It is a relatively objective process.

Verification will help to determine whether the software is of high quality, but it will not ensure that the system is useful. Verification is concerned with whether the system is well-engineered and error-free.

Methods of Verification

Static Testing, a software testing technique in which the software is tested without executing the code. It has two parts as listed below:

1. **Static Review**

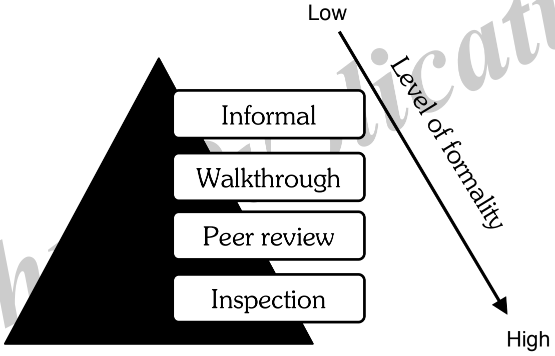
   Typically used to find and eliminate errors or ambiguities in documents such as requirements, design, test cases, etc.

2. **Static Analysis**

   The code written by developers are analysed (usually by tools) for structural defects that may lead to defects.

Types of Static Reviews

The types of reviews can be given by a simple diagram.



1. **Walkthrough**

- It is not a formal process.
- It is led by the authors.
- Author guide the participants through the document according to his or her thought process to achieve a common understanding and to gather feedback.
- Useful for the people if they are not from the software discipline, who are not used to or cannot easily understand software development process.
- Is especially useful for higher level documents like requirement specification, etc.

The Goals of a Walkthrough

- To present the documents both within and outside the software discipline in order to gather the information regarding the topic under documentation.
- To explain or do the knowledge transfer and evaluate the contents of the document
- To achieve a common understanding and to gather feedback.
- To examine and discuss the validity of the proposed solutions.

2. **Peer Review**

- It is less formal review
- It is led by the trained moderator but can also be led by a technical expert
- It is often performed as without management participation
- Defects are found by the experts (such as architects, designers, key users) who focus on the content of the document.
- In practice, technical reviews vary from quite informal to very formal

The Goals of the Peer Review are

- To ensure that an early stage the technical concepts are used correctly
- To access the value of technical concepts and alternatives in the product.
- To have consistency in the use and representation of technical concepts
- To inform participants about the technical content of the document

3. **Inspection**

- It is the most formal review type
- It is led by the trained moderators
- During inspection the documents are prepared and checked thoroughly by the reviewers before the meeting
- It involves peers to examine the product
- A separate preparation is carried out during which the product is examined and the defects are found.
- The defects found are documented in a logging list or issue log.
- A formal follow-up is carried out by the moderator applying exit criteria.

The goals of inspection are

- It helps the author to improve the quality of the document under inspection.
- It removes defects efficiently and as early as possible.
- It improve product quality.
- It create common understanding by exchanging information.
- It learn from defects found and prevent the occurrence of similar defects.

Q11. What is Validation explain with its methods?

Ans :

Definition

The process of evaluating software during or at the end of the development process to determine whether it satisfies specified requirements.

Validation is the process of evaluating the final product to check whether the software meets the customer expectations and requirements. It is a dynamic mechanism of validating and testing the actual product.

Methods of Validation

Dynamic Testing

In Dynamic Testing Technique software is executed using a set of input values and its output is then examined and compared to what is expected. Dynamic execution is applied as a technique to detect defects and to determine quality attributes of the code. Dynamic Testing and Static Testing are complementary methods, as they tend to find different types of defects effectively and efficiently. But as it does not start early in the Software Development Life Cycle hence it definitely increases the cost of fixing defects.

It is done during Validation Process evaluating the finished product.

Dynamic Techniques are subdivided into three more categories:

1. Specification Based Testing : This includes both Functional Testing and Non Functional Testing.
2. Structure Based Testing
3. Experience Based Testing

Specification Based Testing Technique is also known as Behavior Based Testing and Black Box Testing techniques because in this testers view the software as a black-box. As they have no knowledge of how the system or component is structured inside the box. In essence, the tester is only concentrating on what the software does, not how it does it.

Both Functional Testing and Non-Functional Testing is type of Specification Based Testing.

Specification Based Test Design Technique uses the specification of the program as the point of reference for test data selection and adequacy. A specification can be any thing like a written document, collection of use cases, a set of models or a prototype.

Types of Specification Based Testing Techniques

1. **Equivalence Partitioning**

   Software Testing technique that divides the input data of a software unit into partitions of equivalent data from which test cases can be derived.

2. **Boundary Value Analysis**

   Software Testing technique in which tests are designed to include representatives of boundary values in a range.

3. **Decision Tables**

   Software Testing technique in which tests are more focused on business logic or business rules. A decision table is a good way to deal with combinations of inputs.

4. **State Transiting**

   Software Testing technique which is used when the system is defined in terms of a finite number of states and the transitions between the states is governed by the rules of the system.

Structure based testing is also referred to as white-box testing. In this technique, the knowledge of the code and internal architecture of the system is required to carry out the testing.

Structure Based Testing Techniques

The different types of structure based test design or the white box testing techniques are.

- **Statement Testing**

  Statement testing is a white box testing approach in which test scripts are designed to execute code statements. The statement coverage is the measure of the percentage of statements of code executed by the test scripts out of the total code statements in the application. The statement coverage is the least preferred metric for checking test coverage.

- **Decision testing/Branch testing**

  Decision testing or branch testing is a white box testing approach in which test coverage is measured by the percentage of decision points(e.g. if-else conditions) executed out of the total decision points in the application.

- **Condition Testing**

  Testing the condition outcomes(TRUE or FALSE). So, getting 100% condition coverage required exercising each condition for both TRUE and FALSE results using test scripts(For n conditions we will have 2n test scripts).

- **Multiple Condition Testing**

  Testing the different combinations of condition outcomes. Hence for 100% coverage we will have 2 n test scripts. This is very exhaustive and very difficult to achieve 100% coverage.

- **Condition Determination Testing**

  It is an optimized way of multiple condition testing in which the combinations which doesn't affect the outcomes are discarded.

- **Path Testing**

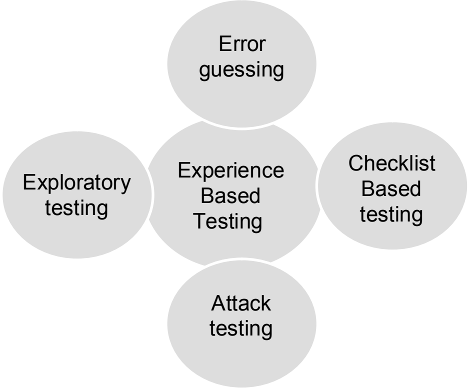
  Testing the independent paths in the system(paths are executable statements from entry to exit points).

- It is the technique of executing testing activities with the help of experience gained through several years of churning. Basically, a tester verifies and validates the software product quality using his/her past experience of testing the similar type of product in the respective domain.

Q12. Explain different Experience Based Testing Techniques.

Ans :

Generally, there are four types of testing techniques where experience of a tester plays the major role in performing testing. These are:



1. **Error Guessing**

   It's a simple technique of guessing and detecting the potential defects that may creep into the software product. In this technique, a tester makes use of his skills, acquired knowledge and past experience to identify the vulnerable areas of the software product that are likely to be affected by the bugs.

   Error guessing technique may be considered as a risk analysis method, where an experienced tester applies his wisdom and gained experience, to spot the areas or functionalities of the software product that are likely to be tainted with the potential defects. What could not be found through formal testing technique may be spotted through error guessing. However, it is preferred that the formal testing technique should be followed by the error guessing technique.

2. **Checklist Based Testing**

   In this technique, experienced tester based on his past experience prepares the checklist, which work as a manual to direct the testing process. The checklist is of high and standard level and consistently reminds the tester of what to be tested. Checklist prepared by a tester is not the static and the final list, i.e. changes may be brought into it in proportion to needs & requirements, occurring during the course of testing.

3. **Exploratory Testing**

   Exploratory testing is a testing technique and simultaneously, a progressive learning approach to perform maximum testing with the minimal planning. During the course of exploratory testing, a tester constantly studies and analyzes the software product and accordingly applies his skills, traits and experience to develop strategy and test cases to perform and carry out the necessary testing.

4. **Attack Testing**

   Attack testing is a testing technique, close enough to negative testing but not similar to it. In attack testing, faults or failures especially those pertaining to security features is introduced in the software product. Generally, all the features and functionalities along with the interfaces, which are responsible for the interaction with the external interfaces, applications, system or any external elements, are being considered during the attack testing. Mainly, external interfaces, database interface, APIs and operating system are being targeted for the attacks.

Q13. Explain the Difference between Validation and Verification.

Ans :

Verification	Validation
1. Verification is a static practice of verifying documents, design, code and program.	Validation is a dynamic mechanism of validating and testing the actual product.
2. It does not involve executing the code.	It always involves executing the code.
3. It is human based checking of documents and files.	It is computer based execution of program.
4. Verification uses methods like inspections, reviews, walkthroughs, and Desk-checking etc.	Validation uses methods like black box (functional) testing, gray box testing, and white box (structural) testing etc.
5. Verification is to check whether the software conforms to specifications.	Validation is to check whether software meets the customer expectations and requirements.
6. It can catch errors that validation cannot catch. It is low level exercise.	It can catch errors that verification cannot catch. It is High Level Exercise.
7. Target is requirements specification, application and software architecture, high level, complete design, and database design etc.	Target is actual product-a unit, a module, a bent of integrated modules, and effective final product.
8. Verification is done by QA team to ensure that the software is as per the specifications in the SRS document.	Validation is carried out with the involvement of testing team.
9. It generally comes first-done before validation.	It generally follows after verification.

5.6 Advantages and Disadvantages of Simulation

Q14. Explain the advantages and dis advantages of simulation?

Ans :

Advantages

1. Simulation is best suited to analyze complex and large practical problems when it is not possible to solve them through a mathematical method.
2. Simulation is flexible, hence changes in the system variables can be made to select the best solution among the various alternatives.
3. In simulation, the experiments are carried out with the model without disturbing the system.
4. Policy decisions can be made much faster by knowing the options well in advance and by reducing the risk of experimenting in the real system.
5. Study the behavior of a system without building it.
6. Results are accurate in general, compared to analytical model.
7. Help to find un-expected phenomenon, behavior of the system.
8. Easy to perform "What-If" analysis.

Disadvantages

1. Simulation does not generate optimal solutions.
2. It may take a long time to develop a good simulation model.
3. In certain cases simulation models can be very expensive.
4. The decision-maker must provide all information (depending on the model) about the constraints and conditions for examination, as simulation does not give the answers by itself.
5. Expensive to build a simulation model.
6. Expensive to conduct simulation.

Sometimes it is difficult to interpret the simulation results.

5.7 Risk Analysis

Q15. Explain about risk analysis with its benefits and uses?

Ans :

Risk analysis is the process of identifying and analyzing potential issues that could negatively impact key business initiatives or critical projects in order to help organizations avoid or mitigate those risks.

Performing a risk analysis includes considering the probability of adverse events caused by either natural processes, like severe storms, earthquakes or floods, or adverse events caused by malicious or inadvertent human activities; an important part of risk analysis is identifying the potential for harm from these events, as well as the likelihood that they will occur.

Enterprises and other organizations use risk analysis to:

1. Anticipate and reduce the effect of harmful results from adverse events.
2. Evaluate whether the potential risks of a project are balanced by its benefits to aid in the decision process when evaluating whether to move forward with the project.
3. Plan responses for technology or equipment failure or loss from adverse events, both natural and human-caused; and
4. Identify the impact of and prepare for changes in the enterprise environment, including the likelihood of new competitors entering the market or changes to government regulatory policy.

Benefits of Risk Analysis

Organizations must understand the risks associated with the use of their information systems to effectively and efficiently protect their information assets.

Risk analysis can help an organization improve its security in a number of ways. Depending on the type and extent of the risk analysis, organizations can use the results to help:

1. Identify, rate and compare the overall impact of risks to the organization, in terms of both financial and organizational impacts.
2. Identify gaps in security and determine the next steps to eliminate the weaknesses and strengthen security.
3. Enhance communication and decision-making processes as they relate to information security.
4. Improve security policies and procedures and develop cost-effective methods for implementing these information security policies and procedures.
5. Put security controls in place to mitigate the most important risks.
6. Increase employee awareness about security measures and risks by highlighting best practices during the risk analysis process; and
7. Understand the financial impacts of potential security risks.

Done well, risk analysis is an important tool for managing costs associated with risks, as well as for aiding an organization's decision-making process.

Q16. Explain the steps risk analysis process.

Ans :

The risk analysis process usually follows these basic steps:

1. **Conduct a Risk Assessment Survey**

   This first step, getting input from management and department heads, is critical to the risk assessment process. The risk assessment survey is a way to begin documenting specific risks or threats within each department.

2. **Identify the Risks**

   The reason for performing risk assessment is to evaluate an IT system or other aspect of the organization and then ask: What are the risks to the software, hardware, data and IT employees? What are the possible adverse events that could occur, such as human error, fire, flooding or earthquakes? What is the potential that the integrity of the system will be compromised or that it won't be available?

3. **Analyze the Risks**

   Once the risks are identified, the risk analysis process should determine the likelihood that each risk will occur, as well as the consequences linked to each risk and how they might affect the objectives of a project.

4. **Develop a Risk Management Plan**

   Based on an analysis of which assets are valuable and which threats will probably affect those assets negatively, the risk analysis should produce control recommendations that can be used to mitigate, transfer, accept or avoid the risk.

5. **Implement the Risk Management Plan**

   The ultimate goal of risk assessment is to implement measures to remove or reduce the risks. Starting with the highest-priority risk, resolve or at least mitigate each risk so it's no longer a threat.

6. **Monitor the Risks**

   The ongoing process of identifying, treating and managing risks should be an important part of any risk analysis process.

The focus of the analysis, as well as the format of the results, will vary depending on the type of risk analysis being carried out.

Q17. Discuss in detail various techniques of risk analysis.

Ans :

Some of the techniques used for analyzing risk in small and medium sized projects are,

1. Break-even analysis
2. Monte-Carlo simulation
3. Decision tree analysis
4. Sensitivity analysis
5. Game theory

1. **Break-Even Analysis**

   Break-even Analysis refers to the study of cost-volume profit analysis. In the true sense, it refers to the analysis of costs and their possible impact on revenues and volume of the firm. In other words break-even analysis is concerned with the determination of particular volume at which firm's cost will be equal to its revenue/profits.

2. **Monte-Carlo Simulation**

   Monte-Carlo Simulation technique involves conducting continuous experiments on the model with known probability distribution in order to draw random samples using random numbers. If the model cannot be described by a probability distribution, then an empirical probability distribution can be constructed. In general, the problem is solved by simulating the data with the generated random numbers. This involves use of two things. One is the model that represents the system under consideration and two the mechanism to simulate the model.

3. **Decision Tree Analysis**

   The risky investment proposals can be ascertain with the help of a technique known as 'Decision Tree Approach'. A decision tree is an analytical technique and a diagrammatic representation which is in the form of a tree. It represents the importance, possibilities and interrelationship of all the possible outcomes.

4. **Sensitivity Analysis**

   Before investing in any project investor first think about the returns on his/her investment.

As future is always uncertain, investor estimate the cash flows. But, it is not possible for investor to forecast accurately. In order to avoid the estimation errors, sensitivity analysis can be used. Sensitivity analysis avoid estimation errors by using possible outcomes in evaluating a project.

The outcomes associated with the project are classified into different cash flows by sensitivity analysis.

(i) Optimistic (the best)

(ii) Most likely (the expected)

(iii) Pessimistic (the worst).

Eventhough, sensitivity analysis is able to forecast variability of returns, it lacks behind in providing the chance of occurrence of those returns. Therefore, probabilities are assigned to the cash flows to provide more accurate calculations of the variability of cash flows. Usually probability of getting cash flow estimates is between zero and one (i.e., 0 and 1). For example, if expected cash flow has probability of occurrence as 0.7, then chances of getting that cash flow is 70%.

The evaluation of variability of returns involve two steps.

Step-1: Assignment of Probabilities

Probabilities can be assigned either by an objective way or by subjective way. When a same situation has occurred in the past then by considering that experience, probabilities are assigned, it is called as objective probability. When the assignment is done by considering personal views but not by the experience which has already occurred, it is known as subjective probability.

Step-2: Estimation of Expected Return

The probabilities which are assigned are multiplied with estimated cash flows to get expected value of a project i.e., weighted average return. The returns are expressed in monetary terms.

5. Game Theory

Game theory is one of the mathematical theories having general characteristics of a competitive situations. It is also called as 'Theory of games' or 'Competitive strategies'. The game theory helps the individuals or organizations having conflicting objectives to make effective decisions. It is suitable for the situations in which two players are trying to win the game. For example, chess, candidates fighting for an election, two enemies planning war tactics etc. In these cases, the decision taken by one decision maker influences the decision taken by the other decision makers and final results are mainly influenced by the decisions of all the parties.

Properties of Game Theory

The properties of game theory are,

(i) Chance or Strategy

In a game, if activities are determined by skill, it is said to be a "game of strategy". If activities are determined by chance, it is a game of chance.

(ii) Number of Persons

A game is called an n-person game, if the number of persons playing in a game is *n*. The person means an individual or a group aiming at a particular objective.

(iii) Number of Activities

The number of activities in a game may be finite or infinite.

(iv) Number of Alternatives

The number of alternatives in a game may be finite or infinite.

(v) Information to the Players about the Past Activities of Other Players

Information to the players about the past activities of other payers may be completely available, partly available or not available at all.

(vi) Pay-off

A quantitative measure of satisfaction a person gets at the end of each play is called as "pay-off'.

5.8 Decision Tree Analysis

Q18. Explain Decision tree analysis with an example.

Ans :

A Decision Tree Analysis is a graphic representation of various alternative solutions that are available to solve a problem. The manner of illustrating often proves to be decisive when making a choice. A Decision Tree Analysis is created by answering a number of questions that are continued after each affirmative or negative answer until a final choice can be made.

Decision Making Process

A Decision Tree Analysis is a scientific model and is often used in the **decision making process** of organizations. When making a decision, the management already envisages alternative ideas and solutions. By using a decision tree, the alternative solutions and possible choices are illustrated graphically as a result of which it becomes easier to make a well-informed choice. This graphic representation is characterized by a tree-like structure in which the problems in decision making can be seen in the form of a flowchart, each with branches for alternative choices.

What if

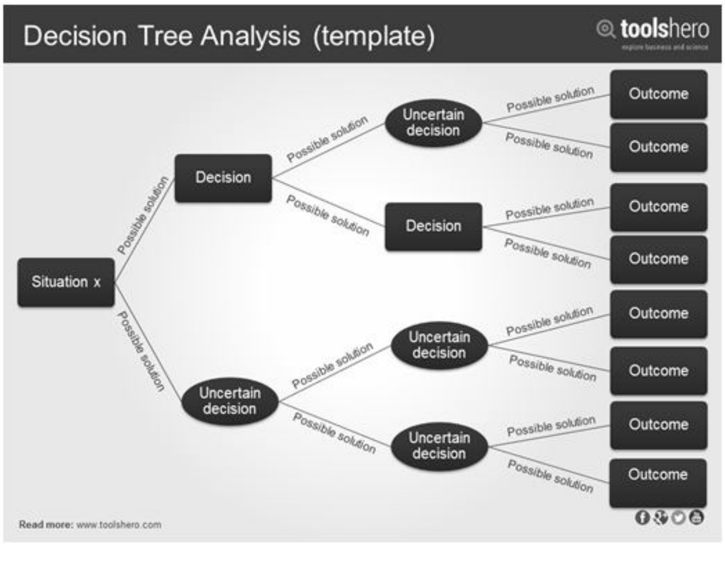
The Decision Tree Analysis makes good use of the 'what if' thought. There are several alternatives that consider both the possible risks and benefits that are brought about by certain choices. The possible alternatives are also made clearly visible and therefore the decision tree provides clarity with respect to the consequences of any decisions that will be made.

Representation

There are several ways in which a decision tree can be represented. The Decision Tree Analysis is commonly represented by lines, squares and circles. The squares represent decisions, the lines represent consequences and the circles represent uncertain outcomes. By keeping the lines as far apart as possible, there will be plenty of space to add new considerations and ideas.

The representation of the decision tree can be created in four steps:

1. Describe the decision that needs to be made in the square.
2. Draw various lines from the square and write possible solutions on each of the lines.
3. Put the outcome of the solution at the end of the line. Uncertain or unclear decisions are put in a circle. When a solution leads to a new decision, the latter can be put in a new square.
4. Each of the squares and circles are reviewed critically so that a final choice can be made.



Example

Suppose a commercial company wishes to increase its sales and the associated profits in the next year.

The different alternatives can then be mapped out by using a decision tree. There are two choice for both increase of sales and profits: 1- expansion of advertising expenditure and 2- expansion of sales activities. This creates two branches. Two new choices arise from choice 1, namely 1-1 a new advertising agency and 1-2 using the services of the existing advertising agency. Choice 2 presents two follow-up choices in turn; 2-1-working with agents or 2-2- using its own sales force.

The branching continues.

The following alternatives from 1-1 are:

1-1-1 The budget will increase by 10% -> end result: sales up 6%, profits up 2%

1-1-2 The budget will increase by 5% -> end result: sales up 4%, profits up 1.5 %

The alternatives that arise from 1.2:

1-2-1 The budget increases by 10% -> end result: sales up 5%, profits up 2.5%

1-2-2 The budget increases by 5 % -> end result: a sales up 4%, profits up 12%

From 2.1 possibly follows:

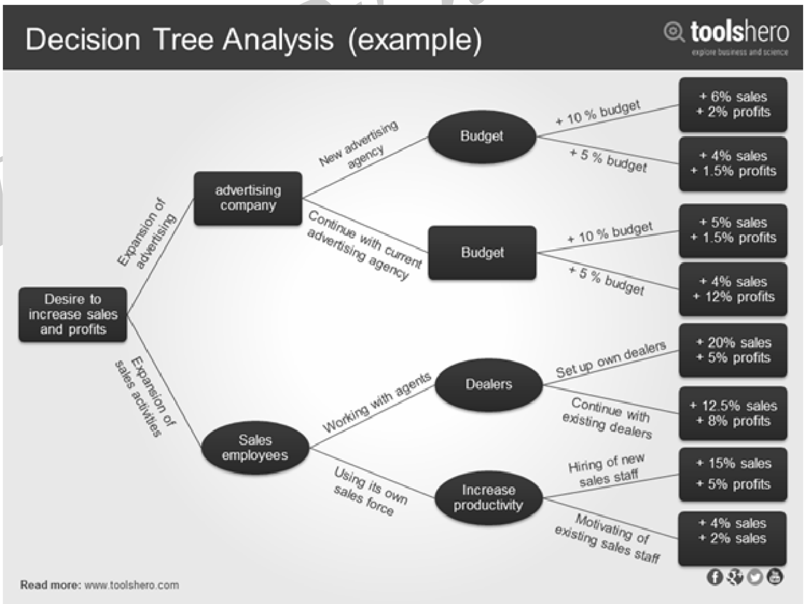
2-1-1 Set up with own dealers -> end result: sales up 20%, profits up 5%

2-1-2 Working with existing dealers -> end result: sales up 12.5%, profit up 8%

From 2.2 possibly follows:

2-2-1 Hiring of new sales staff -> end result: sales up 15%, profits up 5%

2-2-2 Motivating of existing sales staff -> end result: sales up 4%, profits up 2%.



The above example illustrates that, in all likelihood, the company will opt for 1-2-2, because the forecast of this decision is that profits will increase by 12%.

The Decision Tree Analysis is particularly useful in situations in which it is considered desirable to develop various alternatives of decisions in a structured manner as this will present a clear substantiation. This method is increasingly used by medical practitioners and technicians as it enables them to make a diagnosis or determine car problems.

PROBLEMS

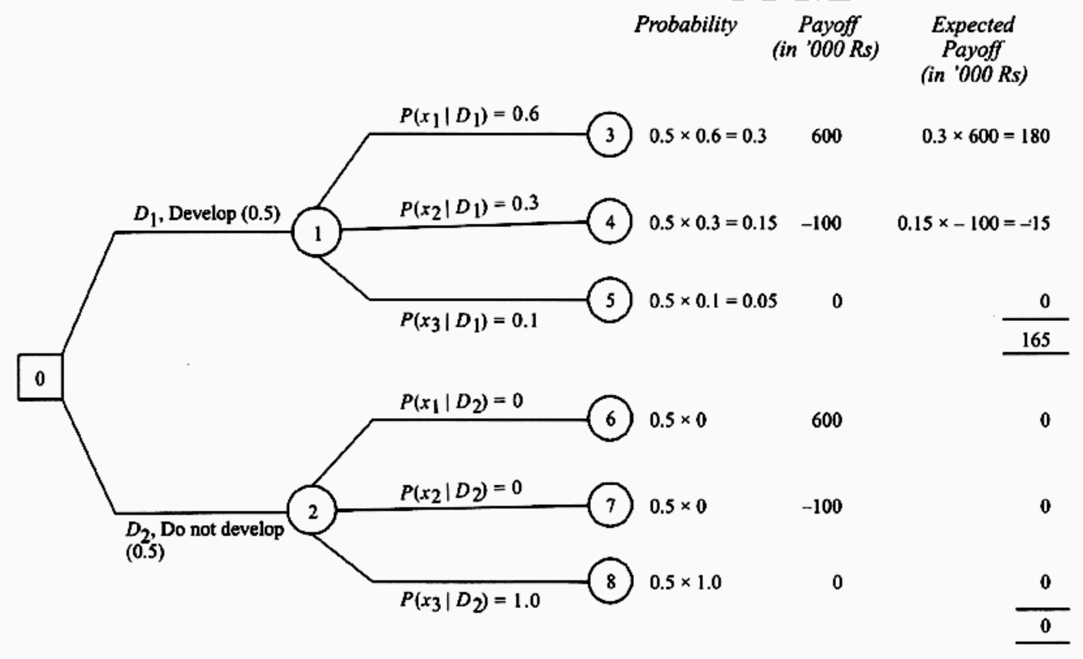
1. You are given the following estimates concerning a Research and Development programme :

Decisions D_i	Probability of Decision D_i Given Research R $P(D_i \setminus R)$	Outcome Number	Probability of Outcome x_i Given D_i $P(x_i \setminus D_i)$	Payoff Value of Outcome, x_i (Rs '000)
Develop	0.5	1	0.6	600
		2	0.3	-100
		3	0.1	0
Do not develop	0.5	1	0.0	600
		2	0.0	-100
		3	1.0	0

Construct and evaluate the decision tree diagram for the above data. Show your workings for evaluation.

Solution :

The decision tree of the given problem along with necessary calculations is shown in Fig. below:



2. A glass factory that specializes in crystal is developing a substantial backlog and for this the firm's management is considering three courses of action: To arrange for subcontracting (S_1), to begin overtime production (S_2), and to construct new facilities (S_3). The correct choice depends largely upon the future demand, which may be low, medium, or high. By consensus, management ranks the respective probabilities as 0.10, 0.50 and 0.40. A cost analysis reveals the effect upon the profits. This is shown in the table below :

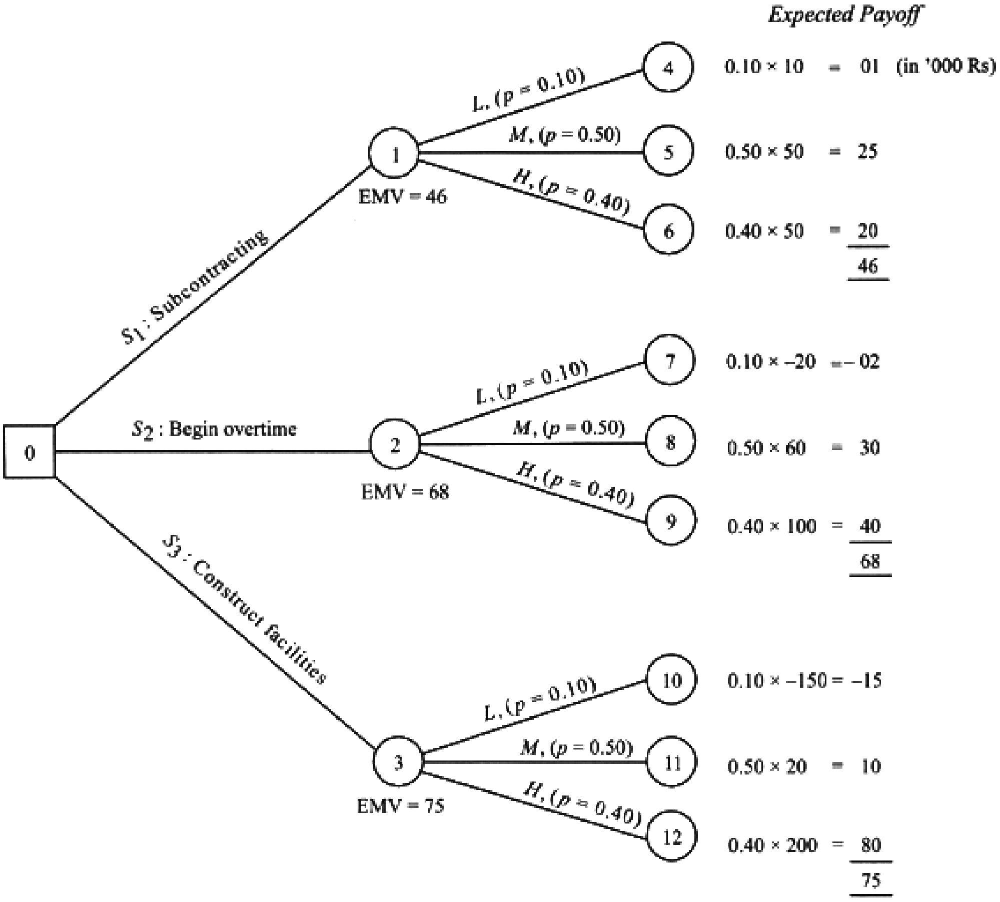
Demand	Probability	Course of Action		
		S_1 (Subcontracting)	S_2 (Begin Overtime)	S_3 (Construct Facilities)
Low (L)	0.10	10	-20	-150
Medium (M)	0.50	50	60	20
High (H)	0.40	50	100	200

Show this decision situation in the form of a decision tree and indicate the most preferred decision and its corresponding expected value.

Solution :

A decision tree that represents possible courses of action and states of nature is shown in Fig. below. In order to analyze the tree, we start working backwards from the end branches.

The most preferred decision at the decision node 0 is found by calculating the expected value of each decision branch and selecting the path (course of action) that has the highest value.



Since node 3 has the highest EMV, therefore, the decision at node 0 will be to choose the course of action S_3, i.e. construct new facilities.

3. A businessman has tow independent investment portfolios A and B, available to him, but he lacks the capital to undertake both of them simultaneously. He can either choose A first and then stop, or if A is not successful, then take, B or vice versa. The probability of success of A is 0.6, while for B it is 0.4. Both investment schemes require an initial capital outlay of Rs. 10,000 and both return nothing if the venture proves to be unsuccessful. Successful completion of A will return Rs. 20,000 (over cost) and successful completion of B will return Rs 24,000 (over cost). Draw a decision tree in order to determine the best strategy.

Solution :

The decision tree based on the given information is shown in figure. The evaluation of each chance node and decision is given in table.

Decision point	Outcomes (Rs)	Probability	Conditional value	Expected value
D_3 (i) Accept A	Success	0.6	20,000	12,000
	Failure	0.4	– 10,000	– 4,000
				8,000
(ii) Stop	-	-	-	0
D_2 (i) Accept B	Success	0.4	24,000	9,600
	Failure	0.6	– 10,000	– 6,000
				3,600
(ii) Stop	-	-	-	0
D_1 (i) Accept B	Success	0.6	20,000 + 3,600 = 23,600	14,160
	Failure	0.4	– 10,000	– 4,000
				10,160
(ii) Accept B	Success	0.4	24,000 + 8,000 = 32,000	12,800
	Failure	0.4	– 10,000	– 6,000
				6,800
(iii) Do nothing	-	-	-	0

Table : Evaluation of Decision and Chance Nodes

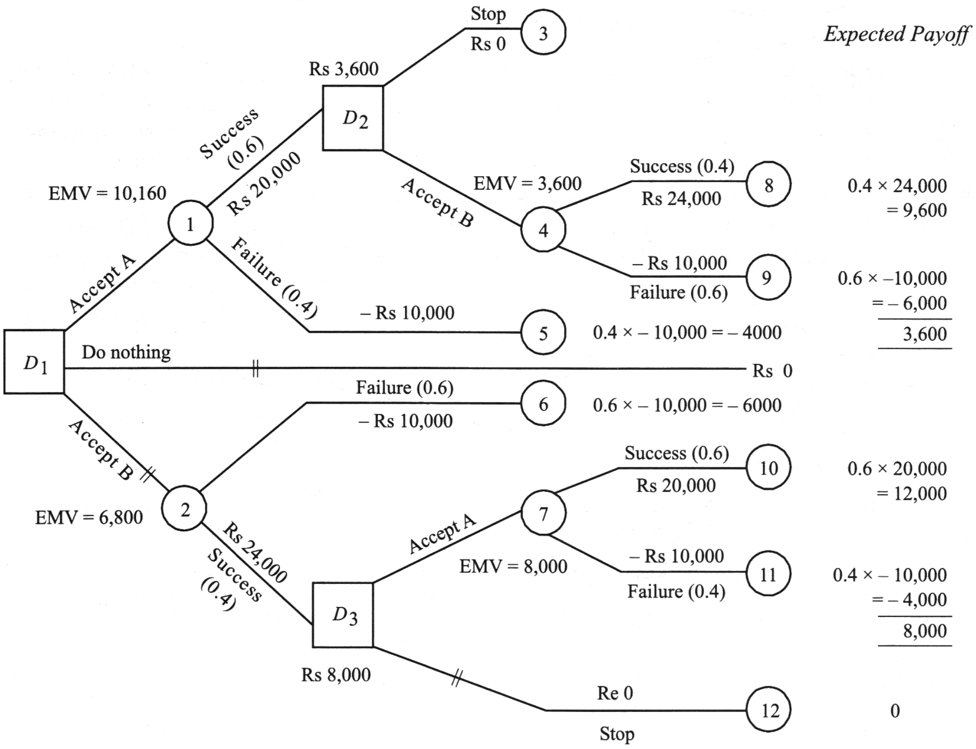


Figure : Decision Tree

Since the EMV = Rs 10,160 at node D1 is highest, therefore the best strategy is to accept course of action A first and if A is successful, then accept B.

4. The Oil India Corporation (OIC) is wondering whether to go for an offshore oil drilling contract that is to be awarded in Bombay High. If OIC bid, value would be Rs 600 million with a 65 per cent change of gaining the contract. The OIC may set up a new drilling operation or move the already existing operation, which has already proved successful for a new site. The probability of success and expected returns are as follows:

Outcome	New Drilling Operation		Existing Operation	
	Probability	Expected revenue (Rs. million)	Probability	Expected revenue (Rs. million)
Success	0.75	800	0.85	700
Failure	0.25	200	0.15	350

If the Corporation do not bid or lose the contract, they can use Rs. 600 million to modernize their operation. This would result in a return of either 5 per cent or 8 per cent on the sum invested with probabilities 0.45 and 0.55. (Assume that all costs and revenue have been discounted to present value).

(a) Construct a decision tree for the problem shown clearly the courses of action

(b) By applying an appropriate decision criterion recommend whether or not the Oil India Corporation should bid the contract.

Solution :

The decision tree based on the given information is shown in figure. the evaluation of each chance node and decision node is given the table.

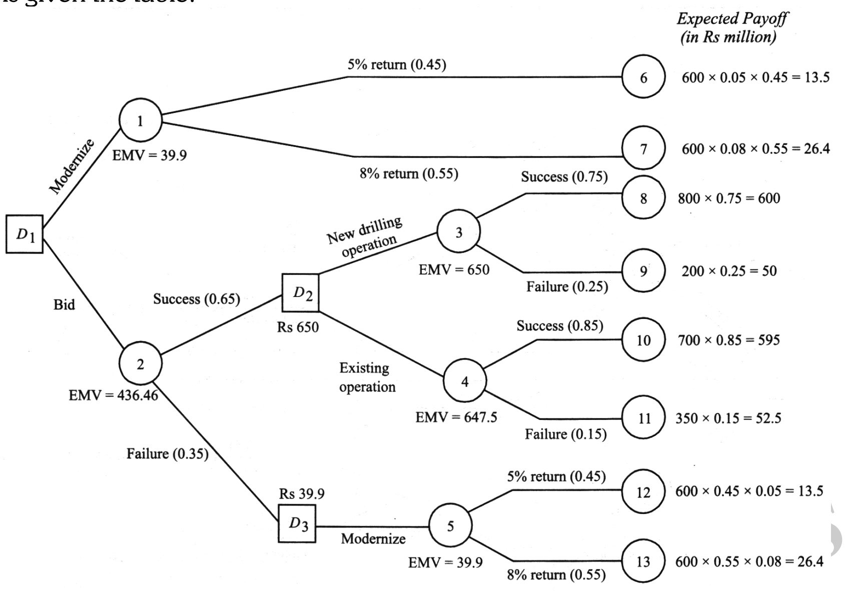


Figure : Decision Tree

Decision point	Outcomes	Probability	Conditional value	Expected value (Rs)
D_3 (i) Modernize	5% return	0.45	600 × 0.05 = 30	30 × 0.45 = 135
	8% return	0.55	600 × 0.08 = 48	48 × 0.55 = 26.4
				39.9
D_2 (i) Undertake new	Success	0.75	800	600
	Failure	0.25	200	50
				650
(ii) Move existing operation	Success	0.85	700	595
	Failure	0.15	350	52.5
				647.5
D_1 (i) Modernize	5% return	0.45	600 × 0.05 = 30	30 × 0.45 = 135
	8% return	0.55	600 × 0.08 = 48	48 × 0.55 = 26.4
				39.9
(ii) Bid	Success	0.65	650	422.50
	Failure	0.35	39.9	13.96
				436.46

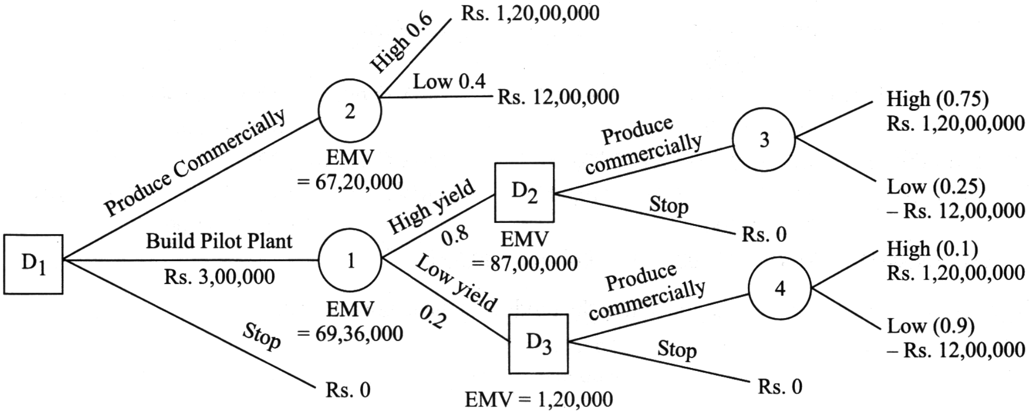
Table : Evaluation of Decision and Chance Nodes

Since EMV, Rs 436.46 at event node 2 is highest, therefor the best decision node D_1 is to decide for bid and if successful establish a new drilling operation.

5. A large steel manufacturing company has three options with regard to production (i) produce commercially (ii) build pilot plant (iii) stop producing steel. The management has estimated that their pilot plant, if built, has 0.8 chance of high yield and 0.2 chance of low yield. If the pilot plant does show a high yield, management assigns a probability of 0.75 that the commercial plant will also have a high yield. If the pilot plant shows a low yield, there is only a 0.1 chance that the commercial plant will show a high yield. Finally, management's best assessment of the yield on a commercial-size plant without building a pilot plant first has a 0.6 chance of high yield. A pilot plant will cost Rs. 3,00,000. The profits earned under high and low yield conditions are Rs. 1,20,000 and Rs. 12,00,000 respectively. Find the optimum decision for the company.

Solution :

A decision tree representing possible courses of action and states of nature are shown in fig. In order to analyse the tree, we proceed backward from the end branches.



Figure

EMV (Node 3) = 0.75 × 1,20,00,000 – 0.25 × 12,00,000 = Rs. 87,00,000

EMV (Node 4) = 0.1 × 1,20,00,000 – 0.9 × 12,00,000

= 12,00,000 – 10,80,000 = Rs. 1,20,000

EMV (Node 1) = 0.8 × 87,00,000 – 0.2 × 1,20,000 = Rs. 69,36,000

EMV (Node 2) = 0.6 × 1,20,00,000 – 0.4 × 12,00,000

= 72,00,000 – 4,800,000 = Rs. 67,20,000

EMV (Node D_2) = Rs. 87,00,000

EMV (Node D_3) = Rs. 1,20,000

EMV (Node D_1) = 69,36,000 – 3,00,000 = Rs. 87,00,000

Since at decision node D_1 the production cost of Rs. 67,20,000, associated with course of action – Build pilot plant is least, the company should build pilot plant.

6. A businessman has an option of selling a product in domestic market or in export market The available relevant data given below.

Iterms	Export Market	Domestic Market
Probability of selling	0.6	1.0
Probability of keeping delivery schedule (Rs.)	0.8	0.9
Penalty of not meeting delivery schedule (Rs.)	50,000	10,000
Selling price (Rs.)	9,00,000	8,00,000
Cost of third party inspection (Rs.)	30,000	Nill
Probability of collection of sale amount	0.9	0.9

If the product is not sold in export market, it can always be sold in domestic market. There are no other implication like interest and time.

(a) Draw the decision tree using the data given above.

(b) Should the businessman go for selling the product in the export market? Justify your answer.

Solution :

(i) The decision tree representing possible curses of action and states of nature is shown in figure.

The revenue generated form the stage of product in export market is equal to the selling price minus inspection cost (i.e, Rs. 30,000). The tree is analyzed by moving backward from the end branches.

EMV (Node 9) = Rs. 0.9 × 8,00,000

= 0.9 × 8,00,000 = Rs. 7,20,000

EMV (Node 7) = Rs. 0.9 × 8,70,000 – 0.1 × 30,000

= Rs. 7,80,000

EMV (Node 4) = Rs. 0.9 × 7,20,000 + 0.1 × 7,10,000

= Rs. 7,19,000

EMV (Node 5) = Rs. 0.9 × 8,00,000 = Rs. 72,20,000

EMV (Node 1) = Rs. 0.6 × 7,70,000 + 0.4 × 7,19,000

= Rs. 7,49,600

EMV (Node D_1) = Max. {7,49,600; 7,19,000} = Rs. 7,49,600

EMV (Node 10) = Rs. 7,20,000

EMV (Node 8) = Rs. 7,80,000

EMV (Node 3) = Rs. 0.8 × 7,80,000 + 0.2 × 7,30,000

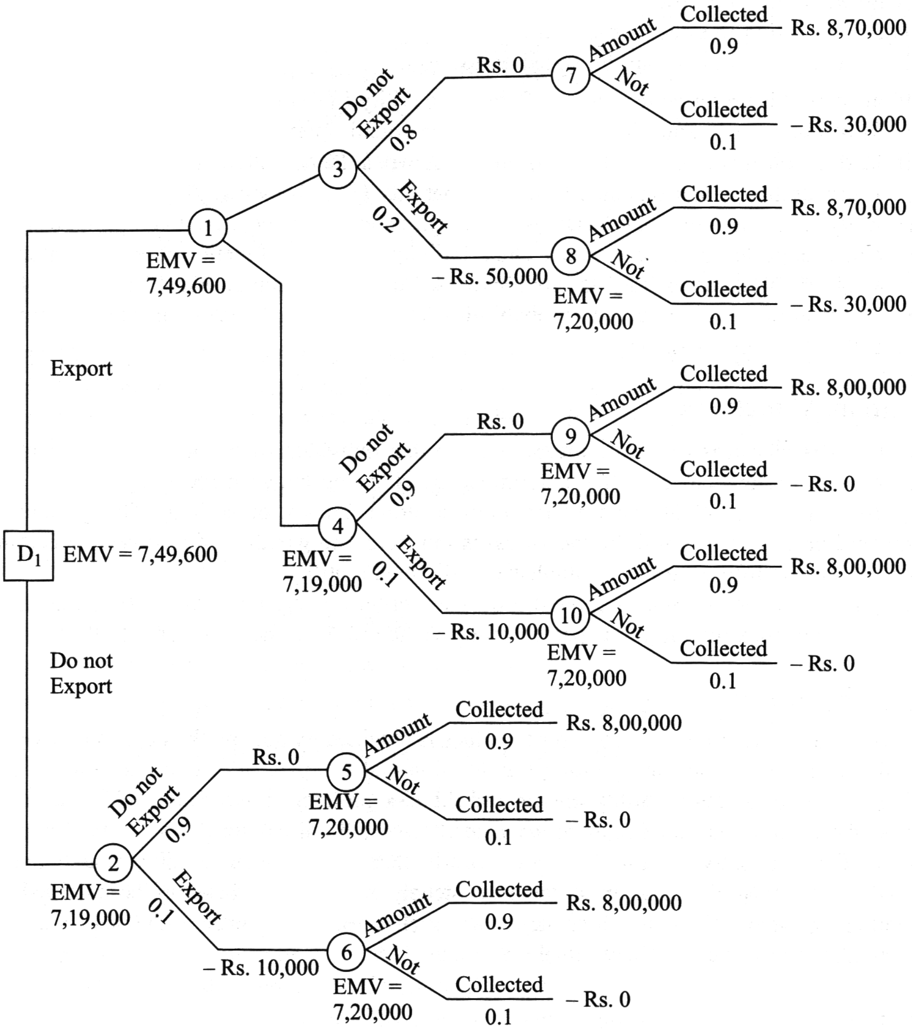
= Rs. 7,70,000

EMV (Node 6) = 7,20,00

EMV (Node 2) = 0.9 × 7,20,000 + 0.1 × 7,10,000

= Rs. 7,19,000

Hence, the businessman should go for selling the product in the export market.



Figure

Short Question and Answers

1. Simulations?

Ans :

Simulation is an imitation of the operation of a real-world process or system. The act of simulating something first requires that a model be developed; this model represents the key characteristics, behaviors and functions of the selected physical or abstract system or process. The model represents the system itself, whereas the simulation represents the operation of the system over time.

Simulation is used in many contexts, such as simulation of technology for performance optimization, safety engineering, testing, training, education, and video games. Often, computer experiments are used to study simulation models. Simulation is also used with scientific modelling of natural systems or human systems to gain insight into their functioning, as in economics. Simulation can be used to show the eventual real effects of alternative conditions and courses of action. Simulation is also used when the real system cannot be engaged, because it may not be accessible, or it may be dangerous or unacceptable to engage, or it is being designed but not yet built, or it may simply not exist.

2. Types of Simulation

Ans :

1. Physical Simulation

It refers to simulation in which physical objects are substituted for the real thing (some circles use the term for computer simulations modelling selected laws of physics, but this article does not). These physical objects are often chosen because they are smaller or cheaper than the actual object or system.

2. Interactive Simulation

It is a special kind of physical simulation, often referred to as a human in the loop simulation, in which physical simulations include human operators, such as in a flight simulator or a driving simulator.

3. Continuous Simulation

It is a simulation where time evolves continuously based on numerical integration of Differential Equations.

4. Discrete Event Simulation

It is a simulation where time evolves along events that represent critical moments, while the values of the variables are not relevant between two of them or result trivial to be computed in case of necessity.

5. Stochastic Simulation

It is a simulation where some variable or process is regulated by stochastic factors and estimated based on Monte Carlo techniques using pseudo-random numbers, so replicated runs from same boundary conditions are expected to produce different results within a specific confidence band.

3. Practical applications of simulation.

Ans :

1. Queueing Problem

The optimal solution for queueing problems can be determined by reducing the effectiveness of queue length, average waiting time, service time, idle time and so on.

In case of uncertain events like bulk arrival, jockeying, balking, reneging etc., Simulation is the only method found to be suitable for optimization. It is also helpful to find approximate waiting time in case of uncertain working conditions. Depending on such information effective decisions can be made.

2. Inventory Problem

In many inventory problems, the analytical method will not be useful to solve and to find the optimal solution because of the complexities, for example, in case of storage problem, due to the existence of uncertain demand and supply distribution tends to be complex. In such situation, with the help of simulation model policy of reorder and order quantity can be designed. The main purpose of simulation method is to reduce the total cost of inventory thereby reducing carrying, holding, storage costs etc. By reducing such costs, optimization can be obtained.

3. **Capital Budgeting**

According to prof. David B. Hertz, simulation can also be used in capital budgeting i.e., to estimate the proposal yielding the best return for its investment. By adopting simulation model, different alternatives of financial evaluation can be determined and out of that, the best optimal solution can be selected which has less risk due to various uncertain factors like Selling Price (SP), market growth rate, market size etc., on financial parameters and yields maximum return on investment.

4. **Financial Planning Problems**

Due to varying situations and different stages of development, the companies are facing problems in financial decision making. If inappropriate decisions are made, then businesses will suffer huge losses. In order to avoid such losses, simulation methods are used. Hence the management needs to be careful while taking decisions during financial planning of the firm. With simulation model, the managers can simulate the overall behavior of the system under given situations if all inputs and decisions had actually occurred and must select those projects which yields maximum returns to the firm.

4. Random number generation

Ans :

Random numbers are numbers that occur in a sequence such that two conditions are met: (1) the values are uniformly distributed over a defined interval or set, and (2) it is impossible to predict future values based on past or present ones. Random numbers are important in statistical analysis and probability theory.

The most common set from which random numbers are derived is the set of single-digit decimal numbers {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}. The task of generating random digits from this set is not trivial. A common scheme is the selection (by means of a mechanical escape hatch that lets one ball out at a time) of numbered ping-pong balls from a set of 10, one bearing each digit, as the balls are blown about in a container by forced-air jets. This method is popular in lotteries. After each number is selected, the ball with that number is returned to the set, the balls are allowed to blow around for a minute or two, and then another ball is allowed to escape.

Sometimes the digits in the decimal expansions of irrational numbers are used in an attempt to obtain random numbers. Most whole numbers have irrational square roots, so entering a string of six or eight digits into a calculator and then hitting the square root button can provide a sequence of digits that seems random. Other algorithms have been devised that supposedly generate random numbers. The problem with these methods is that they violate condition (2) in the definition of randomness. The existence of any number-generation algorithm produces future values based on past and/or current ones. Digits or numbers generated in this manner are called pseudorandom.

5. Monte Carlo simulation

Ans :

Monte Carlo simulation is a computerized mathematical technique that allows people to account for risk in quantitative analysis and decision making. The technique is used by professionals in such widely disparate fields as finance, project management, energy, manufacturing, engineering, research and development, insurance, oil & gas, transportation, and the environment.

Monte Carlo simulation furnishes the decision-maker with a range of possible outcomes and the probabilities they will occur for any choice of action. It shows the extreme possibilities - the outcomes of going for broke and for the most conservative decision - along with all possible consequences for middle-of-the-road decisions.

The technique was first used by scientists working on the atom bomb; it was named for Monte Carlo, the Monaco resort town renowned for its casinos. Since its introduction in World War II, Monte Carlo simulation has been used to model a variety of physical and conceptual systems.

6. "What-if Analysis"

Ans :

By using What-If Analysis tools in Excel, you can use several different sets of values in one or more formulas to explore all the various results.

For example, you can do What-If Analysis to build two budgets that each assumes a certain level of revenue. Or, you can specify a result that you want a formula to produce, and then determine what sets of values will produce that result. Excel provides several different tools to help you perform the type of analysis that fits your needs.

What-If Analysis in **Excel** allows you to try out different values (scenarios) for formulas. The following example helps you master what-if analysis quickly and easily.

What-if analysis is a useful way of being able to test out various scenarios in Excel. You can look at these things two different ways.

The first way is to change the input variables and see what impact that has on the output. The scenario manager and data tables work in this way and can be used to answer questions like, what would happen to our profits if the number of units we sell doubled, or what would happen to our profits if the cost price of each of our units sold increased by 10%.

The second way is to say what outcome you would like to have and ask Excel to calculate what change in the inputs would be required to achieve this. The goal seek feature works this way and can answer questions such as how many units of a product need to be sold in order to reach a desired profit level.

7. Methods of Verification

Ans :

Static Testing, a software testing technique in which the software is tested without executing the code. It has two parts as listed below:

1. **Static Review**

   Typically used to find and eliminate errors or ambiguities in documents such as requirements, design, test cases, etc.

2. **Static Analysis**

   The code written by developers are analysed (usually by tools) for structural defects that may lead to defects.

8. Types of Specification Based Testing Techniques

Ans :

1. **Equivalence Partitioning**

   Software Testing technique that divides the input data of a software unit into partitions of equivalent data from which test cases can be derived.

2. **Boundary Value Analysis**

   Software Testing technique in which tests are designed to include representatives of boundary values in a range.

3. **Decision Tables**

   Software Testing technique in which tests are more focused on business logic or business rules. A decision table is a good way to deal with combinations of inputs.

4. **State Transiting**

   Software Testing technique which is used when the system is defined in terms of a finite number of states and the transitions between the states is governed by the rules of the system.

9. Advantages and dis advantages of simulation?

Ans :

Advantages

1. Simulation is best suited to analyze complex and large practical problems when it is not possible to solve them through a mathematical method.

2. Simulation is flexible, hence changes in the system variables can be made to select the best solution among the various alternatives.
3. In simulation, the experiments are carried out with the model without disturbing the system.
4. Policy decisions can be made much faster by knowing the options well in advance and by reducing the risk of experimenting in the real system.
5. Study the behavior of a system without building it.

Disadvantages

1. Simulation does not generate optimal solutions.
2. It may take a long time to develop a good simulation model.
3. In certain cases simulation models can be very expensive.
4. The decision-maker must provide all information (depending on the model) about the constraints and conditions for examination, as simulation does not give the answers by itself.

10. Benefits of Risk Analysis

Ans :

1. Identify, rate and compare the overall impact of risks to the organization, in terms of both financial and organizational impacts.
2. Identify gaps in security and determine the next steps to eliminate the weaknesses and strengthen security.
3. Enhance communication and decision-making processes as they relate to information security.
4. Improve security policies and procedures and develop cost-effective methods for implementing these information security policies and procedures.
5. Put security controls in place to mitigate the most important risks.
6. Increase employee awareness about security measures and risks by highlighting best practices during the risk analysis process; and
7. Understand the financial impacts of potential security risks.

11. Decision tree analysis

Ans :

A Decision Tree Analysis is a graphic representation of various alternative solutions that are available to solve a problem. The manner of illustrating often proves to be decisive when making a choice. A Decision Tree Analysis is created by answering a number of questions that are continued after each affirmative or negative answer until a final choice can be made.

Decision Making Process

A Decision Tree Analysis is a scientific model and is often used in the **decision making process** of organizations. When making a decision, the management already envisages alternative ideas and solutions. By using a decision tree, the alternative solutions and possible choices are illustrated graphically as a result of which it becomes easier to make a well-informed choice. This graphic representation is characterized by a tree-like structure in which the problems in decision making can be seen in the form of a flowchart, each with branches for alternative choices.

What if

The Decision Tree Analysis makes good use of the ‘what if’ thought. There are several alternatives that consider both the possible risks and benefits that are brought about by certain choices. The possible alternatives are also made clearly visible and therefore the decision tree provides clarity with respect to the consequences of any decisions that will be made.

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

MBA III-Semester Examinations

October / November - 2020

R17

DATA ANALYTICS

Time : 2 Hours] [Max. Marks : 75

Note: Answer any **FIVE** questions

All question carry equal marks

ANSWERS

1. How to prepare a Pivot Table by using typical sales data and brief on data visualization tools . **(Unit-I, Q.No.19,15)**
2. What is meant by Big-Data and Business Analytics in practice ? **(Unit-I, Q.No.5,9)**
3. Explain about data modeling approaches and brief on measures of association. **(Unit-II, Q.No.14,5)**
4. Discuss the assumptions of normal distribution and measures of variability. **(Unit-II, Q.No.4)**

Ans :

Normal distribution is the most essential distribution that is used in statistics. It is a continuous distribution which is described by using a bell-shaped curve. It contains two parameters, mean μ and standard deviation σ. When the value of μ varies, the position of distribution on x-axis also varies. Similarly, when the value of μ increases/decreases the distribution either becomes broad or gets concised.

Properties

1. The normal distribution curve can be symmetric and its skewness found to be zero.
2. It contains mean, median and mode equally, Therefore, the half area is found to be above the mean and the other half area is found to be below it.
3. The range of x is found to be unbounded i.e., the curve of the normal distribution is moved from negative to positive and infinity.

Normal Distribution Curve

There exists several shapes of normal distribution curves differentiated only by the parameters such as means and the standard deviation of the sample.

Few shapes of the curve are shown below,

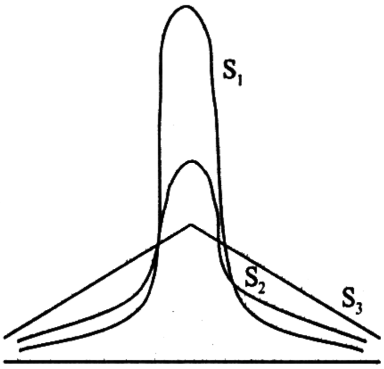


Fig.: Normal Probability Curves with Different Standard Deviation and Same Mean Values

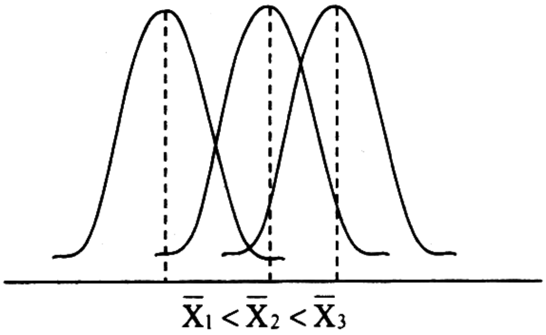
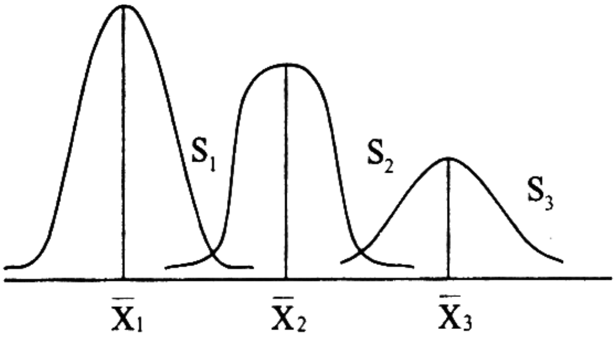


Fig.: Normal Probability Curves with Different Mean and Same Standard Values



Fig,: Normal Probability Curves with Different Mean and Same Standard Deviation Values Area Under Normal Distribution

The distribution of area under normal curve with respect to mean and standard deviation is:

1. The following figure depicts that approximately 68% of all the values in a normally distributed population that lie within 1 standard deviation (S) from the mean.

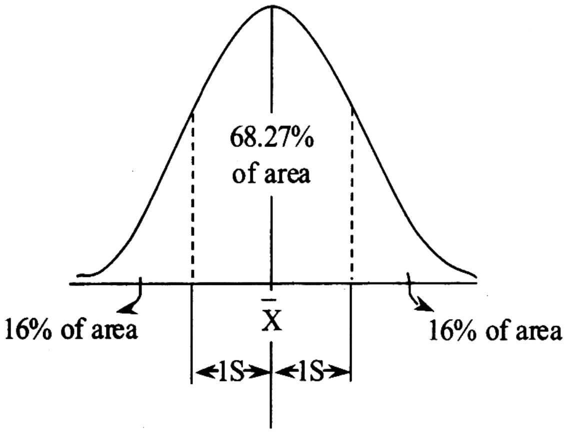


Fig.: Area Under Normal Distribution with 1 Standard Deviation

$\bar{X} \pm 15$ = 68.27% of area

Area covered on both side of the mean = 34.14%

2. The following figure depicts that approximately 95.45% of all the values in a normally distributed population that lie within 2 standard deviation (S) from the mean.

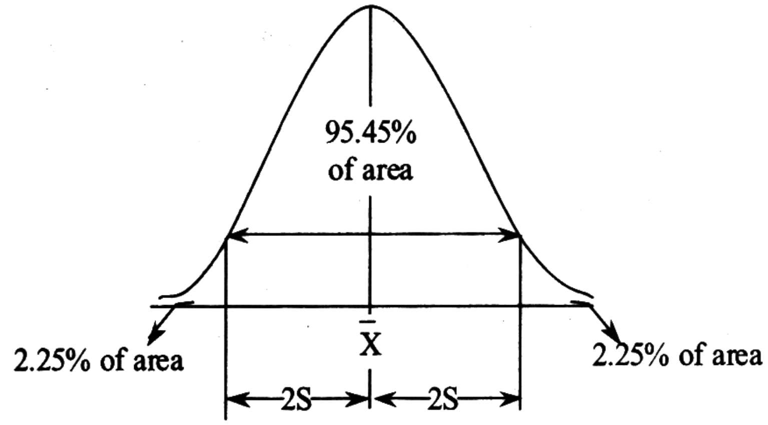


Fig.: Area Under Normal Distribution with 2 Standard Deviations

$\bar{X} \pm 25 = 95.45\%$ of area

Area covered on both side of the mean = 47.73%

3. The following figure depicts that approximately 99.73% of all the values in a normally distributed population lie within 3 standard deviations (*S)* from the mean.

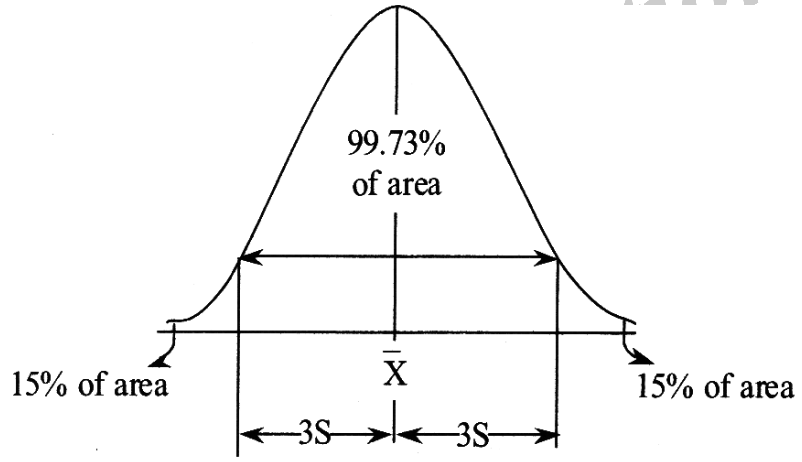


Fig.: Area Under Normal Distribution with 3 Standard Deviations

$\bar{X} \pm 3S = 99.73\%$ of area

Area covered on both sides of the mean = 49.87%

NORM.DISTfx, mean, standard-deviation, cumulative) is the excel function used to determine normal distribution. Cumulative probability F(x) = P(X < x) for a particular mean and deviation.

5. Explain the multiple regression by least squares with an example. **(Unit-III, Q.No.15)**

6. The alibaba Traders company wishes to test whether its three salemen Salim, Basha and Vikram tend to make sales of the same size (or) whether they differ in their selling ability as measured by the average size of their sales. During the last week, there have been 14 sales calls. Salim made 5 calls, Basha made 4 calls and Vikram made 5 calls. The following are the weekly sales records of the three salemen :

Salim Rs.	Basha Rs.	Vikram Rs.
300	600	700
400	300	300
300	300	400
500	400	600
000	---	500

Ans :

Anova : Two-Factor without Replication

Summary	Count	Sum	Average	Variance
Row 1	3	1500	500	70000
Row 2	3	1200	400	10000
Row 3	3	1100	366.6667	3333.333
Row 4	3	1400	466.6667	23333.33
Row 5	3	500	166.6667	83333.33
Column 1	5	1400	280	37000
Column 2	5	1800	360	53000
Column 3	5	2500	500	25000

Source of Variation	SS	Degree of freedom	MS	F	P-value	Critical Ratio
Rows	204000	4	51000	1.59375	0.266086	3.837853
Columns	124000	2	62000	1.9375	0.20598	4.45897
Error	256000	8	32000			
Total	584000	14				

7. Distinguish between unsupervised and supervised learning in Data mining.

Ans :

Parameters	Unsupervised learning	Supervised learning
Process	In a supervised learning model, input and output variables will be given.	In unsupervised learning model, only input data will
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data which is not labeled
Algorithms used	Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees.	Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc.
Computational Complexity	Supervised learning is a simpler method.	Unsupervised learning is computationally complex

8. Discuss the advantages and disadvantages of Simulation in decision making. **(Unit-V, Q.No.14)**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERBAD

M.B.A III - Semester Examination

December - 2019

R17

DATA ANALYTICS

Time : 3 Hours] [Max. Marks : 75

Note : This question paper contains two parts A and B.

Part A is compulsory which carries 25 marks. Answer all questions in Part - A.

Part - B contains of 5 Units. Answer any one full question from each unit.

Each question carries 10 marks and may have a, b, c as sub questions.

PART - A (5 × 5 = 25 Marks)

ANSWERS

1. (a) Explain about the importance of business analytics. **(Unit - I, Q.No. 9)**

(b) Describe about Measures of Variability. **(Unit - II, Q.No. 4)**

(c) What is simple regression ? **(Unit - III, Q.No. 12)**

(d) What are all the association rules ? **(Unit - IV, Q.No. 7)**

(e) What is Monte Carlo Simulation ? **(Unit - V, SQA - 5)**

PART - B (5 × 10 = 50 Marks)

(Essay Type Questions)

2. Explain in detail about statistical methods for summarizing data. **(Unit - I, Q.No. 18)**

(OR)

3. Describe about Pivot Tables. How data is explored using pivot tables? **(Unit - I, Q.No. 19)**

4. Explain in detail about,

(a) Probability distribution **(Unit - II, Q.No. 6)**

(b) Data modeling **(Unit - II, Q.No. 14)**

(c) Continuous probability distribution. **(Unit - II, Q.No. 12)**

(OR)

5. Explain about data modeling and distribution fitting in detail. **(Unit - II, Q.No. 14, 15)**

6. How regression by the method of least squares technique is used ? **(Unit - III, Q.No. 15)**

(OR)

7. Explain in detail about regression with categorical independent variables. **(Unit - III, Q.No. 19)**

8. Describe briefly about,

   (a) Partition data **(Unit - IV, Q.No. 10)**

   (b) Classification accuracy **(Unit - IV, Q.No. 11)**

   (c) Prediction accuracy **(Unit - IV, Q.No. 12)**

(OR)

9. What is cluster analysis ? What are its applications ? **(Unit - IV, Q.No. 6)**

10. Analyze in detail about advantages and disadvantages of simulation techniques. **(Unit - V, Q.No. 14)**

(OR)

11. Explain in detail about

   (a) Risk Analysis **(Unit - V, Q.No. 15)**

   (b) Decision Tree Analysis. **(Unit - V, Q.No. 18)**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERBAD

M.B.A III - Semester Examination

April / May - 2019

R17

DATA ANALYTICS

Time : 3 Hours] [Max. Marks : 75

Note : This question paper contains two parts A and B.

Part A is compulsory which carries 25 marks. Answer all questions in Part - A.

Part - B contains of 5 Units. Answer any one full question from each unit.

Each question carries 10 marks and may have a, b, c as sub questions.

PART - A (5 × 5 = 25 Marks)

ANSWERS

1. (a) What is data ? Explain the importance of Analytics ? **(Unit - I, Q.No. 1,3)**

   (b) What is the best measure of location ? **(Unit - II, Q.No. 2)**

   (c) What is simple and multiple Regressions ? **(Unit - III, Q.No. 12, 14)**

   (d) Define Data mining? Explain the scope of Data mining ? **(Unit - IV, SQA 10)**

   (e) What is simulation ? **(Unit - V, SQA 6)**

PART - B (5 × 10 = 50 Marks)

(Essay Type Questions)

2. Explain the different types of data visualization tools ? **(Unit - I, Q.No. 15)**

(OR)

3. What is business Analytics and its types. **(Unit - I, Q.No. 9)**

4. A random variable X has the following probability function :

X	0	1	2	3	4	5	6	7
P(X)	0	k	2k	2k	3k	k2	2k2	7k2+k

a) Find k

b) Evaluate p[x<6], p[x>=6]

c) If p[x<=c]>1/2 find the minimum value of c.

Sol :

i) Since $\sum_{x=0}^{7} P(x) = 1$, we have

$$k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$10k^2 + 9k - 1 = 0$$

$10k^2 + 10k - k - 1 = 0$

$10k(k + 1) - 1(k + 1) = 0$

$10k(k + 1) - 1(k + 1) = 0$

$(10k-1)(k+1) = 0$

$k = \frac{1}{10} = 0.1 (P(x) \geq 0, \text{ So } k \neq -1)$

ii) $P(x < 6) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) + P(x = 5)$

$0 + k + 2k + 2k + 3k + k^2$

$8k + k^2 \ (k = 0.1)$

$0.8 + 0.01 = 0.81$

$P(x \geq 6) = 1 - P(x<6)$

$= 1 - 0.81 = 0.19$

iii) The required minimum value of k is obtained

$P(x \leq 1) = [P(x=0) + P(x=1)]$

$0 + k = \frac{1}{10} = 0.1$

$P(x \leq 2) = [P(x=0) + P(x=1) + P(x=2)]$

$= \frac{1}{10} + \frac{2}{10} + \frac{3}{10}$

$= 0.3$

$P(x \leq 3) = [P(x=0) + P(x=1) + P(x = 2) + P(x=3)]$

$= 0.3 + 0.2 = 0.5$

$P(x \leq 4) = P(x \leq 3) + P(x=4)$

$0.5 + \frac{3}{10}$

$0.5 + 0.3$

$= 0.8$

$0.8 > 0.5 = \frac{1}{2}$

The minimum value of c for which $P(x \leq c) > \frac{1}{2}$

$c = 4$

(OR)

5. Explain about Random sampling methods with merits and demerits.

Ans :

- **Simple random sampling.** - This method refers to a method having following properties:
  - The population have N objects.
  - The sample have n objects.
  - All possible samples of n objects have equal probability of occurrence.

  One example of simple random sampling is lottery method. Assign each population element a unique number and place the numbers in bowl. Mix the numbers thoroughly. A blind-folded researcher is to select n numbers. Include those population element in the sample whose number has been selected.

- **Stratified sampling** - In this type of sampling method, population is divided into groups called strata based on certain common characteristic like geography. Then samples are selected from each group using simple random sampling method and then survey is conducted on people of those samples.

- **Cluster sampling** - In this type of sampling method, each population member is assigned to a unique group called cluster. A sample cluster is selected using simple random sampling method and then survey is conducted on people of that sample cluster.

- **Multistage sampling** - In such case, combination of different sampling methods at different stages. For example, at first stage, cluster sampling can be used to choose clusters from population and then sample random sampling can be used to choose elements from each cluster for the final sample.

- **Systematic random sampling** - In this type of sampling method, a list of every member of population is created and then first sample element is randomly selected from first k elements. Thereafter, every kth element is selected from the list.

Advantages of Random Sampling

1. **It offers a chance to perform data analysis that has less risk of carrying an error.**

   Random sampling allows researchers to perform an analysis of the data that is collected with a lower margin of error. This is allowed because the sampling occurs within specific boundaries that dictate the sampling process. Because the whole process is randomized, the random sample reflects the entire population and this allows the data to provide accurate insights into specific subject matters.

2. **There is an equal chance of selection.**

   Random sampling allows everyone or everything within a defined region to have an equal chance of being selected. This helps to create more accuracy within the data collected because everyone and everything has a 50/50 opportunity. It is a process that builds an inherent “fairness” into the research being conducted because no previous information about the individuals or items involved are included in the data collection process.

3. **It requires less knowledge to complete the research.**

   A researcher does not need to have specific knowledge about the data being collected to be effective at their job. In random sampling, a question is asked and then answered. An item is reviewed for a specific feature. If the researcher can perform that task and collect the data, then they’ve done their job.

4. It is the simplest form of data collection.

This type of research involves basic observation and recording skills. It requires no basic skills out of the population base or the items being researched. It also removes any classification errors that may be involved if other forms of data collection were being used. Although the simplicity can cause some unintended problems when a sample is not a genuine reflection of the average population being reviewed, the data collected is generally reliable and accurate.

5. Multiple types of randomness can be included to reduce researcher bias.

There are two common approaches that are used for random sampling to limit any potential bias in the data. The first is a lottery method, which involves having a population group drawing to see who will be included and who will not. Researchers can also use random numbers that are assigned to specific individuals and then have a random collection of those number selected to be part of the project.

Disadvantages of Random Sampling

1. No additional knowledge is taken into consideration.

Although random sampling removes an unconscious bias that exists, it does not remove an intentional bias from the process. Researchers can choose regions for random sampling where they believe specific results can be obtained to support their own personal bias. No additional knowledge is given consideration from the random sampling, but the additional knowledge offered by the researcher gathering the data is not always removed.

2. It is a complex and time-consuming method of research.

With random sampling, every person or thing must be individually interviewed or reviewed so that the data can be properly collected. When individuals are in groups, their answers tend to be influenced by the answers of others. This means a researcher must work with every individual on a 1-on-1 basis. This requires more resources, reduces efficiencies, and takes more time than other research methods when it is done correctly.

3. Researchers are required to have experience and a high skill level.

A researcher may not be required to have specific knowledge to conduct random sampling successfully, but they do need to be experienced in the process of data collection. There must be an awareness by the researcher when conducting 1-on-1 interviews that the data being offered is accurate or not. A high skill level is required of the researcher so they can separate accurate data that has been collected from inaccurate data. If that skill is not present, the accuracy of the condusions produced by the offered data may be brought into question.

4. There is an added monetary cost to the process.

Because the research must happen at the individual level, there is an added monetary cost to random sampling when compared to other data collection methods. There is an added time cost that must be included with the research process as well. The results, when collected accurately, can be highly beneficial to those who are going to use the data, but the monetary cost of the research may outweigh the actual gains that can be obtained from solutions created from the data.

5. No guarantee that the results will be universal is offered.

Random sampling is designed to be a representation of a community or demographic, but there is no guarantee that the data collected is reflective of the community on average. In US politics, a random sample might collect 6 Democrats, 3 Republicans, and 1 Independents, though the actual population base might be 6 Republicans, 3 Democrats, and 1 Independent for every 10 people in the community. Asking who they want to be their President would likely have a Democratic candidate in the lead when the whole community would likely prefer the Republican.

6. It requires population grouping to be effective.

If the population being surveyed is diverse in its character and content, or it is widely dispersed, then the information collected may not serve as an accurate representation of the entire population. These issues also make it difficult to contact specific groups or people to have them included in the research or to properly catalog the data so that it can serve its purpose.

6. The 3 samples given below have been obtained from a normal population with equal variance. Test the hypothesis that sample means are equal.

A	8	10	7	14	11
B	7	5	10	9	9
C	12	9	13	12	14

Sol :

A	B	C
8	7	12
10	5	9
7	10	13
14	9	12
11	9	14
50	40	60

Grand Total = 50 + 40 + 60 = 150

Step - I

Null Hypotnesis (H_0) = The sample means are equal

$$H_0 = \mu = \mu_2 = \mu_3$$

Alternative hypotnesis (H_1) = The sample means are not equal.

$$H_1 = \mu_1 \neq \mu_2 \neq \mu_3$$

Level of significance (α) = 0.05 (Assume)

i) Correction factor (C.F) $= \dfrac{(\text{Grand total})^2}{\text{N}}$

$$= \frac{(150)^2}{2} = 1500$$

ii) Total sum of square (T.s.s) $= \varepsilon_x \times \varepsilon_j x_{ij}^2 - \text{CF}$

$(8^2 + 7^2 + 12^2 + 10^2 + 5^2 + 9^2 + 7^2 + 10^2 + 13^2 + 19^2 + 9^2 + 11^2 + 9^2 + 11^2 + 9^2 + 19^2) - \text{CF}$

$$1600 - 1500 = 100$$

iii) Sum of square between samples $\varepsilon_j \dfrac{T_j^2}{nj} - CF$

$$\left(\frac{50^2}{5}+\frac{40^2}{5}+\frac{60^2}{5}\right) - 1500$$

Sum of squares between samples = 1540 –1500 = 40

iv) Sum of squares within samples

(SSW) = TSS – SSD

100 – 40 = 60

Sources of variation	Degree of freedom	Sumof squares	meansum of squares	F. Rate
SSB	k – 1 = 2	40	$^{40}\!/\!_{2}$ = 20	
SSW	n – k = 12	60	$^{60}\!/\!_{12}$ = 5	$^{20}\!/\!_{5}$ =4
TSS	n – 1 = 14	100		

Conclusion

F_{cal} = 4, F_{tab} at 5 level of significance (k–1, n–k) =7 (2,12) Degree of freedom

F_{tab} = 3.89 $\therefore F_{tab} < F_{cal}$

So H_0 is rejected

H_1 is accepted

(OR)

7. Obtain the regression lines associated with the following data by the method of least squares.

X	1	2	3	4	5
Y	166	184	142	180	338

Sol :

x	y	x^2	xy
1	166	1	166
2	184	4	368
3	142	9	426
4	180	16	720
5	338	25	1690
15	1010	55	3370

Least square $y_c = a + bx$

$a = \frac{\varepsilon y}{n}$ $b = \frac{\varepsilon xy}{\varepsilon x^2}$

The two normal equations are

$\varepsilon y = na + b\varepsilon x$ ....(1)

$\varepsilon xy = a\varepsilon x + b\varepsilon x^2$ ....(2)

substitute the values we get

$1010 = 5a + 15b - (1) \times 3$

$3370 = 15a + 55\,b - (2) \times 1$

$$\begin{array}{l} 3030 = \cancel{15a} + 45b \\ 3370 = \cancel{15a} + 55b \\ \quad - \quad - \quad - \\ \hline -340 = -5b \\ \hline \end{array}$$

$b = \frac{340}{5} = 68$

substitute the value in equation (4)

$5a + 15\,(68) = 1010$

$5a = 1020 - 1010$

$a = {}^{10}\!/\!_{2} = 5$

So the equation of straight line $y = 2 + 68b$

8. Explain about the cluster Analysis with an example. **(Unit - IV, Q.No. 6)**

(OR)

9. Explain about different types of learning and brief on data exploration and reduction. **(Unit - IV, Q.No. 3)**

10. Explain the steps involved in Monte - Carlo simulation. **(Unit - V, Q.No. 2)**

(OR)

11. The occurrence of rain in a city on a day is dependent upon whether or not it rained on the previous day. If it rained on the previous day, the rain distribution is :

Event	No rain	1cm.rain	2cm.rain	3cm.rain	4cm.rain	5cm.rain
Probability	0.50	0.25	0.15	0.05	0.03	0.02

If it did not rain on the previous day the rain distribution is :

Event	No rain	1cm.rain	2cm.rain	3cm.rain
Probability	0.75	0.15	0.06	0.04

Simulate the city's weather for 10 days and determine by rainfall during the period.

Use the following random number for simulation: 67, 63, 39, 55, 29, 78, 70, 06, 78, 76

Assume that for the first day of the simulation it had not rained the day before.

Sol :

The numbers 00 – 99 are associated in proportion to the probabilities associated with each event if it rained on the previous day. The rain distribution and the random number allocated are given below

Event	Probability	Cumulative probability	Random Number
No rain	0.50	0.50	00 – 49
1cm .rain	0.25	0.75	50 – 74
2cm . rain	0.15	0.90	75 – 89
3cm . rain	0.05	0.95	90 – 94
4cm . rain	0.03	0.98	95 – 97
5cm. rain	0.02	1.00	98 – 99

Similarly, if it did not rain the previous day. The necessary distribution and the random number allocation is given below.

Event	Probability	Cumulative Probability	Random Number Interval
No rain	0.75	0.75	00 – 74
1cm . rain	0.15	0.90	75 – 89
2cm . rain	0.06	0.96	90 – 95
3 cm . rain	0.04	1.00	96 – 99

Let us now simulate the rainfall for 10 days using the given random numbers. For the first day it is given that it had not rained the day before.

Day	Random	Event	Remark
1	67	No rain	(From table 2)
2	63	No rain	(From table 2)
3	39	No rain	(From table 2)
4	55	No rain	(From table 2)
5	29	No rain	(From table 2)
6	78	1cm rain	(From table 1)
7	70	1 cm rain	(From table 1)
8	06	No rain	(From table 2)
9	78	1 cm rain	(From table 2)
10	76	2 cm rain	(From table 2)

Conclusion : Hence during 5 cm rain the simulated period, it did not rain 6 days of 10. The total rainfall during the period was 5 cm.

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERBAD

M.B.A III - Semester Examination

December - 2018

R17

DATA ANALYTICS

Time : 3 Hours] [Max. Marks : 75

Note : This question paper contains two parts A and B.

Part A is compulsory which carries 25 marks. Answer all questions in Part - A.

Part - B contains of 5 Units. Answer any one full question from each unit.

Each question carries 10 marks and may have a, b, c as sub questions.

PART - A (5 × 5 = 25 Marks)

Answers

1. (a) What is the importance of Data Analytics ? **(Unit - I, Q.No.3)**

(b) Discuss about measures of association. **(Unit - II, Q.No.5)**

(c) Explain the correlation techniques. **(Unit - III, Q.No.3)**

(d) What is logistic regression ? **(Unit - IV, SQA 10)**

(e) Explain about what-if analysis. **(Unit - V, SQA 6)**

PART - B (5 × 10 = 50 Marks)

(Essay Type Questions)

2. Write about Data Visualisation Techniques. **(Unit - I, Q.No. 15)**

(OR)

3. Explore the importance of Data Analytics in business decision making. **(Unit - I, Q.No. 9)**

4. How to measure association and variability of data ? **(Unit - II, Q.No. 5)**

(OR)

5. Enumerate the sampling techniques.

Ans :

Sampling techniques often depend on research objectives of a research work. Generally there are two types of sampling techniques that are widely deployed. These techniques are:

(a) Probability Sampling

This sampling technique includes sample selection which is based on random methods. The techniques that are based in this category are random sampling, stratified sampling, systematic sampling and cluster sampling.

(b) Non-probablity Sampling

This sampling techniques is not based on random selection. Some examples are quota sampling, purposive sampling and convenience sampling.

Probability Sampling

The techniques in probality sampling are as follows:

(a) Random Sampling

Random sampling is used to increase the probability of the sample selected. By deploying this technique, each member of a population stands a chance to be selected. Let's say you are interested to survey the usage of ecommerce application in business-to-consumer (B2C).

The sample you select needs to represent the types of e-commerce application and its usage. Due to financial and time constraints you are unable to survey the usage of all types of e-commerce application across the Malaysian network (N= 100,000). Therefore you decide to confine the study to e-commerce application for merchandise products in Malaysia (n=10,000) which is called the accessible population.

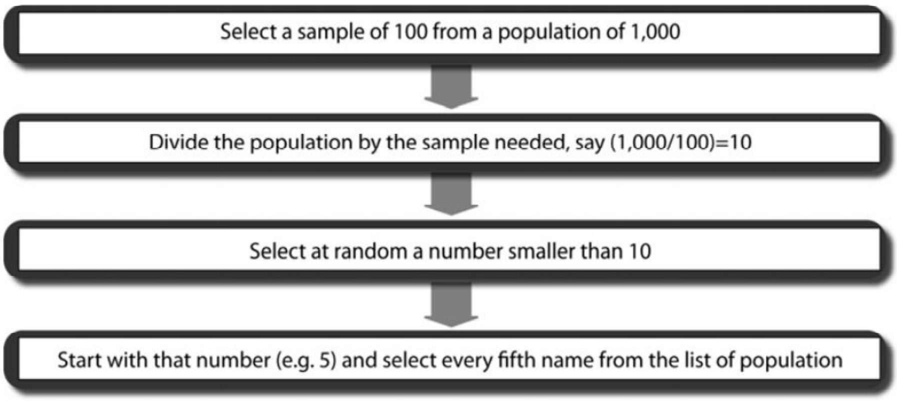
From this accessible population, a sample of 100 e-commerce application is retrieved. How do we randomly select sample? It is understood that random sample is a procedure in which all individuals in the defined population have an equal and independent chance to be selected in the sample design. In the above example, the number of e-commerce application on merchandise products across Malaysian network is 10,000 and you may intend to draw a sample of 100. When you select the first application, it has 1:10,000 chances of being selected. Once the first application selected, the remaining will be 9,999 so that each application has 1:9,999 of being selected as second case. Therefore, once each case is selected, the probability of being selected next changes because the population of selection has become one case smaller each time.

(b) Stratified Sampling

In some IT surveys, a researcher may want to ensure individuals with certain characteristics are included in the sample to be studied. For such cases, stratified sampling is used. In this sampling design, a researcher will attempt to stratify population in such a way that the population within a stratum is homogeneous with respect to the characteristics on the basis of which it is being stratified. You must bear in mind that it is important for the characteristics chosen as the basis of stratification, are clearly identifiable in the population. For example, it is much easier to stratify the population on the basis of gender rather than age or income group.

(c) Systematic Sampling

Systematic sampling also known as 'mixed sampling' category since it has both random and non-random sampling designs. A researcher has to begin by having a list names of members in the population, in random approach.



Select a sample of 100 from a population of 1,000

Divide the population by the sample needed, say (1,000/100)=10

Select at random a number smaller than 10

Start with that number (e.g. 5) and select every fifth name from the list of population

Fig. : Example of systematic sampling

This sampling method is good as long as the list does not contain any hidden order. Systematic sampling is frequently used in ICT research and survey, especially in selecting specified number of records from computer documents.

(d) Cluster Sampling

In cluster sampling, the unit of sampling is not referring to an individual entity but rather a group of entities. For example, in an organisation there are 25 departments and in each department there are an estimated 20 IT administrators. You need a sample of about 100 staff but this would mean going to many departments if random sampling approach is used. Using cluster sampling, you may select 5 departments randomly from a total of 25 departments. You study all the staff in the 5 departments you chose. The advantage that can be highlighted here is: it saves cost and time especially if the population is scattered. The disadvantage is that it is less accurate compared to other techniques of sampling discussed.

Non-Probability Sampling

In some research scenarios, it is not possible to ensure that the sample will be selected based on random selection. Non-probability sampling is based on a researcher‘s judgement and there is possibility of bias in sample selection and distort findings of the study. Nonetheless, this sampling technique is used because of its practicality. It can save time and cost, and at the same time, it is a feasible method given the spread and features of a population. Some common sampling methods are quota sampling, purposive sampling and convenience sampling.

(a) Quota Sampling

The main reason directing quota sampling is the researcher‘s ease of access to the sample population. Similar to stratified sampling, a researcher needs to identify the subgroups and their proportions as they are represented in the population. Then, the researcher will select subjects based on his/ her convenience and judgement to fill each subgroup. A researcher must be confident in using this method and firmly state the criteria for selection of sample especially during results summarisation.

(b) Purposive Sampling

This sampling method is selected on the basis that members conform to certain stipulated criteria. You may need to use your own judgement to select cases to answer certain research questions. This sampling method is normally deployed if the sample population is small and when the main objective is to choose cases that are informative to the research topic selected. Purposive sampling is very useful in the early stages of an exploratory study. One of the disadvantages of this technique is that the sample may have characteristics different from population characteristics.

(c) Convenience Sampling

Using this sampling method, a researcher is free to use anything that they could find in the research outline. The sample is selected based on preferences and ease of sampling respondents. This sampling is easier to conduct and less expensive. However, it has poor reliability due to its high incidence of bias. In ICT, convenience sampling seems to be dominant especially in cases of organisations that conduct web surveys, mail their responses to a survey questions and SMS their opinions to a question. Although convenience sampling can cater to a lot of data, it is not reliable in terms whether the sample represents the real population or not.

6. Explain about Linear Discriminant Analysis with an example. **(Unit - III, Q.No. 20)**

(OR)

7. Realtors are often interested bin seeing how the appraised value of a home varies according to the size of the home. Some data on area (in thousands of square feet) and appraised value (in thousands of Dollars) for a sample of 11 homes follow.

Area	1.1	1.5	1.6	1.6	1.4	1.3	1.1	1.7	1.9	1.5	1.3
Value	75	95	110	102	95	87	82	115	122	98	90

Estimate the Least SQUARES TO PREDICT APPRAISED VALUE FROM SIZE.

Sol :

Least square $y_c = a + bx$

$$a = \frac{\Sigma y}{n} \quad b = \frac{\Sigma xy}{\Sigma x^2}$$

In the given problem area is considered as a ‘x’ and value is considered as ‘y’

x	y	x^2	xy
1.1	75	1.21	82.5
1.5	95	2.25	142.5
1.6	110	2.56	176
1.6	102	2.56	163.2
1.4	95	1.96	133
1.3	87	1.69	113.1
1.1	82	1.21	90.2
1.7	115	2.89	195.5
1.9	122	3.61	231.8
1.5	98	2.25	147
1.3	90	1.69	117
	107	23.88	1591.8

$$a = \frac{1071}{11} \quad b = \frac{1591.8}{23.88}$$

$$a = 97.36 \quad b = 66.65$$

8. Explain data reduction techniques in Data Mining. **(Unit - IV, Q.No. 4)**

(OR)

9. Explain why cluster analysis is called as unsupervised learning. **(Unit - IV, Q.No. 5)**

10. Explain Monte-Carlo simulation. **(Unit - V, Q.No. 6)**

(OR)

11. Briefly explain risk analysis techniques in decision making. **(Unit - V, Q.No. 17)**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

MBA III-Semester Examinations

July/August - 2021

R19

DATA ANALYTICS

Time : 3 Hours] [Max. Marks : 75

Answer any **FIVE** questions
All question carry equal marks

ANSWERS

1. Explain the different types of data visualization tools and state the statistical tools for summarizing data. **(Unit-I, Q.No. 15, 18)**

2. Define Business Analytics in detail with its types? **(Unit-I, Q.No. 4, 9)**

3. The probability distribution for the random variable X follows.

X	20	25	30	35
F(X)	0.20	0.15	0.25	0.40

(a) Is this probability distribution valid? Explain.

(b) What is the probability that X = 30?

(c) What is the probability that X is less than or equal to 25?

(d) What is the probability that X is greater than 30?

Ans :

Given that,

X	20	25	30	35
F(X)	0.20	0.15	0.25	0.40

(a) Yes, it is valid. Because, the sum of probabilities is equal to 1 i.e.,

$\Rightarrow$ 0.20 + 0.15 + 0.25 + 0.40 = 1

Also, the probability of each event lies between 0 and 1.

(b) P(X = 30)

The probability that X is equal to 30 is,

From the given distribution table,

P(X = 30) = 0.25

(c) P(X < = 25)

The probability that X less than or equal to 25 is,

P(X < = 25) = P(X = 20) + P(X = 25)

= 0.20 + 0.15

P(X < = 25) = 0.35.

Therefore, that probability that X less than or equal to 25 is '0.35'.

(d) $P(X > 30)$

The probability that X greater than 30 is,

$$\begin{aligned} P(X > 30) &= 1 - P(X < = 30) \\ &= 1 - (P(X = 20) + P(X = 25) + P(X = 30) \\ &= 1 - (0.20 + 0.15 + 0.25) \\ &= 1 - (0.60) \end{aligned}$$

$P(X > 30) = 0.40$

Therefore, that probability that X greater than 30 is '0.40'.

4. In a study about viral fever, the numbers of people affected in a town were noted as:

Age in years	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of people affected	3	5	16	18	12	7	4

Find its Standard deviation.

Ans :

Given that,

Age in years	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of people affected	3	5	16	18	12	7	4

Assume that,

Mean, A = 35

Age (in years)	Number of People Affected (f_i)	Mid Value ((x_i))	$d_i = x_i - A$ (Here, A = 35)	fidi	fi d_i^2
0 - 10	3	5	$5 - 35 = -30$	$-30 \times 3 = -90$	2700
10 - 20	5	15	$15 - 35 = -20$	$-20 \times 5 = -100$	2000
20 – 30	16	25	$25 - 35 = -10$	$-10 \times 16 = -160$	1600
30 – 40	18	35	$35 - 35 = 0$	$0 \times 18 = 0$	0
40 – 50	12	45	$45 - 35 = 10$	$10 \times 12 = 120$	1200
50 – 60	7	55	$55 - 35 = 20$	$20 \times 7 = 140$	2800
60 – 70	4	65	$65 - 35 = 30$	$30 \times 4 = 120$	3600

Now,

$$\begin{aligned} \Sigma f_i &= 3 + 5 + 16 + 18 + 12 + 7 + 4 \\ \Sigma f_i &= 65 \\ \Sigma f_i d_i &= 90 + (-100) + (-160) + 0 + 120 + 140 + 120 \\ \Sigma f_i d_i &= 30 \\ \Sigma f_i d_i^2 &= 2700 + 2000 + 1600 + 0 + 1200 + 2800 + 3600 \\ \Sigma f_i d_i^2 &= 13{,}900 \end{aligned}$$

We know that

$$\text{Standard deviation} \quad s = \frac{\Sigma fd^2}{N}\left(\frac{\Sigma f_i d_i}{N}\right)^2$$

$$= \sqrt{\frac{13,900}{65} - \left(\frac{30}{65}\right)^2}$$

$$= \sqrt{213.84 - (0.46)^2}$$

$$= \sqrt{213.84 - 0.21}$$

$$= \sqrt{213.627}$$

$$= 14.615$$

5. Write down the characteristics of correlation coefficient. Also state how does rank correlation differ from Pearson correlation coefficient? **(Unit-III, Q.No. 1)**

Ans :

Comparison of Pearson and Spearman Coefficients

(i) The fundamental difference between the two correlation coefficients is that the Pearson coefficient works with a linear relationship between the two variables whereas the Spearman Coefficient works with monotonic relationships as well.

(ii) One more difference is that Pearson works with raw data values of the variables whereas Spearman works with rank-ordered variables.

6. Obtain a regression line associated with the following data by the method of least squares. **(May-19, Prob. 7)**

X	1	2	3	4	5
Y	166	184	142	180	338

7. (a) What is Classification and Regression Trees (CART)? **(Unit-IV, Q.No. 14)**

(b) Explain about different types of learning and brief on data exploration and reduction. **(Unit-IV, Q.No. 3)**

8. (a) Analyze in detail about advantages and disadvantages of simulation technique. **(Unit-V, Q.No. 14)**

(b) Explain the steps involved in Monte-Carlo simulation. **(Unit-V, Q.No. 2)**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERBAD

M.B.A III Semester Examination **R19**

MODEL PAPER - I

DATA ANALYTICS

Time : 3 Hours] [Max. Marks : 75

Note : This question paper contains two parts A and B.
Part A is compulsory which carries 25 marks. Answer all questions in Part - A.
Part - B contains of 5 Units. Answer any one full question from each unit.
Each question carries 10 marks and may have a, b, c as sub questions.

PART - A (5 × 5 = 25 Marks)
(Short Answer Questions)

ANSWERS

1. (a) Data visualization **(Unit-I, SQA - 6)**
(b) Sample **(Unit-II, SQA - 2)**
(c) Karl Pearson's of correlation **(Unit-III, SQA - 1)**
(d) Cluster Analysis **(Unit-IV, SQA - 5)**
(e) "What-if Analysis" **(Unit-V, SQA - 6)**

PART - B (5 × 10 = 50 Marks)
(Essay Type Questions)

2. a) Explain the importance of Analytics. **(Unit-I, Q.No.3)**

(OR)

b) How to explore data in pivot tables ? **(Unit-I, Q.No.19)**

3. a) Explain about Measures of Location. **(Unit-II, Q.No.2)**

(OR)

b) Explain in detail about distribution fitting? **(Unit-II, Q.No.15)**

4. a) Explain the multiple correlation between more than two variable. **(Unit-III, Q.No.7)**

(OR)

b) Explain One Way ANOVA with example ? **(Unit-III, Q.No.22)**

5. a) What are the methods of correlation ? **(Unit-IV, Q.No.3)**

(OR)

b) What is regression by the method of lease square? **(Unit-IV, Q.No.15)**

6. a) What is simulation and explain different types of simulations? **(Unit-V, Q.No.1)**

(OR)

b) Explain about risk analysis with its benefits and uses? **(Unit-V, Q.No.15)**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERBAD

M.B.A III Semester Examination R19

MODEL PAPER - II

DATA ANALYTICS

Time : 3 Hours] [Max. Marks : 75

Note : This question paper contains two parts A and B.
Part A is compulsory which carries 25 marks. Answer all questions in Part - A.
Part - B contains of 5 Units. Answer any one full question from each unit.
Each question carries 10 marks and may have a, b, c as sub questions.

PART - A (5 × 5 = 25 Marks)

(Short Answer Questions)

ANSWERS

1. (a) Dimensions of Big Data **(Unit-I, SQA - 4)**

(b) Random Sample **(Unit-II, SQA - 3)**

(c) Regression Analysis **(Unit-III, SQA - 5)**

(d) Data Mining **(Unit-IV, SQA -1)**

(e) Decision tree analysis **(Unit-V, SQA -11)**

PART - B (5 × 10 = 50 Marks)

(Essay Type Questions)

2. a) What is Big Data? Explain the various technologies in big data? **(Unit-I, Q.No.5)**

(OR)

b) Explain the different statistical methods for summerizing data **(Unit-I, Q.No.18)**

3. a) What is Data Modeling and Explain how data models deliver benefit? **(Unit-II, Q.No.7)**

(OR)

b) What is discrete probability distribution and explain what are the most common applications used ? **(Unit-II, Q.No.11)**

4. a) Explain the concept of regression analysis. **(Unit-III, Q.No.9)**

(OR)

b) Describe briefly the systematic approach for building good regression models. **(Unit-III, Q.No.18)**

5. a) What is data reduction and explain data reduction techniques? **(Unit-IV, Q.No.4)**

(OR)

b) What is *k*-nearest neighbors and explain about KNN classifier ? **(Unit-IV, Q.No.13)**

6. a) Explain how Monte Carlo simulation can be developed in Ms.Excel ? **(Unit-V, Q.No.7)**

(OR)

b) What is verification explain with its methods ? **(Unit-V, Q.No.10)**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERBAD

M.B.A III Semester Examination **R19**

MODEL PAPER - III

DATA ANALYTICS

Time : 3 Hours] [Max. Marks : 75

Note : This question paper contains two parts A and B.
Part A is compulsory which carries 25 marks. Answer all questions in Part - A.
Part - B contains of 5 Units. Answer any one full question from each unit.
Each question carries 10 marks and may have a, b, c as sub questions.

PART - A (5 × 5 = 25 Marks)

(Short Answer Questions)

			Answers
1.	(a)	Various Challenges in Business Analytics	**(Unit-I, SQA - 1)**
	(b)	Data Modelling	**(Unit-II, SQA - 10)**
	(c)	Assumptions of ANOVA	**(Unit-III, SQA - 10)**
	(d)	k-nearest neighbors	**(Unit-IV, SQA - 9)**
	(e)	Simulation	**(Unit-V, SQA -1)**

PART - B (5 × 10 = 50 Marks)

(Essay Type Questions)

2. a) What are the Data Visualization Techniques? **(Unit-I, Q.No.15)**

(OR)

b) Discuss briefly about role of business analytics in current business environment. **(Unit-I, Q.No.9)**

3. a) Explain the various ways of Measure of variability ? **(Unit-II, Q.No.4)**

(OR)

b) What is Data Modelling? **(Unit-II, Q.No.14)**

4. a) Explain the properties of Karl Pearson's Coefficient of Correlation. **(Unit-III, Q.No.4)**

(OR)

b) Explain the Two Way ANOVA with an example. **(Unit-III, Q.No.23)**

5. a) What is data mining and explain the steps involved in this process ? **(Unit-IV, Q.No.1)**

(OR)

b) What is logistic regression and explain how to enter data in it ? **(Unit-IV, Q.No.17)**

6. a) Explain the random number generation in excel. **(Unit-V, Q.No.5)**

(OR)

b) Explain different Experience Based Testing Techniques. **(Unit-V, Q.No.12)**