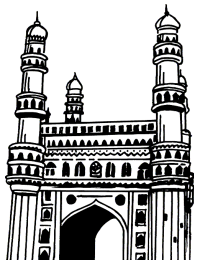
Rahul's ✔
Topper's Voice

R22
New Syllabus



JNTU (H)
MBA

Latest Edition

II Year III Semester

BUSINESS ANALYTICS

- Study Manual
- Internal Assessment
- FAQ's and Important Questions
- Short Questions & Answers
- Choose the correct Answers
- Fill in the blanks
- Very Short Questions & Answers
- Solved Model Papers
- Solved Previous Question Papers

Price
₹. 209-00

- by -
WELL EXPERIENCED LECTURER

Since 1986
Rahul Publications™
Hyderabad. Cell : 9391018098, 9505799122.

All disputes are subjects to Hyderabad Jurisdiction only

Subjects List

- Production & Operations Management
- Management Information Systems
- Business Analytics

Marketing

- Digital Marketing
- Sales and Promotion Management
- Consumer Behavior

Finance

- Security Analysis and Portfolio Management
- Risk Management and Financial Derivatives
- Strategic Cost and Management Accounting

Human Resource

- Talent and Performance Management Systems
- Learning and Development
- Employee Relations

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.

© *No part of this publications should be reporduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording and/or otherwise without the prior written permission of the publisher*

Sole Distributors : **Cell : 9391018098, 9505799122**

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.

BUSINESS ANALYTICS

STUDY MANUAL

FAQ's and Important Questions	IV - VIII
Unit - I	1 - 28
Unit - II	29 - 60
Unit - III	61 - 104
Unit - IV	105 - 134
Unit - V	135 - 168
Internal Assessment	169 - 174

SOLVED MODEL PAPERS

Model Paper - I	175 - 176
Model Paper - II	177 - 178
Model Paper - III	179 - 180

SOLVED PREVIOUS QUESTION PAPERS

December - 2018	181 - 182
April / May - 2019	183 - 184
December - 2019	185 - 186
July / August - 2021	187 - 188
April / May - 2022	189 - 190
March / April - 2023	191 - 192
February - 2024	193 - 197
July / August - 2024	198 - 200

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

Karl Pearson Correlation Technique: 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

1.1 Introduction to Data 1

1.1.1 Importance of analytics 1

1.2 Data for Business Analytics 3

1.3 Big Data 4

1.4 Business Analytics in Practice 8

1.5 Data Visuliztion 10

1.5.1 Tools 11

1.6 Data Queries 17

1.7 Statistical Method for Summaraizing Data 17

1.8 Exploring Data using Pivot Tables 19

- Short Questions and Answers 24
- Choose the Correct Answers 26
- Fill in the Blanks 27
- Very Short Questions and Answers 28

UNIT - II

2.1 Descriptive Statistical Measures 29

2.1.1 Population and Samples 29

2.1.2 Measures of Location 29

2.1.3 Measures of Dispersion 32

2.1.4 Measures of Variability 34

2.1.5 Measures of Association 36

2.2 Probability Distribution and Data Modeling 36

2.2.1 Discrete Probability Distribution, Continuous Probability Distribution 36

2.2.2 Random Sampling from Probability Distribution, Data modeling and Distribution Fitting 46

- Short Questions and Answers 53
- Choose the Correct Answers 57
- Fill in the Blanks 59
- Very Short Questions and Answers 60

Topic **Page No.**

UNIT - III

3.1 Karl Pearson Correlation Technique 61

3.2 Multiple Correlation 68

3.3 Spearman's Rank Correlation 71

3.4 Simple and Multiple Regression 77

3.5 Regression by the Method of Least Squares 82

3.6 Building Good Regression Models 83

3.7 Regression with Categorical Independent Variables 84

3.8 Linear Discriminant Analysis 87

3.9 One-Way and Two-Way ANOVA 88

- Short Questions and Answers 99
- Choose the Correct Answers 101
- Fill in the Blanks 103
- Very Short Questions and Answers 104

UNIT - IV

4.1 Scope of Data Mining 105

4.2 Data Exploration and Reduction 108

4.3 Unsupervised Learning 110

4.3.1 Cluster Analysis 110

4.4 Association Rules 113

4.5 Supervised Learning 116

4.5.1 Partition Data 117

4.5.2 Classification Accuracy 118

4.5.3 Prediction Accuracy 120

4.5.4 K-Nearest Neighbors 121

4.5.5 Classification and Regression Trees 123

4.5.6 Logistics Regression 124

- Short Questions and Answers 129
- Choose the Correct Answers 131
- Fill in the Blanks 131
- Very Short Questions and Answers 134

Topic **Page No.**

UNIT - V

5.1 Simulation 135

5.1.1 Advantages and Disadvantages of Simulation 137

5.2 Random Number Generation 138

5.3 Monte Carlo Simulation 144

5.4 What If Analysis 151

5.5 Verification and Validation 155

5.6 Risk Analysis 157

5.7 Decision Tree Analysis 159

- Short Questions and Answers 163
- Choose the Correct Answers 166
- Fill in the Blanks 167
- Very Short Questions and Answers 168

Frequently Asked & Important Questions

UNIT - I

1. Why classification of data is considered hierarchical in nature?

Ans : **(July-24, May - 23, Imp.)**

Refer Unit-I, Page No. 1, Q.No. 1

2. Enumerate in detail about Data analytics and the importance of data analytics in the current scenario.

Ans : **(July-24, May-22, Oct.-22, Dec.-19, Dec.-18, Imp.)**

Refer Unit-I, Page No. 2, Q.No. 2

3. What is big data. Explain the evolution of big data?

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

Refer Unit-I, Page No. 4, Q.No. 5

4. Define business analytics. Discuss briefly the role of business analytics in current business environment.

Ans : **(Feb.-24, Oct.-22, Dec.-19, May-19, Dec.-18, Imp.)**

Refer Unit-I, Page No. 8, Q.No. 8

5. Explain the use of bubble chart for comparing stock characteristics.

Ans : **(Feb.-24, April-23, Oct.-22, May.-22, Aug.-21, May-19, Imp.)**

Refer Unit-I, Page No. 11, Q.No. 12

6. What do you understand by frequency distributions for categorical data?

Ans : **(April - 23, Dec.- 19,Imp.)**

Refer Unit-I, Page No. 17, Q.No. 14

7. Describe about Pivot Tables. How data is explored using pivot tables?

Ans : **(April-22, Oct.-20, Dec.-19, Imp.)**

Refer Unit-I, Page No. 19, Q.No. 15

UNIT - II

1. Explain briefly about Measures of location.

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

Refer Unit-II, Page No. 29, Q.No. 2

2. Explain the concept of Measures of Dispersion.

Ans : **(Feb.-24, Oct.-22, Imp.)**

Refer Unit-II, Page No. 32, Q.No. 3

3. Describe about Measures of Variability.

Ans : **(Feb.-24, Oct.-20, Dec.-19, Dec.-18, Imp.)**

Refer Unit-II, Page No. 34, Q.No. 4

4. Define Probability Distribution. Explain briefly about Discrete Probability Distribution.

Ans : **(May-22, Dec.-19, Imp.)**

Refer Unit-II, Page No. 36, Q.No. 6

5. What is binomial distribution. Give its properties.

Ans : **(Imp.)**

Refer Unit-II, Page No. 37, Q.No. 7

6. Explain briefly about continuous probability distribution.

Ans : **(July-24, May-22, Imp.)**

Refer Unit-II, Page No. 41, Q.No. 9

7. Enumerate the various sampling techniques.

Ans : **(July-24, Dec.-18, Imp.)**

Refer Unit-II, Page No. 46, Q.No. 13

8. What is data modeling? Explain different types of data modeling.

Ans : **(Oct.-20, Imp.)**

Refer Unit-II, Page No. 50, Q.No. 15

9. Explain about data modeling and distribution fitting in detail.

Ans : **(Feb.-24, Oct.-22, Dec.-19, Imp.)**

Refer Unit-II, Page No. 51, Q.No. 17

UNIT - III

1. Define correlation. Explain the significance of correlation.

Ans : **(Imp.)**

Refer Unit-III, Page No. 61, Q.No. 1

2. Define Multiple Correlation. Explain the steps involved in generating correlation coefficient in MS Excel.

Ans : **(May-22, Imp.)**

Refer Unit-III, Page No. 68, Q.No. 9

3. Define simple regression. Explain the concept of simple linear regression with Excel.

Ans : **(July-24, Dec.-19, May-19, Imp.)**

Refer Unit-III, Page No. 78, Q.No. 15

4. Explain the multiple regression by least squares with an example.

Ans : **(July-24, Oct.-20, Imp.)**

Refer Unit-III, Page No. 82, Q.No. 16

5. Briefly explain the process of constructing regression line using the method of least squares.

Ans : **(Feb.-24, April - 23, Oct.-22, Dec.-19, Imp.)**

Refer Unit-III, Page No. 82, Q.No. 17

6. Explain in detail about regression with categorical independent variables.

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

Refer Unit-III, Page No. 84, Q.No. 19

7. What is two way Anova? Mention merits and demerits

Ans : **(May-22, Imp.)**

Refer Unit-III, Page No. 89, Q.No. 24

8. 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 : **(May-19, Imp.)**

Refer Unit-III, Page No. 93, Prob. 6

UNIT - IV

1. Define Data Mining. Explain the process of Data Mining.

Ans : **(Oct.-22, May-19, Dec.-18, Imp.)**

Refer Unit-IV, Page No. 105, Q.No. 1

2. Explain the Scope and Architecture of Data Mining.

Ans : **(July-24, Oct.-22, May-19, Imp.)**

Refer Unit-IV, Page No. 106, Q.No. 2

3. Discuss about techniques of data mining.

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

Refer Unit-IV, Page No. 107, Q.No. 3

4. Explain data reduction techniques in Data Mining.

Ans : **(Feb.-24, April-23, Dec.-18, Imp.)**

Refer Unit-IV, Page No. 109, Q.No. 6

5. Define Cluster Analysis. Explain the properties of clustering.

Ans : **(July-24, Dec.-19, Imp.)**

Refer Unit-IV, Page No. 111, Q.No. 10

6. Describe briefly about Association Rules.

Ans : **(July-24, Dec.-19, Imp.)**

Refer Unit-IV, Page No. 113, Q.No. 13

7. How data can be partioned ? Explain.

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

Refer Unit-IV, Page No. 117, Q.No. 16

8. Explain briefly about K-Nearest Neighbors.

Ans : **(Feb.-24, May-22, Imp.)**

Refer Unit-IV, Page No. 121, Q.No. 21

9. What is Classification and Regression Trees (CART) ?

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

Refer Unit-IV, Page No. 123, Q.No. 22

10. Explain why cluster analysis is called as unsupervised learning.

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

Refer Unit-IV, Page No. 127, Q.No. 24

UNIT - V

1. Define Simulation. Explain different types of Simulation.

Ans : **(April-23, May-19, Imp.)**

Refer Unit-V, Page No. 135, Q.No. 1

2. Analyze in detail about advantages and disadvantages of simulation technique.

Ans : **(July-24, April-23, Aug.-21, Nov.-20, Dec.-19, Imp.)**

Refer Unit-V, Page No. 137, Q.No. 3

3. Define Random Number. Discuss the various methods of Random Number Generations.

Ans : **(July-24, April-23, Imp.)**

Refer Unit-V, Page No. 138, Q.No. 5

4. Define Monte Carlo Simulation. State the Advantages and Disadvantages of Monte Carlo Simulation.

Ans : **(July-24, Oct.-22, Dec.-19, Dec.-18, Imp.)**

Refer Unit-V, Page No. 144, Q.No. 7

5. Explain the process of Monte Carlo Stimulation.

Ans : **(Feb.-24, May-22, Aug.-21, Imp.)**

Refer Unit-V, Page No. 150, Q.No. 9

6. Define What If Analysis. What are the basic options available in excel for performing What If Analysis?

Ans : **(July-24, Dec.-18, Imp.)**

Refer Unit-V, Page No. 151, Q.No. 10

7. Define verification. Explain various methods of verification.

Ans : **(Imp.)**

Refer Unit-V, Page No. 155, Q.No. 12

8. What is Risk Analysis? Explain the benefits of Risk Analysis.

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

Refer Unit-V, Page No. 157, Q.No. 15

9. Briefly explain risk analysis techniques in decision making.

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

Refer Unit-V, Page No. 158, Q.No. 16

10. Define Decision Tree Analysis. Explain the steps involved in construction of Decision Tree Analysis.

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

Refer Unit-V, Page No. 159, Q.No. 17

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

1.1.1 Importance of analytics

Q1. Define data. Explain the classification of data

(OR)

Why classification of data is considered hierarchical in nature?

Ans : **(May - 23)**

Data

A collection of raw facts that can be used to conclude some information

Classification

Data can be broadly classified into 3 types.

1. Structured Data

Structured data is created using a fixed schema and is maintained in tabular format. The elements in structured data are addressable for effective analysis. It contains all the data which can be stored in the SQL database in a tabular format. Today, most of the data is developed and processed in the simplest way to manage information.

Examples :

Relational data, Geo-location, credit card numbers, addresses, etc.

Consider an example for Relational Data like you have to maintain a record of students for a university like the name of the student, ID of a student, address, and Email of the student. To store the record of students used the following relational schema and table for the same.

S_ID	S_Name	S_Address	S_Email
1001	A	hyd	A@gmail.com
1002	B	Bangalore	B@gmail.com

2. Unstructured Data

It is defined as the data in which is not follow a pre-defined standard or you can say that any does not follow any organized format. This kind of data is also not fit for the relational database because in the relational database you will see a pre-defined manner or you can say organized way of data. Unstructured data is also very important for the big data domain and To manage and store Unstructured data there are many platforms to handle it like No-SQL Database. 0 seconds of 0 seconds Volume 0%

Examples :

Word, PDF, text, media logs, etc.

Rahul Publications

3. Semi-Structured Data

Semi-structured data is information that does not reside in a relational database but that have some organizational properties that make it easier to analyze. With some process, you can store them in a relational database but is very hard for some kind of semi-structured data, but semi-structured exist to ease space.

Example –

XML data.

Q2. Define Data Analytics. Explain the importance.

(OR)

Define Data Analytics and explain its importance in an organization?

(OR)

Enumerate in detail about Data analytics and the importance of data analy tics in the current scenario.

(OR)

What is Data ? Explain the Importance of Analytics ?

Ans : **(May-22, Oct.-22, Dec.-19, Dec.-18, Imp.)**

Data analytics

Data or information is in raw format. The increase in size of the data has led to a rise in need for carrying out inspection, data cleaning, transformation as well as data modeling to gain in sights from the data in order to derive conclusions for better decision making process. This process known as data analysis.

Importance

The following points highlights the importance of data analytics,

1. Business to Consumer Applications (B2C)

Organizations gather data which has been collected through various sources like customers, economy, businesses and practical experience. The collected data is then processed and categorized based on the requirement and the data is analyzed for studying the purchase patterns.

2. Growth of the Business

Data analytics enables companies to find the insights of the value chains which can be done by a data analytics expert. It offers comprehensive knowledge of industry and economy. It also helps in understanding the business aspects of the near future and obtain benefits before the competitors. Data analytics helps in seeing opportunities at the right time by providing analyzed data.

3. Data Analytics for Students

Data analytics offers the students an immense collection of tech tools to study. These tech tools considered as the key to industrial growth. It also helps student to clearly understand the institutional data which enables further innovation. The scope for institutional research and exploration led to the emergence of data analytics as a high class or fashion courses.

4. Data Analytics in Educational Sector

In educational sector, new reforms are needed based on the necessities of the student and patterns of previous students Analysts are more helpful than researcher in this context as they have tools which help in analyzing the information in short period of time.

5. Data Analytics for Professionals

Professionals use data analytics for understanding work patterns. They require analytical skills for prioritizing their work based upon the ongoing scenario. Analytical skills are required for various work process and organizations are recruiting several analysts to improve their workability.

6. 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.

Q3. Explain different types of data analytics.

Ans :

There are four main types of data analytics: descriptive, diagnostic, predictive, and prescriptive. Each has its own set of goals and roles in the data analytics process.

Rahul Publications

1. **Descriptive Analytics**

Descriptive analytics answers the "what" questions in the data analytics process. It helps stakeholders understand large datasets by summarizing them. The descriptive analysis tracks the organization's past performance. It includes the following steps:

- Data collection
- Data processing
- Data analysis and
- Data visualization

2. **Diagnostic Analytics**

Diagnostic analytics answers the "why" questions in the data analytics process. It analyzes the results from the descriptive analysis and then further evaluates it to find the cause. The diagnostic analysis process takes place in three steps:

Identifying any unexpected changes in the data

Data related to the changes is collected.

Statistical techniques help find relationships and trends related to the changes.

3. **Predictive Analytics**

The purpose of predictive analytics is to answer questions about the future of the data analytics process. The past data identifies the trends. The techniques used in the process include statistical and machine learning techniques. A few of them are neural networks, decision trees, and regression.

4. **Prescriptive Analytics**

Prescriptive analysis helps businesses make well-informed decisions and predict the analytics. This type of data analytics uses machine learning strategies that are capable of finding patterns in large datasets.

1.2 Data for Business Analytics

Q4. Explain in detail about how data is used in business analytics.

Ans :

Meaning

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 perfor- mance 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

Rahul Publications

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

(i) 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 deploy- ments.

(ii) 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.

(iii) 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.

(iv) 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 evolution of big data ?

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

Meaning

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 technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

Evolution

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.

Rahul Publications

(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.

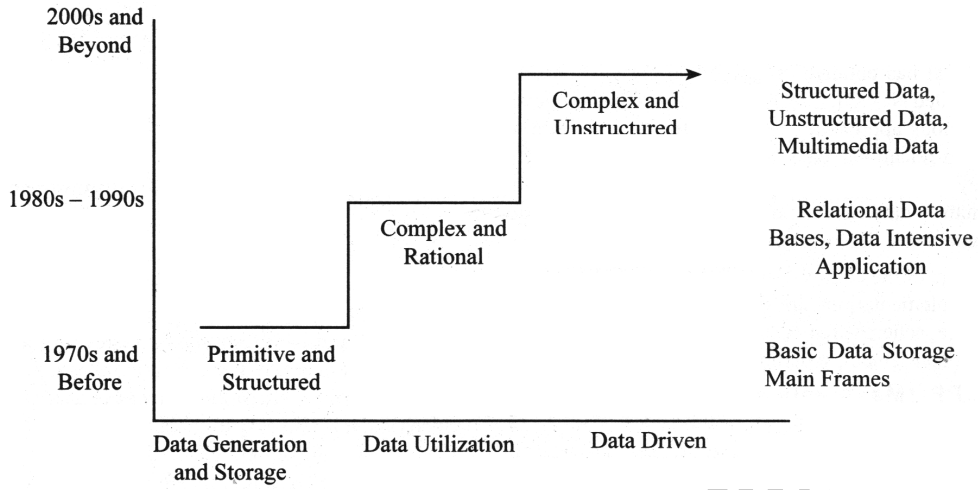


Fig.: The Evolution of Big Data

Q6. What are the characteristics of Big Data.

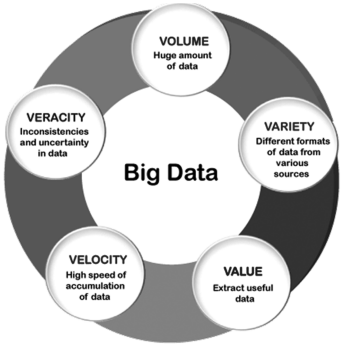
Ans :

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many multinational companies to process the data and business of many organizations. The data flow would exceed 150 exa bytes per day before replication.

There are five v's of Big Data that explains the characteristics.

5 V's of Big Data

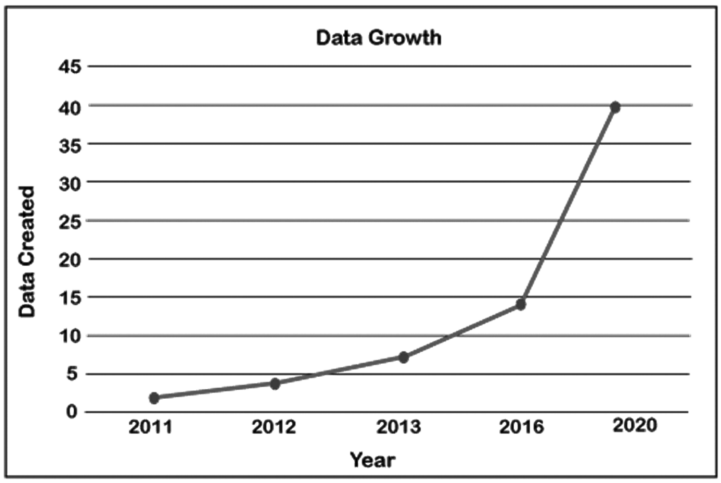
1. Volume
2. Veracity
3. Variety
4. Value
5. Velocity



1. Volume

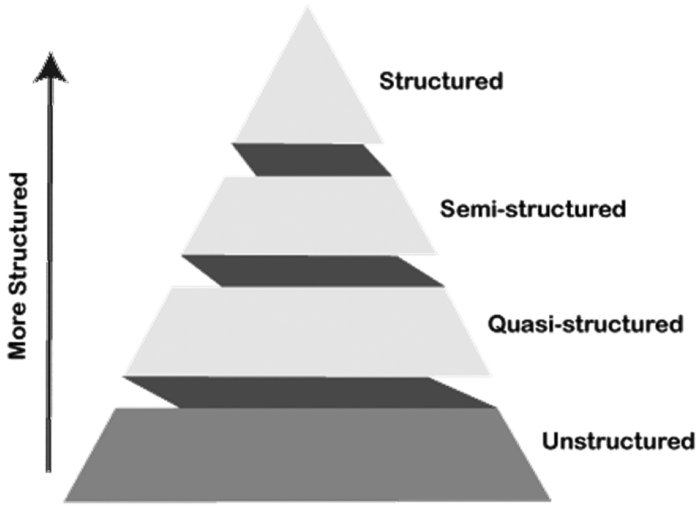
The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as business processes, machines, social media platforms, networks, human interactions, and many more.

Facebook can generate approximately a billion messages, 4.5 billion times that the "Like" button is recorded, and more than 350 million new posts are uploaded each day. Big data technologies can handle large amounts of data.



2. Variety

Big Data can be structured, unstructured, and semi-structured that are being collected from different sources. Data will only be collected from databases and sheets in the past, But these days the data will comes in array forms, that are PDFs, Emails, audios, SM posts, photos, videos, etc.



The data is categorized as below:

(i) Unstructured Data

All the unstructured files, log files, audio files, and image files are included in the unstructured data. Some organizations have much data available, but they did not know how to derive the value of data since the data is raw.

(ii) Quasi-structured Data

The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

Example :

Web server logs, i.e., the log file is created and maintained by some server that contains a list of activities.

(iii) Semi-structured Data

In Semi-structured, the schema is not appropriately defined, e.g., JSON, XML, CSV, TSV, and email. OLTP (Online Transaction Processing) systems are built to work with semi-structured data. It is stored in relations, i.e., tables.

(iv) Structured data

In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

3. Veracity

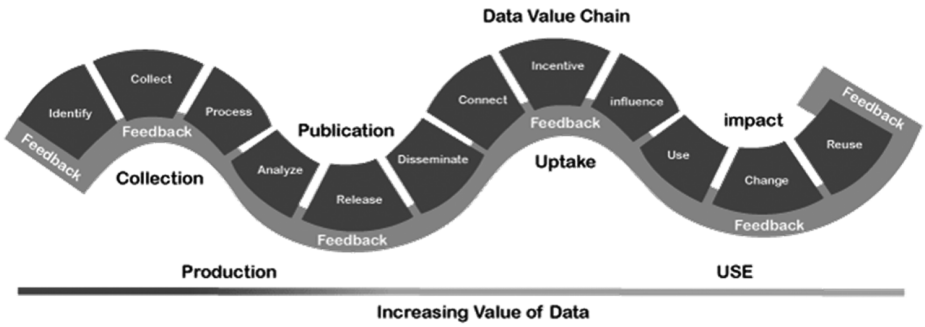
Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development.

For example, Facebook posts with hashtags.

4. Value

Value is an essential characteristic of big data. It is not the data that we process or store. It is valuable and reliable data that we store, process, and also analyze.

Rahul Publications



5. Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in real-time. It contains the linking of incoming data sets speeds, rate of change, and activity bursts. The primary aspect of Big Data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like application logs, business processes, networks, and social media sites, sensors, mobile devices, etc.

Q7. Explain various challenges of bigdata.

Ans :

Many companies get stuck at the initial stage of their Big Data projects. This is because they are neither aware of the challenges of Big Data nor are equipped to tackle those challenges. The challenges of conventional systems in Big Data need to be addressed.

Below are some of the major Big Data challenges and their solutions.

1. Lack of proper understanding of Big Data

Companies fail in their Big Data initiatives due to insufficient understanding. Employees may not know what data is, its storage, processing, importance, and sources. Data professionals may know what is going on, but others may not have a clear picture.

For example, if employees do not understand the importance of data storage, they might not keep the backup of sensitive data. They might not use databases properly for storage. As a result, when this important data is required, it cannot be retrieved easily.

2. Data growth issues

One of the most pressing challenges of Big Data is storing all these huge sets of data properly. The amount of data being stored in data centers and databases of companies is increasing rapidly. As these data sets grow exponentially with time, it gets extremely difficult to handle.

Most of the data is unstructured and comes from documents, videos, audios, text files and other sources. This means that you cannot find them in databases. This can pose huge Big Data analytics challenges and must be resolved as soon as possible, or it can delay the growth of the company.

3. Lack of data professionals

To run these modern technologies and Big Data tools, companies need skilled data professionals. These professionals will include data scientists, data analysts and data engineers who are experienced in working with the tools and making sense out of huge data sets.

Companies face a problem of lack of Big Data professionals. This is because data handling tools have evolved rapidly, but in most cases, the professionals have not. Actionable steps need to be taken in order to bridge this gap.

4. Securing data

Securing these huge sets of data is one of the daunting challenges of Big Data. Often companies are so busy in understanding, storing and analyzing their data sets that they push data security for later stages. But, this is not a smart move as unprotected data repositories can become breeding grounds for malicious hackers.

5. Integrating data from a variety of sources

Data in an organization comes from a variety of sources, such as social media pages, ERP applications, customer logs, financial reports, e-mails, presentations and reports created by employees. Combining all this data to prepare reports is a challenging task.

This is an area often neglected by firms. But, data integration is crucial for analysis, reporting and business intelligence, so it has to be perfect.

1.4 Business Analytics in Practice

Q8. Define business analytics. Discuss briefly the role of business analytics in current business environment.

(OR)

Explore the importance of Data Analytics in business decision making.

(OR)

What is business Analytics and its types ?

Ans : **(Oct.-22, Dec.-19, May-19, Dec.-18, Imp.)**

Meaning

Business analytics refers to the broad category of skills, tools and techniques employed for gathering, storing and analyzing of business related data and information by any business enterprise. The main aim of business analytics is to help the management to arrive at an objective and accurate business decisions.

Types

The role of business analytics in current business environment is as follows.

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

Rahul Publications

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 of 20-30% 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

Con Agra 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

Rahul Publications

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. 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.

Q9. Distinguish between Business Intelligence and Business Analytics.

Ans :

S.No.	BUSINESS INTELLIGENCE	BUSINESS ANALYTICS
1.	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.
2.	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.

Rahul Publications

1.5 DATA VISULIZTION

Q10. Define Data Visualization. What are the benefits of data visualization.

Ans : **(Imp.)**

Meaning

Data visualization is the representation of information and data using charts, graphs, maps, and other visual tools. These visualizations allow us to easily understand any patterns, trends, or outliers in a data set.

Data visualization also presents data to the general public or specific audiences without technical knowledge in an accessible manner. For example, the health agency in a government might provide a map of vaccinated regions.

The purpose of data visualization is to help drive informed decision-making and to add colorful meaning to an otherwise bland database.

Benefits of data visualization

Data visualization can be used in many contexts in nearly every field, like public policy, finance, marketing, retail, education, sports, history, and more. Here are the benefits of data visualization:

(i) Storytelling

People are drawn to colors and patterns in clothing, arts and culture, architecture, and more. Data is no different colors and patterns allow us to visualize the story within the data.

(ii) Accessibility

Information is shared in an accessible, easy-to-understand manner for a variety of audiences.

(iii) Visualize relationships

It's easier to spot the relationships and patterns within a data set when the information is presented in a graph or chart.

(iv) Exploration

More accessible data means more opportunities to explore, collaborate, and inform actionable decisions.

Q11. What are the applications of data visualization ?

Ans : **(Imp.)**

Using data visualization tools, different types of charts and graphs can be created to illustrate important data.

These are a few examples of data visualization in the real world:

(i) Data science

Data scientists and researchers have access to libraries using programming languages or tools such as Python or R, which they use to understand and identify patterns in data sets. Tools help these data professionals work more efficiently by coding research with colors, plots, lines, and shapes.

(ii) Marketing

Tracking data such as web traffic and social media analytics can help marketers analyze how customers find their products and whether they are early adopters or more of a laggard buyer. Charts and graphs can synthesize data for marketers and stake-holders to better understand these trends.

(iii) Finance

Investors and advisors focused on buying and selling stocks, bonds, dividends, and other commodities will analyze the movement of prices over time to determine which are worth purchasing for short- or long-term periods. Line graphs help financial analysts visualize this data, toggling between months, years, and even decades.

(iv) Health policy

Policymakers can use choropleth maps, which are divided by geographical area (nations, states, continents) by colors. They can, for example, use these maps to demonstrate the mortality rates of cancer or ebola in different parts of the world.

Rahul Publications

1.5.1 Tools

Q12. Explain different tools of data Visualization.

(OR)

Examine the types of data visualization tools.

(OR)

Explain the different types of data visualization tools ?

(OR)

Write about Data Visualization Techniques.

(OR)

Explain the use of bubble chart for comparing stock characteristics.

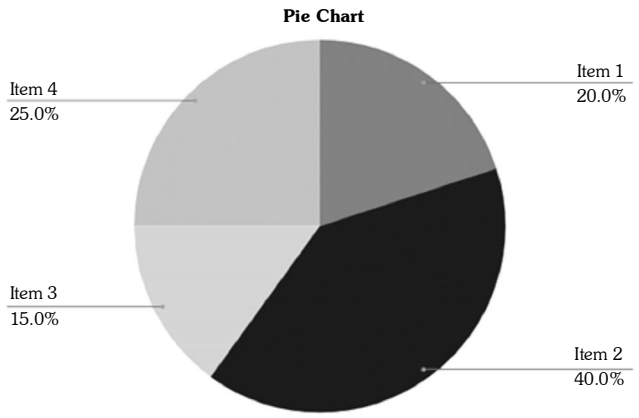
Ans : **(April-23, Oct.-22, May.-22, Aug.-21, May-19, Imp.)**

The type of data visualization technique you leverage will vary based on the type of data you're working with.

Here are some important data visualization techniques to know:

1. Pie Chart

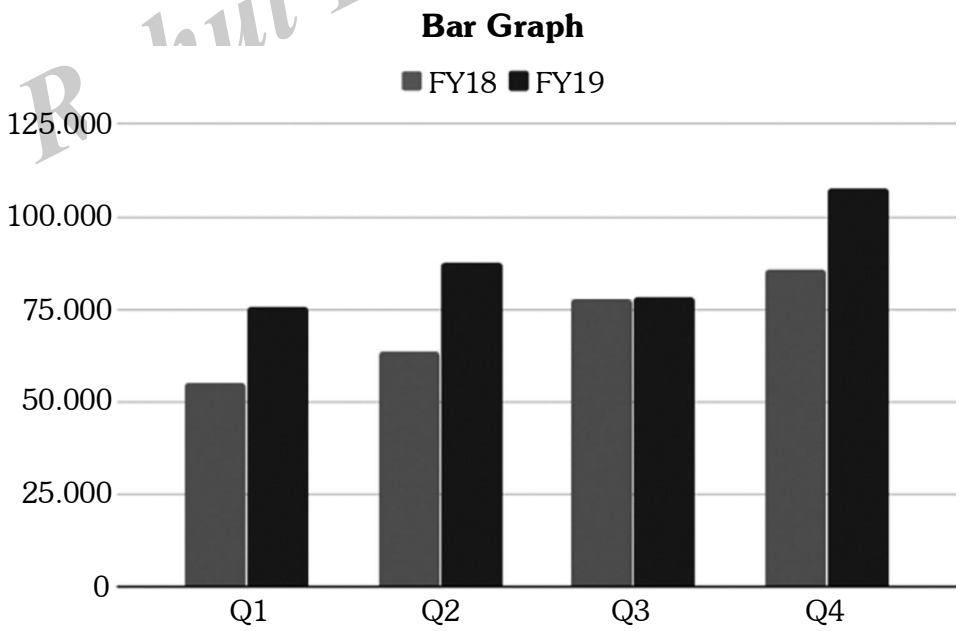
Pie charts are one of the most common and basic data visualization techniques, used across a wide range of applications. Pie charts are ideal for illustrating proportions, or part-to-whole comparisons.



Because pie charts are relatively simple and easy to read, they're best suited for audiences who might be unfamiliar with the information or are only interested in the key take aways. For viewers who require a more thorough explanation of the data, pie charts fall short in their ability to display complex information.

2. Bar Chart

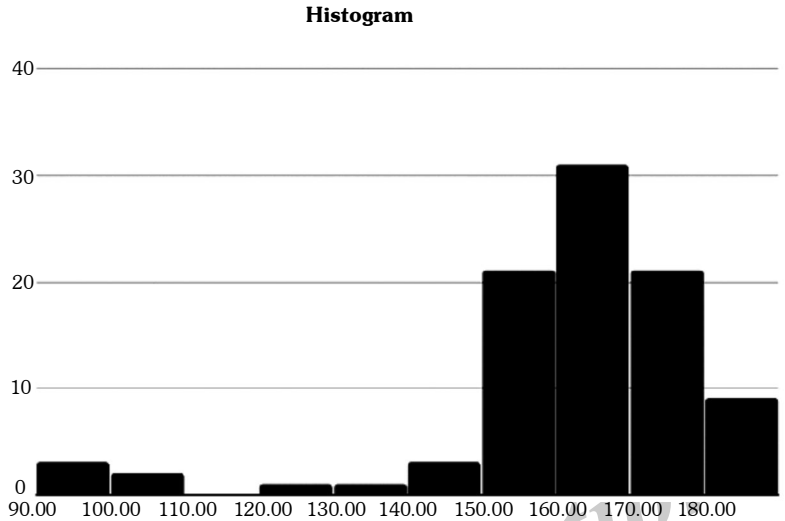
The classic bar chart, or bar graph, is another common and easy-to-use method of data visualization. In this type of visualization, one axis of the chart shows the categories being compared, and the other, a measured value. The length of the bar indicates how each group measures according to the value.



One drawback is that labeling and clarity can become problematic when there are too many categories included. Like pie charts, they can also be too simple for more complex data sets.

3. Histogram

Unlike bar charts, histograms illustrate the distribution of data over a continuous interval or defined period. These visualizations are helpful in identifying where values are concentrated, as well as where there are gaps or unusual values.

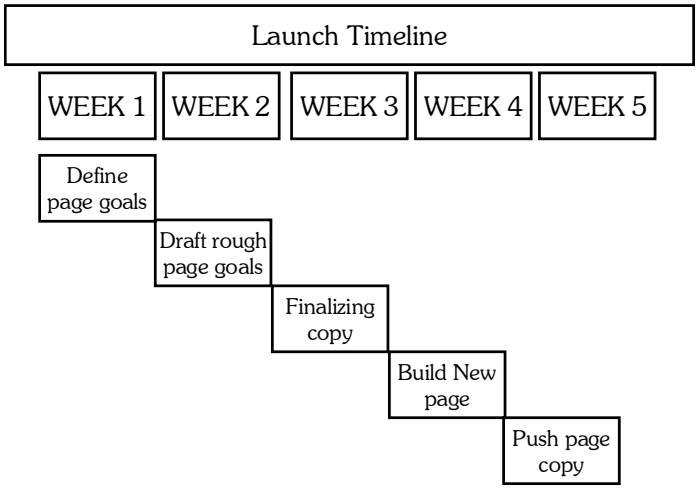


Histograms are especially useful for showing the frequency of a particular occurrence. For instance, if you'd like to show how many clicks your website received each day over the last week, you can use a histogram. From this visualization, you can quickly determine which days your website saw the greatest and fewest number of clicks.

4. Gantt Chart

Gantt charts are particularly common in project management, as they're useful in illustrating a project timeline or progression of tasks. In this type of chart, tasks to be performed are listed on the vertical axis and time intervals on the horizontal axis. Horizontal bars in the body of the chart represent the duration of each activity.

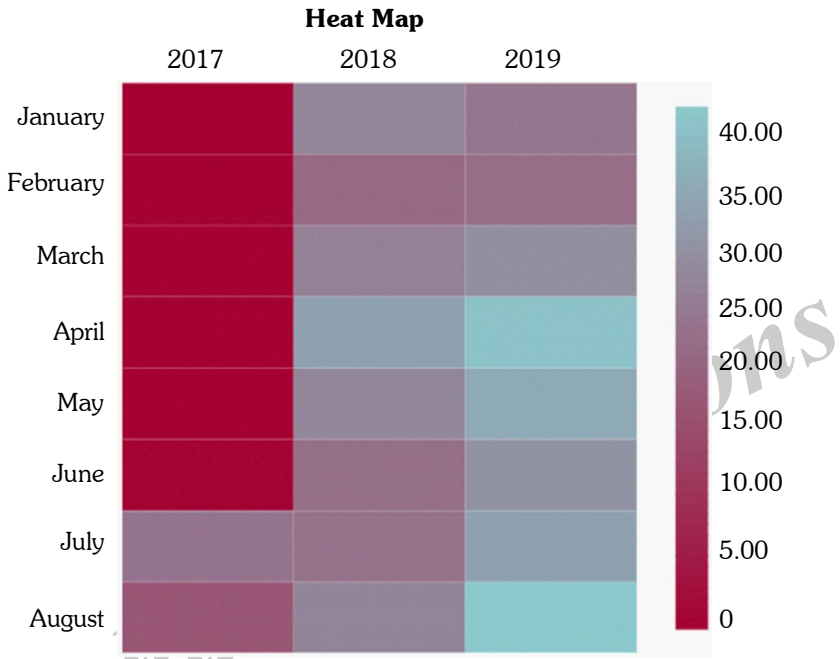
Rahul Publications



Utilizing Gantt charts to display timelines can be incredibly helpful, and enable team members to keep track of every aspect of a project. Even if you're not a project management professional, familiarizing yourself with Gantt charts can help you stay organized.

5. Heat Map

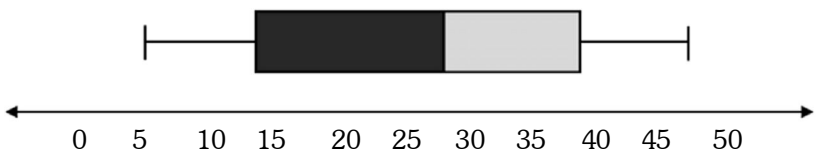
A heat map is a type of visualization used to show differences in data through variations in color. These charts use color to communicate values in a way that makes it easy for the viewer to quickly identify trends. Having a clear legend is necessary in order for a user to successfully read and interpret a heatmap.



There are many possible applications of heat maps. For example, if you want to analyze which time of day a retail store makes the most sales, you can use a heat map that shows the day of the week on the vertical axis and time of day on the horizontal axis. Then, by shading in the matrix with colors that correspond to the number of sales at each time of day, you can identify trends in the data that allow you to determine the exact times your store experiences the most sales.

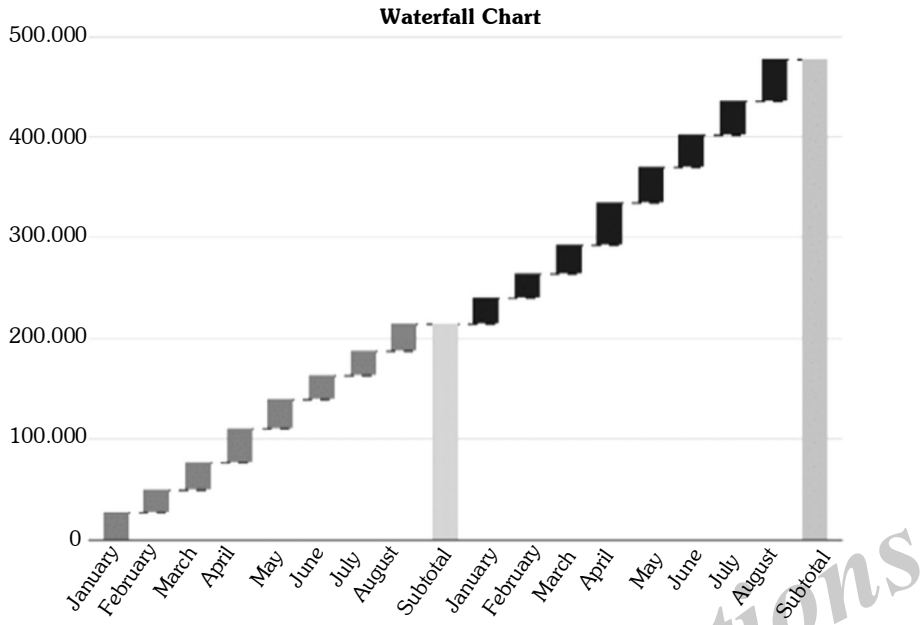
6. A Box and Whisker Plot

A box and whisker plot, or box plot, provides a visual summary of data through its quartiles. First, a box is drawn from the first quartile to the third of the data set. A line within the box represents the median. "Whiskers," or lines, are then drawn extending from the box to the minimum (lower extreme) and maximum (upper extreme). Outliers are represented by individual points that are in-line with the whiskers.



This type of chart is helpful in quickly identifying whether or not the data is symmetrical or skewed, as well as providing a visual summary of the data set that can be easily interpreted.

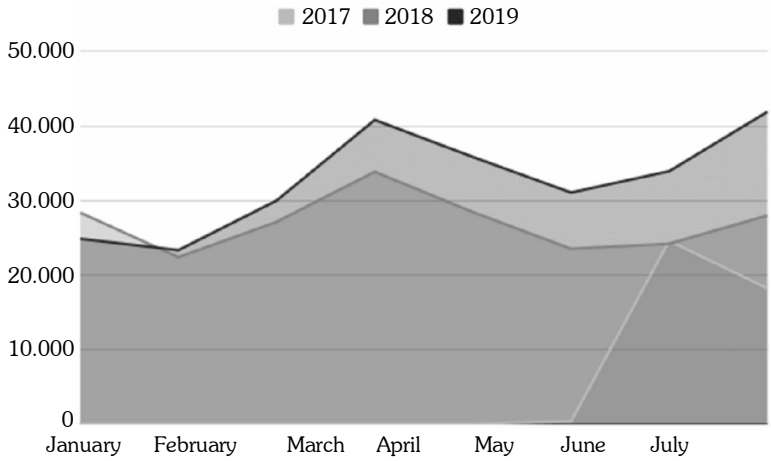
7. Waterfall Chart



A waterfall chart is a visual representation that illustrates how a value changes as it's influenced by different factors, such as time. The main goal of this chart is to show the viewer how a value has grown or declined over a defined period. For example, waterfall charts are popular for showing spending or earnings over time.

8. Area Chart

An area chart, or area graph, is a variation on a basic line graph in which the area underneath the line is shaded to represent the total value of each data point. When several data series must be compared on the same graph, stacked area charts are used.

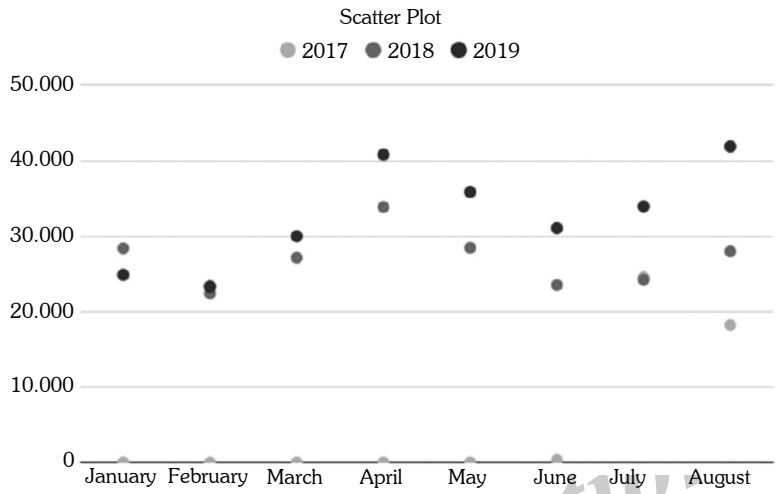


This method of data visualization is useful for showing changes in one or more quantities over time, as well as showing how each quantity combines to make up the whole. Stacked area charts are effective in showing part-to-whole comparisons.

Rahul Publications

9. Scatter Plot

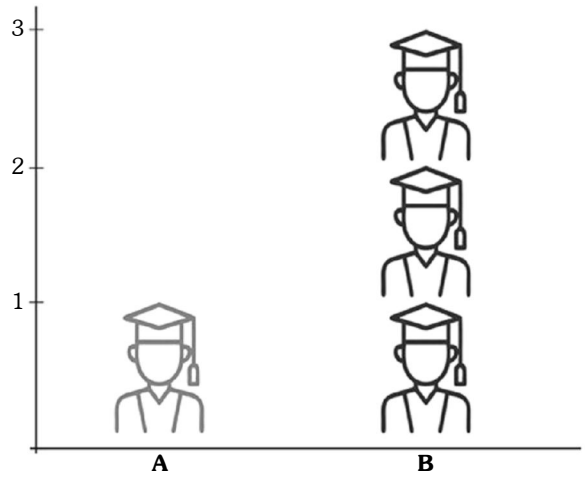
Another technique commonly used to display data is a scatter plot. A scatter plot displays data for two variables as represented by points plotted against the horizontal and vertical axis. This type of data visualization is useful in illustrating the relationships that exist between variables and can be used to identify trends or correlations in data.



Scatter plots are most effective for fairly large data sets, since it's often easier to identify trends when there are more data points present. Additionally, the closer the data points are grouped together, the stronger the correlation or trend tends to be.

10. Pictogram Chart

Pictogram charts, or pictograph charts, are particularly useful for presenting simple data in a more visual and engaging way. These charts use icons to visualize data, with each icon representing a different value or category. For example, data about time might be represented by icons of clocks or watches. Each icon can correspond to either a single unit or a set number of units (for example, each icon represents 100 units).



In addition to making the data more engaging, pictogram charts are helpful in situations where language or cultural differences might be a barrier to the audience's understanding of the data.

1.6 Data Queries

Q13. Explain briefly about Data queries.

Ans :

In an organization, numerous date Queries arise. It canbe addressed accurately if the data queries.

(i) 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 Query Language (SQL) or as pictorials, graphs or complex results, e.g., trend analyses from data-mining tools.

(ii) 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.

(iii) 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's request. For instance, if you have an employee table, you might issue a SQL statement that returns the employee who is paid the most.

(iv) 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.

Rahul Publications

1.7 Statistical Method for Summaraizing Data

Q14. Explain different Statistical Method for Summaraizing Data.

OR

Explain in detail about statistical methods for summarizing data.

(OR)

What do you understand by frequency distributions for categorical data ?

(OR)

Explain the process of computing percentiles.

Ans : **(April - 23, Dec.- 19,Imp.)**

1. Frequency Distributions for Categorical Data

Frequency distribution refers to the tabular arrangement of data, when arranged into groups or categories according to conveniently established divisions of the range of the observarions.

Types

There are four types of frequency distributions:

(i) Ungrouped frequency distributions:

The number of observations of each value of a variable.

(ii) Grouped frequency distributions

The number of observations of each class interval of a variable. Class intervals are ordered groupings of a variable's values.

(iii) Relative frequency distributions

The proportion of observations of each value or class interval of a variable.

(iv) Cumulative frequency distributions

The sum of the frequencies less than or equal to each value or class interval of a variable.

Categorical Frequency Distribution

A categorical frequency distribution is a table to organize data that can be placed in specific categories, such as nominal- or ordinal-level data.

2. Relative Frequency Distributions

In order to evaluate data, sometimes it becomes necessary to disclose the percentage of observations which may come under each distribution class rather than disclosing the actual frequencies of class. However, for converting frequency distribution into corresponding relative frequency distribution one should use the following formula,

$$\text{Relative Frequency Distribution} = \frac{\text{Frequency Observation}}{\text{Total Number of Observations}}$$

3. Frequency Distributions for Numerical Data

Frequency distributions for numerical data involves COUNTIF to count the frequencies of each discrete value as the numerical data consists of small number of discrete values.

4. Excel Histogram Tool

A histogram is a graphical representation of frequency distribution for a quantitative data in the form of a column chart. In Excel, Frequency distributions and histograms are framed using the 'Analysis Toolpak'.

5. Cumulative Relative frequency Distributions

A cumulative relative frequency distribution is a summary of cumulative relative frequencies in tabular form. Cumulative relative frequencies are the frequencies which represent the proportion of the total observations number which appear at or below every upper limit of each group. The lower limit is always equal to zero.

6. Percentiles and Quartiles

Percentile refers to the percent of dividing the given series into 100 equal parts. It is a value at which atleast 'K' percent of the observations lie.

In Excel, percentile is calculated as PERCENTILE.INC which computes the Kth percentile of data with specified range in array field. The range of k is 0 to 1.

The following are the steps to calculate the kth percentile (where k is any number between zero and one hundred).

Rahul Publications

Step 1:

Arrange all data values in the data set in ascending order.

Step 2:

Count the number of values in the data set where it is represented as 'n'.

Step 3:

Calculate the value of k/100, where k = any number between zero and one hundred.

Step 4:

Multiply 'k' percent by 'n'.The resultant number is called an index.

Step 5:

If the resultant index is not a whole number then round to the nearest whole number, then go to Step 7. If the index is a whole number, then go to Step 6.

Step 6:

Count the values in your data set from left to right until you reach the number. Then find the mean for that corresponding number and the next number. The resultant value is the kth percentile of your data set.

Step 7:

Count the values in your data set from left to right until you reach the number. The obtained value will be the kth percentile of your data set.

Quartiles refers to the values dividing the given series into four equal parts. In excel quartiles are calculated as QUARTILE.INC which specifies the range of the data.

7. Cross Tabulation

Cross tabulation is a method of representing the number of observations in a data set belonging to various categories of two categorical variables in tabular form. It is a widely used statistical tool which offers insights into various characteristics of different market segments in marketing.

Rahul Publications

1.8 Exploring Data using Pivot Tables

Q15. Explain about pivot tables.

(OR)

Describe about Pivot Tables. How data is explored using pivot tables ?

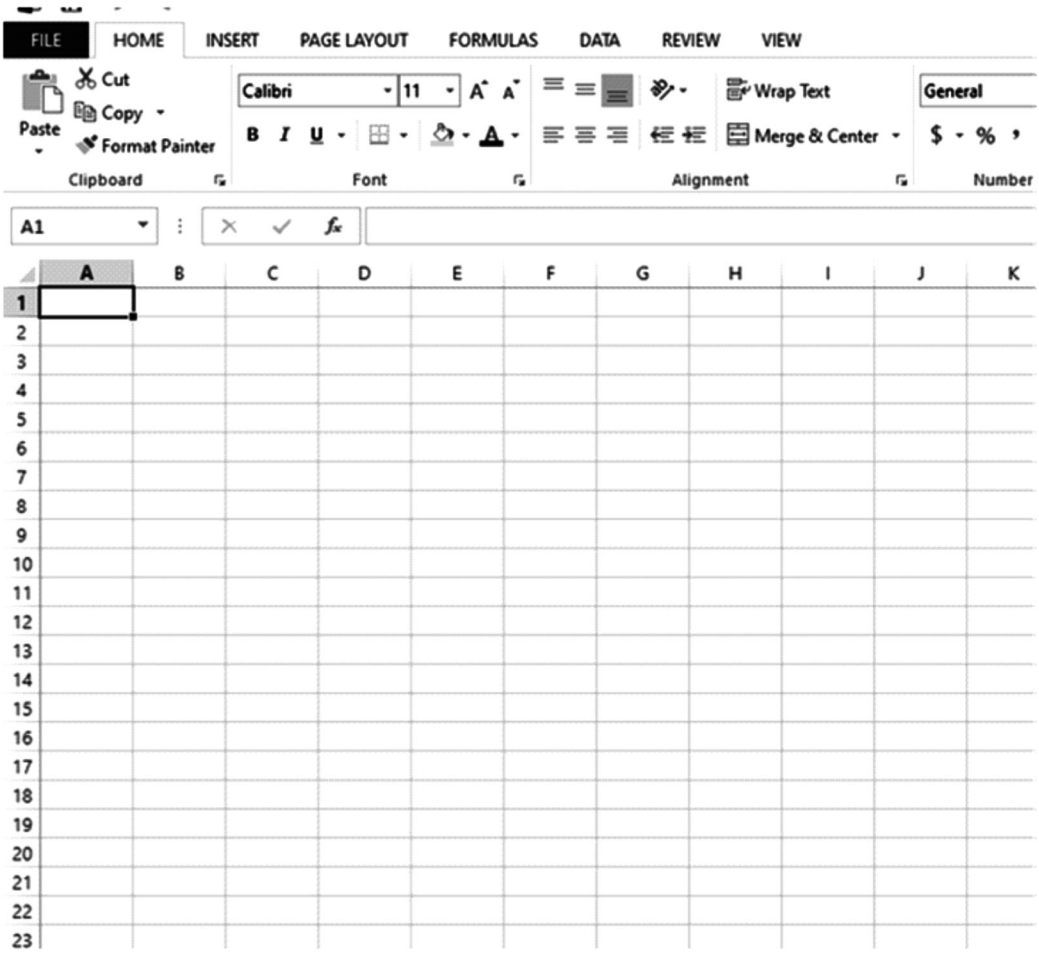
Ans : **(April-22, Oct.-20, Dec.-19, Imp.)**

Broad information examination can be done utilizing PivotTables and produce wanted reports. The joining of the Data Model with PivotTable improves how the information is examined, associated, summed up, and detailed. You can import tables from outside information sources and make a PivotTable with the imported tables. This works with programmed updations of the qualities in the PivotTable at whatever point the information in the associated information sources is refreshed.

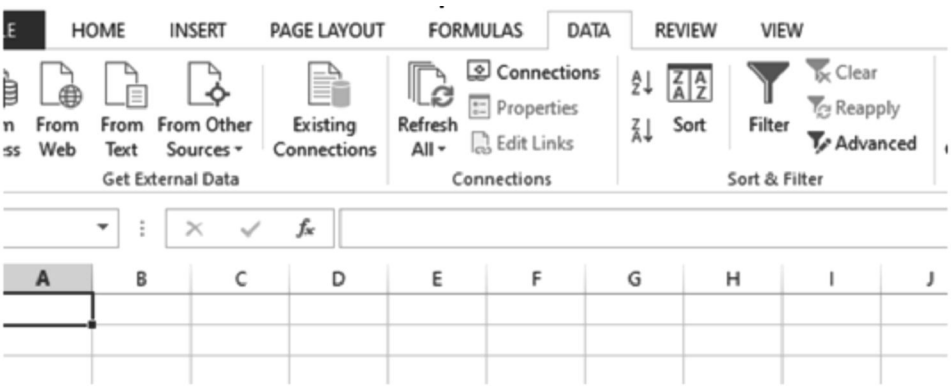
Creating a PivotTable to analyze External Data

To make a PivotTable to investigate outer information,

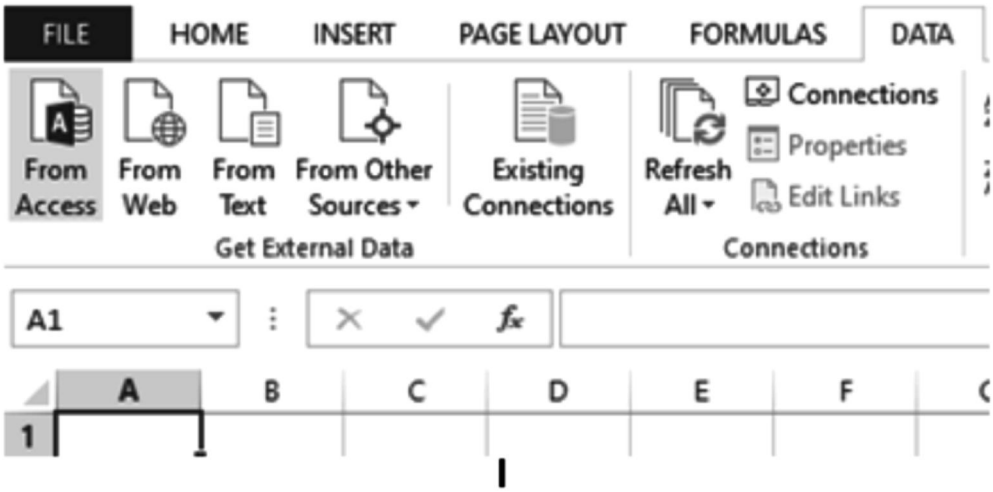
- Open another clear exercise manual in Excel.



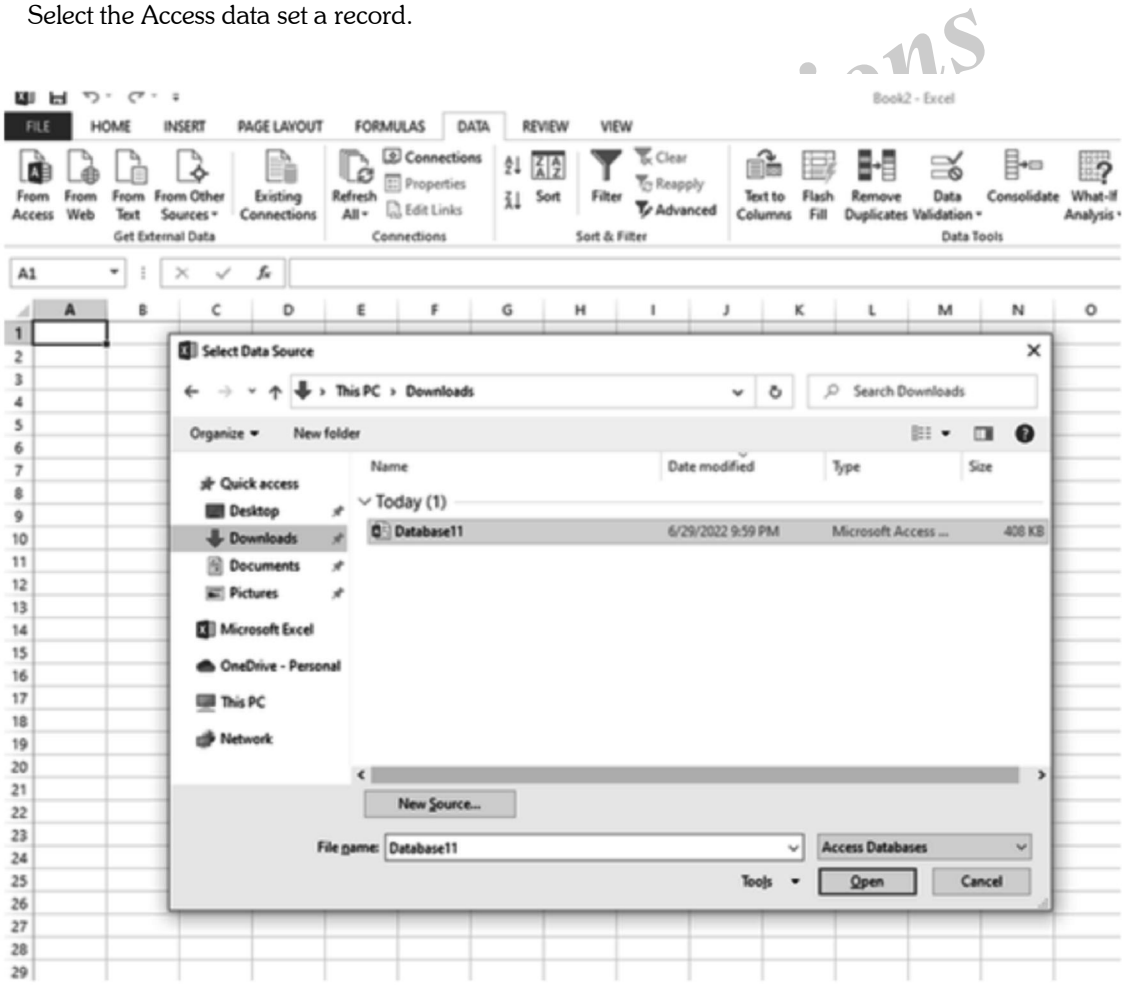
- Click the DATA tab on the Ribbon.



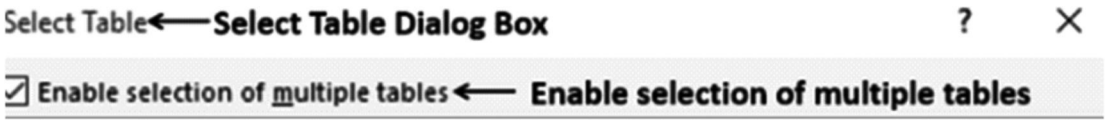
➢ Click From Access in the Get External Data bunch. The Select Data Source discourse box shows up.



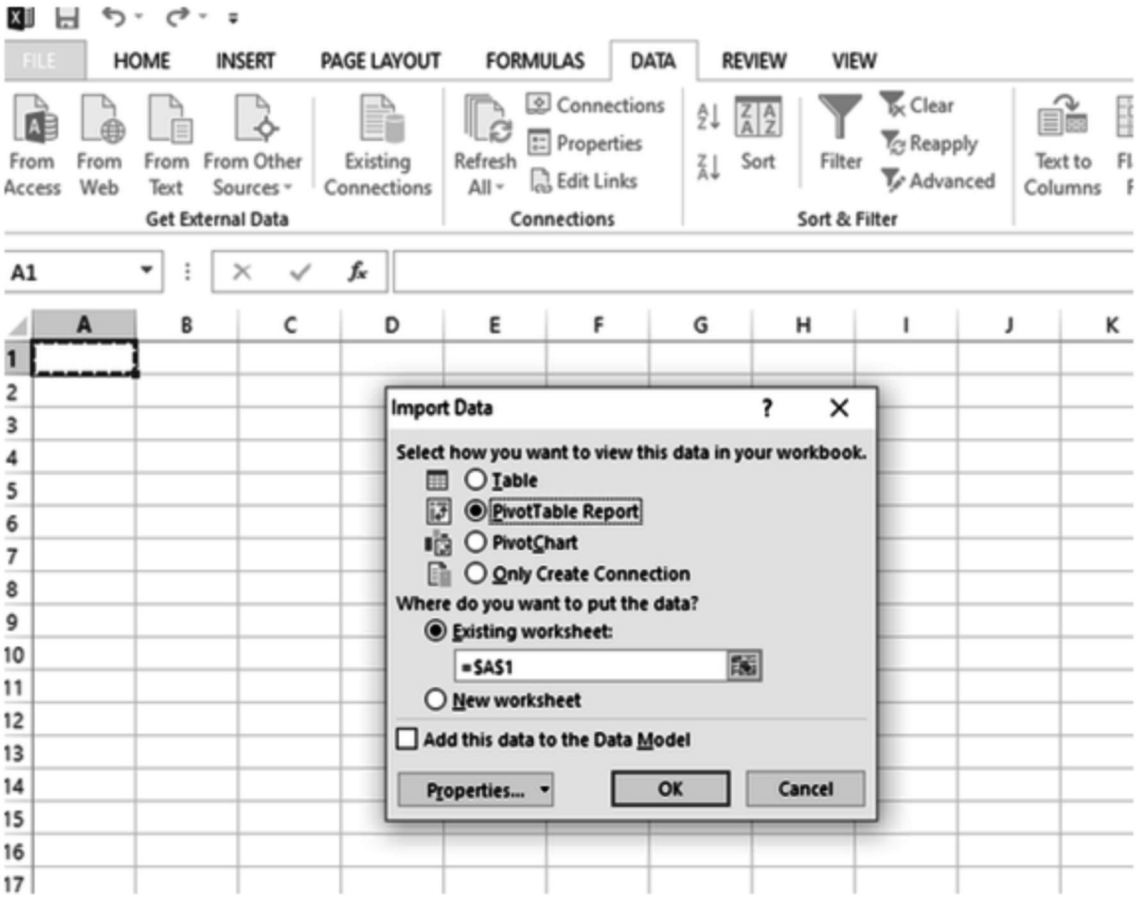
➢ Select the Access data set a record.



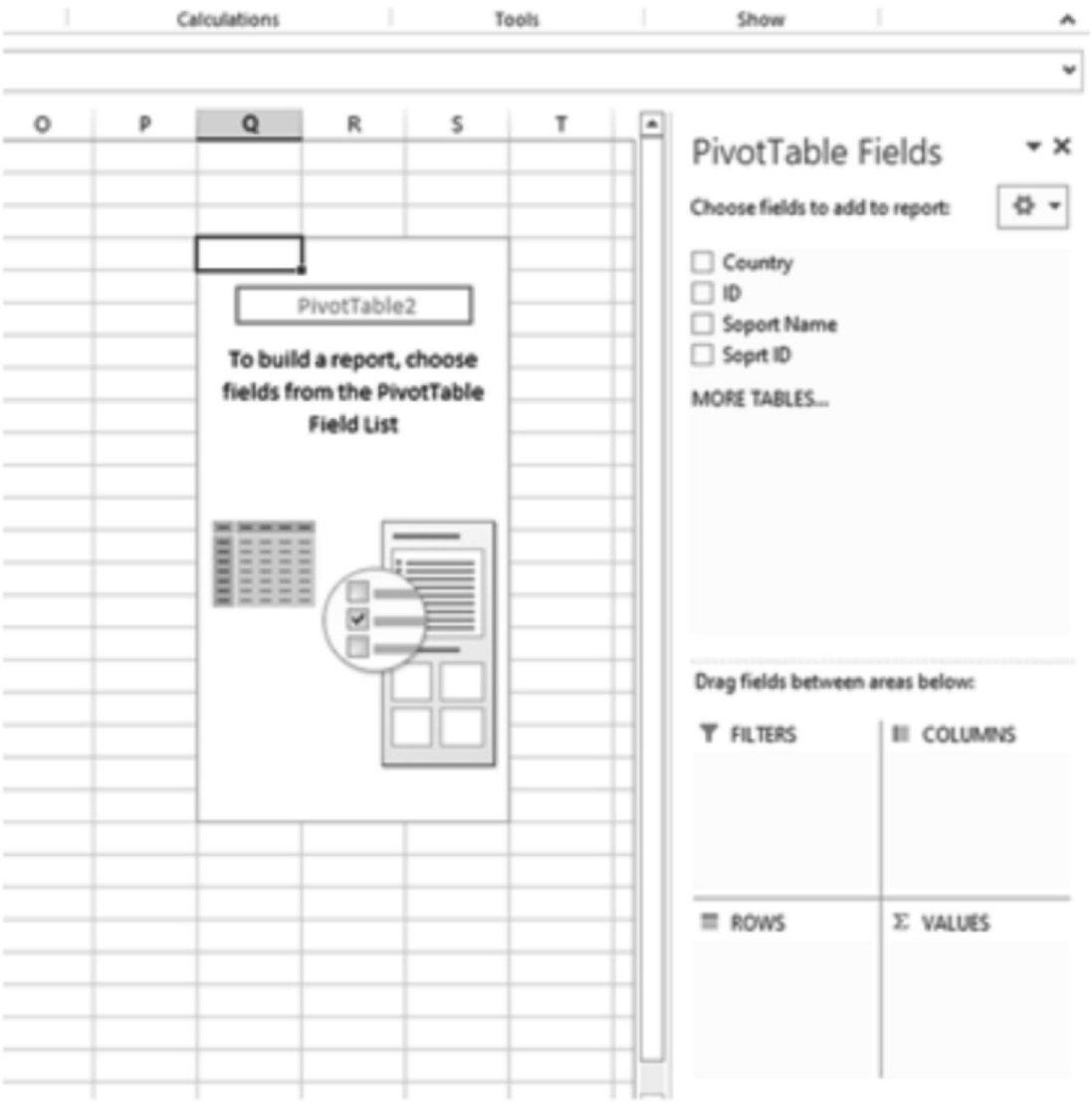
- Click the Open button. The Select Table talk put away shows, showing the tables in the informational collection. The access database is a social informational collection and the tables will resemble Excel tables, with the exclusion that associations exist among those tables.
- Check the case Enable the choice of different tables.
- Select all of the tables. Click OK.



- The Import Data trade put away shows. Select PivotTable Report. This decision brings the tables into your Excel practice manual and makes a PivotTable for separating the imported tables.



- As you notice, the checkbox Add this data to the Data Model is picked and debilitated, exhibiting that the tables will be added to the Data Model normally. The data will be imported and a void PivotTable will be made. The imported tables appear in the PivotTable Fields list.



Exploring Data in Multiple Tables

- You can look at the data from the imported various tables with PivotTable and appear at the specific report you really want in just two or three stages. This is possible because of the past associations among the tables in the source informational collection. As you imported all of the tables from the informational collection together at the same time, Excel duplicates the associations in its Data Model.
- In the PivotTable Fields show, you will find all of the tables that you imported and the fields in all of them. In case the fields are not evident for any table.

Short Questions and Answers

1. Data

Ans :

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

Two aspects of data can be identified.

(i) Record objective facts which will be understood in exactly the same way by everyone.

(ii) Record absolutely any type of concept, with no guarantees as to its accuracy or validity.

2. Data Analytics

Ans :

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.

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 technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

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. Importance of Analytics

Ans :

1. Business to Consumer Applications (B2C)

Organizations gather data which has been collected through various sources like customers, economy, businesses and practical experience. The collected data is then processed and categorized based on the requirement and the data is analyzed for studying the purchase patterns.

2. Growth of the Business

Data analytics enables companies to find the insights of the value chains which can be done by a data analytics expert. It offers comprehensive knowledge of industry and economy. It also helps in understanding the business aspects of the near future and obtain benefits before the competitors. Data analytics helps in seeing opportunities at the right time by providing analyzed data.

3. Data Analytics for Students

Data analytics offers the students an immense collection of tech tools to study. These tech tools considered as the key to industrial growth. It also

Rahul Publications

helps student to clearly understand the institutional data which enables further innovation. The scope for institutional research and exploration led to the emergence of data analytics as a high class or fashion courses.

4. Data Analytics in Educational Sector

In educational sector, new reforms are needed based on the necessities of the student and patterns of previous students Analysts are more helpful than researcher in this context as they have tools which help in analyzing the information in short period of time.

6. Define business analytics.

Ans :

Business analytics refers to the broad category of skills, tools and techniques employed for gathering, storing and analysing of business related data and information by any business enterprise. The main aim of business analytics is to help the management to arrive at an objective and accurate business decisions.

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. Percentiles and Quartiles

Ans :

Percentile refers to the percent of dividing the given series into 100 equal parts. It is a value at which atleast 'K' percent of the observations lie.

In Excel, percentile is calculated as PERCENTILE.INC which computes the Kth percentile of data with specified range in array field. The range of k is 0 to 1.

The following are the steps to calculate the kth percentile (where k is any number between zero and one hundred).

9. Cross Tabulation

Ans :

Cross tabulation is a method of representing the number of observations in a data set belonging to various categories of two categorical variables in tabular form. It is a widely used statistical tool which offers insights into various characteristics of different market segments in marketing.

10. Relative Frequency Distributions.

Ans :

In order to evaluate data, sometimes it becomes necessary to disclose the percentage of observations which may come under each distribution class rather than disclosing the actual frequencies of class. However, for converting frequency distribution into corresponding relative frequency distribution one should use the following formula,

Relative Frequency Distribution

$$= \frac{\text{Frequency Observation}}{\text{Total Number of Observations}}$$

Rahul Publications

Choose the Correct Answers

1. Data Analytics uses _____ to get insights from data. [c]
   (a) Statistical figures (b) Numerical aspects
   (c) Statistical methods (d) None of the above
2. Amongst which of the following is / are the branch of statistics which deals with the development of statistical methods is classified as _____. [c]
   (a) Industry statistics (b) Economic statistics
   (c) Applied statistics (d) None of the above
3. Linear Regression is the supervised machine learning model in which the model finds the best fit _____ between the independent and dependent variable. [a]
   (a) Linear line (b) Nonlinear line
   (c) Curved line (d) All of the above
4. Amongst which of the following is / are the types of Linear Regression, [c]
   (a) Simple Linear Regression (d) Multiple Linear Regression
   (c) Both A and B (d) None of the above
5. Amongst which of the following is / are the true about regression analysis? [b]
   (a) Describes associations within the data
   (b) Modeling relationships within the data
   (c) Answering yes/no questions about the data
   (d) All of the above
6. A Linear Regression model's main aim is to find the best fit linear line and the _____ of intercept and coefficients such that the error is minimized. [a]
   (a) Optimal values (b) Linear line
   (c) Linear polynomial (d) None of the above
7. The process of quantifying data is referred to as _____. [c]
   (a) Decoding (b) Structure
   (c) Enumeration (d) Coding
8. _____ are used when we want to visually examine the relationship between two quantitative variables. [b]
   (a) Bar graph (b) Scatterplot
   (c) Line graph (d) Pie chart
9. A graph that uses vertical bars to represent data is called a _____. [a]
   (a) Bar graph (b) Line graph
   (c) Scatterplot (d) All of the above
10. Data Analysis is a process of, [d]
   (a) Inspecting data (b) Data Cleaning
   (c) Transforming of data (d) All of the above

Rahul Publications

Fill in the blanks

1. _____ 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.
2. _____ organizations utilize analytics to provide more effective treatment and control cost. They use descriptive, prescriptive and predictive.
3. _____ Charts are one of the common yet popular techniques. It also comes under data visualization techniques in excel.
4. _____ chart basically displays the relationship between two patterns.
5. The _____ charts can display the data in a horizontal way or in a vertical way.
6. _____ graph is similar to a line chart but provides graphically quantitative data.
7. Excel _____ Table helps in exploring and extracting important data from an Excel table or a range of data.
8. _____ data is helpful when you have large amounts of data in a PivotTable or PivotChart.
9. To focus on a smaller portion of a large amount of your PivotTable data for in-depth analysis, you can _____ the data.
10. SPL stands for _____

Rahul Publications

ANSWERS

1. Big Data
2. Health care
3. Pie
4. Line
5. Bar
6. Area
7. Pivot
8. Sorting
9. Filter
10. Structured Query Language

Very Short Questions and Answers

1. Data.

Ans :

A collection of raw facts that can be used to conclude some information

2. Descriptive Analytics.

Ans :

Descriptive analytics answers the "what" questions in the data analytics process. It helps stakeholders understand large datasets by summarizing them.

3. Diagnostic Analytics.

Ans :

Diagnostic analytics answers the "why" questions in the data analytics process. It analyzes the results from the descriptive analysis and then further evaluates it to find the cause.

4. Unstructured Data.

Ans :

It is defined as the data in which is not follow a pre-defined standard or you can say that any does not follow any organized format.

5. Structured Data.

Ans :

Structured data is created using a fixed schema and is maintained in tabular format. The elements in structured data are addressable for effective analysis.

Rahul Publications

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 Descriptive Statistical Measures

2.1.1 Population and Samples

Q1. Define the terms Population and Samples.

Ans :

1. 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.

Examples

i) All registered voters in Crawford County

ii) All members of the International Machinists Union

iii) All Americans who played golf at least once in the past year

2. Samples

i) 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.

ii) A random sample is one in which every member of a population has an equal chance of being selected.

iii) 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.

iv) The field of inferential statistics enables you to make educated guesses about the numerical characteristics of large groups.

Examples

i) Budget of AP

ii) MBA students of XYZ college.

2.1.2 Measures of Location

Q2. Explain briefly about Measures of location.

(OR)

What is the best measure of location ?

(OR)

Brief on measures of location.

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

Measures of location can be used to estimate about the cluster of values or observations in the central part of the distribution. These are also known as 'Averages'. Averages are the values that lie between the smallest and the largest observations. It is the Mean of the given data. The four important measures of location are Arithmetic mean, Median, Mode and Mid Range.

1. Arithmetic Mean

The centre of data set for symmetric data is computed by a conventional, simple and most effectual approach known as 'Arithmetic mean'.

Rahul Publications

This approach is based on numeric quantities which are used to analyze the co-relation between different data values.

Mathematically, the mean of a sample for '*k*' observations is computed using the formula,

$$\bar{a} = \frac{a_1 + a_2 + a_3 + \ldots + a_1}{k}$$

$$\bar{a} = \sum_{i=1}^{k} \frac{a_i}{k}$$

Where, *a.* represents the set of *k* data values.

Calculating mean is similar to that of average calculation by using built-in function AVERAGE (data range) present in Excel. It operates based on the property that sum of the deviations for every single observation obtained from the mean is equal to 0.

$$i.e., \sum_i (d_i - \bar{a}) = 0$$

The mean of a population is given by,

$$\mu = \sum_{i=1}^{k} \frac{a_i}{k}$$

Rahul Publications

	A	B	C	D	E	F
1	Item	Status	Amount			
2	Cherry	Delivered	$100			
3	Banana	Delivered	$70			
4	Apple	Delivered	$130			
5	Banana	Delivered	$250			
6	Apple	Cancelled	$90			
7	Cherry	In transit	$115			
8	Banana	In transit	$90			
9						
10	Average	$121	=AVERAGE(C2:C8)			
11	(all items)					
12						
13	Average	$136.67	=AVERAGEIF(A2:A8,A14,C2:C8)			
14	Banana					
15						
16	Average	$160.00	=AVERAGEIFS(C2:C8,A2:A8,A17,B2:B8,A18)			
17	Banana					
18	Delivered					

2. Median

Another approach for estimating centre of data set for asymmetric data is 'median' which is an example of holistic measure. There are two cases for calculating median for fc-data values which are distinctive and sequentially ordered.

(i) When k is Odd

The median is the center value of entire sorted data set.

(ii) When k is Even

The median is the mean of two center data value.

MEDIAN (data range) is a EXCEL function which is used for ratio, ordinal and interval data.

	A	B	C	D
1	Item	Status	Amount	
2	Banana	Delivered	$70	
3	Apple	Cancelled	$90	
4	Banana	In transit	$90	
5	Cherry	Delivered	$100	
6	Cherry	In transit	$115	
7	Apple	Delivered	$130	
8	Banana	Delivered	$250	
9				
10	Median	$100	=MEDIAN(C2:C8)	

3. Mode

Mode measure is another way of computing central tendency. If a data value occurs more number of times than any other values in data set, then that data value is referred as mode. If a data set contains values which are distinctive then there is no mode in that data set. However, there are situations, where more than one data value may be occurring frequently in the same data set because of which there is a possibility of getting more than one mode. MODE.SNGL (data range) is an another excel function that also serves as mode. A data set with at least two modes are referred to as multimode).

	A	B	C	D
1	Item	Status	Amount	
2	Banana	Delivered	$70	
3	Apple	Cancelled	$90	
4	Banana	In transit	$90	
5	Cherry	Delivered	$100	
6	Cherry	In transit	$115	
7	Apple	Delivered	$130	
8	Banana	Delivered	$250	
9				
10	Mode	$90	=MODE(C2:C8)	

2.1.3 Measures of Dispersion

Q3. Explain the concept of Measures of Dispersion.

(OR)

Briefly on measures of Dispersion.

Ans : **(Oct.-22, Imp.)**

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

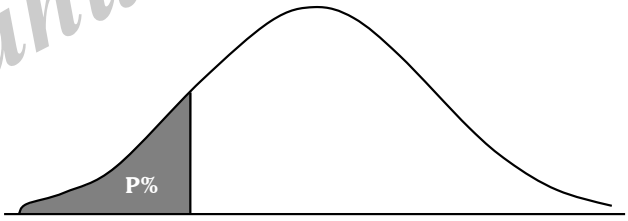
(b) Very much affected by extreme values (ranges is not a resistant measure of variability)

(ii) Interquartile Range (IQR)

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

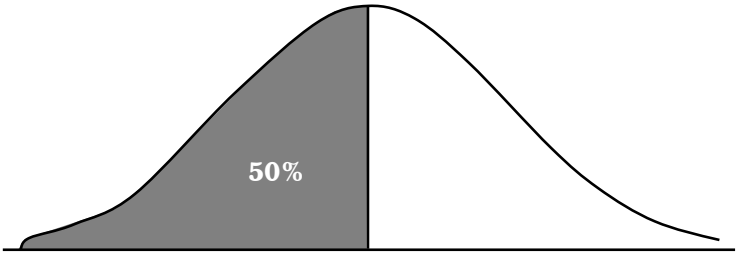
The path percentile of the data set is a measurement such that after the data are ordered from smallest to largest, at most, p% of the data are at or below this value and at most, (100 – p)% at or above it.

Rahul Publications



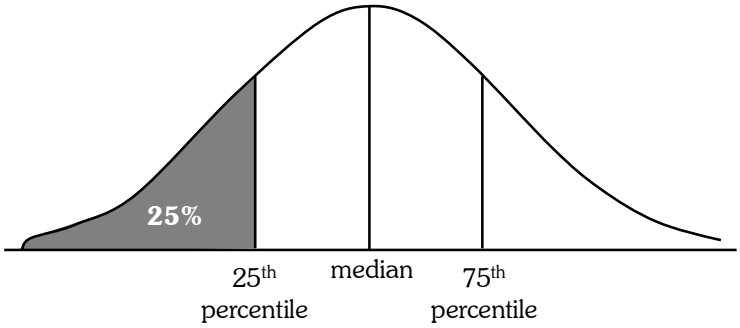
p^{th} percentile

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



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, 5, 4

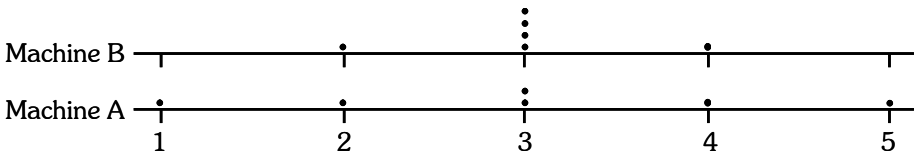
Mean = 3, Median = 3, Mode = 3

Pieces of candy from vending machine B:

2, 3, 3, 3, 4

Mean = 3, Median = 3, Mode = 3

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



(iv) Standard Deviation

The standard deviations defined s the square root of a variance. The standard deviation for a population can be computed by the formula,

$$s = \sqrt{\frac{\sum_{i=1}^{k}(a_i - \mu)^2}{k}}$$

The standard deviation for a samples is given by,

$$s = \sqrt{\frac{\sum_{i=1}^{k}(a_i - \bar{a})^2}{k-1}}$$

STDEV.P (data range) is the excel function used to determine the standard deviation for a populations (σ).

STDEV.S (data range) is the excel function used to determine the standard deviation for a sample(S).

2.1.4 Measures of Variability

Q4. Discuss about Measures of Variability.

(OR)

Describe about Measures of Variability.

Ans : **(Oct.-20, Dec.-19, Dec.-18)**

The following are the different Measures of Variability :

1. Range

Refer to Unit-II, Q.No. 3

2. Variance

Refer to Unit-II, Q.No. 3

3. Standard Deviation

Refer to Unit-II, Q.No. 3

4. Coefficient variation

The coefficient of variation caters the relative measure of dispersion in data corresponding to the mean.

As, the standard deviation is an absolute measure of dispersion, it cannot be used in carrying out a comparison that includes different units.

Coefficient of variation is generally used to compare the consistency or variability of different samples. When the coefficient of variation is greater, the consistency of the sample is smaller. When the coefficient of variation is smaller, the consistency of the sample is greater.

Coefficient of variation, $CV = \frac{S_d}{\bar{a}} \times 100$

Where, S_d = Standard deviation

$\bar{a}$ = Mean.

Rahul Publications

PROBLEMS

1. 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.

Sol. : **(Aug.-21)**

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 × 3 = – 90	2700
10 - 20	5	15	15 – 35 = –20	–20 × 5 = –100	2000
20 – 30	16	25	25 – 35 = –10	–10×16 =–160	1600
30 – 40	18	A 35	35 – 35 = 0	0 ×18 = 0	0
40 – 50	12	45	45 – 35 = 10	10 × 12 = 120	1200
50 – 60	7	55	55 – 35 = 20	20 × 7 = 140	2800
60 – 70	4	65	65 – 35 = 30	30 × 4 = 120	3600

Rahul Publications

Now,

Σf_i=Q 2000 + 1600 + 0 + 1200 + 2800 + 3600

$\Sigma f_i d_i^2 = 13{,}900$

We know that

$$\text{Standard deviation r} = \sqrt{\frac{\Sigma fd^2}{N}\left(\frac{\Sigma f_i d_i}{N}\right)^2}$$

$$= \sqrt{\frac{13{,}900}{65} - \left(\frac{30}{65}\right)} = \sqrt{213.84 - (0.46)^2}$$

$$= \sqrt{213.84 - 0.21}$$

$$= \sqrt{213.627}$$

$$= 14.615$$

The standard deviations = 14.615.

2. Find the Value of Range and the Coefficient for the following data:

7, 9, 6, 8, 11, 10, 4

Sol : **(Oct.-22, Imp.)**

Range = Highest Value – Lowest Value

$$11 - 4 = 7$$

$$\text{Coefficient of Range} = \frac{\text{Highest value} - \text{Lowest value}}{\text{Highest value} + \text{Lowest value}}$$

$$= \frac{11-4}{11+4} = \frac{7}{15} = 0.46$$

2.1.5 Measures of Association

Q5. Explain briefly about Measures of Association.

(OR)

Discuss about measures of association.

(OR)

What are measures of Variability.

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

The following are the different measures of variability.

i) Range

For answer refer Unit-II, Q.No. 3.

ii) Variance

For answer refer Unit-II, Q.No. 3.

iii) Standard Deviation

For answer refer Unit-II, Q.No. 3.

iv) Coefficeint of variation

For answer refer Unit-II, Q.No. 4.

2.2 Probability Distribution and Data Modeling

2.2.1 Discrete Probability Distribution, Continuous Probability Distribution

Q6. Define Probability Distribution. Explain briefly about Discrete Probability Distribution.

Ans : **(May-22, Dec.-19)**

Meaning

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.

Rahul Publications

Discrete Probability Distribution

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.

Discrete probability distribution is the probability distribution of a discrete random variable 'X' which takes only a finite number of variables. Random variable 'X' as function y(x) satisfies the following conditions,

(i) $f(x) \geq 0$

(ii) $\Sigma f(x) = 1$

(iii) $P(X=x)=fx$

Probability Function (or) Probability Mass Function

Let X be discrete random variable. The discrete probability function f(x) for X is given $f(x) = p(X = x)$ for all real x.

Since, probabilities cannot be negative, a probability function f(x) cannot assume negative values.

Note that the probability associated with a sample space is 1. Thus, If the values of f(x) are added over all possible values of X, then the total should be 1.

Infact the following two properties completely characterize the probability function of a discrete random variable.

Properties

1. $f(x) \geq 0$

2. $\sum_{\text{all } x} f(x) = 1$

Now, the discrete probability distribution, probability function or probability mass function of a discrete random variable X as the function f(x) satisfying the following condition,

(i) $f(x) \geq 0$

(ii) $\sum_{\text{all } x} f(x) = 1$

(iii) $P(X = x)=f(x)$

Cumulative Distribution Function

The probability F(x) is used to determine whether the value of a random variable is less than or equal to x. This is given by

$$\boxed{F(x) = P(X \leq x);}$$

Where X is a random variable.

Q7. Define binomial distribution. Explain the properties of binomial distribution.

(OR)

What is binomial distribution. Give its properties.

Ans : **(Imp.)**

Meaning

In binomial probability distribution, the number of 'Success' in a sequence of n experiments, where each time a question is asked for yes-no, then the boolean-valued outcome is represented either with success/yes/true/one (probability p) or failure/no/false/zero (probability q = 1 - p). A single success/failure test is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process.

For n = 1, i.e. a single experiment, the binomial distribution is a Bernoulli distribution. The binomial distribution is the base for the famous binomial test of statistical importance.

Binomial Distribution Formula

The binomial distribution formula is for any random variable X, given by;

$$P(x:n, p) = {}^nC_x\, p^x\, (1-p)^{n-x}$$

OR

$$P(x:n, p) = {}^nC_x\, p^x\, (q)n-x$$

Where,

n = the number of experiments

x = 0, 1, 2, 3, 4, …

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

The binomial distribution formula can also be written in the form of n-Bernoulli trials, where

$^nC_x = n!/x!(n-x)!$. Hence,

$P(x:n, p) = n!/[x!(n-x)!].p^x.(q)^{n-x}$

Properties of Binomial Distribution

The properties of the binomial distribution are:

i) There are two possible outcomes: true or false, success or failure, yes or no.

ii) There is 'n' number of independent trials or a fixed number of n times repeated trials.

iii) The probability of success or failure remains the same for each trial.

iv) Only the number of success is calculated out of n independent trials.

v) Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

Formula for Binomial Distribution

= **BINOM.DIST(number_s,trials,probability_s,cumulative)**

The BINOM.DIST utilizes the accompanying contentions:

1. Number_s (required argument): This is the number of accomplishments in preliminaries.
2. Trials (required argument): This is the number of autonomous preliminaries. It should be more noteworthy than or equivalent to 0.
3. Probability_s (required argument): This is the likelihood of progress in every preliminary.
4. Cumulative (required argument): This is a legitimate worth that decides the type of capability. It can either be:

   i) TRUE: Utilizes the aggregate conveyance capability.

   ii) FALSE: Utilizes the likelihood mass capability.

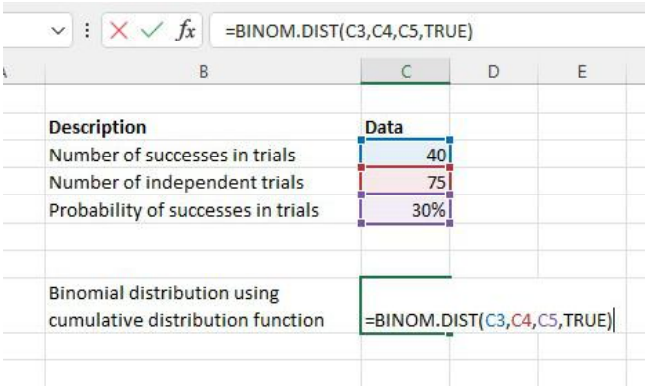
   iii) Example: Assume we are given the accompanying information,

Description	Data
Number of successes in trials	40
Number of independent trials	75
Probability of successes in trials	30%

Rahul Publications

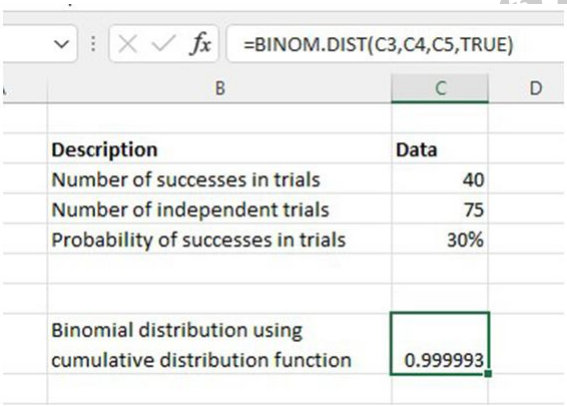
Step 1

The equation for ascertaining binomial distribution utilizing the total distribution capability is displayed underneath,



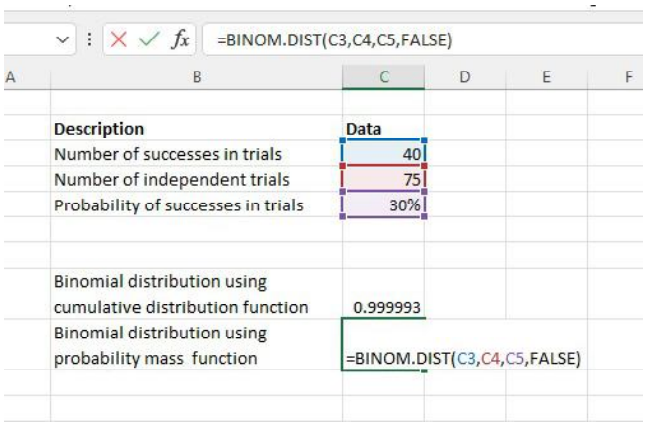
Step 2

We come by the outcome underneath.



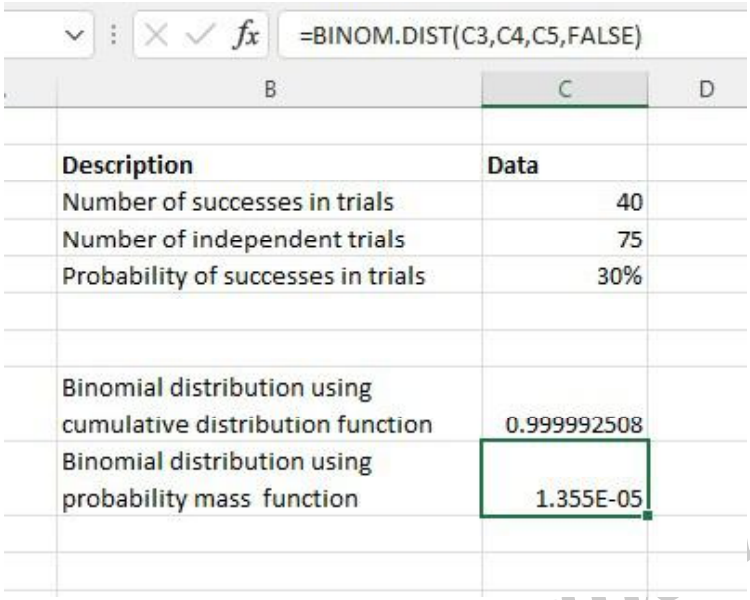
Step 3

The recipe for working out binomial distribution utilizing the likelihood mass capability is displayed beneath:



Step 4

We come by the outcome beneath.



=BINOM.DIST(C3,C4,C5,FALSE)

	B	C	D
	Description	**Data**	
	Number of successes in trials	40	
	Number of independent trials	75	
	Probability of successes in trials	30%	
	Binomial distribution using cumulative distribution function	0.999992508	
	Binomial distribution using probability mass function	1.355E-05	

Q8. Explain briefly about Poisson distribution.

Ans :

Definition

The Poisson distribution is a discrete probability function that means the variable can only take specific values in a given list of numbers, probably infinite. A Poisson distribution measures how many times an event is likely to occur within "x" period of time. In other words, we can define it as the probability distribution that results from the Poisson experiment. A Poisson experiment is a statistical experiment that classifies the experiment into two categories, such as success or failure. Poisson distribution is a limiting process of the binomial distribution.

A Poisson random variable "x" defines the number of successes in the experiment. This distribution occurs when there are events that do not occur as the outcomes of a definite number of outcomes. Poisson distribution is used under certain conditions. They are:

The number of trials "n" tends to infinity

Probability of success "p" tends to zero

np = 1 is finite

Poisson Distribution Formula

The formula for the Poisson distribution function is given by:

$$\boxed{f(x) = (e^{-\gamma}\lambda^{x}) / x!}$$

Where,

e is the base of the logarithm

x is a Poisson random variable

λ is an average rate of value

Q9. Explain briefly about continuous probability distribution.

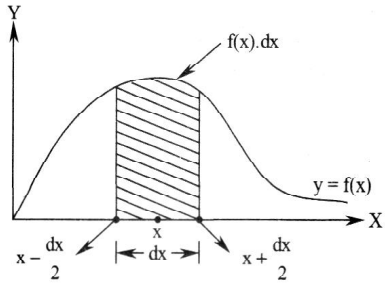
Ans : **(May-22)**

Meaning

Continuous probability distribution is a probability distribution of a continuous random variable. A continuous random variable is a variable which assumes infinite and uncountable set of values. Example of continuous probability distribution is 'Normal Distribution'.

Probability Density Function (p.d.f)

Consider the small interval (x, x + dx) of length dx around the point x. Let f(x) be any continuous function of x so that f(x)dx represents the probability that X falls in the interval (x, x + dx). Symbolically,



In the figure, f(x)dx represents the area bounded by the curve y = f(x), x-axis and the ordinates at the points x and x + dx. The function f(x) so defined is known as probability density function or simply density function of random variable X.

Note

1. The expression, f(x)dx is known as the probability differential.
2. The curve y = f(x) is known as the probability density curve or simply probability curve.

The probability for a variate value to lie in the interval dx is f(x) dx and hence, the probability for a variate value to fall in the finite interval [a, b] is,

$$P(a \le X \le b) = \int_a^b f(x)dx$$

which represents the area between the curve y = f(x), x-axis and the ordinates at x = a and x = b.

Since, the total probability is unit $\int_a^b f(x)dx = 1$

Properties

The following are the properties of probability density function.

(i) $f(x) \ge 0, -\infty < x < \infty$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1.$

(iii) The probability P(E) given by,

$$P(E) = \int_E f(x)dx$$

(iv) $P(a < X < b) = \int_a^b f(x)dx$ = Area under f(x) between ordinates x = a and x = b.

Rahul Publications

Q10. Define normal distribution. Explain the features of normal distribution.

(OR)

What is the nature of Gaussian distribution? How is it, unique?

Ans : **(April-20)**

Meaning

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.

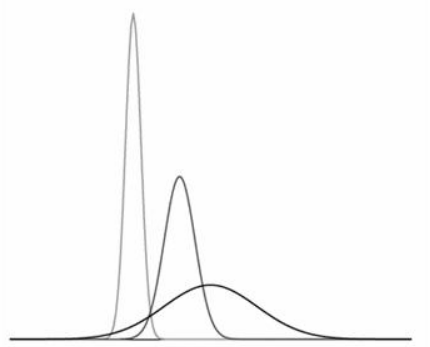


Fig. 1: Normal distributions differing in mean and standard deviation

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.

Features

Seven features of normal distributions are listed below.

i) Normal distributions are symmetric around their mean.

ii) The mean, median, and mode of a normal distribution are equal.

iii) The area under the normal curve is equal to 1.0.

iv) Normal distributions are denser in the center and less dense in the tails.

v) Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).

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

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

Q11. Discuss the assumptions and properties of normal distribution.

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

Assumptions

i) Normal distribution is the most essential distribution that is used in statistics.

ii) It is a continuous distribution which is described by using a bell-shaped curve.

iii) 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.

iv) Similarly, when the value of μ increases/decreases the distribution either becomes broad or gets concised.

Properties

i) The normal distribution curve can be symmetric and its skewness found to be zero.

ii) 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.

iii) 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.

Q12. Example explain multiplication law of probability.

Ans : **(April-23)**

If 'A' and 'B' are two independent events then the probability of occurrence of both the events is equal to the product of their individual probabilities.

For independent events,

$$\boxed{P(A \cap B) = P(A) \cdot P(B)}$$

Similarly,

$$\boxed{P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)} \text{ and so on}$$

If 'A' and 'B' are two dependent events, in such a case multiplication theorem is altered and is given as follows. For dependent events,

$$\boxed{\begin{aligned} P(A \cap B) &= P(A/B) \cdot P(B) \\ &= P(B/A) \cdot P(A) \end{aligned}}$$

Where, P(A/B) is a conditional probability of A given that B has occurred (The probability of occurrence of event A when event B has already occurred is the conditional probability of A given B).

PROBLEMS

3. Solve the below given problem. Let X be a discrete random variable with the following PMF.

$$P_x(k) = \begin{cases} 0.1 & \text{for } k = 0 \\ 0.4 & \text{for } k = 1 \\ 0.3 & \text{for } k = 2 \\ 0.2 & \text{for } k = 3 \\ 0 & \text{otherwise} \end{cases}$$

(i) Find EX

(ii) Find Var(X)

(iii) If Y = (X – 2)², find EY

Sol : **(May. 22)**

Given that,

k	0	1	2	3	Otherwise
$P_x(k)$	0.1	0.2	0.3	0.4	0

(i) EX

$EX = \Sigma.k.P_x(k)$

$= 0 \times 0.1 + 1 \times 0.2 + 2 \times 0.3 + 3 \times 0.4 + 0$

$= 0 + 0.2 + 0.6 + 1.2$

$= 2$

$EX = 2$

(ii) Var(X)

$Var(X) = E(X^2) - [E(x)]^2$

$= [0(0.1) + 1^2(0.2) + 2^2(0.3) + 3^2(0.4) + 0] - 2^2$

$= [0 + 0.2 + 1.2 + 3.6] - 4$

$= 5 - 4 = 1$

$\therefore Var(X) = 1$

(iii) EY

$EY = E(X - 2)^2$

$= (0-2)^2(0.1) + (1-2)^2(0.2) + (2-2)^2(0.3) + (3-2)^2(0.4) + 0$

$= (4 \times 0.1) + (1 \times 0.2) + (0 \times 0.3) + (1 \times 0.4)$

$= 0.4 + 0.2 + 0 + 0.4$

$= 1$

$\therefore EY = 1$

4. 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?

Rahul Publications

Sol : **(Aug.-21, Imp.)**

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,

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)

P(X >30) = 0.40

Therefore, that probability that X greater than 30 is '0.40'.

5. 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.

Rahul Publications

Sol : **(May-19, Imp.)**

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 $P(x \leq c) > \frac{1}{2}$

$c = 4$

2.2.2 Random Sampling from Probability Distribution, Data modeling and Distribution Fitting

Q13. Enumerate the various sampling techniques.

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

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:

1. 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.

2. Non-probability Sampling

This sampling techniques is not based on random selection. Some examples are quota sampling, purposive sampling and convenience sampling.

1. Probability Sampling

The techniques in probability 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).

(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.

Rahul Publications

(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.

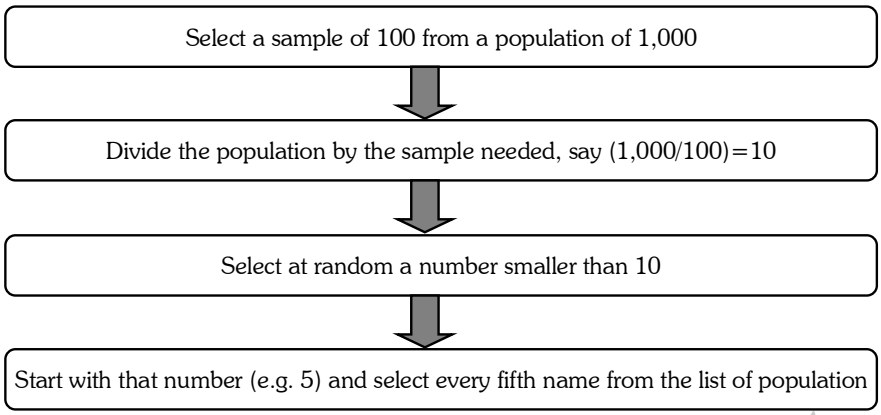


Fig. : Example of systematic sampling

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.

2. 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.

Rahul Publications

(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.

Q14. Explain the merits and demerits of random sampling methods.

Ans : **(May-19)**

Merits

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

Rahul Publications

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.

Demerits

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 conclusions 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.

Rahul Publications

6. The students of a class have elected live candidates to represent them on the college management council.

S.No.	Gender	Age in years
1.	Male	18
2.	Male	19
3.	Female	22
4.	Female	20
5.	Male	23

This group decides to elect a spokesperson by randomly drawing a name from a hat. Calculate the probablity of the spokesperson being either female or over 21 years.

Sol : **(April-23)**

No. of ways of selecting one member n(s) = 5

The probability of the spokes person is female $P(A) = \frac{n(A)}{n(S)} = 2/5$

Members overs 21 year of age are n(B) = 2

The probability that the spokesperson is over 21 years.

$$P(B) = \frac{n(B)}{n(S)} = \frac{2}{5}$$

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

$$= \frac{2}{5} + \frac{2}{5} - \frac{1}{5}$$

$$= \frac{4-1}{5} = \frac{3}{5}$$

Rahul Publications

Q15. What is data modeling? Explain different types of data modeling.

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

Meaning

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.

Types

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.

Q16. Explain the advantages and disadvantages of Data Model?

Ans :

Advantages and Disadvantages of Data Model:

Advantages

i) The main goal of a designing data model is to make certain that data objects offered by the functional team are represented accurately.

ii) The data model should be detailed enough to be used for building the physical database.

iii) The information in the data model can be used for defining the relationship between tables, primary and foreign keys, and stored procedures.

iv) Data Model helps business to communicate the within and across organizations.

v) Data model helps to documents data mappings in ETL process

vi) Help to recognize correct sources of data to populate the model

Disadvantages

i) To developer Data model one should know physical data stored characteristics.

ii) This is a navigational system produces complex application development, management. Thus, it requires a knowledge of the biographical truth.

iii) Even smaller change made in structure require modification in the entire application.

iv) There is no set data manipulation language in DBMS.

Rahul Publications

Q17. Write about Data Modeling and Distribution Fitting?

(OR)

Explain about data modeling and distribution fitting in detail.

(OR)

Brief on data modeling and data fitting.

Ans : **(Oct.-22, Dec.-19, Imp.)**

i) Most of the business analytics applications requires information about variables such as customer demand, purchase behavior, machine failure times and service activity times for generating the distributions.

ii) The probability distribution for a sample data can be obtained by fitting.

iii) The theoretical distribution and validating its goodness.

iv) An appropriate theoretical distribution can be selected by analysing the histograms for that specific distribution.

v) However, this approach does not perform well incase of small sample size. Summary statistics is another approach that helps in analyzing the nature of distribution.

vi) The information about nature of distribution is obtained by using the mean, median, standard deviation and coefficient of variation. For example, in a normally distributed data mean and median are approximately similar.

vii) Whereas, in exponentially distributed data the value of mean will be greater than the median and equal to the standard deviation.

viii) Therefore, by examining histograms and summary statistics the user can get only some idea of the appropriate distribution.

ix) So the better approach would be to fit the data analytically to the best type of probability distribution.

Goodness of Fit

i) Goodness of fit is the statistical method used for fitting the data into a probability distribution. This is used to conclude the nature of distribution.

ii) The fitting of data into a distribution is determined by any one of the statistic method such as chi-square, Kolmogorov-Smimov and Anderson-Darling statistics.

iii) The comparison of data of histogram with a particular theoretical probability distribution can be measured or given by those statistics.

iv) The chi-square method is used to divide the theoretical distribution into areas of equal probability and also used for comparing the data points in every area to the number which is expected for that distribution.

v) The Kolmogarov-Smimov method is used for comparing the cumulative distribution of the data with the theoretical distribution and concluded by finding the maximum vertical distance among them.

vi) The Anderson-Darling method is used for providing more weight for the differences between the rails of the distributions.

vii) Analytic solver platform is used to fit the probability distribution to the given data by using any one of the methods of goodness-of-fit. This is used to perform for analyzing and passing inputs to the simulation Models.

Rahul Publications

Short Questions 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:

i) All registered voters in Crawford County

ii) All members of the International Machinists Union

iii) All Americans who played golf at least once in the past year.

2. Sample

Ans :

i) Often, researchers want to know things about populations but do not have data for every person or thing in the population.

ii) 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.

iii) 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.

iv) A random sample is one in which every member of a population has an equal chance of being selected.

v) 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.

vi) 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.

Rahul Publications

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 :

i) Offers Organisation-wide coverage of the business concepts.

ii) This type of Data Models are designed and developed for a business audience.

iii) 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. Discrete 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 experi-ment 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.

Rahul Publications

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

i) Data objects

ii) The associations between different data objects

iii) The rules.

Data modeling helps in the visual represen-tation 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.

Rahul Publications

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 occur-rence 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.

12. Measures of association.

Ans :

The following are the different measures of variability.

i) Range

ii) Variance

iii) Standard Deviation

iv) Coefficeint of variation

13. Measures of location.

Ans :

Measures of location can be used to estimate about the cluster of values or observations in the central part of the distribution. These are also known as 'Averages'. Averages are the values that lie between the smallest and the largest observations. It is the Mean of the given data. The four important measures of location are Arithmetic mean, Median, Mode and Mid Range.

14. Measures of Variability.

Ans :

The following are the different Measures of Variability :

1. Range
2. Variance
3. Standard Deviation
4. Coefficient variation

Rahul Publications

Choose the Correct Answers

1. The weight of the earth is _____. [c]

   (a) Discrete variable (b) Qualitative variable

   (c) Quantitative variable (d) Difficult to tel

2. If $\bar{x}$ is 4 and the distribution is 2, 3, 4, 5, 6, the sum of squared deviations from the $\bar{x}$ will be: [b]

   (a) 8 (b) 10

   (c) 6 (d) 12

3. Given the N values in a series, the geometric mean is [d]

   (a) The third root of the product of N values

   (b) The square root of the product of N values

   (c) The fourth root of the product of N values

   (d) The Nth root of the product of N values

4. A statistics which is not measurable is called [b]

   (a) A constant (b) An attribute

   (c) A variable (d) A parameter

5. A contractor employs 20 male, 15 female and 5 children in his factory. Male wages are Rs. 10 per day, female Rs. 8 per day, and children Rs. 3 per day. The weighted $\bar{x}$ of wages paid per day will be

   (a) 3.86 (b) 8.37

   (c) 9.21 (d) 10.63

6. The number 143.9500 rounded off to the nearest tenth (one decimal place) is [b]

   (a) 143.9 (b) 144.0

   (c) 143.0 (d) 144

7. When all the values in a series occur the same number of times, then one must not compute the [c]

   (a) Mean (b) Median

   (c) Mode (d) Weighted mean

8. The process of arranging data into rows and columns is called [c]

   (a) Frequency distribution (b) Classification

   (c) Tabulation (d) Array

Rahul Publications

9. The average monthly production of a factory for the first 8 months is 2,500 units, and for the next 4 months, the production was 1,200 units. The average monthly production of the year will be [c]

(a) 2066.55 units (b) 5031.10 units

(c) 4021.12 units (d) 3012.11 units

10. In a frequency distribution the last cumulative frequency is 300, Median shall lie in: [d]

(a) 140th item (b) 130th item

(c) 160th item (d) 150th item

Rahul Publications

Fill in the Blanks

1. _____ is concerned with how each variable is related to the other variable(s).
2. A _____ distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.
3. _____ 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.
4. There are _____ schemas of data modeling.
5. _____ is the process of developing data model for the data to be stored in a Database.
6. The _____ distribution represents the number of failures before you get a success in a series of Bernoulli trials.
7. Data modeling is the process of creating a data model for the data to be stored in a _____ .
8. UML stands for _____.
9. MLE stands for _____.
10. The aim of distribution fitting is to predict the _____.

ANSWERS

1. Association
2. Probability
3. Data
4. Three
5. Data modeling
6. Geometric
7. Database.
8. Unified Modelling Language
9. Maximum Likelihood Estimation
10. Probability

Rahul Publications

Very Short Questions and Answers

1. What is range?

Ans :

Range for k-data values $a_1, a_2, \ldots\ldots a_k$ can be computed by finding out the difference between the maximum value and minimum value i.e., subtracting maximum data value from minimum data value.

Range = MAX(data range) – MIN (data range).

2. What is Chebyshev's theorem?

Ans :

Chebyshev's theorem states that the probability of obtaining values that lies with in K standard deviations of mean can be at least $1 - 1/K^2$ for any number or amount of data.

3. What is meant by coefficient of variation?

Ans :

The coefficient of variation provides the relative measure of dispersion in data corresponding to the mean. It can be given as, CV = Standard deviation/ mean.

As, the standard deviation is an absolute measure of dispersion, it cannot be used in carrying out a comparison that includes different units.

4. What is covariance?

Ans :

Covariance is defined as the measure that determines the association among two variables X and Y linearly. Covariance of a population is given by the average of products of deviations for every observation with their respective means.

$$\text{Cov}(X, Y) = \sum_{i=1}^{k} \frac{(x_i - \mu_x)(y_i - \mu_y)}{k}$$

5. What is the use of cumulative distribution function?

Ans :

The cumulative distribution function F(x) is used to determine whether the value of a random variable is less than or equal to x. This is given by,

$$\boxed{F(x) = P(X \leq x);}$$

Where X is a random variable.

Rahul Publications

UNIT III

Karl Pearson Correlation Technique: 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 Pearson Correlation Technique

Q1. Define correlation. Explain the significance of correlation.

Ans : **(Imp.)**

Meaning

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.

Definitions

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

i) **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".

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

Significance

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. 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.
3. The coefficient of correlation is a relative measure and we can compare the relationship between variables, which are expressed in different units.
4. 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.
5. Sampling error can also be calculated.
6. Correlation is the basis for the concept of regression and ratio of variation.
7. The decision making is heavily felicitated by reducing the range of uncertainty and hence empowering the predictions.

Q2. Explain different types correlation.

Ans :

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

Rahul Publications

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.

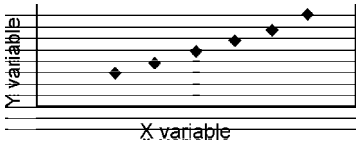


Fig. : Perfect Positive Correlation (r = +1)

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.

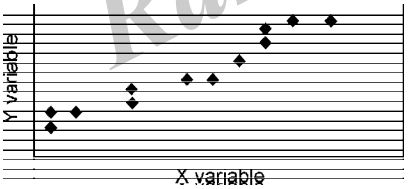


Fig. : Very High Positive Correlation (r = nearly + 1)

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.

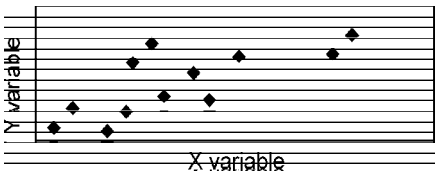


Fig.: Very Low Positive Correlation (r = near to +0)

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.

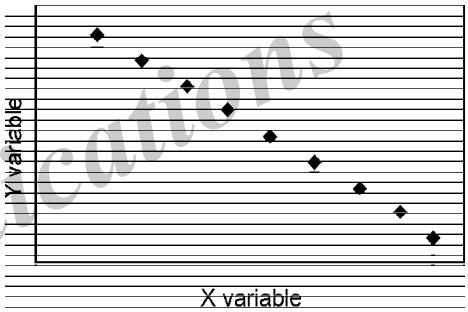


Fig. : Perfect Negative Correlation (r = –1)

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.

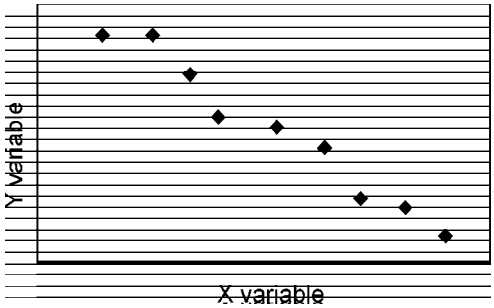


Fig. : Very High Negative Correlation (r = near to –1)

iii) Very low Negative Correlation : If the variables X and Y are related to each other

Rahul Publications

with a very low degree of negative relationship then we can notice a graph as in fig. below.

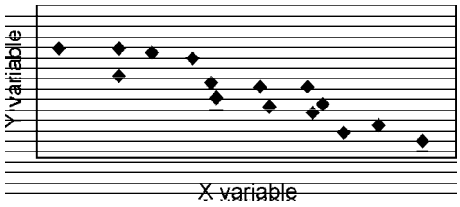


Fig.: Very Low Negative Correlation (r = near to 0 but negative)

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.

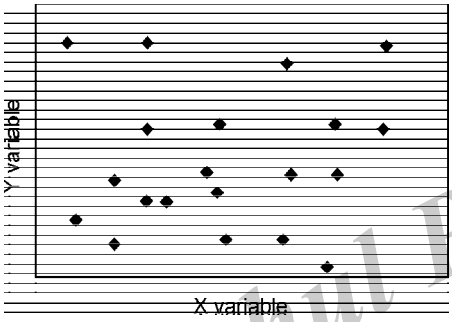


Fig. : No Correlation (r = 0)

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.

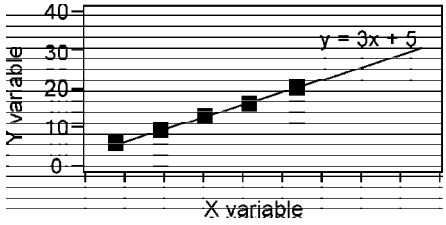


Fig.: Linear Correlation

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

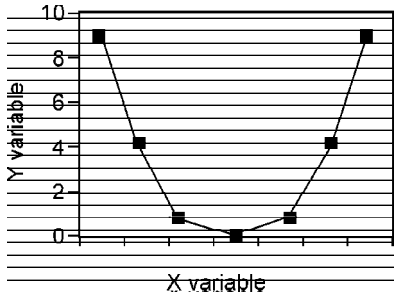


Fig.: Non Linear Correlation

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

Rahul Publications

(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. Explain the properties of correlation

Ans :

Properties

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 = 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.

Q4. State the merits and demerits of correlation.

Ans :

Merits

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.

Rahul Publications

Demerits

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

$$x = X - \bar{X},$$

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 :

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

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

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

Q7. Explain the computation of correlation coefficient using MS Excel.

Ans :

Step 1

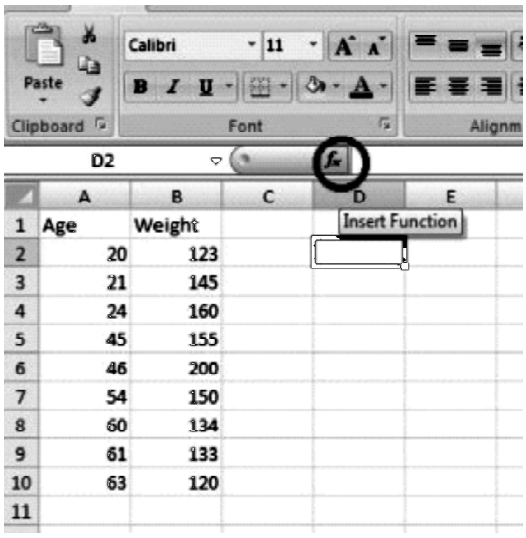
Type your data into two columns in Excel. For example, type your "x" data into column A and your "y" data into column B.

Step 2

Select any empty cell.

Step 3

Click the function button on the ribbon.

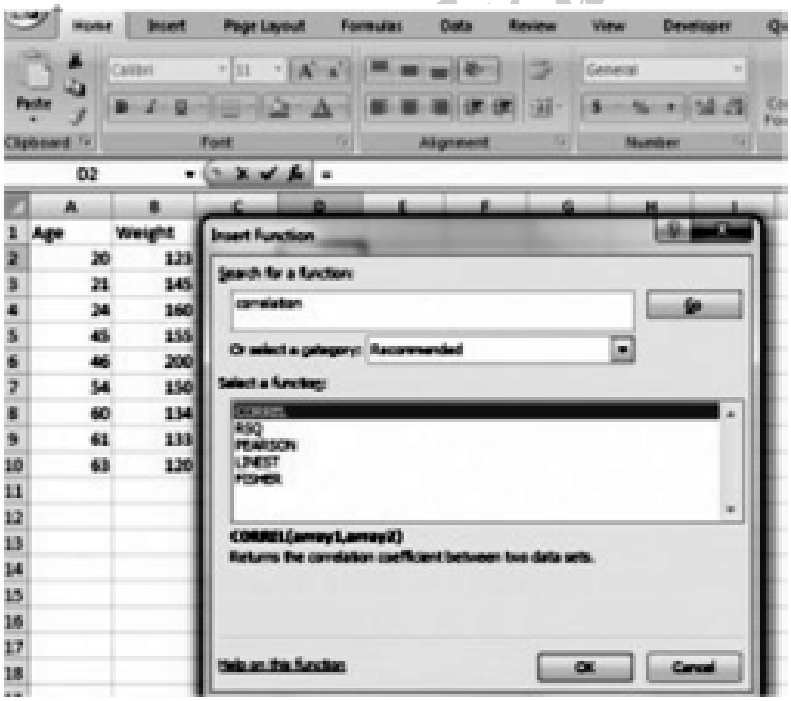


Step 4

Type "correlation" into the 'Search for a function' box.

Step 5

Click "Go." CORREL will be highlighted.

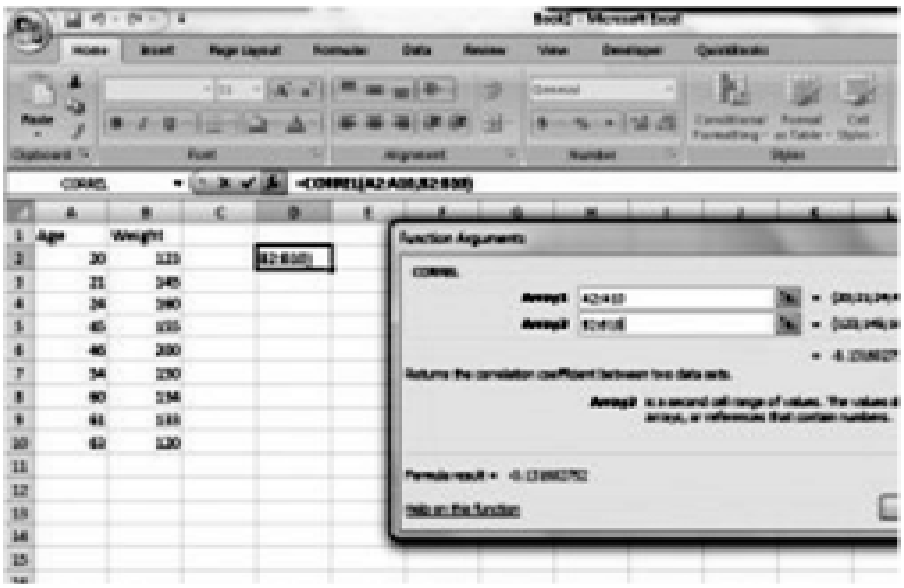


Step 6

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.



Q8. Explain the use of scatter chart for determining covariance.

Ans : **(April-22)**

Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis. These plots are often called scatter graphs or scatter diagrams.

A scatter plot is also called a scatter chart, scattergram, or scatter plot, XY graph. The scatter diagram graphs numerical data pairs, with one variable on each axis, show their relationship. Now the question comes for everyone:

Scatter plots are used in either of the following situations.

(i) When we have paired numerical data

(ii) When there are multiple values of the dependent variable for a unique value of an independent variable

(iii) In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems, checking whether two products that appear to be related both occur with the exact cause and so on.

3.2 Multiple Correlation

Q9. Define Multiple Correlation. Explain the steps involved in generating correlation coefficient in MS Excel.

(OR)

Explain about multiple correlation.

Rahul Publications

Ans : **(May-22)**

Meaning

When more than two variables are identified, then the relationship between a single dependable variable and multiple independent variables are considered. Therefore, this type of correlation is said to be 'Multiple Correlation'.

The multiple correlation coefficient between x_1 and x_2, x_3 is given by,

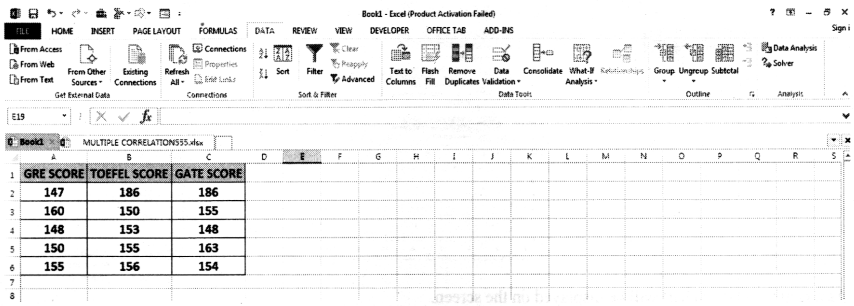
$$R_{1,23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Consider the following data to illustrate the concept of multiple correlation using excel.

GRE SCORE	TOEFEL SCORE	GATE SCORE
147	186	186
160	150	155
148	153	148
150	155	163
155	156	154

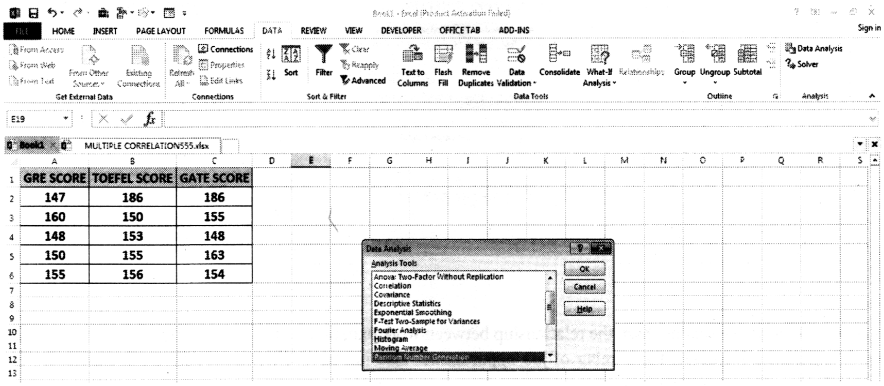
The steps to be followed for generating correlation coefficients between multiple variables are as follows,

1. Click on the 'Data' tab present on the excel ribbon.

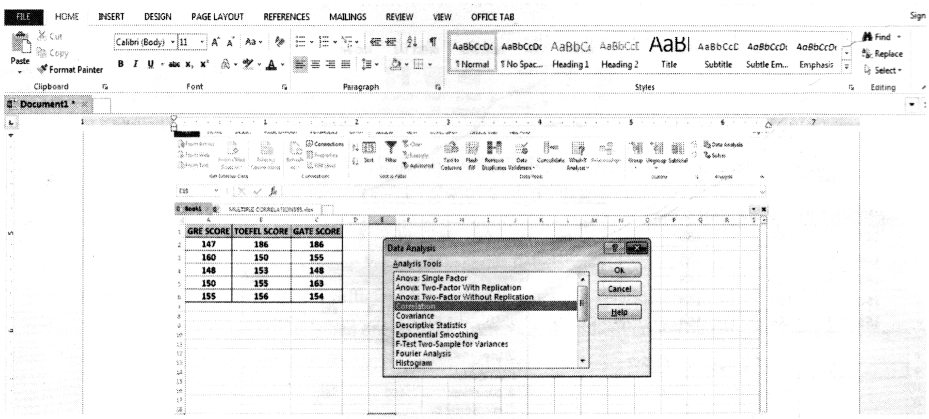


2. Select 'Data Analysis' command present under 'Analysis' group.

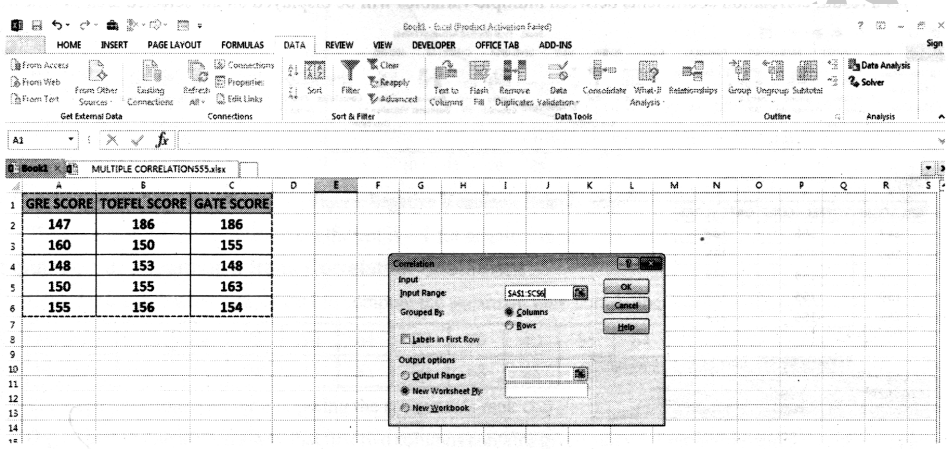
   A 'Data Analysis' dialog box will be displayed.



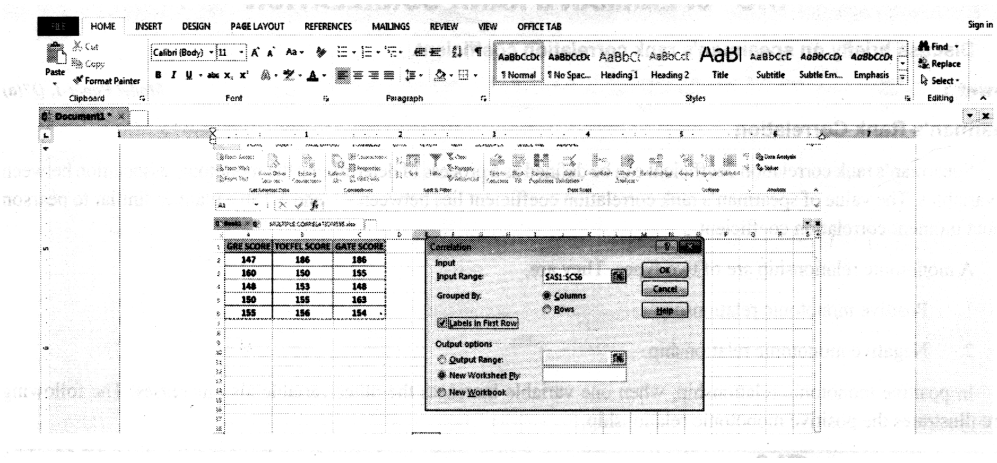
3. Goto 'Analysis Tools' section and select 'Correlation' option from drop down menu and then click on "OK" button.



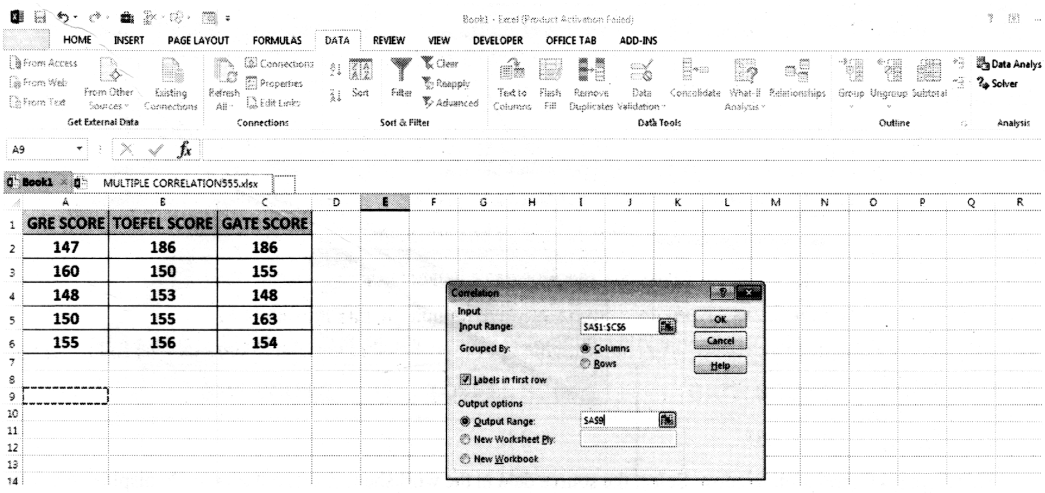
4. This displays Correlation window. Select the source data in the worksheet. Excel automatically displays the selected range in the 'Input Range' field of 'Correlation' dialog box.



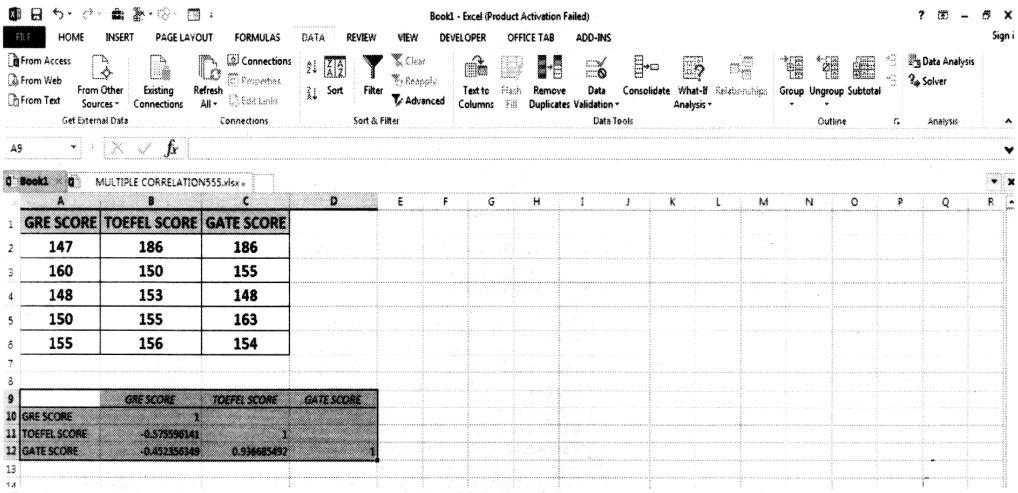
5. Checkmark the checkbox beside 'Labels in First Row' option.



6. Specify the Output Range in 'Output Option' field by clicking on any one of the empty all in the worksheet and then



7. As a result, correlation coefficients between multiple variables will be displayed on the selected area as follows,



3.3 Spearman's Rank Correlation

Q10. Define Rank correlation. Explain the properties of rank Correlation.

Ans :

Meaning

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.

$$R = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \text{ or } 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Where

R denotes rank coefficeint of correlation and

D refers to the diffeence of rank betwee paired items in two series.

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.

Q11. State the Merits and Demerits of rank correlation

Ans :

Merits

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

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.

Rahul Publications

Q12. Explain the computation of rank correlation using MS Excel.

Ans :

	A	B	C
1	Name	Physical activity (min)	Blood presure (mm Hg)
2	Alan	60	118
3	Carl	55	117
4	David	25	120
5	Don	50	121
6	John	40	119
7	Matt	45	122
8	Mike	35	123
9	Neal	10	124
10	Rick	30	125
11	Rob	20	126

To find the Spearman correlation coefficient in Excel, perform these steps:

1. Rank your data

Because the Spearman correlation evaluates the associations between two variables based on their ranks, you need to rank your source data. This can be quickly done by using the Excel RANK.AVG function.

To rank the first variable (physical activity), enter the below formula in D2 and then drag it down to D11:

=RANK.AVG(B2,$B $2: $ B $11,0)

To rank the second variable (blood pressure), put the following formula in cell E2 and copy it down the column:

=RANK.AVG(C2,$C$2:$C$11,0)

For the formulas to work correctly, please be sure to lock the ranges with absolute cell references.

At this point, your source data should look similar to this:

D2 =RANK.AVG(B2,$B$2:$B$11,0) =RANK.AVG(C2,$C$2:$C$11,0)

	A	B	C	D	E	F	G
1	Name	Physical activity (min)	Blood presure (mm Hg)	Physical activity (rank)	Blood presure (rank)		
2	Alan	60	118	1	9		
3	Carl	55	117	2	10		
4	David	25	120	8	7		
5	Don	50	121	3	6		
6	John	40	119	5	8		
7	Matt	45	122	4	5		
8	Mike	35	123	6	4		
9	Neal	10	124	10	3		
10	Rick	30	125	7	2		
11	Rob	20	126	9	1		

1. Find Spearman correlation coefficient

With the ranks established, we can now use the Excel CORREL function to get Spearman's rho:

=CORREL(D2:D11, E2:E11)

The formula returns a coefficient of -0.7576 (rounded to 4 digits), which shows a fairly strong negative correlation and allows us to conclude that the more a person exercises, the lower their blood pressure.

The Pearson correlation coefficient for the same sample (-0.7445) indicates a bit weaker correlation, but still statistically significant:

	A	B	C	D	E
1	Name	Physical activity (min)	Blood presure (mm Hg)	Physical activity (rank)	Blood presure (rank)
2	Alan	60	118	1	9
3	Carl	55	117	2	10
4	David	25	120	8	7
5	Don	50	121	3	6
6	John	40	119	5	8
7	Matt	45	122	4	5
8	Mike	35	123	6	4
9	Neal	10	124	10	3
10	Rick	30	125	7	2
11	Rob	20	126	9	1
12					
13	Spearman correlation		-0.757575758	=CORREL(D2:D11, E2:E11)	
14	Pearson correlation		-0.744475477	=CORREL(B2:B11, C2:C11)	

Calculate Spearman correlation coefficient in Excel with traditional formula

If you are not quite sure that the CORREL function has computed Spearman's rho right, you can verify the result with the traditional formula used in statistics. Here's how:

1. Find the difference between each pair of ranks (d) by subtracting one rank from the other:

   = D2-E2

   This formula goes to F2 and is then copied down the column.

2. Raise each rank difference to the power of two (d^2):

   = F2 ^ 2

   This formula goes to column G.

3. Add up the squared differences:

   = SUM(G2:G11)

   This formula can go to any blank cell, G12 in our case.

From the following screenshot, you will probably gain better understanding of the data arrangement:

	A	B	C	D	E
1	Name	Physical activity (min)	Blood presure (mm Hg)	Physical activity (rank)	Blood presure (rank)
2	Alan	60	118	1	9
3	Carl	55	117	2	10
4	David	25	120	8	7
5	Don	50	121	3	6
6	John	40	119	5	8
7	Matt	45	122	4	5
8	Mike	35	123	6	4
9	Neal	10	124	10	3
10	Rick	30	125	7	2
11	Rob	20	126	9	1
12					
13	Spearman correlation		-0.757575758	=CORREL(D2:D11, E2:E11)	
14	Pearson correlation		-0.744475477	=CORREL(B2:B11, C2:C11)	

Rahul Publications

Calculate Spearman correlation coefficient in Excel with traditional formula

If you are not quite sure that the CORREL function has computed Spearman's rho right, you can verify the result with the traditional formula used in statistics. Here's how:

1. Find the difference between each pair of ranks (d) by subtracting one rank from the other:

   =D2-E2

   This formula goes to F2 and is then copied down the column.

2. Raise each rank difference to the power of two (d^2):

   =F2^2

   This formula goes to column G.

3. Add up the squared differences:

   =SUM(G2:G11)

   This formula can go to any blank cell, G12 in our case.

From the following screenshot, you will probably gain better understanding of the data arrangement:

F2 =D2-E2 =F2^2

	A	B	C	D	E	F	G
1	Name	Physical activity (min)	Blood presure (mm Hg)	Physical activity (rank)	Blood presure (rank)	d	d_2
2	Alan	60	118	1	9	-8	64
3	Carl	55	117	2	10	-8	64
4	David	25	120	8	7	1	1
5	Don	50	121	3	6	-3	9
6	John	40	119	5	8	-3	9
7	Matt	45	122	4	5	-1	1
8	Mike	35	123	6	4	2	4
9	Neal	10	124	10	3	7	49
10	Rick	30	125	7	2	5	25
11	Rob	20	126	9	1	8	64
12						=SUM(G2:G11)	290

1. Depending on whether your data set has any tied ranks or not, use one of these formulas to calculate the Spearman correlation coefficient.

   In our example, there are no ties, so we can go with a simpler formula:

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

With d^2 equal to 290, and n (number of observations) equal to 10, the formula undergoes the following transformations:

$$= 1 - \frac{6 * 290}{10\left(10^2 - 1\right)}$$

$$= 1 - \frac{1740}{990}$$

As the result, you get –0.757575758, which perfectly agrees with the Spearman correlation coefficient calculated in the previous example.

PROBLEMS

1. Calculate the coefficient of correlation between X and Y from the following data.

Marks in English (X)	2	5	4	6	9
Marks in Mathematics (Y)	3	4	4	8	9

Sol : **(April -23)**

X	Y	$x = (x - \bar{x})$	$y = (y - \bar{y})$	x^2	y^2	xy
2	3	–3.2	–2.6	10.24	6.76	8.32
5	4	–0.2	–1.6	0.04	2.56	0.32
4	4	–1.2	–1.6	1.44	2.56	1.92
6	8	0.8	2.4	0.64	5.76	1.92
9	9	3.8	3.4	14.44	11.56	12.92
26	28		11.6	26.8	29.2	25.4

$$\bar{X} = \frac{26}{5} = 5.2$$

$$\bar{Y} = \frac{28}{5} = 5.6$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{25.4}{\sqrt{26.8 \times 29.2}}$$

$$= \frac{25.4}{\sqrt{782.56}} = \frac{25.4}{27.974} = 0.9079.$$

Rahul Publications

2. Find the Karl Pearson's Coefficient of Correlation from the following data between height of the father (x) and son (y).

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Sol : **(Oct.-22)**

X	Y	$X = x - \bar{x}$	$Y = y - \bar{y}$	x^2	y^2	xy
64	66	–3	–2	9	4	6
65	67	–2	–1	4	1	2
66	65	–1	–3	1	9	3
67	68	0	0	0	0	0
68	70	+1	+2	1	4	2
69	68	+2	0	4	0	0
70	72	+3	4	9	16	12
469	476			28	34	25

$$\bar{X} = \frac{469}{7} = 67$$

$$\bar{Y} = \frac{476}{7} = 68$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{25}{\sqrt{28 \times 34}} = \frac{25}{\sqrt{952}}$$

$$= \frac{25}{30.85} = 0.810$$

3.4 Simple and Multiple Regression

Q13. Define Regression analysis. Explain different types of regression analysis.

Ans :

Meaning

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

(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.

Rahul Publications

(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.

Q14. Distinguish between 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 does not indicate the cause and effect relationship between the variables.	Regression clearly indicates the cause and effect relationship between the variables.
3.	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.
4.	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.
5.	Correlation coefficients are symmetric i.e., $r_{yx} = r_{xy}$.	Regression coefficients are asymmetric i.e., $b_{xy} \neq b_{yx}$
6.	The range of r is +1 to –1	The range of b_{xy} and b_{yx} is not restricted.
7.	Correlation coefficient can be calculated from regression coefficients.	Regression coefficients cannot be directly compared from correlation coefficient.
8.	There may be non-sense correlation between variables due to chance.	There is no such thing in regression.

Rahul Publications

Q15. Define simple regression. Explain the concept of simple linear regression with Excel.

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

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

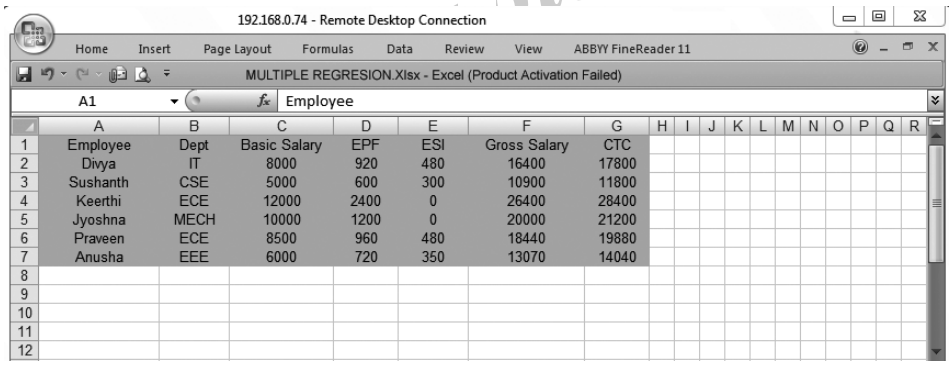
In the above table, the multiple regression model can be written as,

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

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.

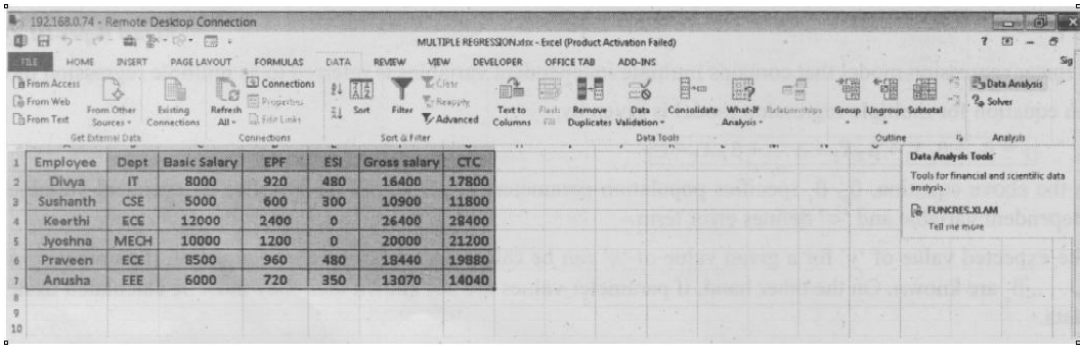
The steps to be followed for generating regression analysis output in case of multiple linear regression are given below,

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



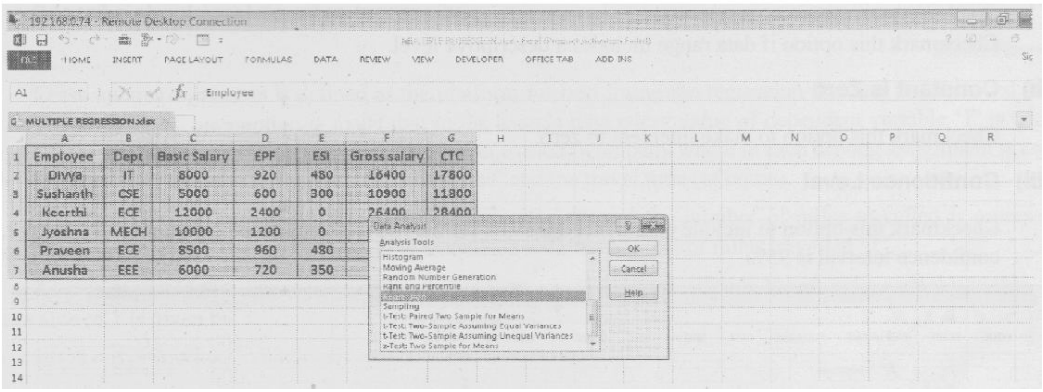
	A	B	C	D	E	F	G
1	Employee	Dept	Basic Salary	EPF	ESI	Gross Salary	CTC
2	Divya	IT	8000	920	480	16400	17800
3	Sushanth	CSE	5000	600	300	10900	11800
4	Keerthi	ECE	12000	2400	0	26400	28400
5	Jyoshna	MECH	10000	1200	0	20000	21200
6	Praveen	ECE	8500	960	480	18440	19880
7	Anusha	EEE	6000	720	350	13070	14040

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



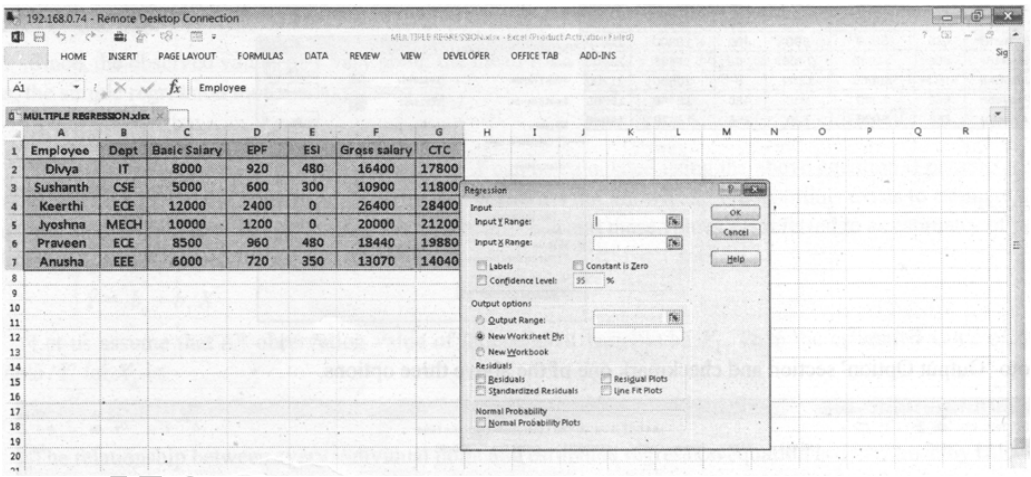
	Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
2	Divya	IT	8000	920	480	16400	17800
3	Sushanth	CSE	5000	600	300	10900	11800
4	Keerthi	ECE	12000	2400	0	26400	28400
5	Jyoshna	MECH	10000	1200	0	20000	21200
6	Praveen	ECE	8500	960	480	18440	19880
7	Anusha	EEE	6000	720	350	13070	14040

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.



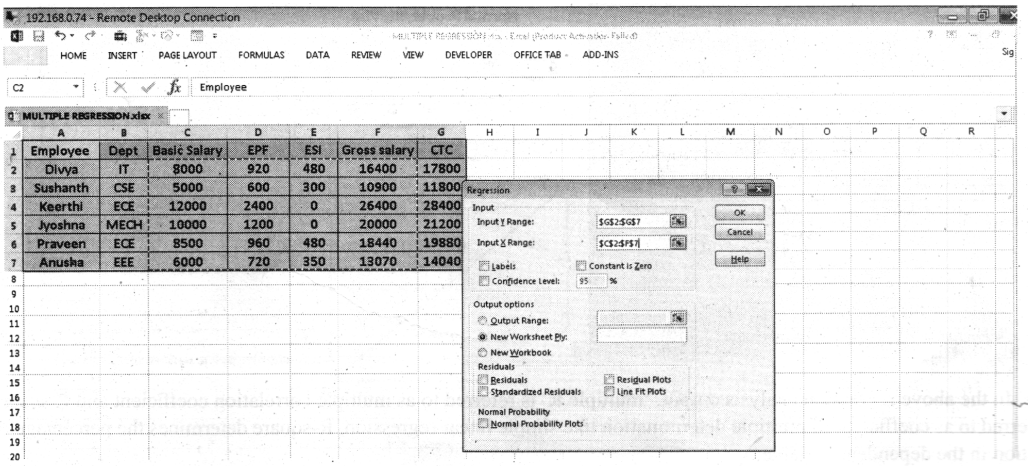
Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0		
Praveen	ECE	8500	960	480		
Anusha	EEE	6000	720	350		

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



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

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 JC.



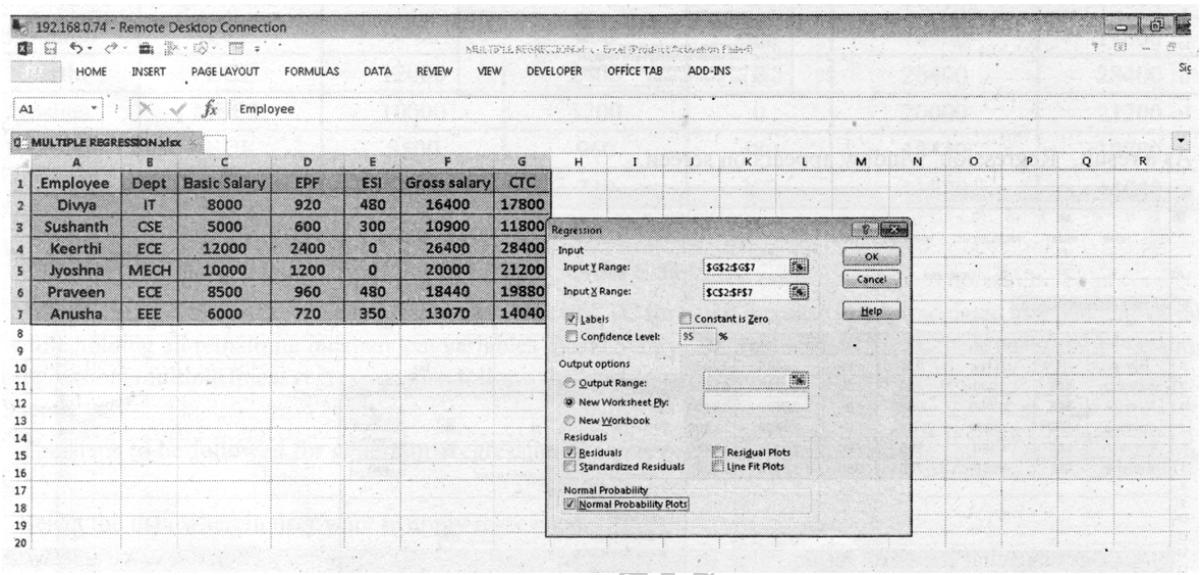
Employee	Dept	Basic Salary	EPF	ESI	Gross salary	CTC
Divya	IT	8000	920	480	16400	17800
Sushanth	CSE	5000	600	300	10900	11800
Keerthi	ECE	12000	2400	0	26400	28400
Jyoshna	MECH	10000	1200	0	20000	21200
Praveen	ECE	8500	960	480	18440	19880
Anusha	EEE	6000	720	350	13070	14040

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.

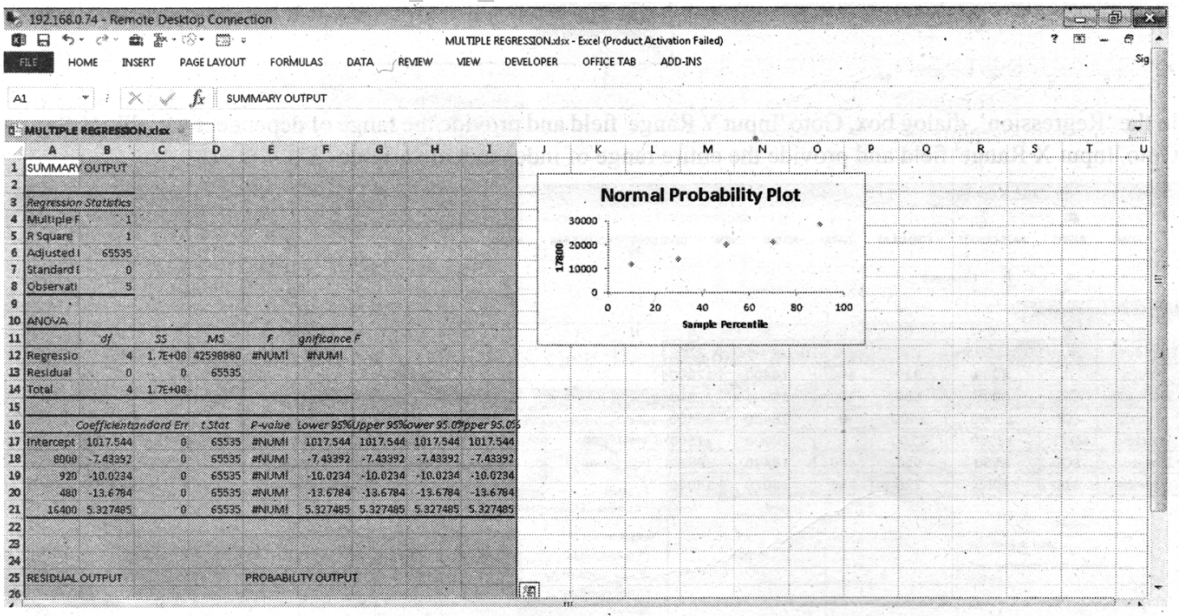
   (ii) Constant is Zero: Checkmark this option to make intercept to zero.

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



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

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



SUMMARY OUTPUT								
Regression Statistics								
Multiple F	1							
R Square	1							
Adjusted I	65535							
Standard E	0							
Observati	5							
ANOVA								
	df	SS	MS	F	gnificance F			
Regressio	4	1.7E+08	42598980	#NUM!	#NUM!			
Residual	0	0	65535					
Total	4	1.7E+08						
	Coefficient	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	1017.544	0	65535	#NUM!	1017.544	1017.544	1017.544	1017.544
8000	-7.43392	0	65535	#NUM!	-7.43392	-7.43392	-7.43392	-7.43392
920	-10.0234	0	65535	#NUM!	-10.0234	-10.0234	-10.0234	-10.0234
480	-13.6784	0	65535	#NUM!	-13.6784	-13.6784	-13.6784	-13.6784
16400	5.327485	0	65535	#NUM!	5.327485	5.327485	5.327485	5.327485
RESIDUAL OUTPUT				PROBABILITY OUTPUT				

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.

Q16. Explain the multiple regression by least squares with an example.

Ans : **(Oct.-20, Imp.)**

A linear regression model that contains multiple independent variables is referred to as multiple regression models'

An equation for multiple regression model is given by,

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \in$$

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 'x' 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_q in β_0, β_1, ... ,β_q

The estimated regression equation in multiple regression model is,

$$\hat{y} = b_0, + b_1 x_1 + b_1 x_2 + ... + b_j, x_p$$

In the above equation, prefers 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.

3.5 Regression by the Method of Least Squares

Q17. Briefly explain the process of constructing regression line using the method of least squares.

(OR)

Explain about method of Least squares.

(OR)

How regression by the method of least squares technique is used ?

Ans : **(April - 23, Oct.-22, Dec.-19, Imp.)**

The Least Square Method is a mathematical regression analysis used to determine the best fit for processing data while providing a visual demonstration of the relation between the data points. Each point in the set of data represents the relation between any known independent value and any unknown dependent value. Also known as the Least Squares approximation, it is a method to estimate the true value of a quantity-based on considering errors either in measurements or observations.

In other words, the Least Square Method is also the process of finding the curve that is best fit for data points through reduction of the sum of squares of the offset points from the curve. During finding the relation between variables, the outcome can be quantitatively estimated, and this process is known as regression analysis.

In regression analysis, an assumption is made that in the sample data every value of dependent variable 'Y' is taken from independent variable 'X. For instance, consider a butter trucking company where in each driving assignment is associated with two parameters namely 'number of miles travelled' and the travel time (in hours).

The travel time is represented by dependent variable 'Y' and the number of miles travelled is represented by independent variable 'X'. The travel time 'Y' is associated with the number of miles travelled 'X'.

As it is assumed that linear relationship exists between dependent variable Y and independent variable 'X'. The mean value of 7 is given by,

$$E(Y) = \beta_0 + \beta_1 X$$

In the above equation, P_0 and P, refers to population parameters representing the intercept and slope, respectively. If X = 0, the intercept 'β_0' refers to expected value of Y and the slope 'β_1' refers to change in the expected value of Y as X changes by one unit.

Since, the observed values of Y vary about the mean for a given value of X, an error term '$\in$' is added to the mean. Thus, the simple regression model is expressed as,

Rahul Publications

$$\boxed{Y = \beta_0 + \beta_1 X + \in}$$

The expected value of 'Y' for a given value of X can be calculated using the above equation if parameter values of β_0 and β_1 are known. On the other hand, if parameter values are not known and if constraints exists to estimate the values using only sample data then sample statistics b_0 and b_1 are used and these values are referred to as estimates of population parameters β_0 and β_1 respectively.

$$\hat{Y} = b_0 + b_1 X \qquad \text{... (1)}$$

Let us assume that K^{th} observation value of independent variable be X_K. Then the estimated value of dependent variable 'Y' for X_K is,

$$\hat{Y} = b_0 + b_1 X_k \qquad \text{... (2)}$$

The relationship between every individual point and estimated regression equation is determined by calculating the vertical distance between them. These differences which are generally known as residuals ek is dependent on the estimated value of dependent variable using the regression line. Thus, for K^{th} observation, the error term 'e_k' is,

$$e_k = y_k - \hat{y}_k$$

3.6 Building Good Regression Models

Q18. Explain briefly about 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 analyzing 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.

Rahul Publications

3.7 Regression with Categorical Independent Variables

Q19. Explain the concept of Regression with Categorical Independent variables.

(OR)

Explain in detail about regression with categorical independent variables.

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

Meaning

Categorical variables are the independent variables that contain values associated with multiple categories e.g., Gender [Male or Female], These type of variables can be incorporated into regression analysis by coding the independent variables with numerical values. For instance, if a variable represents gender of an employee is female then coding can be done by defining No as '0' and Yes as '1'. The following example illustrates how categorical variables are handled in regression analysis.

Example

Consider a regression study involving a dependent variable y, a quantitative independent variable x1, and a categorical independent variable with three possible levels (level 1, level 2, and level 3).

1. How many dummy variables are required to represent the categorical variable?
2. Write a multiple regression equation relating x1 and the categorical variable to y.
3. Interpret the parameters in your regression equation.

The possible level of categorical independent variable is n=3.

The quantitative independent variable is x1.

We need to find the number of dummy variables to represent the categorical data, the regression equation related to the independent variable and categorical data, and the interpretation of the parameters.

(a) The number of dummy variables to represent the categorical data is :

$3 - 1 = 2$

Dummy variables are represented as D_1 and D_2.

The number of dummy variables to represent the categorical data is calculated using the formula, n – 1, where n is the number of possible levels.

(b) The regression equation related to x_1 and the categorical data to y is :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 D_2$$

We define the dummy variables D_1 and D_2 that represent the categorical data. β_0 , β_1 and β_2 are the intercept, coefficient between level 2 and level 1 and, coefficient between level 3 and level 1 respectively.

(c) The interpretation of the parameters is :

β_0 : This is the expected value of the dependent variable when all the other independent variables are set to zero.

β_1 : This is the coefficient of x_1. It represents the change in the dependent variable y.

β_0 and β_3 : These are the coefficients of D_1 and D_2 . They represent the difference in the expected value between level 2 and level 1, level 3 and level 1 respectively.

Rahul Publications

PROBLEMS

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

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 $y_c = a + bx$

$$a = \frac{\Sigma y}{n} \quad b = \frac{\Sigma xy}{\Sigma x^2}$$

The two normal equations are

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

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

substitute the values we get

$1010 = 5a + 15b - (1) \times 3$

$3370 = 15a + 55\,b - (2) \times 1$

$$\begin{array}{r} 3030 = \cancel{15a} + 45b \\ 3370 = \cancel{15a} + 55b \\ - \quad - \quad - \\ \hline -340 = -10 \\ \hline \end{array}$$

$$b = \frac{340}{10} = 34$$

substitute the value in equation (4)

$5a + 15\,(34) = 1010$

$5a = 10\ 510 - 1010$

$$a = \frac{500}{5} = 100$$

$$\boxed{y = 100 + 34x}$$

So the equation of straight line $y = 100 + 34x$

Rahul Publications

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

The least square estimate is computed byt he straight line,

$$Y = a + bX$$

Where, $a = \bar{Y} - b\bar{X}$

$$b = \frac{\Sigma XY - n\bar{X}\,\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

$$\bar{X} = \frac{\Sigma X}{n}$$

$$\bar{Y} = \frac{\Sigma Y}{n}$$

Area x	Value (Y)	XY	X^2
1.1	75	82.5	1.21
1.5	95	142.5	2.25
1.6	110	176.0	2.56
1.6	102	163.2	2.56
1.4	95	133.0	1.96
1.3	87	113.1	1.69
1.1	82	90.2	1.21
1.7	115	195.5	2.89
1.9	122	231.8	3.61
1.5	98	147.0	2.25
1.3	90	117.0	1.69
$\Sigma X = 16$	$\Sigma Y = 1071$	$\Sigma XY = 1591.8$	$\Sigma X^2 = 23.88$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{16}{11}$$

$$= 1.45455$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{1071}{11} = 97.36364$$

Rahul Publications

$$b = \frac{\Sigma XY - n\bar{X}\,\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

$$= \frac{1591.8 - 11 \times 1.45455 \times 97.36364}{23.88 - 11 \times (1.45455)^2}$$

$$= \frac{1591.8 - 1557.82311}{23.88 - 11 \times 2.11572}$$

$$= \frac{1591.8 - 1557.82311}{23.88 - 23.27292}$$

$$= \frac{33.97689}{0.60708} = 55.96773$$

$\therefore \quad b = 55.96773$

$a = \bar{Y} - b\bar{X}$

$= 97.36364 - 55.96773 \times 1.45455$

$= 97.36364 - 81.40786 = 15.95577$

$Y = 15.95577 + 55.96773X.$

3.8 Linear Discriminant Analysis

Q20. Explain briefly about Linear Discriminant Analysis.

(OR)

Explain about Linear Discriminant Analysis with an example.

Ans : **(April -23, Dec.-18)**

Meaning

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.

Example

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.

Rahul Publications

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.

Q21. Explain how disciminant analysis serves the purpose of classifying new data.

Ans : **(April-23)**

Discriminant analysis classifies a group of action and generates predefined classes. It aims to define the observations class depending upon a set of predictor variables. Certain functions called discriminant functions are created by using training data sets. The general form of them is,

$$L = b_1x_1 + b_2x_2 + ... + b_nx_n + c$$

Here,

'h' are weights (or discriminant coefficients)

'x' are input variables (or predictors) and

'c' is constant (or intercept).

The group variance can be increased to define the weights. The observation that is new is predicted by using the discriminant functions. The n number of discriminant functions need to be created for n number of n categories. And a function will be created for a new observation which will only be assigned to the function that has maximum value. The discriminant analysis needs to make certain assumptions like independent variables normality etc., are to be applied.

Rahul Publications

3.9 One-Way and Two-Way ANOVA

Q22. Define ANOVA. State the assumption and applications of Anova.

Ans :

Meaning

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.

Assumptions

(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.

Applications

The following points shows the applications of ANOVA,

1. Anova is used in education, industry, business, psychology fields mainly in their experiment design.
2. Anova helps to save time and money as several population means can be compared simultaneously.
3. Anova is used to test the linearity of the fitted regression line and correlation ratio significance test statistic of Anova = F(r – 1, n – r).

Q23. Explain briefly about oneway anova.

Ans :

Under this, only one factor is considered and its effect on elementary units is observed i.e., data are classified according to only one criterion.

(a) Calculate the variance between samples,

$$SSB = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$$

(b) Calculate the variance within sample

$$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{c} \frac{T_j^2}{n_j}$$

(c) Calculate the F ratio as a ratio of mean squares between samples and mean square within samples.

(d) Compare the calculated value of F ratio with table value of F for degrees of freedom for between and within sample.

(e) If calculated F < table value of F, the difference is taken as insignificant i.e., due to chance and we accept null hypothesis. If calculated value of F is equal to or more than table value, the difference is said to be significant.

ANOVA Table for One-way Classified Data

Source of variation	Sum of squares	d.f	Mean sum of squares	Variable ratio
Treatment Ratio	S_t^2	$k-1$	$S_t^2 = \frac{S_t^2}{(k-1)}$	$\frac{S_t^2}{S_E^2} = F_{k-1}, N-K$
Error	S_E^2	$N-K$	$S_E^2 = \frac{S_E^2}{(N-K)}$	
Total	S_T^2	$N-1$		

Rahul Publications

Q24. Explain two way ANOVA.

(OR)

What is two way Anova? Mention merits and demerits

Ans : **(May-22)**

Two independent factors have an effect on the response variable of interest.

Procedure use for Two-way ANOVA

(a) Calculate the variance between columns,

$$SSC = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$$

(b) Calculate the variance between rows,

$$SSR = \sum_{i=1}^{r} \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

(c) Compute the total variance,

$$SST = \sum X_{ij}^2 - \frac{T^2}{N}$$

(d) Calculate the variance of residual or error,

$$SSE = TSS - (SSC + SSR)$$

(e) Divide the variances of between columns, between rows and residue by their respective degrees of freedom to get the mean squares.

(f) Compute F ratio as follows,

$$\text{F-ratio concerning variation between columns} = \frac{\text{Mean square between colums}}{\text{Mean squares of residual}}$$

$$\text{F-ratio concerning variation between rows} = \frac{\text{Mean square between rows}}{\text{Mean squares of residual}}$$

(g) Compare F-ratio calculated with F-ratio from table,

If F-ratio (calculated) < F-ratio (table), H_0 accepted,

If F-ratio (calculated) $\geq$ F-ratio (table), H_0 rejected,

H_0 accepted $\Rightarrow$ no significant differences

H_0 rejected $\Rightarrow$ significant differences.

Rahul Publications

ANOVA Table for Two-way Classified Data with m-Observation Per Cell

Source of variation	d.f	S.S	M.S.S	Variance ratio F
Factor A	p – 1	S_A^2	$S_A^2 = \frac{S_A^2}{p-1}$	$F_A = \frac{S_A^2}{S_E^2}$
Factor B	q – 1	S_B^2	$S_B^2 = \frac{S_B^2}{q-1}$	$F_B = \frac{S_B^2}{S_E^2}$
Interaction AB	(p –1) (q –1)	S_{AB}^2	$S_{AB} = \frac{S_{AB}^2}{S_E^2}$	
Factor AB	pq (m – 1)	S_E^2	$S_E^2 = \frac{S_E^2}{pq(m-1)}$	
Total	pqm – 1			

Merits

- Any number of blocks and treatments can be used.
- Number of units in each block should be equal.
- It is the most used design in view of the smaller total sample size since we are studying two variable at a time.

Demerits

- If the number of treatments is large enough, then it becomes difficult to maintain the homogeneity of the blocks.
- If there is a missing value, it cannot be ignored. It has to be replaced with some function of the existing values and certain adjustments have to be made in the analysis. This makes the analysis slightly complex.

Q25. Distinguish between one-way and two-way ANOVA.

Ans : **(April-23)**

S.No.	Nature	One Way ANOVA	Two Way ANOVA
1.	Factor	In one way ANOVA, only one factor is considered and its effect on elementary units is observed.	In two way ANOVA, two factors are considered. These factors have an effect on the response variable of interest.
2.	Alternate name	It is also called as "single factor ANOVA".	It is also called as "two factor ANOVA".
3.	Information	In one way ANOVA, information is classified based on only one criteria.	In two way ANOVA, information is classified based on two criteria.

PROBLEMS

5. The alibaba Traders company wishes to test whether its three salesmen Saleem, 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. Saleem made 5 calls, Basha made 4 calls and Vikram made 5 calls. The following are the weekly sales records of the three salesmen :

Saleem Rs.	Basha Rs.	Vikram Rs.
300	600	700
400	300	300
300	300	400
500	400	600
000	---	500

Perform the analysis of variance test and draw your conclusions.

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

Null Hypothesis, H_0

There is no significant difference between the sales of three salesman i.e., $\mu_1 = \mu_2 = \mu_3$.

Alternative Hypothesis, H_1

There exist significant difference between the sales of salesman.

For simplicity, the observations are divided by 100. The result of this will not effect the hypothesis. Therefore, the given samples can be tabulated as follows,

Salim (A)	Basha (B)	Vikram (C)	Total
3	6	7	16
4	3	3	10
3	3	4	10
5	4	6	15
0	–	5	5
15	16	25	GT= 56

Correction Factor, $CF = \frac{(GT)^2}{N} = \frac{56^2}{14} = \frac{3136}{14} = 224$

Total Sum of Squres,

$$SST = \sum_i \sum_j x_{ij}^2 - \frac{(GT)^2}{N}$$

$= (3^2 + 6^2 + 7^2 + 4^2 + 3^2 + 3^2 + 3^2 + 3^2 + 4^2 + 5^2 + 4^2 + 6^2 + 0^2 + 5^2) - 224$

$= (9 + 36 + 49 + 16 + 9 + 9 + 9 + 9 + 16 + 25 + 16 + 36 + 0 + 25) - 224$

$= 264 - 224 = 40$

$\therefore$ SST = 40

Sum of Squares Between Columns (i.e., Salesmen)

$$SSC = \sum_j \frac{T_j^2}{n_j} - \frac{(GT)^2}{N}$$

$$= \left(\frac{15^2}{5} + \frac{16^2}{4} + \frac{25^2}{5}\right) - 224 = \left(\frac{225}{5} + \frac{256}{4} + \frac{625}{5}\right) - 224$$

$= (45 + 64 + 125) - 224 = 234 - 224 = 10$

$\therefore$ SSC = 10

Sum of Squares Between Rows (Samples)

$$SSR = \sum_i \frac{T_i^2}{n_i} - \frac{GT^2}{N}$$

$$= \left(\frac{16^2}{3} + \frac{10^2}{3} + \frac{10^2}{3} + \frac{15^2}{3} + \frac{5^2}{2}\right) - 224 = \left(\frac{256}{3} + \frac{100}{3} + \frac{100}{3} + \frac{225}{3} + \frac{25}{2}\right) - 224$$

$$= \left(\frac{256 + 100 + 100 + 225}{3} + \frac{25}{2}\right) - 224 = \left(\frac{681}{3} + \frac{25}{2}\right) - 224 = (227 + 12.5) - 224$$

$= 239.5 - 224 = 15.5$

Sum of squares of residual or error

SSE = SST – (SSC + SSR)

$= 40 - (10 + 15.5) = 40 - 25.5$

$= 14.5$

Rahul Publications

ANOVA Table

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-ratio
Between Salesmen	10	3 – 1 = 2	$\frac{10}{2} = 5$	$F = \frac{5}{1.8125} = 2.758$
Between Rows	15.5	5 – 1 = 4	$\frac{15.5}{4} = 3.875$	
Residual or Error	14.5	2×4 = 8	$\frac{14.5}{8} = 1.8125$	
Total	40	14		

Conclusion

Sicne the tabulated value of F = 2.758 is less thn the table value F = 4.46 at $df_1 = 2$ and $df_2 = 8$, the null hypothesis is accepted and there exist no significant difference between the three salesmen.

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

Null Hypothesis (H_0)

The sample means are equal i.e., $\mu_1 = \mu_2 = \mu_3$

Alternative hypothesis (H_1)

All means are not equal

The three samples have size 5 each.

∴ The degrees of freedom in the numerator,

$V_1 = k - 1$

$= 3 - 1$

$V_1 = 2.$

The number of degrees of freedom for the denominator can be determined by the total number of observations of all the three samples i.e., n = 15.

$V_2 = n - k$

$V_2 = 15 - 3$

$V_2 = 12$

Thus, the value of $F_{Critical}$ at 5% level of significance at $V_1 = 2$ and $V_2 = 12$ is 3.88.

Now, the calculations are performed on three samples and the resultant values can be represented in a tabular format as follows,

	A	B	C
	8	7	12
	10	5	9
	7	10	13
	14	9	12
	11	9	14
T_j	50	40	80
n_i	5	5	5
x_j	5	8	12

Where,

$T_j = \Sigma x_i$ for A, B and C

n_j = Number of obervations in eacho f the three smaples.

$$\bar{x}_j = \frac{T_j}{n_j}$$

$$T = \sum_{j=1}^{p} n_j = \sum_{j=1}^{3} T_j = 50 + 40 + 60 = 150$$

$$N = \sum_{j=1}^{p} n_j = \sum_{j=1}^{3} n_j = 5 + 5 + 5 = 15$$

Let,

$$X = \frac{T^2}{N} = \frac{(150)^2}{15} = \frac{22500}{15} = 1500$$

$\boxed{\therefore X = 1500}$

A	A^2	B	B^2	C	C^2
8	64	7	49	12	144
10	100	5	25	9	81
7	49	10	100	13	169
14	196	9	81	12	144
11	121	9	81	14	196
$\Sigma A^2 = 530$		$\Sigma B^2 = 336$		$\Sigma C^2 = 734$	

$$\frac{T_{j_1}^2}{n_{j_1}} = \frac{(50)^2}{5} = \frac{2500}{5} = 500$$

Rahul Publications

$$\frac{T_{j_2}^2}{n_{j_2}} = \frac{(40)^2}{5} = \frac{1600}{5} = 320$$

$$\frac{T_{j_3}^2}{n_{j_3}} = \frac{(60)^2}{5} = \frac{3600}{5} = 720$$

Let,

$Y = \Sigma(\Sigma A^2 + \Sigma B^2 + \Sigma C^2]$

$= 530 + 336 + 734 = 1600$

$\boxed{\therefore Y = 1600}$

Let,

$$Z = \sum\left[\frac{T_{j_1}^2}{n_{j_1}} + \frac{T_{j_2}^2}{n_{j_2}} + \frac{T_{j_3}^2}{n_{j_3}}\right]$$

$= 500 + 320 + 720$

$Z = 1540$

Sum of squares (Between),

$SS_B = Z - X = 1540 - 1500$

$\boxed{\therefore SS_B = 40}$

Sum of squares (Within),

$SS_W = Y - Z = 1600 - 1540$

$\boxed{\therefore SS_W = 60}$

Sum of squares,

Total $= Y - X = 1600 - 1500$

$\boxed{\therefore \text{Total} = 100}$

Source	SS	df	MS(Mean Square)	Fobserved
Between	40	2	$\frac{SS_B}{df} = \frac{40}{2} = 20$	$\frac{20}{5} = 4$
Within	60	12	$\frac{SS_W}{df} = \frac{60}{12} = 5$	
Total	100	14		

$F_{observed} = 4, F_{critical} = 3.88$

$\therefore \quad F_{observed} > F_{critical}$

Hence, the null, hypothesis is rejected.

Therefore, there is a significant difference in sample means.

Rahul Publications

7. A reputed marketing agency in India has three different training programs for its salesmen. The three programs are Method - A, B, C. To assess the success of the programs. 4 salesmen from each of the programs were sent to the field. Their performances in terms of sales are given in the following table. Test whether there is significant difference among methods and among salesmen.

Salesman	Methods		
	A	B	C
1	4	8	5
2	7	9	8
3	10	5	9
4	6	7	8

Sol : **(April-22)**

Step 1 : Hypotheses

Null Hypotheses: $H_{01} : \mu_{M1}\, \mu_{M2} = \mu_{M3}$ (for treatments)

That is, there is no significant difference among the three programs in their mean sales.

$\mathbf{H_{02}}$ **:** $\mu_{S1} = \mu_{S2} = \mu_{S3} = \mu_{S4}$ (for blocks)

Alternative Hypotheses:

H_1 : At least one average is different from the other, among the three programs.

H_2 : At least one average is different from the other, among the four salesmen.

Step 2 : Data

Salesmen	Methods		
	A	B	C
1	4	8	5
2	7	9	8
3	10	5	9
4	6	7	8

Step 3 : Level of Significance $\alpha = 5\%$

Step 4 : Test Statistic

$$F_{0t}(\text{treatment}) = \frac{MST}{MSE}$$

$$F_{0b}\,(\text{block}) = \frac{MSB}{MSE}$$

Rahul Publications

Step 5 : Calculation of the Test Statistic

	Methods				
	A	**B**	**C**	**Total x_1**	**x_i^2**
1	4	8	5	17	289
2	7	9	8	24	576
3	10	5	9	24	576
4	6	7	8	21	441
x_i	27	29	30	86	1882
x_i^2	729	841	900	2470	

Squares		
16	64	25
49	81	64
100	25	81
36	49	64

$$\sum\sum x_{ij}^2 = 654$$

Correction Factor : $CF = \frac{G^2}{n} = \frac{(86)^2}{12} = \frac{7396}{12} = 616$

Total Sum of Squares : $TSS = \sum\sum x_{ij}^2 - C.F$

$= 654 - 616 = 38$

Sum of Squares due to Treatments : $SST = \frac{\sum_{i=1}^{k} x_{ij}^2}{k} - C.F$

$= \frac{1882}{3} - 616$

$= 627 - 616 = 11$

Sum of Square due to Blocks $SSB = \frac{\sum_{i=1}^{k} x_{ij}^2}{k} - C.F$

$= \frac{2470}{4} - 616$

$= 617.5 - 616 = 1.5$

Rahul Publications

Sum of Square due to Error $SSE = TSS - SST - SSB$

$= 38 - 11 - 1.5 = 25.5$

$$MST = \frac{SST}{k-1} = \frac{11}{3} = 3.67$$

$$MSB = \frac{SSB}{m-1} = \frac{1.5}{2} = 0.75$$

$$MSE = \frac{SSE}{(k-1)(m-1)}$$

$$= \frac{25.5}{6} = 4.25$$

TWO WAY ANOVA

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments (Programs)	11	3	3.67	$F_{ot} = \frac{3.67}{4.25} = 0.86$
Between blocks (Salesmen)	1.5	2	0.75	$F_{ob} = \frac{0.75}{4.25} = 0.17$
Error	25.5	6	4.25	
Total		11		

Rahul Publications

Step 6 : Critical values

f(3, 6),0.05 =4.7571 (for treatments)

f(2, 6),0.05 = 5.1456 (for blocks)

Step 7 : Decision

(i) Calculated $F_{0t} = 0.86 < f_{(3,6),0.05} = 4.7571$, the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programs.

(ii) Calculate $F_{0b} = 0.17 < f_{(2,6),0.05} = 5.1456$, the null hypothesis is accepted and conclude that there is exist significant difference in the mean sales among the four salesmen.

Short Questions and Answers

1. Define correlation.

Ans :

Meaning

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.

2. Define simple regression.

Ans :

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,

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

3. Multiple regression.

Ans :

A linear regression model that contains multiple independent variables is referred to as multiple regression models'

An equation for multiple regression model is given by,

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

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.

4. Distinguish between one-way and two-way ANOVA.

Ans :

S.No.	Nature	One Way ANOVA	Two Way ANOVA
1.	Factor	In one way ANOVA, only one factor is considered and its effect on elementary units is observed.	In two way ANOVA, two factors are considered. These factors have an effect on the response variable of interest.
2.	Alternate name	It is also called as "single factor ANOVA".	It is also called as "two factor ANOVA".

Rahul Publications

5. Positive correlation

Ans :

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.

6. Properties 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.

7. Define Multiple Correlation.

Ans :

Meaning

When more than two variables are identified, then the relationship between a single dependable variable and multiple independent variables are considered. Therefore, this type of correlation is said to be 'Multiple Correlation'.

The multiple correlation coefficient between x_1 and x_2, x_3 is given by,

$$R_{1,23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

8. Define Rank correlation.

Ans :

Meaning

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.

9. Define Regression analysis.

Ans :

Meaning

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.

10. Linear Discriminant Analysis.

Ans :

Meaning

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.

Rahul Publications

Choose the Correct Answers

1. Choose the correct option concerning the correlation analysis between 2 sets of data. [c]
   (a) Multiple correlations is a correlational analysis comparing two sets of data.
   (b) A partial correlation is a correlational analysis comparing two sets of data.
   (c) A simple correlation is a correlational analysis comparing two sets of data.
   (d) None of the preceding.
2. The slope of the regression line of Y on X is also referred to as the: [c]
   (a) Regression coefficient of X on Y (b) The correlation coefficient of X on Y
   (c) Regression coefficient of Y on X (d) Correlation coefficient of Y on X.
3. Which of the assertions below is the least accurate? [a]
   (a) When outliers are present in the data series, correlation is a more reliable or relevant measure.
   (b) Two variables having a significant nonlinear relation can still have a relatively low correlation.
   (c) Correlation among two variables can emerge from their relationship with a third variable rather than a direct relationship between them.
   (d) None of the preceding.
4. Choose the least likely assumption of a classic normal linear regression model? [b]
   (a) The independent variable and the dependent variable have a linear relationship.
   (b) The independent variable is normally distributed.
   (c) There is no randomness in the independent variable.
   (d) None of the preceding.
5. Which one of the below statements regarding the regression line is correct? [d]
   (a) The prediction equation is another name for a regression line.
   (b) A regression line is also referred to as the line of the average relationship.
   (c) The estimating equation is another name for a regression line.
   (d) All of the above.
6. The correlation coefficient is? [c]
   (a) The square of the coefficient of determination
   (b) Can never be negative
   (c) The square root of the coefficient of determination.
   (d) The same as r square

Rahul Publications

7. The correlation for the values of two variables moving in the same direction is [c]
   (a) Perfect positive (b) Negative
   (c) Positive (d) No correlation.
8. Who suggested the mathematical approach for determining the magnitude of a linear relationship between two variables, such as X and Y? [c]
   (a) Ya Lun Chou (b) Croxton and Cowden
   (c) Karl Pearson (d) Spearman.
9. Who introduced the term 'regression'? [d]
   (a) Karl Pearson (b) R.A Fischer
   (c) Croxton and Cowden (d) Francis Galton.
10. The correlation coefficient describes [b]
   (a) Only magnitude (b) Both magnitude and direction
   (c) Only direction (d) None of the preceding options.

Rahul Publications

Fill in the Blanks

1. _____ is the study of the linear relationship between two variables.
2. The _____ of the coefficient of correlation helps in interpretation.
3. _____ takes its name from studies made by Sir Francis Galton. He compared the heights of persons to the heights of their parents.
4. The variable whose value is influenced or is to be predicted is called _____ variable.
5. The regression analysis for studying more than two variables at a time is termed as _____ regression.
6. _____ studies 'nature' of relationship between the variables.
7. Regression clearly indicates the _____ and _____relationship between the variables.
8. The regression analysis for studying more than two variables at a time is termed as _____ regression.
9. The _____ 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.
10. _____ 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.

Rahul Publications

Answers

1. Correlation
2. Probable error
3. Regression
4. Dependent
5. Multiple
6. Regression
7. Cause, effect
8. Multiple
9. Least
10. Linear discriminant

Very Short Questions and Answers

1. Independent Variable

Ans :

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.

2. Dependent Variable

Ans :

The variable whose value is influenced or is to be predicted is called dependent variable.

3. Simple Regression

Ans :

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

4. Define ANOVA. State the assumption and applications of Anova.

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.

5. Applications of Anova.

Ans :

The following points shows the applications of ANOVA,

1. Anova is used in education, industry, business, psychology fields mainly in their experiment design.
2. Anova helps to save time and money as several population means can be compared simultaneously.
3. Anova is used to test the linearity of the fitted regression line and correlation ratio significance test statistic of Anova = $F(r - 1, n - r)$.

Rahul Publications

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. Define Data Mining. Explain the process of Data Mining.

Ans : **(Oct.-22, May-19, Dec.-18, Imp.)**

Meaning

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.

Steps

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.

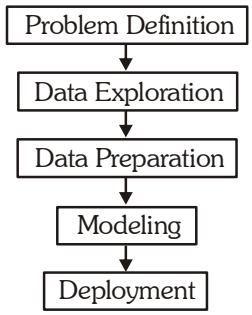


Fig. : Process of Data Mining

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.

Rahul Publications

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.

Q2. Explain the Scope and Architecture of Data Mining.

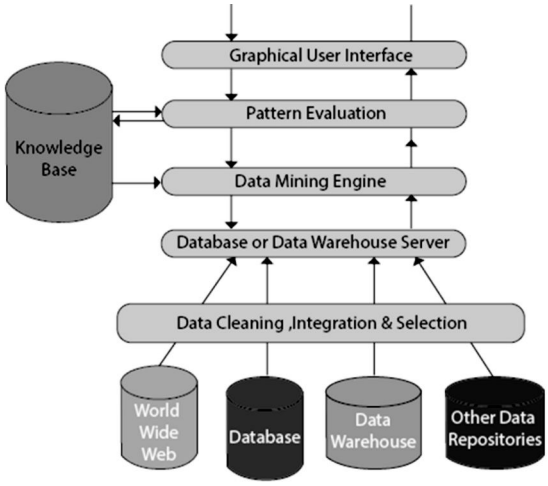
Ans : **(Oct.-22, May-19, Imp.)**

Scope

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.

Architecture

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.



1. Data Source

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

2. Different processes

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it

Rahul Pub

can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server. These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

3. Database or Data Warehouse Server

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

4. Data Mining Engine

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

5. Pattern Evaluation Module

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

6. Graphical User Interface

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

7. Knowledge Base

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable.

Q3. Discuss about techniques of data mining.

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

(i) Classification

This data mining technique is more complex, using attributes of data to move them into discernable categories, helping you draw further conclusions. Supermarket data mining may use classification to group the types of groceries customers are buying, like produce, meat, bakery items, etc. These classifications help the store learn even more about customers, outputs, etc.

(ii) Clustering

This technique is very similar to classification, chunking data together based on their similarities. Cluster groups are less structured than classification groups, making it a more simple option for data mining. In the supermarket example, a simple cluster group could be food and non-food items instead of the specific classes.

(iii) Association rules

Association in data mining is all about tracking patterns, specifically based on linked variables. In the supermarket example, this may mean that many customers who buy a specific item may also buy a second, related item. This is how stores may know how to group certain food items to-

Rahul Publications

gether, or in online shopping they may show "people also bought this" section.

(iv) Regression analysis

Regression is used to plan and model, identifying the likelihood of a specific variable. The supermarket may be able to project price points based on availability, consumer demand, and their competition. Regression helps data mining by identifying the relationship between variables in a set.

(v) Anomaly/outlier detection

For many data mining cases, just seeing the overarching pattern might not be all you need. Data needs to be able to identify and understand the outliers in your data as well. For example, in the supermarket if most of the shoppers are female, but one week in February is mostly men, you'll want to investigate that outlier and understand what is behind it.

4.2 Data Exploration and Reduction

Q4. Define is Data Exploration? Explain the process of Data Exploration.

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

Meaning

Data exploration refers to the initial step in data analysis. Data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, to understand the nature of the data better.

Data exploration techniques include both manual analysis and automated data exploration software solutions that visually explore and identify relationships between different data variables, the structure of the dataset, the presence of outliers, and the distribution of data values to reveal patterns and points of interest, enabling data analysts to gain greater insight into the raw data.

Data without a question is simply information. Asking a question of data turns it into an answer. Data with the right questions and exploration can provide a deeper understanding of how things work and even enable predictive abilities.

R and Python are the most common languages used for exploration; the former works best for statistical learning while the latter lends itself well to machine learning. Coding is not necessary for data exploration through no-code platforms.

The exploration process is also increasingly important to working with Geographic Information Systems (GIS) since so much of today's data is location-enriched.

Process

Data exploration typically follows three steps:

1. Understand the Variables

The basis for any data analysis begins with an understanding of variables. A quick read of column names is a good place to start. A closer look at data catalogues, field descriptions, and metadata can offer insight into to what each field represents and help discover missing or incomplete data.

2. Detect Any Outliers

Outliers or anomalies can derail an analysis and distort the reality of a dataset, so it's important to identify them early on. Data visualization, numerical methods, interquartile ranges, and hypothesis testing are the most common ways of detecting outliers. A boxplot, histogram, or scatterplot, for example, makes it easy to spot points far outside the standard range, while a z-score informs how far from the mean a data point is. Once found, an analyst can investigate, adjust, omit, or ignore the outliers. No matter the choice, the decision should be noted in the analysis.

3. Examine Patterns and Relationships

Plotting a dataset in a variety of ways makes it easier to identify and examine the patterns and relationships among variables. For example, a business exploring data from multiple stores may have information on location, population, temperature, and per capita income. To estimate sales for a new location, they need to decide which variables to include in their predictive model.

Q5. Define is Data Reduction? What are the benefits of data reduction ?

Ans : **(May-19)**

Meaning

Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a

Rahul Publications

reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved, which produces the same analytical results.

Data reduction does not affect the result obtained from data mining. That means the result obtained from data mining before and after data reduction is the same or almost the same.

Benefits

The main benefit of data reduction is simple: the more data you can fit into a terabyte of disk space, the less capacity you will need to purchase. Here are some benefits of data reduction, such as:

- Data reduction can save energy.
- Data reduction can reduce your physical storage costs.
- And data reduction can decrease your data center track.

Q6. Explain briefly about various data reduction methods.

(OR)

What does data exploration and reduction involve ?

(OR)

Explain data reduction techniques in Data Mining.

Ans : **(April-23, Dec.-18, Imp.)**

Data mining introduces certain techniques that explore and reduce data by dividing the huge data into small manageable groups or segments. One example of data exploration is tools and techniques of XLMiner. Various methods are used for exploring and reducing data. They are as follows,

1. Sampling

There are various sampling methods categorized into subjective or probabilistic. They are as follows,

(a) Judgement Sampling: In this method, the choice of sampling items exclusively depend upon the judgement of the investigations. In other words, the investigator exercise his judgements in the choice of sample items and includes those items in the sample which he thinks are most typical of the population with regards to the characteristics under investigation. The success of this method depends on the excellence in judgement. If the individual making decisions is knowledgeable about the population and has a good judgment, then resulting sample may be representative.

(b) Convenience Sampling (Accidental Sampling): Convenience sampling is also called 'chunk'. A chunk is a fraction of one population taken for investigation because of its convenient availability. A sample obtained from readily available lists such as telephone directories, automobile registrations is a convenient sample, even if the sample is drawn at random from the lists.

If the sample enter by 'accident' they just happen to be at the right place and at the right time. Therefore, it is also called accidental sampling.

2. Data Visualization

Number of charts are provided by XLMiner to visualize the data. Examples of it are bar, line, scatter charts and histograms. It can even generate box plots, scatter plot, matrix charts, parallel coordinate charts and variable charts. These options are available in explore button of data analysis group.

(a) Box Plots : Box plots are used to show five statistics of one data set at a time namely minimum, first quartile, median, third quartile and maximum. These are very much useful to identify the shape of distribution as well as data outliers.

(b) Parallel Coordinates Chart : Parallel coordinates chart contains a group of vertical axes for every variable that is selected. A line will be drawn for every observation by connecting vertical axes. Value of a variable would be the point that is crossed by the lines. A multivariate profile is created by parallel coordinates that allows to explore the data and draw certain basic conclusion.

(c) Scatterplot Matrix : A scatterplot matrix will group various scatter charts into a panel by displaying the pair wise relationship among variables to the user.

Rahul Publications

(d) Variable Plot : It plots a matrix of histograms for all the variables that are being selected.

3. Dirty Data

Before analyzing the data sets that have missing values or errors are located so that they can be cleaned. These data sets are called dirty. Various methods are used for this purpose. One method is to eliminate the records containing missing data, calculate reasonable values for the observations that are missing or to make use of a data mining for dealing method. Other than, these XLMiner can handle missing data in transform menu in data analysis group.

Q7. Distinguish between Data Mining and Data Exploration.

Ans :

S.No.	Data Mining	S.No.	Data Exploration
1.	Data mining is also named knowledge web discovery in databases, extraction, data /pattern analysis, and information harvesting.	1.	Data Exploration is used interchangeably with exploration, web scraping, web crawling, data retrieval, data harvesting, etc.
2.	Data mining studies are mostly on structured data.	2.	Data Exploration usually retrieves data out of un structured or poorly structured data sources.
3.	Data mining aims to make available data more useful for generating insights.	3.	Data Exploration is to collect data and gather them into a place where they can be stored or further processed.
4.	Data mining is based on mathematical methods to reveal patterns or trends.	4.	Data Exploration is based on programming laguages or data Exploration tools to crawl the data sources.
5.	The purpose of data mining is to find facts that are previously unknown or ignored,	5.	Data Exploration deals with existing information.
6.	Data mining is much more complicated and requires large investments in staff training.	6.	Data Exploration can be extremely easy and cost-effective when conducted with the right tool.

Rahul Publications

4.3 Unsupervised Learning

4.3.1 Cluster Analysis

Q8. What is Unsupervised Learning? Explain briefly about common cluster algorithm methods.

Ans :

Meaning

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.

Common clustering algorithms include:

(i) **Hierarchical clustering**: builds a multilevel hierarchy of clusters by creating a cluster tree

(ii) **k-Means clustering**: partitions data into k distinct clusters based on distance to the centroid of a cluster

(iii) **Gaussian mixture models**: models clusters as a mixture of multivariate normal density components

(iv) **Self-organizing maps**: uses neural networks that learn the topology and distribution of the data

(v) **Hidden Markov models**: uses observed data to recover the sequence of states.

Q9. Explain about different types of learning.

Ans : **(May-19)**

There are two different types of learning,

1. Supervised Learning

It is also known as classification. This process is supervised because the given input examples are class labeled in the training data set. For instance, consider the postal code recognition problem in which the set of handwritten zip code recognition problem and the respected machine-readable translations are used as training examples.

2. Unsupervised Learning

It is also known as clustering. In this process, the input examples are not class labeled. Thus, this learning process is called as unsupervised learning. Generally, clustering finds the classes in the data. For instance, consider the set of images of handwritten letters. Suppose if a cluster of 26 letters is found then the cluster corresponds to 26 different letters from A to Z, respectively. The semantic meaning of the cluster is unknown because the examples are not class labeled.

Q10. Define Cluster Analysis. Explain the properties of clustering.

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

Meaning

Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups.

This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious.

There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering. The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

Properties

1. Clustering Scalability

Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable, if it is not scalable, then we can't get the appropriate result which would lead to wrong results.

2. High Dimensionality

The algorithm should be able to handle high dimensional space along with the data of small size.

3. Algorithm Usability with multiple data kinds

Different kinds of data can be used with algo-

Rahul Publications

rithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.

4. Dealing with unstructured data

There would be some databases that contain missing values, and noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle unstructured data and give some structure to the data by organising it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.

5. Interpretability

The clustering outcomes should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

Q11. Explain various methods of cluster analysis.

Ans :

The clustering methods can be classified into the following categories:

1. Partitioning Method
2. Hierarchical Method
3. Agglomerative Approach
4. Divisive Approach
5. Density-based Method
6. Grid-Based Method
7. Model-Based Method
8. Constraint-based Method

1. Partitioning Method

It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:

(i) One objective should only belong to only one group.

(ii) There should be no group without even a single purpose.

(iii) In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning

2. Hierarchical Method

In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

3. Agglomerative Approach

The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.

4. Divisive Approach

The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are:

(i) One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.

(ii) One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into microclusters, macro clustering is performed on the microcluster.

5. Density-Based Method

The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in

Rahul Publications

the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster. The radius of a given cluster has to contain at least a minimum number of points.

6. Grid-Based Method

In the Grid-Based method a grid is formed using the object together i.e, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

7. Model-Based Method

In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. Therefore it yields robust clustering methods.

8. Constraint-Based Method

The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.

Q12. State the applications, advantages and disadvantages of cluster analysis.

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

Applications

(i) It is widely used in image processing, data analysis, and pattern recognition.

(ii) It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.

(iii) It can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.

(iv) It also helps in information discovery by classifying documents on the web.

Advantages

(i) It can help identify patterns and relationships within a dataset that may not be immediately obvious.

(ii) It can be used for exploratory data analysis and can help with feature selection.

(iii) It can be used to reduce the dimensionality of the data.

(iv) It can be used for anomaly detection and outlier identification.

(v) It can be used for market segmentation and customer profiling.

Disadvantages

(i) It can be sensitive to the choice of initial conditions and the number of clusters.

(ii) It can be sensitive to the presence of noise or outliers in the data.

(iii) It can be difficult to interpret the results of the analysis if the clusters are not well-defined.

(iv) It can be computationally expensive for large datasets.

(v) The results of the analysis can be affected by the choice of clustering algorithm used.

(vi) It is important to note that the success of cluster analysis depends on the data, the goals of the analysis, and the ability of the analyst to interpret the results.

4.4 Association Rules

Q13. Describe briefly about Association Rules.

(OR)

What are all the association rules ?

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

Meaning

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."

Rahul Publications

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.

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:

1. Basket data analysis

Is to analyze the association of purchased items in a single basket or single purchase as per the examples given above.

2. 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.

3. 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.

Rahul Publications

Q14. Explain the steps for developing Association Rules using XL miner.

Ans :

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.

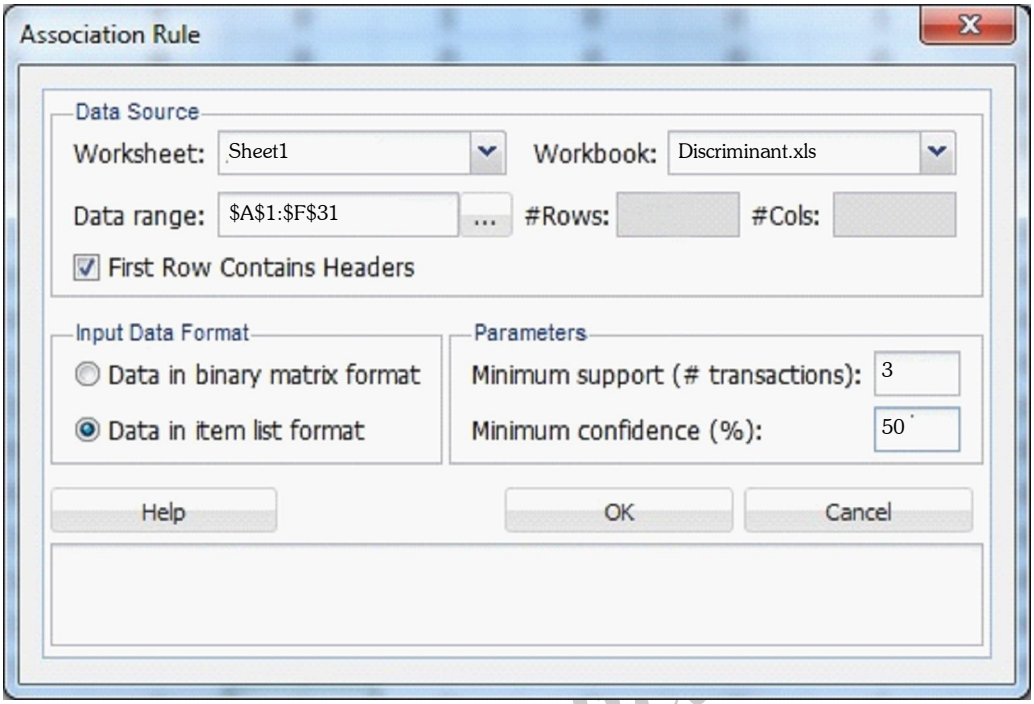


Fig.: Association Rules Dialog

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

Fig.: Association Rules for PC Purchase Data

4.5 Supervised Learning

Q15. Define Supervised Learning. Explain the process of Supervised Learning.

Ans : **(Imp.)**

Meaning

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.

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.

Process

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.

Rahul Publications

4.5.1 Partition Data

Q16. How data can be partioned ? Explain.

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

Partitioning technique 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.

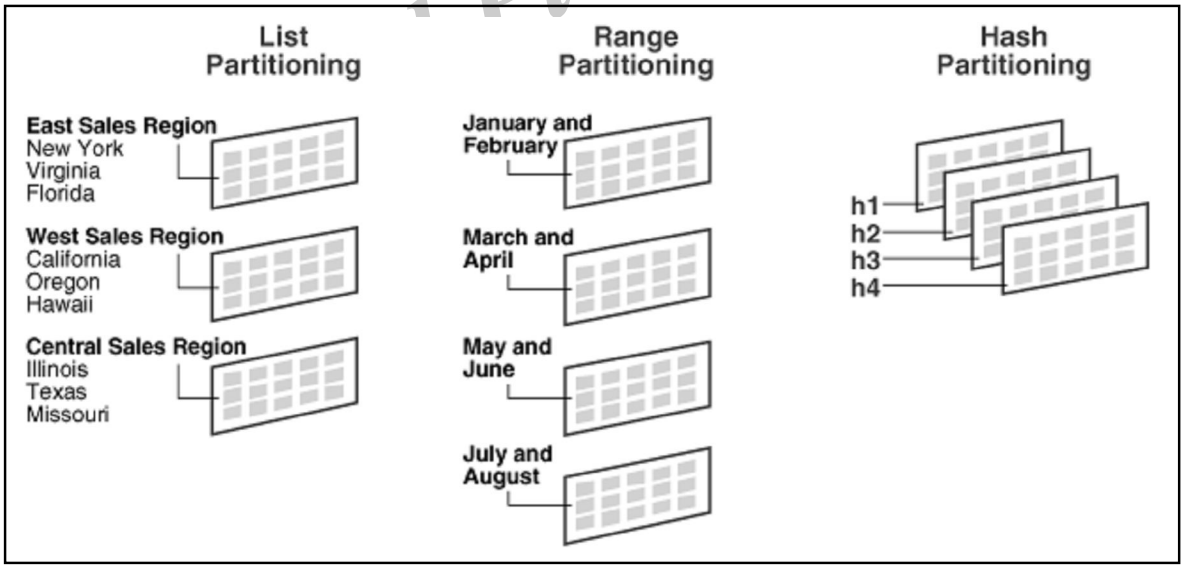
Partitioning Key Extensions

The key extensions assist in signifying the keys used for the partitioning processes. These extensions are

(i) **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.

(ii) **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



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:

i) Hash Partitioning

ii) Range Partitioning

iii) List Partitioning

(i) 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 partitioning is an user-friendly partitioning system, particularly when the information to be detached has no apparent partitioning key.

(ii) Range Partitioning: Range partitioning 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.

(iii) 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.

Rahul Publications

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.

4.5.2 Classification Accuracy

Q17. Explain briefly about Classification Accuracy.

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

Classification problems have observations that are classified into two classes namely loan default or no default. But there are concepts that even have more than two classes. So performance evaluation of a classification method or classifier is done to determine the number of times an observation is predicted to be in wrong class. An accurate measure of the model's classification performance is generated by knowing the classification errors of a large validation set or test set. The classification error is displayed mostly in the form of classification confusion matrix.

The success rate of classification problem solution is measured with classification accuracy. It is defined as the relative frequency of correct classifications.

$$\text{Accuracy} = \frac{n_{corr}}{n}.100\% \text{ n}$$

Here, n indicates the number of possible examples of a problem and n_{corr} indicates the number of correctly classified examples by current theory. Since the exact classifications of the examples are not known, if is not possible virtually to define the exact classification accuracy.

Therefore, classification accuracy is computed from a testing set of examples that are separately solved. The classification accuracy is computed with the following formula,

$$\text{Accuracy} = \frac{n_{t,corr}}{n_t}.100\%.$$

Here, n_t is the number of testing examples. The testing set of examples must not depend on the learning set. The classification accuracy on the learning set is computed as,

Majority class is used to compute the classification accuracy lower bound. The distribution of classes is undefined in example space, so it is computed from the learning set.

The lowest classification accuracy can be computed as,

$$\text{Accuracy}_m = \max_c \left\{\frac{n(c)}{n}\right\}.100\%$$

Here, n(c) indicates the number of learning examples from class c and n indicates the number of learning examples. The below table shows the confusion matrix for three class problem.

Rahul Publications

Q18. What is validation data set? How is it useful ?

Ans : **(April-23)**

(i) Huge sets of data is being used by the data mining projects.

(i) The data is divided into training data set and validation data set before a model is built.

(iii) The training data set contains expected outputs and even it describe about the data mining algorithm.

(iv) If some part of original data set is maintained in validation data set by not using in training process, a real estimate can be obtained about the performance of model with hidden data.

Q19. How do you measure classification performance ?

Ans : **(April-23)**

(i) Classification techniques are used to classify the output into several categories depending upon various data attributes.

(ii) Every record contains a categorical variable and multiple predictor variables.

(iii) A set of predictor variables need the best value of categorical variable to be assigned.

(iv) A database credit approval decision is used for this purpose. Consider the below example, it contains

categorical variable of interest in the form of decision to approve or reject a credit application. Other variables are predictor variables.

(v) There is possibility for errors in any classification rule. This leads to mis-classification.

(vi) So the effectiveness of classification rule can be judged by computing the probability of mis-classification error and. by summarizing the results in the form of classification matrix. This matrix will display the classified cases.'

4.5.3 Prediction Accuracy

Q20. Explain briefly about Prediction Accuracy.

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

Accuracy can be measured in several ways while continuous outcome variable is estimated. Each of it is a function of error to calculate the result of observation i.

Consider a test set T^S such that,

$$T^s = \{(p_1, q_1)\ (p_2, q_2)\ \ldots\ (p_1, q_1)\}$$

Where,

i. represents n-dimensional test tuples related to q. known values (where q represents response variable) and

t denotes number of test tuples in the test set T^S.

Here, it is difficult to guess whether the predicted value q_i' is correct for its associated value P_1. This is because the predictors yields continuous value instead of categorized label. Hence, rather than concentrating on exact value of procedure aims at knowing how far the predicted value is from the actual known value. For this a loss function is used in order to calculate the error between q_i and the predicted value q_i'. The two most common loss functions used in practice are as follows,

Absolute error : $|q_i - q_i'|$

Squared error : $(q_i - q_i')^2$

Depending on the above loss functions the test error rate or generalization error gives the average loss that occurs over the test set. Therefore, the error rates obtained are as follows,

Mean absolute error : $\dfrac{\sum_{i=1}^{t} | q_i - q_i' |}{t}$

Mean squared error : $\dfrac{\sum_{i=1}^{t} | q_i - q_i' |^2}{t}$

In the above error rates, the mean squared error is capable in greatly notifying the occurence of outliers, whereas the mean absolute error cannot notify them. Also, calculating square root of mean squared error would yield error measure known as "root mean squared error". This "root mean squared error" helps in making sure that the magnitude of measured error is equivalent to the magnitude of the predicted value.

Root Mean Squared Error (RMSE) = $\sqrt{\sum_{i=1}^{n} \frac{q_i^2}{t}}$

Furthermore, it is possible that the error relative to the predicted value q should be measured for the mean value q i.e., the total loss can be normalized by dividing it with the total loss incurred from predicting the value of mean.

The two relative error rates available are as follows,

Relative absolute error : $\frac{\sum_{i=1}^{t} | q_i - q_i' |}{\sum_{i=1}^{t} | q_i - \bar{q} |}$

Relative squared error : $\frac{\sum_{i=1}^{t} | q_i - q_i' |^2}{\sum_{i=1}^{t} | q_i - \bar{q} |^2}$

Here $\bar{q}$ represents the mean value for the q_t from the trained data T such that,

$$\bar{q} = \frac{\sum_{i=1}^{u} q_i}{t}$$

Also, calculating the root of relative squared error provides root relative squared error, which helps in obtaining the same magnitude for both, resulting error and predicted value.

4.5.4 K-Nearest Neighbors

Q21. Explain briefly about K-Nearest Neighbors.

(OR)

Explain the algorithm of KNN.

Ans : **(May-22)**

These classifiers compare a particular test tuple with its equivalent training tuples. The description of the training tuples require n-attributes. These tuples can be depicted as points in an n-dimensional pattern space. For an unknown tuple, the K-nearest neighbour classifier looks for the K-neighbouring training tuples that are nearest to that unknown tuple.

The "closeness" between two training tuples $X_1 = (x_{11}, x_{12}, x_{13}, ..., x_{1n})$ and $X_2 = (x_{2l}, x_{22}, x_{23}, ..., x_{2n})$ can be represented as Euclidean distance. It is shown below,

$$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

A value 'V' of a numeric attribute A can be transformed to V' within the [0,1] range, using Min-Max normalization as shown below.

$$V' = \frac{V - \min_A}{\max_A - \min_A}$$

Where,

$\min_A$ = Minimum value of attribute A

$\max_A$ = Maximum value of attribute A

If the attributes describe the category such as color which is not numerical, then the method employed is to compare the attributes of X_1 with the corresponding attributes of X_2. If they match, then the difference between them is '0' otherwise it is 1.

The ideal K-value that represents the number of neighbours which are closer to the test tuple can be determined by starting with K= 1, and then employing a test set which finds the classifier's error rate. The K-value is then incremented one by one until a value with a minimal error rate is obtained.

When the test tuples are being classified, the nearest-neighbour classifiers work very slowly. If there are |D| tuples present in a 'D' training database and if K = 1 then a test tuple can be classified by performing 0(|D|) comparisons.

If the classification is carried out after the tuples have been sorted and arranged in the form of search trees, then a test tuple can be classified by performing O(log |D|) which is comparatively less that O(|D1).

The time for performing classification can be decreased by using these methods.

(i) Partial-distance Method

The distance is calculated that depends on the subset of n-attributes. That is, if the distance is more than the threshold value, then that stored tuple's computation is stopped and the next stored tuple is computed.

(ii) Editing or Pruning or Condensing Method

It deletes all the training tuples which are not useful, thus the stored tuples are minimized.

Rahul Publications

PROBLEMS

1. We have the data from the questionnaire survey and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here are four training samples.

x_1 = Acid Durability (sec)	x_2 = Strength (kg/sqmt)	y = Classification
8	7	Bad
6	6	Bad
4	5	Good
2	4	Good

Now the Factory produces a new paper tissue that pass laboratory test with X_1= 3 and X_2 = 7. Without another expensive survey can we guess what the classification of this new tissue is?

Sol : **(May-22)**

1. Determine parameter K = number of nearest neighbors

   Suppose use K = 3

2. Calculate the distance between the query-instance and all the training samples

Coordinate of query instance is (3, 7), instead of calculating the distance we compute square distance which is faster to calculate (without square root)

x_1 = Acid Durability (seconds)	x_2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)
8	7	$(8-3)^2+(7-7)^2=25$
6	6	$(6-3)^2+(6-7)^2=10$
4	5	$(4-3)^2+(5-7)^2=5$
2	4	$(2-3)^2+(4-7)^2=10$

3. Sort the distance and determine nearest neighbors based on the K-th minimum distance

X_1 = Acid Durability (seconds)	X_2 = Strength (kg/square meter)	Square Distance to query instance (3,7)	Rank minimum distance	Is it included in 3-Nearest neighbours?
8	7	25	4	No
6	6	10	3	Yes
4	5	5	1	Yes
2	4	10	2	Yes

4. Gather the category (Y) of the nearest neighbors. Notice in the second row last column that the category of nearest neighbor (Y) is not included because the rank of this data is more than 3 (=K).

x_1 = Acid Durability (seconds)	x_2 =Strength (kg/square meter)	Square Distance to query instance (3,7)	Rank minimum distance	Is it included in 3-Nearest neighbours?	y = Category of nearest neighbours?
8	7	25	4	No	–
6	6	10	3	Yes	Bad
4	5	5	1	Yes	Good
2	4	10	2	Yes	Good

5. Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

We have 2 good and 1 bad, since 2 >1 then we conclude that a new paper tissue that pass laboratory test with x_1 = 3 and x_2 = 7 is included in Good category.

4.5.5 Classification and Regression Trees

Q22. Explain about Classification and Regression Trees.

(OR)

What is Classification and Regression Trees (CART) ?

Ans : **(Aug.-21)**

CART (Classification and Regression Trees) is one of the most important tools used in modern Data Mining, Machine Learning and Predictive Analytics. Classification and Regression Trees has revolutionised the field of advanced analytics and inaugurated the current era of Data Science. CART model can quickly reveal important

data relationships, automatically searches for patterns and uncover hidden structure even in highly complex data. CART can be used to generate accurate and reliable predictive models for a wide range of applications in all types of industry verticals.

The most common applications include credit scoring, drug discovery, targeted marketing, fraud detection, financial market modelling, manufacturing quality control, engineering, clinical research, length of patient service, predictive maintenance, etc.

Types

1. Classification Tree

Classification Tree is used to create a decision tree for a categorical response (commonly known as target) with many categorical or continuous predictors (factors). The categorical response can be in the form of binomial or multinomial (e.g. Pass/Fail, high, medium & low, etc.). It illustrates important patterns and relationships between a categorical response and important predictors within highly complicated data, without using parametric methods. Also, identify groups in the data with desirable characteristics, and to predict response values for new observations. For e.g., a credit card company can use classification tree to identify customers that will take credit card or not based on several predictors.

2. Regression Tree

Regression Tree is used to create a decision tree for a continuous response (commonly known as target) with many categorical or continuous predictors (factors). The continuous response can be in the form of a real number (e.g. piston diameter, blood pressure level, etc.). It also illustrates the important patterns and relationships between a continuous response and predictors within highly complicated data, without using parametric methods. Also, identify groups in the data with desirable characteristics, and to predict response values for new observations. For example, a pharmaceutical company can use regression tree to identify the potential predictors which are affecting the dissolution rate based on several predictors.

Rahul Pu lications

4.5.6 Logistics Regression

Q23. Define Logistics Regression. Explain the how to enter data in it.

(OR)

What is logistic regression ?

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

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 independent (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:

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

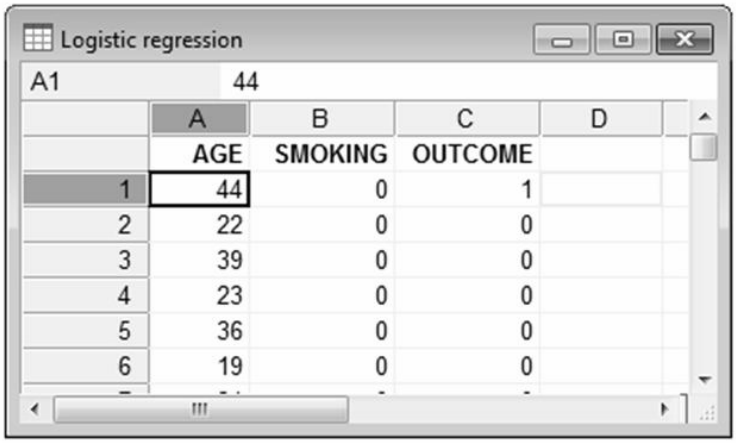
$$\text{odds} = \frac{p}{1-p} = \frac{\text{Probability of presence of characteristic}}{\text{probability of absence of characteristic}}$$

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.

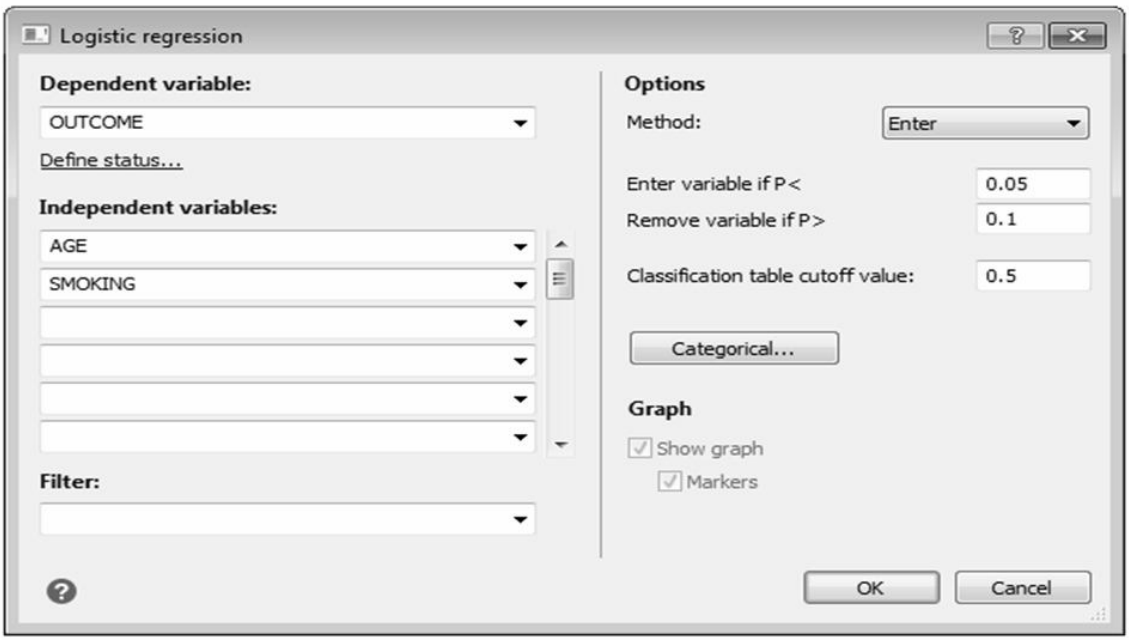
How to enter data

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).



	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



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 record 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:

Rahul Publications

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

Save predicted probabilities - Save residuals

Q24. Explain why cluster analysis is called as unsupervised learning.

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

Unsupervised learning is a type of machine learning where in machine itself has to categorize data/observations by learning the differences and commonalities present in it. Here, machine learning is not provided with any outcome variable for identifying the relation between the observations. For instance, consider an image containing different types of peacocks and crows and the machine is not capable enough to seggregate them based on their features. So, it categorize the image into two groups, peacocks and crows by observing their differences and commonalities.

Ideally clustering can be regarded as unsupervised learning task. Its role is to perform automatic seggregation of data into clusters, sets, groups which includes similar items. Once the cluster is formed it is not necessary that they stay in one shape or size, they keep changing there sizes. Interestingly, the clustering is used for discovering knowledge instead of prediction. It helps in acquiring deep understanding of natural groups present in the data.

The clustering cannot be related to classification, numeric prediction and pattern detection. Here, the result generated is in the form of model which finds similarities of features in outcome or features to other features. The model operates on existing patterns in the data, while clustering generates new data and these newly generated data is in unlabeled form and cluster labels are assigned to them. They are deduced from existed relationships present in the data. For such reasons, the clustering is referred to as unsupervised learning and classification since it classifies unlabeled examples.

Q25. Distinguish between unsupervised learning and supervised learning in data mining.

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

S.No.	Nature	Unsupervised learning	Supervised learning
1.	**Process**	In a supervised learning model, input and output variables will be given.	In unsupervised learning model, only input data will.
2.	**Input Data data**	Algorithms are trained using labeled data.	Algorithms are used against data which is not labeled.
3.	**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.
4.	**Computational Complexity**	Supervised learning is a simpler method.	Unsupervised learning is computationally complex.

Rahul Publications

Short Questions & 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 :

(i) Data mining process the work in such a manner that it allows business to more proactive to grow substantially.

(ii) It optimizes large database within the short time and works business intelligence which is more vital to organizational growth.

(iii) It represents the data in some logical order or may be in pattern to identify the sequential way of processing of data.

(iv) It includes tree-shaped structure to understand the hierarchy of data and representation of the set of information described in the database.

3. What are all the association rules ?

Ans :

Meaning

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.

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.

Rahul Publications

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.

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.

7. Data partitioning.

Ans :

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.

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.

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).

11. What is cause and effect modeling ?

Ans :

Cause-and-effect modeling is the process that develops analytic models to define the metrics relationship that drive the performance. Examples of it are profitability, customer satisfaction or employee satisfaction etc. Performance can be improved by knowing the drivers of it. Key tools of cause and effect modeling are Regression and Correlation.

Rahul Publications

Choose the Correct Answers

1. Which of the following refers to the problem of finding abstracted patterns (or structures) in the unlabeled data? [b]
   (a) Supervised learning (b) Unsupervised learning
   (c) Hybrid learning (d) Reinforcement learning
2. Which one of the following refers to querying the unstructured textual data? [c]
   (a) Information access (b) Information update
   (c) Information retrieval (d) Information manipulation
3. Which of the following can be considered as the correct process of Data Mining? [a]
   (a) Infrastructure, Exploration, Analysis, Interpretation, Exploitation
   (b) Exploration, Infrastructure, Analysis, Interpretation, Exploitation
   (c) Exploration, Infrastructure, Interpretation, Analysis, Exploitation
   (d) Exploration, Infrastructure, Analysis, Exploitation, Interpretation
4. Which of the following is an essential process in which the intelligent methods are applied to extract data patterns? [b]
   (a) Warehousing (b) Data Mining
   (c) Text Mining (d) Data Selection
5. What is KDD in data mining? [a]
   (a) Knowledge Discovery Database (b) Knowledge Discovery Data
   (c) Knowledge Data definition (d) Knowledge data house
6. The adaptive system management refers to: [c]
   (a) Science of making machine performs the task that would require intelligence when performed by humans.
   (b) A computational procedure that takes some values as input and produces some values as the output.
   (c) It uses machine learning techniques, in which programs learn from their past experience and adapt themself to new conditions or situations.
   (d) All of the above
7. For what purpose, the analysis tools pre-compute the summaries of the huge amount of data. [d]
   (a) In order to maintain consistency (b) For authentication
   (c) For data access (d) To obtain the queries response
8. What are the functions of Data Mining? [d]
   (a) Association and correctional analysis classification
   (b) Prediction and characterization
   (c) Cluster analysis and Evolution analysis
   (d) All of the above

Rahul Publications

9. Which of the following statements is incorrect about the hierarchal clustering? [a]
   (a) The hierarchal type of clustering is also known as the HCA
   (b) The choice of an appropriate metric can influence the shape of the cluster
   (c) In general, the splits and merges both are determined in a greedy manner
   (d) All of the above
10. Which one of the following can be considered as the final output of the hierarchal type of clustering? [a]
   (a) A tree which displays how the close thing are to each other
   (b) Assignment of each point to clusters
   (c) Finalize estimation of cluster centroids
   (d) None of the above

Rahul Publications

Fill in the blanks

1. Data mining is also known as _____ and knowledge discovery.
2. _____ experts build the data model for the modeling process.
3. _____ exploration is an informative search used by data consumers to form true analysis from the information gathered.
4. _____ 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.
5. _____ analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters.
6. Any data table is addressed by identifying one of the above data distribution methodologies, using one or more columns as the _____ key.
7. _____ is one metric for evaluating classification models.
8. _____ is nonparametric and therefore does not rely on data belonging to a particular type of distribution.
9. _____ Regressionis the appropriate regression analysis to conduct when the dependent variable is dichotomous.
10. The outcome is measured with a _____ variable.

Rahul Publications

Answers

1. Data discovery
2. Domain
3. Data
4. Unsupervised
5. Cluster
6. Partitioning
7. Accuracy
8. CART
9. Logistic
10. Dichotomous

Very Short Questions and Answers

1. What are association rules ?

Ans :

The i-means is an iterative clustering algorithm in which objects are moved among sets of clusters until the desired set is achieved. It is the most popular and commonly used method. This algorithm is built on the concept of user specified input parameter (k).

2. Define market basket analysis.

Ans :

Market basket analysis is a process of determining groups or set of items which customers are likely to purchase together. It is the process of finding relationships or association between various items which are found in a single transaction.

3. What is k-means clustering ?

Ans :

The k-means is an iterative clustering algorithm in which objects are moved among sets of clusters until the desired set is achieved. It is the most popular and commonly used method. This algorithm is built on the concept of user specified input parameter (k).

4. What is cluster sampling ?

Ans :

Cluster method of sampling is done for small groups or units. Whole population is taken into consideration according to the given problem and is divided into sub-units which are known as 'clusters'. From this sample of sub-units each unit can be easily measured in the selected cluster group.

5. What is cluster sampling ?

Ans :

Cluster method of sampling is done for small groups or units. Whole population is taken into consideration according to the given problem and is divided into sub-units which are known as 'clusters'. From this sample of sub-units each unit can be easily measured in the selected cluster group.

For example, obtaining the income in a city, the whole city can be divided into N different blocks or localities and simple random sample of n blocks is drawn. Cluster sample is determined by the individuals of the selected blocks.

Rahul Publications

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. Define Simulation. Explain different types of Simulation.

(OR)

Write a short notes on Stochastic Simulation.

Ans : **(April-23, May-19, Imp.)**

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.

Types

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.

Rahul Publications

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.

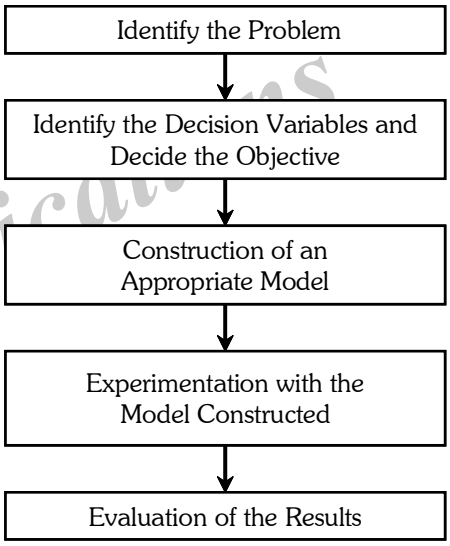


Fig. : Process of Simulation.

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.

Rahul Publications

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.

5.1.1 Advantages and Disadvantages of Simulation

Q3. Explain the Advantages and Disadvantages of Simulation.

(OR)

Analyze in detail about advantages and disadvantages of simulation technique.

Ans : **(April-23, Aug.-21, Nov.-20, Dec.-19, Imp.)**

Advantages

1. Solve Cumbersome Problems

The main advantage of simulation technique is its ability to solve the most difficult problems which are impossible to handle mathematically using quantitative methods.

Many important managerial decision problems are too intricate to be solved by mathematical programming and experimentation with the actual system, even if possible, is too costly and risky.

Simulation offers by allowing experimentation with model of the system without interfering with the real system. Simulation is thus often a bypass for complex mathematical analysis.

2. Foresee Difficulties

Through simulation management can foresee the difficulties and bottlenecks which may come up due to the introduction of new machines, equipment or process. It, thus eliminates the need of costly trial and error methods of trying out the new concept on real methods and equipment.

3. Study the Long Term Effect

It enables the manager to visualize he long term effects in a quick manner.

4. Modification

Simulation models are comparatively flexible and can be modified to accommodate the changing environments of the real situation.

5. Time Saving

All results can be obtained with one model only. For example, the effects of ordering consumer behavior or other policies of many years can be obtained by simulation in a short time.

6. No Interference

Simulation model never interferes with the real world system. It may be too disruptive because experiments are done with the model and not in the system itself. So simulation do not interfere with the real world system.

7. To Study Log Term Effect

It enables managers to visualize the long term effects in a quick manner.

8. It can be used for Training Purposes

Simulation has advantageously been used for training the operating and managerial staff in the operation of complex plans. It is always beneficial to train people on simulated models before putting into their hands the real system.

Disadvantages

1. Does not Produce Optimum Results

When the model deals with uncertainties, the results of simulation are only reliable approximations subject to statistical errors.

2. Expensive

Developing a simulation model can be very expensive because it is long and complicated process to develop a model. Sometimes, it may take years to develop a model. Hence huge expenditure are involved.

3. Not Precise

Simulation is not precise and it does not yield an answer but provide a set of the system in different conditions.

4. Quantification of the Variables is Another Difficulty

In a number of situations, it is not possible to quantify all the variables that affect the behavior of the system.

Rahul Publications

5. Non-transferrable Solution

Each solution model is unique and its solution and inferences cannot be transferable to other problems.

6. Adhoc

Each application of simulation is adhoc to a great extent, which indicates that simulation does not require standard probability distribution.

Q4. Distinguish between solutions derived from simulation models and solutions from analytical models. Highlight some of the problem areas of application, where simulation is considered most preferable.

Ans : **(April-23, Imp.)**

S.No.	Simulation Model	S.No.	Analytical Model
1.	The solutions derived from simulation models are riot accurate or precise and truely not optimal.	1.	The solutions derived from analytical models are accurate or precise and also reflect the systems true stated
2.	These solutions are generated quickly with no complex (complicated) calculations.	2.	These solutions are generated after reasonable amount of time because of complex (complicated) calculations are involved.
3.	It is not accurate but it provides fair overview of the systems behaviour.	3.	It can be easily implemented on small scale systems.
4.	It involves only statistical calculation.	4.	It probably derives factors such as MTTF/ MTBF from the result.

Applications of Simulation

(i) Simulation is used in the field of manufacturing application.

(ii) It is used in construction engineering and project management.

(iii) To develop Military application and software. And also used in transportation model in traffic

(iv) Complex and logistics process can be optimized with simulation techniques. And also used in supply chain and distribution applications.

(v) Simulation is used to optimize learning environment. And to optimize business process.

(vi) To optimize health care and networks. And risk analysis are optimized using simulation techniques.

(vii) Simulation is used in semiconductor manu-facturing and computer to optimize them by simulation techniques.

5.2 Random Number Generation

Q5. Define Random Number. Discuss the various methods of Random Number Generations.

(OR)

Write short notes on Random Number Generation.

Ans : **(April-23)**

Random numbers are numbers that occur in a sequence such that two conditions are met:

(i) The values are uniformly distributed over a defined interval or set, and

(ii) It is impossible to predict future values based on past or present ones. Random numbers are important in statistical analysis and probability theory.

Methods

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.

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 to more formal pseudo randomness based on important properties

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.

Q6. 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.

Rahul Publications

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	

Rahul Publications

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

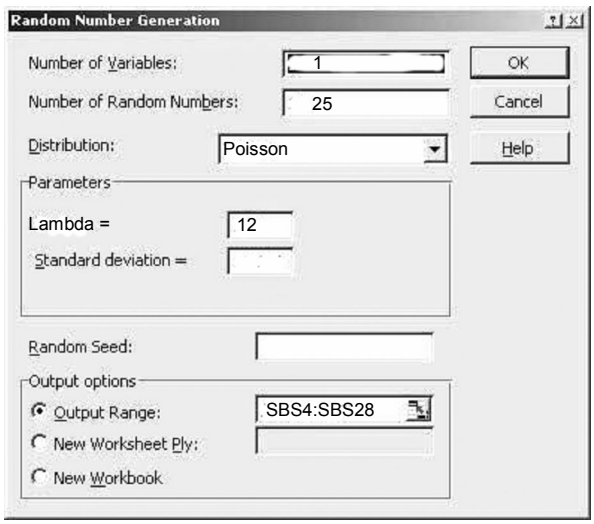
(i) Enter '1' in the 'Number of Variables' field (This option creates number of columns).

(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.

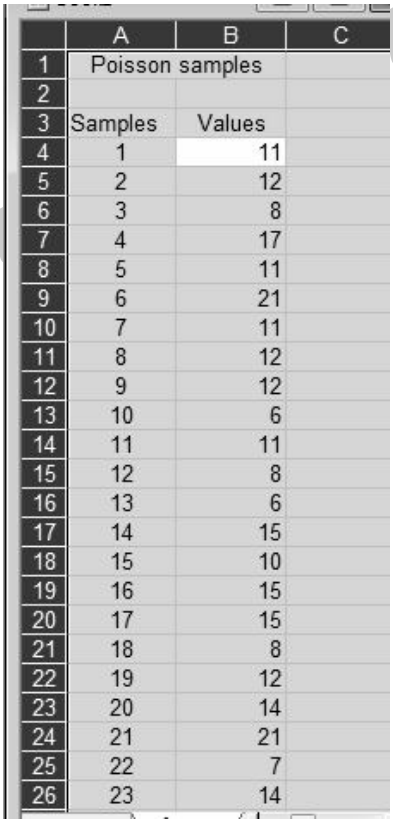
(v) Select 'Output Range under Output options. 'Here, the random numbers will be generated under value'. Enter $B$4:$B$28 or you can select cell B4 to cell B25. After entering all details, the final dialog will appear as shown below.



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.

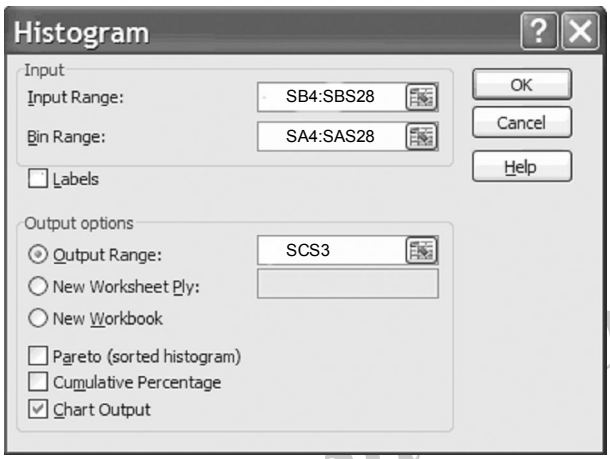
Step- 7

(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.

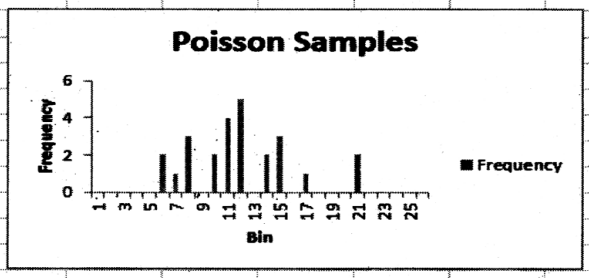
(iv) Tick mark 'chart output'. The final dialog box will appear after entering details is shown below,



Step-8

Press Enter. The histogram for the generated random numbers is shown below.

	A	B	C	D
1	Poisson Samples			
2				
3	Sample	Value	Bin	Frequency
4	1	11	1	0
5	2	12	2	0
6	3	8	3	0
7	4	17	4	0
8	5	11	5	0
9	6	21	6	2
10	7	11	7	1
11	8	12	8	3
12	9	12	9	0
13	10	6	10	2
14	11	11	11	4
15	12	8	12	5
16	13	6	13	0
17	14	15	14	2
18	15	10	15	3
19	16	15	16	0
20	17	15	17	1
21	18	8	18	0
22	19	12	19	0
23	20	14	20	0
24	21	21	21	2
25	22	7	22	0
26	23	14	23	0



PROBLEMS

1. 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 : **(May-19)**

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

Rahul Publications

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 (1 + 1 + 1 + 2).

5.3 Monte Carlo Simulation

Q7. Define Monte Carlo Simulation. State the Advantages and Disadvantages of Monte Carlo Simulation.

(OR)

Explain the Monte Carlo Simulation.

Ans : **(Oct.-22, Dec.-19, Dec.-18)**

Meaning

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.

Advantages

Some of the advantages/benefits of Monte Carlo technique are as follows,

1. The part experiences of the firms using Monte Carlo technique shows that the solutions obtained by using Monte Carlo technique are very close to the real situations.
2. The support of computer software has made Monte Carlo technique a significant component of risk assessment in different stages of the important projects.
3. Monte Carlo Technique is one of the commonly used simulation techniques applied in academic research and industrial applications. The most important problem faced in operations research is with regard to the handling of the complicated situation in which it is very difficult to create an analytical equation. The main objective of Monte Carlo Technique is to simulate the operation in a systematic manner so that the main factors and interactions among these factors are analyzed easily.

Rahul Publications

Disadvantages

Inspite of the above mentioned advantages, Monte Carlo technique also has certain disadvantages which are as follows,

1. The support of computer software to Monte Carlo technique is very essential. In the absence of computer software, iterations have to be done through manual process which consumes lot of time and energy. The manual process may also leads to errors.
2. Several practitioners avoid using Monte Carlo technique as they feel that "it is good to be nearly right than being completely wrong".
3. The practitioners considered the methodology and algorithm followed in Monte Carlo technique as very complicated.

Q8. Explain how Monte Carlo simulation can be developed in Ms.Excel?

Ans :

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.

Rahul Publications

Game of Dice

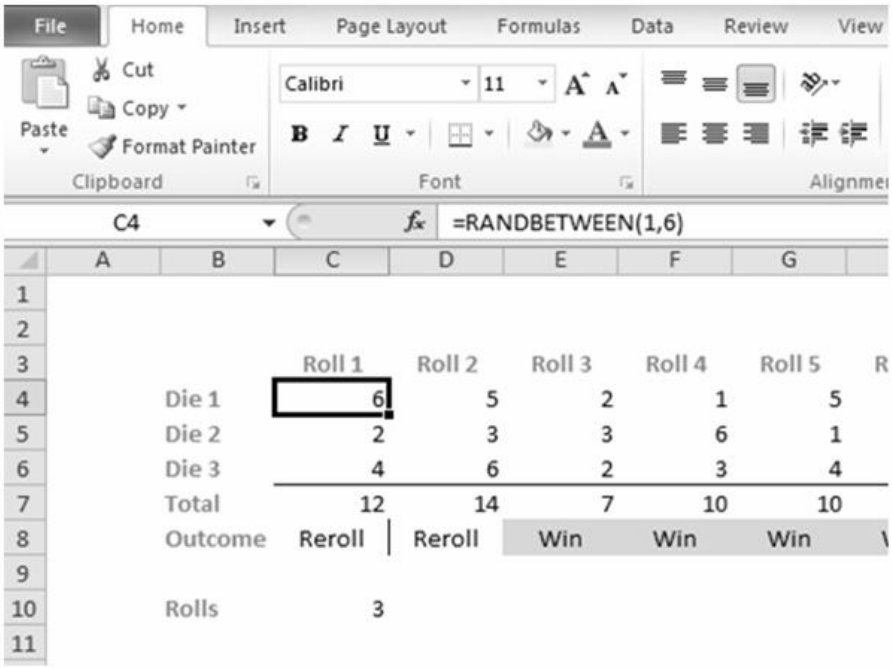
Here's how the dice game rolls:

- 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.



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

Rahul Publications

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)

	Roll 1	Roll 2	Roll 3	Roll 4	Roll 5
Die 1	3	5	5	5	
Die 2	3	4	5	5	
Die 3	1	6	5	6	
Total	7	15	15	16	
Outcome	Win	Win	Win	Win	Win
Rolls	1				

	1st Roll	2nd Roll
1	=IF($C$7=C14,"Win",D14)	
2	N IF(logical_test, [value_if_true], [value_if_	
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

Rahul Publications

Step 3 : Conclusions

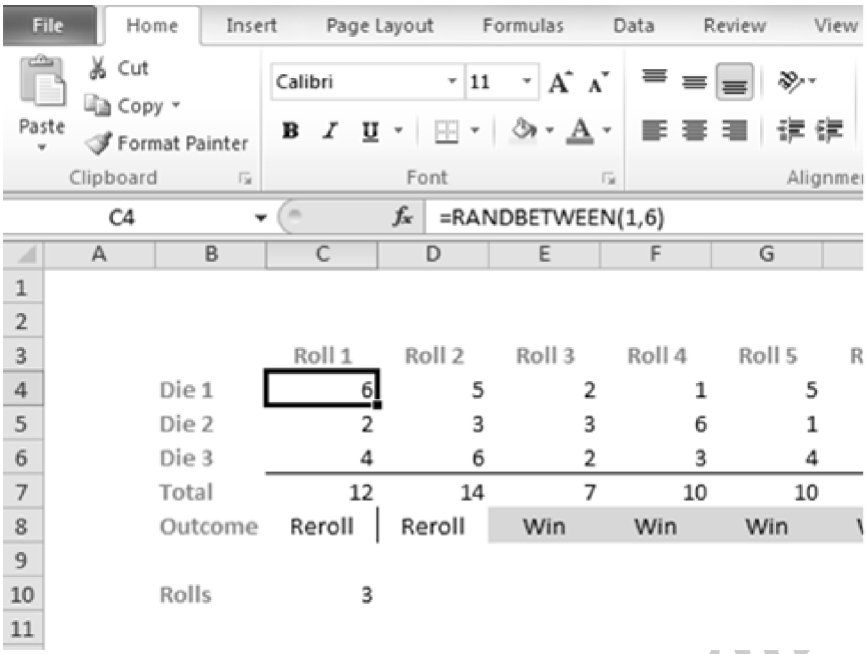
In this step, we identify the outcome of the 50 dice rolls. The first conclusion can be obtained with an index function. This function searches the possible results of the first round, the conclusion corresponding to the result obtained. For example, when obtaining 6, as is the case in the picture below, we play again.

C7 =SUM(C4:C6)

	Roll 1	Roll 2	Roll 3	Roll 4
Die 1	1	5	6	2
Die 2	4	4	3	1
Die 3	1	3	2	5
Total	6	12	11	8
Outcome	Reroll	Lose	Lose	Lose
Rolls	3			

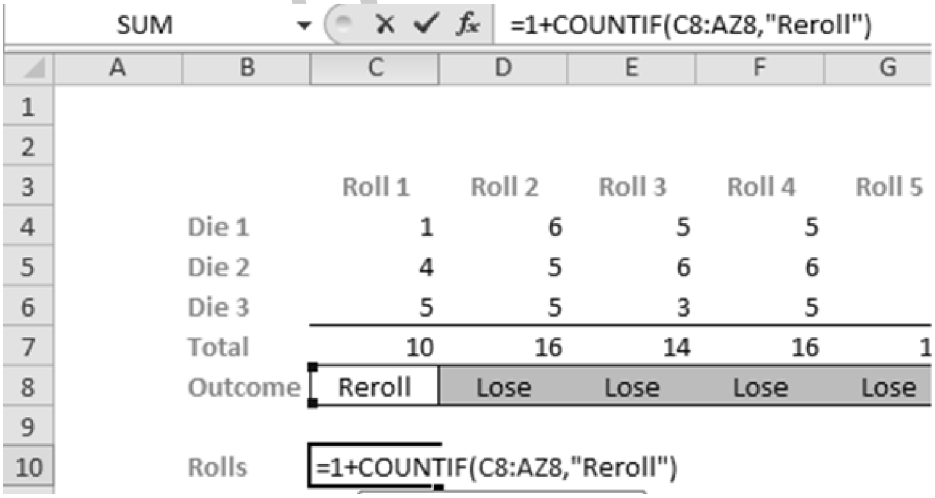
	1st Roll	2nd Roll
1	N/A	N/A
2	N/A	N/A
3	Lose	Lose
4	Lose	Lose
5	Lose	Lose
6	Reroll	Reroll
7	Win	Win
8	Lose	Reroll

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.



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).



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)

	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	Reroll	Reroll
9	Reroll	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

	Win	1
1	Win	2
2	Win	2
3	Win	1
4	Lose	1
5	Lose	3
6	Win	2
7	Win	4
8	Win	1
9	Lose	1
10	Lose	8
11	Win	6
12	Lose	7
13	Win	1
14	Lose	1
15	Win	3
16	Win	4
17	Lose	3
18	Lose	5

Win	3671	73.4%
Lose	1329	26.6%
Reroll	0	0.0%
	5000	100.0%
Rolls	2.8382	

Rahul Publications

Q9. Explain the process of Monte Carlo Stimulation.

Ans : **(May-22, Aug.-21)**

The production volume (Demand) was known but now we will assume that production volume is uncertain, therefore the demand is modelled as a random variable consisting of some probability distribution. Let us assume that demand or production volume is normally distributed with a mean of 1,000 and a standard deviation of 100.

Step-1

In cell B12, remove the previous entered value and enter = ROUND (NORM.INV (RAND(), 1,000,100)0).

Step-2

Enter the following in spread sheet as shown below,

D	E	F	G
Monte Carlo Stimulation			
	Demand	**Cost Difference**	**Decision**
Trials			
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Rahul Publications

Step-3

In order to give references to the cells associated with demand, cost difference and decision, enter = B12 in cell E3, = B19 in cell F3, = B20 in cell G3.

Step-4

Select the cells range from D3 to G23, click data tab, click 'what-if analysis' select data table, leave blank in row input cell field, click in the column input cell field and enter any blank cell in the spread sheet and click 'OK'.

Step-5

Type 'Average' in cell E24. Enter=Average (F4:F23) to calculate average.

Step-6

Type 'manufacture' in cell F25. Enter = COUNT IF ($G$4:$G23, F25)/ COUNT A($G$4.$G$24) in cell G25, to calculate percentage. It will calculate in decimals to calculate in %, click on home tab and click on '%'.

Type 'Outsource' in cell F26. Enter = COUNT IF ($G$4:$G23,F26) ACOUNT A ($G$4:$G$24) in cell G26 and click on'%'.

5.4 What If Analysis

Q10. Define What If Analysis. What are the basic options available in excel for performing What If Analysis?

(OR)

Explain about what-if analysis.

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

A technique which aims to assess the behaviour of a complex system under different scenarios is referred as 'What If Analysis'. It considers alternative values for various random variables like direct labour cost, demand of first year etc. This will also helps to determine the actual profit or output.

Some of the benefits that can be derived while implementing what-if analysis are,

1. It facilitates in taking better and informed decisions.
2. It helps in predicting the results of decisions.
3. It helps in taking quicker decisions.

Basic Options Available in Excel

Excel offers the following basic options to perform what if analysis,

(i) Manual What if Analysis

This option allows the user to notice the affect generated on the formula cells while working on the input cells. The user can edit the value in the input cell to see its affect on the formula cell. This is a very common method.

(ii) Data Tables

This option creates a special table that summarizes the formula cells while varying the input cells.

However, these tables donot support more than two input cells. Data tables are of two types one is one-input data table and the other is two-input data table.

Rahul Publications

(iii) Scenario Manager

A 'scenario manager' is a what-if analysis tool that allows users to save multiple set of values in the same cells of worksheet. The cells saving different input values for different variables are called changing cells. Each data set containing different set of values is referred to as scenario. It is identified by its name and given during its creation. In a scenario, the number of changing cells can be maximum 32 i.e., the number of variables can't exceed 32.

Q11. 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,

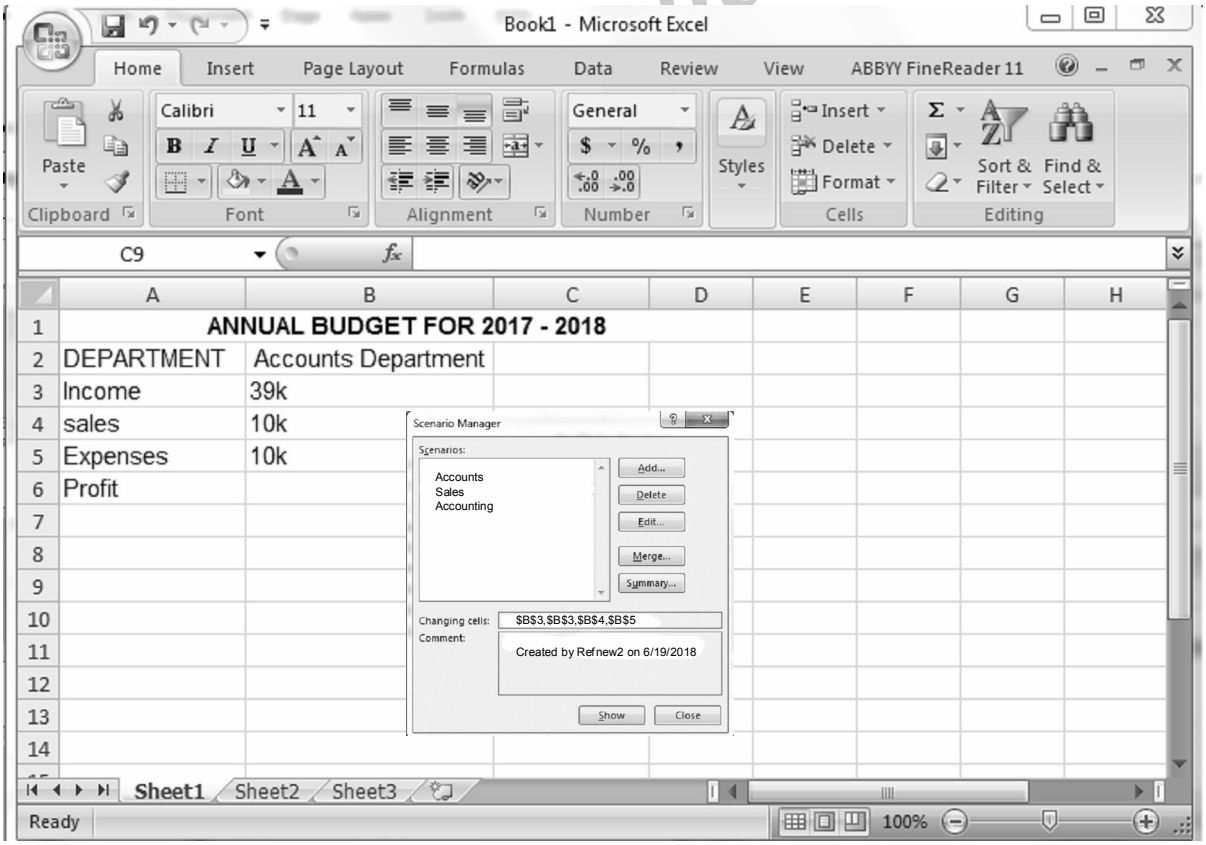
(i) Scenario summary report

(ii) Scenario pivot table report.

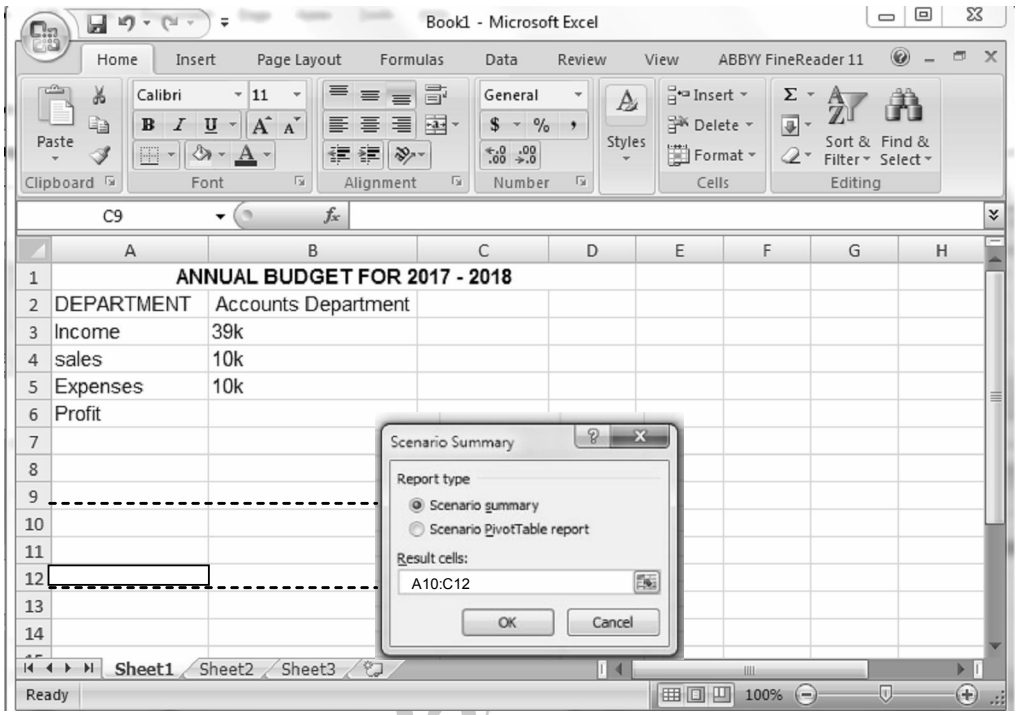
The former one follows the structure of worksheet outline and it is used in cases involving less complexity while the latter follows the structure of pivot table and it is used in cases where scenarios are defined with numerous result cells. The following steps are involved in generating scenario report,

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.

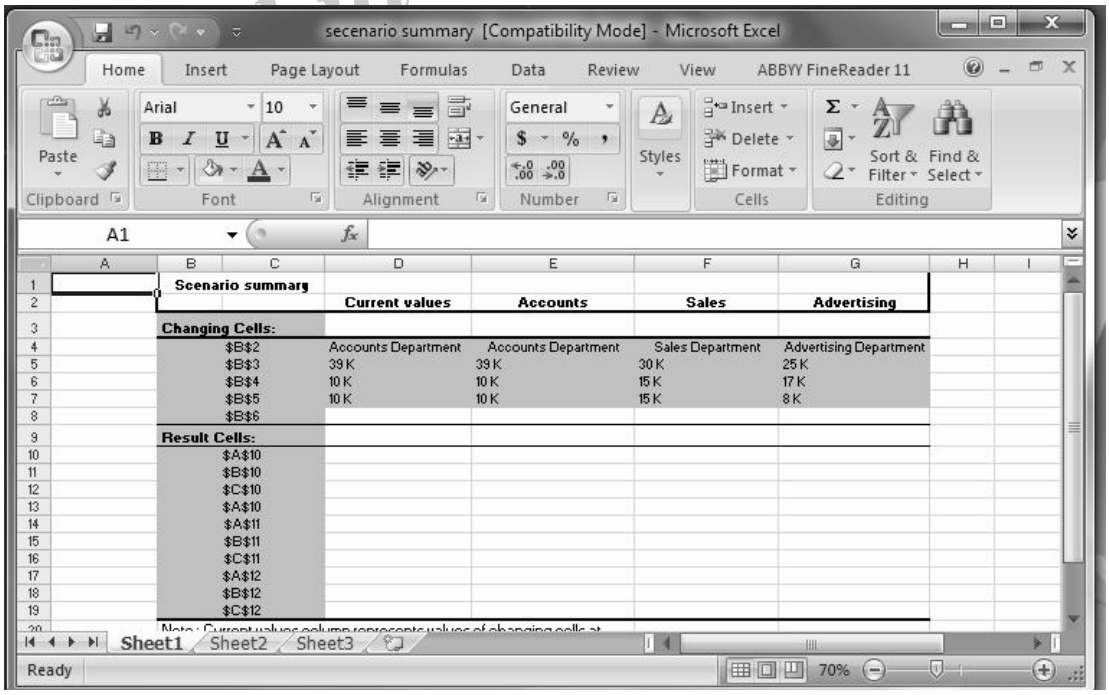
ications



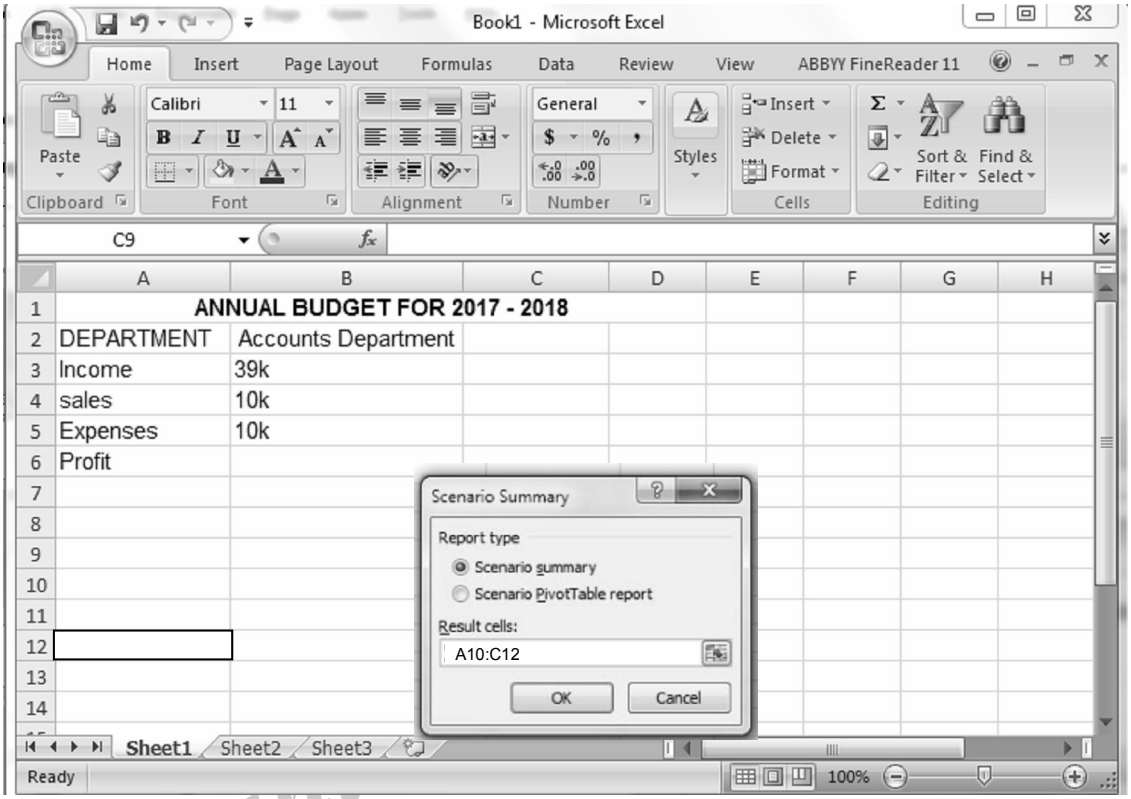
3. Click on 'Summary' button in scenario manager dialog box. Which display's a 'Scenario summary' dialog box. Go to 'Report Type' group and check mark the option beside 'Scenario Summary' option to generate scenario summary report or check mark the option beside 'Scenario Pivot Table Report' to generate scenario pivot table report.

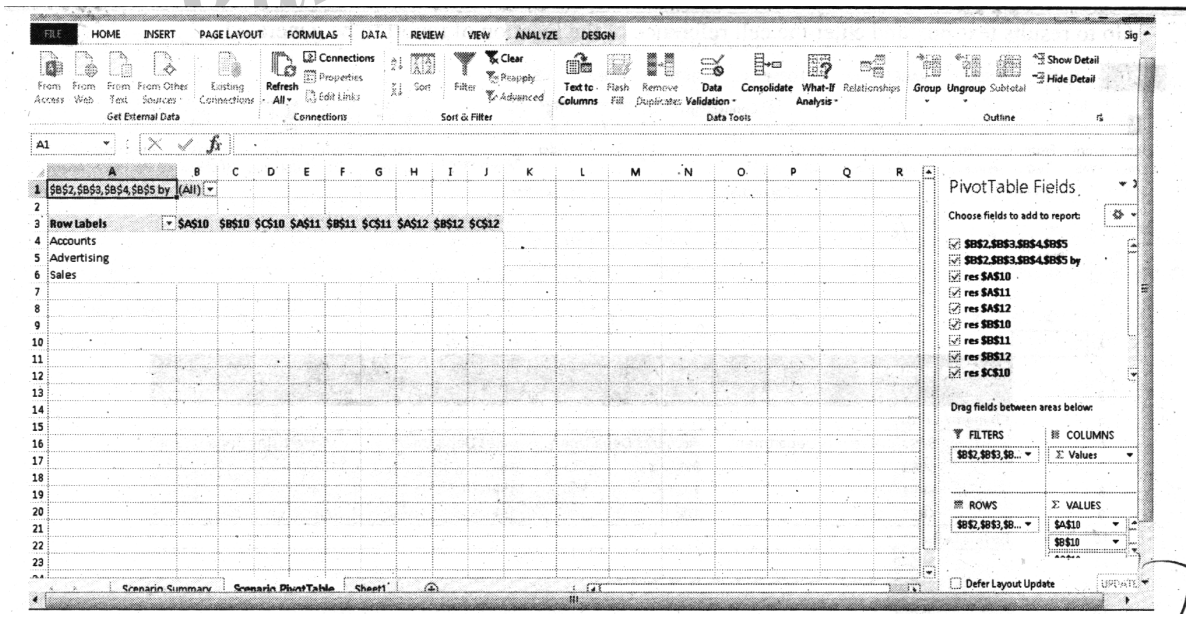


4. Go to results text box and enter the cell reference of the location of report to be generated.



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.





5.5 Verification and Validation

Q12. Define verification. Explain various methods of verification.

Ans : **(Imp.)**

Meaning

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.

Methods

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.

Rahul Publications

Q13. Define validation. Explain various methods of validation.

Ans :

Meaning

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

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

1. Specification 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

(i) 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.

(ii) Boundary Value Analysis: Software Testing technique in which tests are designed to include representatives of boundary values in a range.

(iii) 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.

(iv) 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.

2. Structure Based Testing Techniques

The different types of structure based test design or the white box testing techniques are.

(i) 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.

(ii) 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.

(iii) 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).

Rahul Publications

(iv) 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.

(v) Condition Determination Testing

It is an optimized way of multiple condition testing in which the combinations which doesn't affect the outcomes are discarded.

(vi) Path Testing

Testing the independent paths in the system (paths are executable statements from entry to exit points).

Q14. Distinguish between Verification and Validation.

Ans : **(Imp.)**

S.No.	Verification	S.No.	Validation
1.	Verification is a static practice of verifying documents, design, code and program.	1.	Validation is a dynamic mechanism of validating and testing the actual product.
2.	It does not involve executing the code.	2.	It always involves executing the code.
3.	It is human based checking of documents and files.	3.	It is computer based execution of program.
4.	Verification uses methods likes inspections, reviews, walkthroughts, and Desk-checking etc.	4.	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.	5.	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.	6.	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.	7.	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.	8.	Validation is carried out with the involvement of testing team.
9.	It generally comes first-done before validation.	9.	It generally follows after verification.

Rahul Publications

5.6 Risk Analysis

Q15. What is Risk Analysis? Explain the benefits of Risk Analysis.

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

Meaning

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.

Benefits

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.

Q16. Discuss in detail various techniques of risk analysis.

(OR)

Briefly explain risk analysis techniques in decision making.

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

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.

Rahul Publications

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).

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.

Properties

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'.

5.7 Decision Tree Analysis

Q17. Define Decision Tree Analysis. Explain the steps involved in construction of Decision Tree Analysis.

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

Meaning

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.

Steps

1. The first step is understanding and specifying the problem area for which decision making is required.
2. The second step is interpreting and chalking out all possible solutions to the particular issue as well as their consequences.
3. The third step is presenting the variables on a decision tree along with its respective probability values.
4. The fourth step is finding out the outcomes of all the variables and specifying it in the decision tree.
5. The last step is highly crucial and backs the overall analysis of this process. It involves calculating the EMV values for all the chance nodes or options, to figure out the solution which provides the highest expected value.

Q18. Briefly explain the use of decision tree analysis in problems involving a measure of probability, of success or failure.

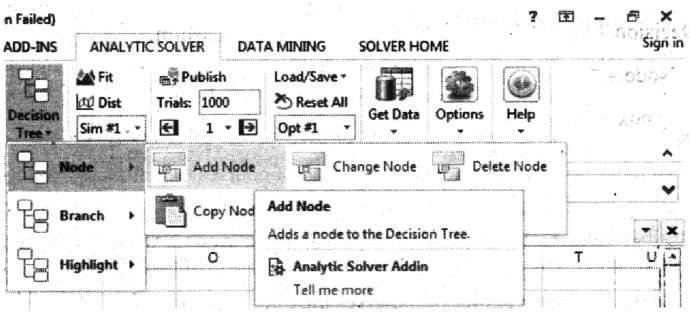
Ans : **(April-23)**

In excel 2013, decision tree can be created with the help of analytics solver platform option. Under this option 'Decision Tree' button will be available. By using analytic solver platform you can draw a decision tree with the help of node creation. In order to add nodes, following steps are to be followed,

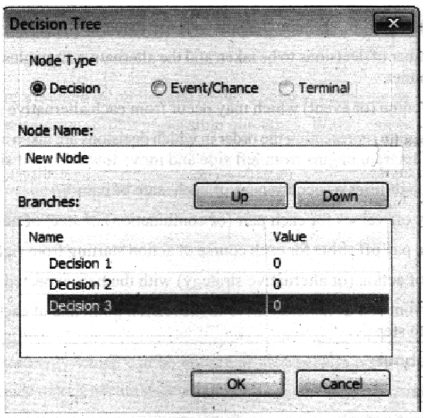
Rahul Publications

Step 1: Click on decision tree button in analytic solver platform.

Step 2: Select node and then click on add node.



Step 3: In this step, you can give any name to the node. By default it will 'New Node' under node name as shown in the below screen. In this the user can add or rename the branches of nodes like event, task etc. By default, it will be mentioned as Decision 1 and Decision 2.



Step 4: Click on OK button.

Rahul Publications

PROBLEMS

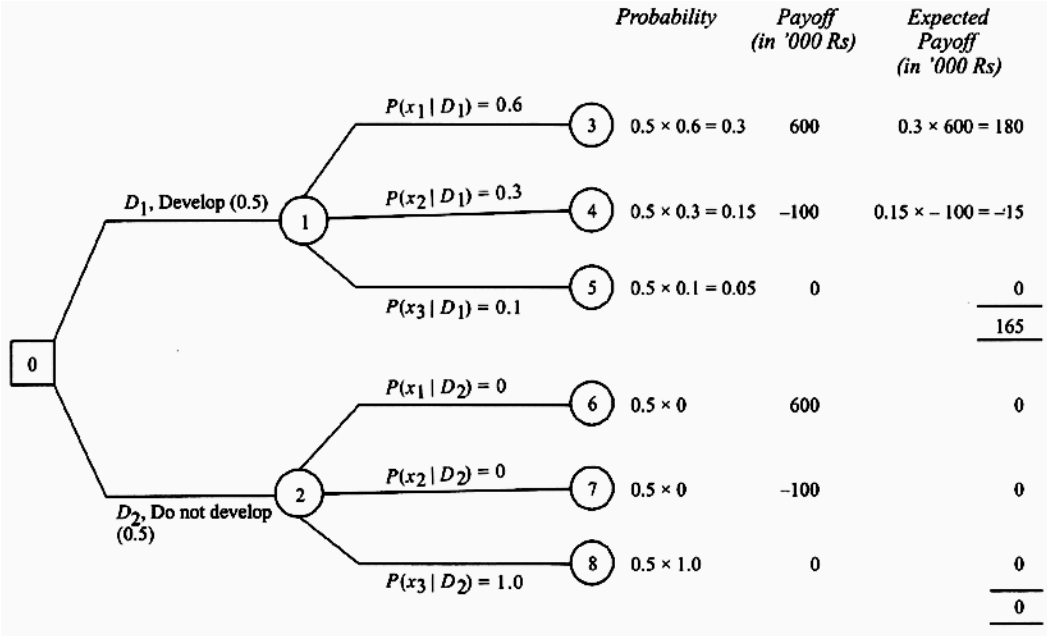
2. You are given the following estimates concerning a Research and Development programme :

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.

Sol :

The decision tree of the given problem along with necessary calculations is shown in Fig. below:



3. 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 :

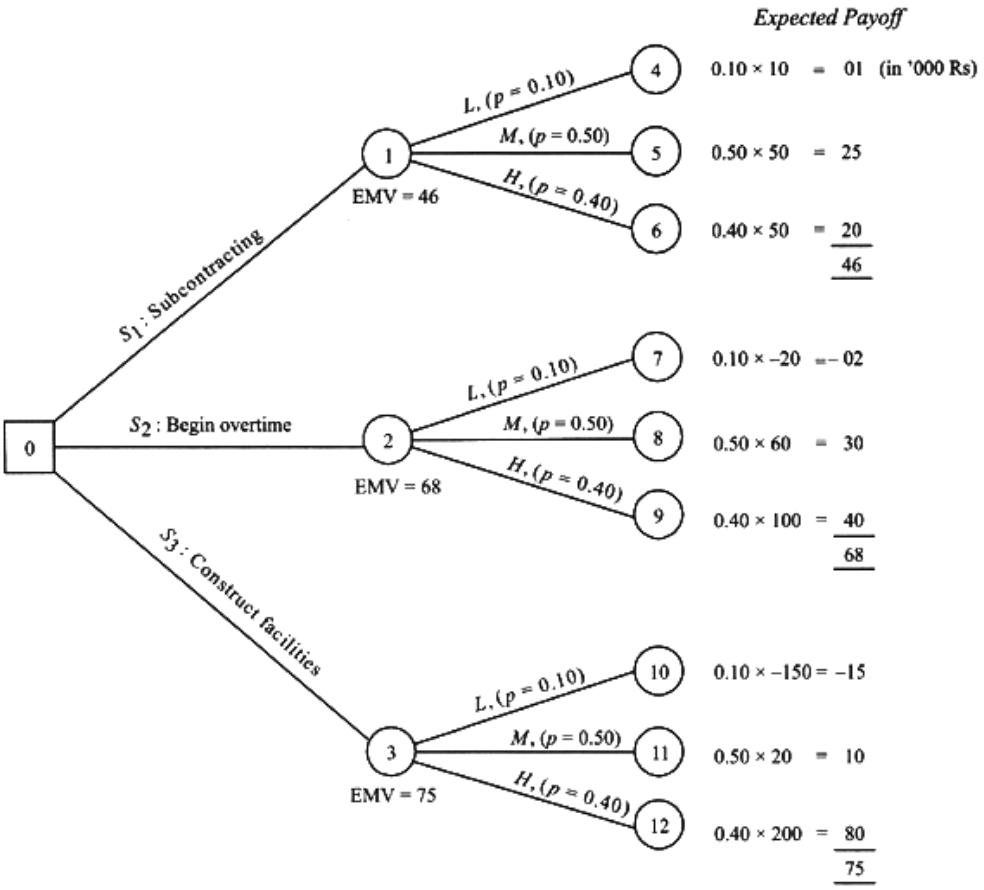
Demand	Probability	Course of Action		
		S_1 (Subcontracting)	S_2 (Begin Overtime)	S_3 (Construct Facilities)
Low (L)	0.10	10	-20	-150
Medium (M)	0.50	50	60	20
High (H)	0.40	50	100	200

Show this decision situation in the form of a decision tree and indicate the most preferred decision and its corresponding expected value.

Sol :

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.

Short Questions & Answers

1. Simulation

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.

2. Types of Simulation

Ans :

(i) 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.

(ii) 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.

(iii) Continuous Simulation

It is a simulation where time evolves continuously based on numerical integration of Differential Equations.

(iv) 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.

(v) 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. 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.

4. 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.

Rahul Publications

5. "What-if Analysis"

Ans :

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.

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.

6. Advantages and disadvantages of simulation?

Ans :

Advantages

(i) Simulation is best suited to analyze complex and large practical problems when it is not possible to solve them through a mathematical method.

(ii) Simulation is flexible, hence changes in the system variables can be made to select the best solution among the various alternatives.

(iii) In simulation, the experiments are carried out with the model without disturbing the system.

(iv) 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.

(v) Study the behavior of a system without building it.

Disadvantages

(i) Simulation does not generate optimal solutions.

(ii) It may take a long time to develop a good simulation model.

(iii) In certain cases simulation models can be very expensive.

(iv) 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.

7. Benefits of Risk Analysis

Ans :

(i) Identify, rate and compare the overall impact of risks to the organization, in terms of both financial and organizational impacts.

(ii) Identify gaps in security and determine the next steps to eliminate the weaknesses and strengthen security.

(iii) Enhance communication and decision-making processes as they relate to information security.

(iv) Improve security policies and procedures and develop cost-effective methods for implementing these information security policies and procedures.

(v) Put security controls in place to mitigate the most important risks.

(vi) Increase employee awareness about security measures and risks by highlighting best practices during the risk analysis process; and

(vii) Understand the financial impacts of potential security risks.

Rahul Publications

8. 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.

9. Stochastic Simulation

Ans :

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.

10. Applications of Simulation

Ans :

(i) Simulation is used in the field of manufacturing application.

(ii) It is used in construction engineering and project management.

(iii) To develop Military application and software. And also used in transportation model in traffic

(iv) Complex and logistics process can be optimized with simulation techniques. And also used in supply chain and distribution applications.

Rahul Publications

Choose the Correct Answers

1. Which of the following statements are NOT true of simulation? [b]
   (a) Simulation model cannot prescribe what should be done about a problem
   (b) Simulation models can be used to study alternative solutions to a problem
   (c) Simulation models the behaviour of a system
   (d) The equations describing the operating characteristics of the system are known
2. Monte Carlo simulation gets its name from which of the following? [c]
   (a) Data collection (b) Model formulation
   (c) Random-number assignment (d) Analysis
3. Simulation models can be used to obtain operating characteristic estimates in less time than with the real system using a feature of simulation called: [c]
   (a) Microseconds (b) Warp speed
   (c) Time compression (d) None of the above
4. Which of the following statistical methods are commonly used to analyze simulation results? [d]
   (a) Regression analysis (b) t-tests
   (c) Analysis of variance (d) All of the above
5. A _____ is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. [a]
   (a) Decision tree (b) Graphs
   (c) Trees (d) Neural Networks
6. Choose from the following that are Decision Tree nodes? [d]
   (a) Decision Nodes (b) End Nodes
   (c) Chance Nodes (d) All of the mentioned
7. Decision Nodes are represented by _____ [b]
   (a) Disks (b) Squares
   (c) Circles (d) Triangles
8. Chance Nodes are represented by _____ [c]
   (a) Disks (b) Squares
   (c) Circles (d) Triangles
9. End Nodes are represented by [d]
   (a) Disks (b) Squares
   (c) Circles (d) Triangles
10. Which of the following are the advantage/s of Decision Trees? [d]
   (a) Possible Scenarios can be added
   (b) Use a white box model, If given result is provided by a model
   (c) Worst, best and expected values can be determined for different scenarios
   (d) All of the mentioned

Rahul Publications

Fill in the blanks

1. _____ is an imitation of the operation of a real-world process or system.
2. _____ simulation is a computerized mathematical technique that allows people to account for risk in quantitative analysis and decision making.
3. _____ is a useful way of being able to test out various scenarios in Excel.
4. _____ is a static practice of verifying documents, design, code and program.
5. _____ is the process of evaluating the final product to check whether the software meets the customer expectations and requirements.
6. _____ testing is a testing technique and simultaneously, a progressive learning approach to perform maximum testing with the minimal planning.
7. _____ is carried out with the involvement of testing team.
8. _____ 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.
9. _____ Analysis refers to the study of cost- volume profit analysis.
10. A _____ Analysis is a graphic representation of various alternative solutions that are available to solve a problem.

Rahul Publications

ANSWERS

1. Simulation
2. Monte Carlo
3. What-if analysis
4. Verification
5. Validation
6. Exploratory
7. Validation
8. Risk analysis
9. Break-even
10. Decision Tree

Very Short Questions and Answers

1. Define Decision Tree.

Ans :

A decision tree is a schematic representation of a sequential and multidimensional decision problems, it is made up of nodes, branches, probabilities estimates and payoffs.

2. What are the types of simulation models?

Ans :

The different types of simulation models are as follows,

1. Continuous models
2. Discrete models
3. Analog simulation/environmental simulation
4. Computer simulation/system simulation
5. Deterministic and probabilistic simulation

3. State the objectives of simulation.

Ans :

The primary objectives of the simulation include,

(a) Simulations are used to change the attitudes of the individuals towards the change through the creation of suitable environment.

(b) It is also used to develop the skills of managers.

(c) It is also involved in the identification of needs and problems of the organisation.

4. What is Risk Analysis?

Ans :

Risk analysis plays a crucial role in project appraisal. Every project is exposed to some kind of risk. Evaluation of these projects can be done by making certain assumptions. In project appraisal, assumptions are unavoidable as there is no identical projects. There may be some similarities in project but they are not identical exactly. A project includes numerous elements which involves assumed values.

5. What are random numbers?

Ans :

A random number is a number obtained from sequential numbers whose probability is same as compared to the list of sequential numbers. When random numbers are generated from some deterministic process then they are called as "Pseudorandom numbers".

Rahul Publications

Internal Assessment (Mid Examinations)

In CIE, for theory subjects, during a semester, there shall be two mid-term examinations. Each MidTerm examination consists of two parts i) Part – A for 10 marks, ii) Part – B for 20 marks with a total duration of 2 hours as follows:

1. Mid-Term Examination for 30 marks:

   (a) Part - A: Objective/quiz paper/Short Note questions for 10 marks.

   (b) Part - B: Descriptive paper for 20 marks.

   The objective/quiz paper is set with multiple choice, fill-in the blanks and match the following type of questions for a total of 10 marks. The descriptive paper shall contain 6 full questions out of which, the student has to answer 4 questions, each carrying 5 marks. The average of the two Mid Term Examinations shall be taken as the final marks for Mid Term Examination (for 30 marks). The remaining 10 marks of Continuous Internal Evaluation are distributed as:

2. Assignment for 5 marks. (Average of 2 Assignments each for 5 marks)

3. PPT/Poster Presentation/ Case Study/Video presentation/Survey/Field Study/Group discussion /Role Play on a topic in the concerned subject for 5 marks before II Mid-Term Examination.

   While the first mid-term examination shall be conducted on 50% of the syllabus, the second mid-term examination shall be conducted on the remaining 50% of the syllabus.

   Five (5) marks are allocated for assignments (as specified by the subject teacher concerned). The first assignment should be submitted before the conduct of the first mid-term examination, and the second assignment should be submitted before the conduct of the second mid-term examination. The average of the two assignments shall be taken as the final marks for assignment (for 5 marks).

   PPT/Poster Presentation/ Case Study/Video presentation/Survey/Field Study/Group discussion /Role Play on a topic in the concerned subject for 5 marks before II Mid-Term Examination.

Rahul Publications

UNIT - I

Part - A

Multiple Choice Questions

1. A Linear Regression model's main aim is to find the best fit linear line and the _____ of intercept and coefficients such that the error is minimized. [a]

   (a) Optimal values (b) Linear line

   (c) Linear polynomial (d) None of the above

2. Amongst which of the following is / are the types of Linear Regression, [c]

   (a) Simple Linear Regression (d) Multiple Linear Regression

   (c) Both A and B (d) None of the above

3. Amongst which of the following is / are the true about regression analysis? [b]
   (a) Describes associations within the data
   (b) Modeling relationships within the data
   (c) Answering yes/no questions about the data
   (d) All of the above

Fill in the Blanks

4. Excel _____ Table helps in exploring and extracting important data from an Excel table or a range of data. **(Pivot)**
5. _____ Charts are one of the common yet popular techniques. It also comes under data visualization techniques in excel. **(Pie)**
6. _____ organizations utilize analytics to provide more effective treatment and control cost. They use descriptive, prescriptive and predictive. **(Health care)**

Short Notes

7. Data Analytics **(Unit-I, SQA -2)**
8. Dimensions of Big Data **(Unit-I, SQA - 4)**
9. Define business analytics. **(Unit-I, SQA - 6)**
10. Percentiles and Quartiles **(Unit-I, SQA - 8)**

Part - B

1. Define data. Explain the classification of data **(Unit-I, Q.No.1)**
2. Define Data Analytics. Explain the importance. **(Unit-I, Q.No. 2)**
3. Define Data Visualization. What are the benefits of data visualization. **(Unit-I, Q.No.10)**
4. Explain the use of bubble chart for comparing stock characteristics. **(Unit-I, Q.No.12)**
5. Explain briefly about Data queries. **(Unit-I, Q.No.13)**
6. What do you understand by frequency distributions for categorical data? **(Unit-I, Q.No. 14)**

Rahul Publications

UNIT - II

Part - A

Multiple Choice Questions

1. A statistics which is not measurable is called [b]
   (a) A constant (b) An attribute
   (c) A variable (d) A parameter
2. The number 143.9500 rounded off to the nearest tenth (one decimal place) is [b]
   (a) 143.9 (b) 144.0
   (c) 143.0 (d) 144

3. When all the values in a series occur the same number of times, then one must not compute the [c]

(a) Mean (b) Median

(c) Mode (d) Weighted mean

Fill in the Blanks

4. _____ is concerned with how each variable is related to the other variable(s). **(Association)**

5. UML stands for _____. **(Unified Modelling Language)**

6. MLE stands for _____. **(Maximum Likelihood Estimation)**

Short Notes

7. Probability Distribution **(Unit-II, SQA - 4)**

8. Discrete Probability Distribution **(Unit-II, SQA - 6)**

9. Measures of association. **(Unit-II, SQA - 12)**

10. Measures of Variability. **(Unit-II, SQA - 14)**

Part - B

1. Brief on measures of location. **(Unit-II, Q.No. 2)**

2. Explain the concept of Measures of Dispersion. **(Unit-II, Q.No. 3)**

3. Describe about Measures of Variability. **(Unit-II, Q.No.4)**

4. What is binomial distribution. Give its properties. **(Unit-II, Q.No.7)**

5. What is the nature of Gaussian distribution? How is it, unique? **(Unit-II, Q.No. 10)**

6. Explain the merits and demerits of random sampling methods. **(Unit-II, Q.No.14)**

Rahul Publications

UNIT - III

Part - A

Multiple Choice Questions

1. Choose the correct option concerning the correlation analysis between 2 sets of data. [c]

(a) Multiple correlations is a correlational analysis comparing two sets of data.

(b) A partial correlation is a correlational analysis comparing two sets of data.

(c) A simple correlation is a correlational analysis comparing two sets of data.

(d) None of the preceding.

2. The correlation for the values of two variables moving in the same direction is [c]

(a) Perfect positive (b) Negative

(c) Positive (d) No correlation.

3. Who suggested the mathematical approach for determining the magnitude of a linear relationship between two variables, such as X and Y? [c]

(a) Ya Lun Chou (b) Croxton and Cowden

(c) Karl Pearson (d) Spearman.

Fill in the Blanks

4. The _____ of the coefficient of correlation helps in interpretation. **(Probable error)**

5. Regression clearly indicates the _____ and _____relationship between the variables. **(Cause, effect)**

6. The _____ 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)**

Short Notes

7. Distinguish between one-way and two-way ANOVA. **(Unit-III, SQA - 4)**

8. Properties of correlation **(Unit-III, SQA - 6)**

9. Define Rank correlation. **(Unit-III, SQA - 8)**

10. Define correlation. **(Unit-III, SQA - 1)**

Rahul Publications

Part - B

1. State the merits and demerits of correlation. **(Unit-III, Q.No. 4)**

2. Explain the Computation of Karl Pearson's Coefficient of Correlation. **(Unit-III, Q.No. 5)**

3. Define Multiple Correlation. Explain the steps involved in generating correlation coefficient in MS Excel. **(Unit-III, Q.No.9)**

4. Define Rank correlation. Explain the properties of rank Correlation. **(Unit-III, Q.No.10)**

5. Define Regression analysis. Explain different types of regression analysis. **(Unit-III, Q.No.13)**

6. Define simple regression. Explain the concept of simple linear regression with Excel. **(Unit-III, Q.No.15)**

7. Explain in detail about regression with categorical independent variables. **(Unit-III, Q.No. 19)**

UNIT - IV

Part - A

Multiple Choice Questions

1. Which of the following is an essential process in which the intelligent methods are applied to extract data patterns? [b]

(a) Warehousing (b) Data Mining

(c) Text Mining (d) Data Selection

2. Which of the following can be considered as the correct process of Data Mining? [a]

(a) Infrastructure, Exploration, Analysis, Interpretation, Exploitation

(b) Exploration, Infrastructure, Analysis, Interpretation, Exploitation

(c) Exploration, Infrastructure, Interpretation, Analysis, Exploitation

(d) Exploration, Infrastructure, Analysis, Exploitation, Interpretation

3. Which of the following refers to the problem of finding abstracted patterns (or structures) in the unlabeled data? [b]

(a) Supervised learning

(b) Unsupervised learning

(c) Hybrid learning

(d) Reinforcement learning

Fill in the Blanks

4. Any data table is addressed by identifying one of the above data distribution methodologies, using one or more columns as the _____ key. **(Partitioning)**

5. _____ is one metric for evaluating classification models. **(Accuracy)**

6. _____ experts build the data model for the modeling process. **(Domain)**

Rahul Publications

Short Notes

7. Data Reduction. **(Unit-IV, SQA - 4)**

8. Association Rules in Data Mining. **(Unit-IV, SQA - 6)**

9. Data partitioning. **(Unit-IV, SQA - 7)**

10. Scope of Data Mining. **(Unit-IV, SQA - 2)**

Part - B

1. Define Data Mining. Explain the process of Data Mining. **(Unit-IV, Q.No. 1)**

2. Define is Data Reduction? What are the benefits of data reduction ? **(Unit-IV, Q.No. 5)**

3. What is Unsupervised Learning? Explain briefly about common cluster algorithm methods. **(Unit-IV, Q.No.8)**

4. State the applications, advantages and disadvantages of cluster analysis. **(Unit-IV, Q.No. 12)**

5. Explain the steps for developing Association Rules using XL miner. **(Unit-IV, Q.No. 14)**

UNIT - V

Part - A

Multiple Choice Questions

1. Which of the following statements are NOT true of simulation? [b]
   (a) Simulation model cannot prescribe what should be done about a problem
   (b) Simulation models can be used to study alternative solutions to a problem
   (c) Simulation models the behaviour of a system
   (d) The equations describing the operating characteristics of the system are known
2. Chance Nodes are represented by _____ [c]
   (a) Disks (b) Squares
   (c) Circles (d) Triangles
3. Choose from the following that are Decision Tree nodes? [d]
   (a) Decision Nodes (b) End Nodes
   (c) Chance Nodes (d) All of the mentioned

Fill in the Blanks

4. _____ simulation is a computerized mathematical technique that allows people to account for risk in quantitative analysis and decision making. **(Monte Carlo)**
5. _____ 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. **(Risk analysis)**
6. _____ Analysis refers to the study of cost- volume profit analysis. **(Break-even)**

Short Notes

7. Types of Simulation **(Unit-V, SQA - 2)**
8. Monte Carlo simulation **(Unit-V, SQA - 4)**
9. Stochastic Simulation **(Unit-V, SQA - 9)**
10. Applications of Simulation **(Unit-V, SQA -10)**

Part - B

1. Write a short notes on Stochastic Simulation. **(Unit-V, Q.No.1)**
2. Analyze in detail about advantages and disadvantages of simulation technique. **(Unit-V, Q.No.3)**
3. Distinguish between solutions derived from simulation models and solutions from analytical models. Highlight some of the problem areas of application, where simulation is considered most preferable. **(Unit-V, Q.No.4)**
4. Define Monte Carlo Simulation. State the Advantages and Disadvantages of Monte Carlo Simulation. **(Unit-V, Q.No. 7)**
5. Define verification. Explain various methods of verification. **(Unit-V, Q.No. 12)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

M.B.A III Semester Examinations

R22

MODEL PAPER - I

BUSINESS ANALYTICS

Time : 3 Hours] **[Max. Marks : 60**

Note : This question paper contains two parts **A** and **B**.

Part A is compulsory which carries 10 marks. Answer all questions in **Part A.**

Part B consists 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 (10 × 1 = 10 Marks)

ANSWERS

1. (a) Data Analytics. **(Unit - I, SQA-2)**

   (b) Cross Tabulation. **(Unit - I, SQA-9)**

   (c) Random Sample. **(Unit - II, SQA-3)**

   (d) Measures of association. **(Unit - II, SQA-12)**

   (e) Define simple regression. **(Unit - III, SQA-2)**

   (f) Distinguish between one-way and two-way ANOVA. **(Unit - III, SQA-4)**

   (g) Data Mining. **(Unit - IV, SQA-1)**

   (h) Cluster Analysis. **(Unit - IV, SQA-5)**

   (i) Types of Simulation. **(Unit - V, SQA-2)**

   (j) Stochastic Simulation. **(Unit - V, SQA-9)**

PART - B (5 × 10 = 50 Marks)

2. Enumerate in detail about Data analytics and the importance of data analytics in the current scenario. **(Unit - I, Q.No.2)**

OR

3. Describe about Pivot Tables. How data is explored using pivot tables? **(Unit - I, Q.No.15)**

4. Explain briefly about Measures of location. **(Unit - II, Q.No.2)**

OR

5. Find the Value of Range and the Coefficient for the following data:

   7, 9, 6, 8, 11, 10, 4 **(Unit - II, Prob.5)**

6. Define Multiple Correlation. Explain the steps involved in generating correlation coefficient in MS Excel. **(Unit - III, Q.No.9)**

Rahul Publications

OR

7. The alibaba Traders company wishes to test whether its three salesmen Saleem, 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. Saleem made 5 calls, Basha made 4 calls and Vikram made 5 calls. The following are the weekly sales records of the three salesmen :

Saleem Rs.	Basha Rs.	Vikram Rs.
300	600	700
400	300	300
300	300	400
500	400	600
000	---	500

Perform the analysis of variance test and draw your conclusions. **(Unit - III, Prob.5)**

8. Define Data Mining. Explain the process of Data Mining. **(Unit - IV, Q.No.1)**

OR

9. Describe briefly about Association Rules. **(Unit - IV, Q.No.13)**

10. Explain how Monte Carlo simulation can be developed in Ms.Excel? **(Unit - V, Q.No.8)**

OR

11. Define What If Analysis. What are the basic options available in excel for performing What If Analysis? **(Unit - V, Q.No.10)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

M.B.A III Semester Examinations

R22

MODEL PAPER - II

BUSINESS ANALYTICS

Time : 3 Hours] [Max. Marks : 60

Note : This question paper contains two parts **A** and **B**.

Part A is compulsory which carries 10 marks. Answer all questions in **Part A.**

Part B consists 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 (10 × 1 = 10 Marks)

ANSWERS

1. (a) Define business analytics. **(Unit - I, SQA-6)**
   (b) Dimensions of Big Data. **(Unit - I, SQA-4)**
   (c) Distribution Fitting **(Unit - II, SQA-11)**
   (d) Measures of location. **(Unit - II, SQA-13)**
   (e) Properties of correlation **(Unit - III, SQA-6)**
   (f) Define Rank correlation. **(Unit - III, SQA-8)**
   (g) Data Reduction. **(Unit - IV, SQA-4)**
   (h) Logistic Regression. **(Unit - IV, SQA-10)**
   (i) Random number generation. **(Unit - V, SQA-3)**
   (j) Decision tree analysis. **(Unit - V, SQA-8)**

PART - B (5 × 10 = 50 Marks)

2. Explain the different types of data visualization tools ? **(Unit - I, Q.No.12)**

OR

3. Define business analytics. Discuss briefly the role of business analytics in current business environment. **(Unit - I, Q.No.8)**

4. Explain briefly about Measures of Association. **(Unit - II, Q.No.5)**

OR

5. Explain about data modeling and distribution fitting in detail. **(Unit - II, Q.No.17)**

6. Define simple regression. Explain the concept of simple linear regression with Excel. **(Unit - III, Q.No.15)**

OR

Rahul Publications

7. Calculate the coefficient of correlation between X and Y from the following data.

Marks in English (X)	2	5	4	6	9
Marks in Mathematics (Y)	3	4	4	8	9

(Unit - III, Prob.1)

8. Explain briefly about various data reduction methods. **(Unit - IV, Q.No.6)**

OR

9. Explain briefly about K-Nearest Neighbors. **(Unit - IV, Q.No.21)**

10. Analyze in detail about advantages and disadvantages of simulation technique. **(Unit - V, Q.No.3)**

OR

11. Distinguish between Verification and Validation. **(Unit - V, Q.No.14)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

M.B.A III Semester Examinations

R22

MODEL PAPER - III

BUSINESS ANALYTICS

Time : 3 Hours] [Max. Marks : 60

Note : This question paper contains two parts **A** and **B**.

Part A is compulsory which carries 10 marks. Answer all questions in **Part A.**

Part B consists 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 (10 × 1 = 10 Marks)

ANSWERS

1. (a) Importance of data visualization. **(Unit - I, SQA-7)**
   (b) Data. **(Unit - I, SQA-1)**
   (c) Probability Distribution **(Unit - II, SQA-6)**
   (d) Probability Distribution **(Unit - II, SQA-4)**
   (e) Define correlation. **(Unit - III, SQA-1)**
   (f) Define Regression analysis. **(Unit - III, SQA-9)**
   (g) Predictive Accuracy. **(Unit - IV, SQA-8)**
   (h) What is cause and effect modeling ? **(Unit - IV, SQA-11)**
   (i) Simulation. **(Unit - V, SQA-1)**
   (j) Applications of Simulation **(Unit - V, SQA-10)**

PART - B (5 × 10 = 50 Marks)

2. What do you understand by frequency distributions for categorical data? **(Unit - I, Q.No.14)**

OR

3. Why classification of data is considered hierarchical in nature? **(Unit - I, Q.No.1)**

4. Define binomial distribution. Explain the properties of binomial distribution. **(Unit - II, Q.No.7)**

OR

5. The students of a class have elected live candidates to represent them on the college management council.

Rahul Publications

S.No.	Gender	Age in years
1.	Male	18
2.	Male	19
3.	Female	22
4.	Female	20
5.	Male	23

This group decides to elect a spokesperson by randomly drawing a name from a hat.

Calculate the probablity of the spokesperson being either female or over 21 years. **(Unit - II, Prob.6)**

6. Briefly explain the process of constructing regression line using the method of least squares. **(Unit - III, Q.No.17)**

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. **(Unit - III, Prob.4)**

8. Define Cluster Analysis. Explain the properties of clustering. **(Unit - IV, Q.No.10)**

OR

9. Explain why cluster analysis is called as unsupervised learning. **(Unit - IV, Q.No.24)**

10. Define Simulation. Explain different types of Simulation. **(Unit - V, Q.No.1)**

OR

11. Briefly explain risk analysis techniques in decision making. **(Unit - V, Q.No.16)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERBAD

M.B.A III - Semester Examination **R17**

December - 2018

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, SQA- 5)**

   (b) Discuss about measures of association. **(Unit - II, SQA- 12)**

   (c) Explain the correlation techniques. **(Unit - III, SQA- 1)**

   (d) What is logistic regression ? **(Unit - IV, SQA- 10)**

   (e) Explain about what-if analysis. **(Unit - V, SQA-5)**

PART - B (5 × 10 = 50 Marks)

(Essay Type Questions)

2. Write about Data Visualization Techniques. **(Unit - I, Q.No.12)**

(OR)

3. Explore the importance of Data Analytics in business decision making. **(Unit - I, Q.No. 8)**

4. How to measure association and variability of data ? **(Unit - II, Q.No. 5)**

(OR)

5. Enumerate the sampling techniques. **(Unit - II, Q.No. 13)**

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. **(Unit - III, Prob. 4)**

Rahul Publications

8. Explain data reduction techniques in Data Mining. **(Unit - IV, Q.No. 6)**

(OR)

9. Explain why cluster analysis is called as unsupervised learning. **(Unit - IV, Q.No. 24)**

10. Explain Monte-Carlo simulation. **(Unit - V, Q.No. 7)**

(OR)

11. Briefly explain risk analysis techniques in decision making. **(Unit - V, Q.No. 16)**

Rahul Publications

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, SQA-1,5)**

   (b) What is the best measure of location ? **(Unit - II, SQA- 13)**

   (c) What is simple and multiple Regressions ? **(Unit - III, SQA- 2,3)**

   (d) Define Data mining? Explain the scope of Data mining ? **(Unit - IV, SQA- 1, 2)**

   (e) What is simulation ? **(Unit - V, SQA-1)**

PART - B (5 × 10 = 50 Marks)

(Essay Type Questions)

2. Explain the different types of data visualization tools ? **(Unit - I, Q.No.12)**

(OR)

3. What is business Analytics and its types. **(Unit - I, Q.No.8)**

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. **(Unit - II, Prob.5)**

(OR)

5. Explain about Random sampling methods with merits and demerits. **(Unit - II, Q.No. 14)**

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

(Unit - III, Prob. 6)

Rahul Publications

(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

(Unit - III, Prob. 3)

8. Explain about the cluster Analysis with an example. **(Unit - IV, Q.No. 10)**

(OR)

9. Explain about different types of learning and brief on data exploration and reduction.

(Unit - IV, Q.No. 4,5)

10. Explain the steps involved in Monte - Carlo simulation. **(Unit - V, Q.No. 9)**

(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.

(Unit - V, Prob.1)

Rahul Publications

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, SQA- 5)**

   (b) Describe about Measures of Variability. **(Unit - II, SQA- 14)**

   (c) What is simple regression ? **(Unit - III, SQA- 2)**

   (d) What are all the association rules ? **(Unit - IV, SQA- 3)**

   (e) What is Monte Carlo Simulation ? **(Unit - V, SQA - 4)**

PART - B (5 × 10 = 50 Marks)

(Essay Type Questions)

2. Explain in detail about statistical methods for summarizing data. **(Unit - I, Q.No. 14)**

(OR)

3. Describe about Pivot Tables. How data is explored using pivot tables? **(Unit - I, Q.No.15)**

4. Explain in detail about,

   (a) Probability distribution **(Unit - II, Q.No. 6)**

   (b) Data modeling **(Unit - II, Q.No. 17)**

   (c) Continuous probability distribution. **(Unit - II, Q.No. 9)**

(OR)

5. Explain about data modeling and distribution fitting in detail. **(Unit - II, Q.No. 17)**

6. How regression by the method of least squares technique is used ? **(Unit - III, Q.No. 17)**

(OR)

7. Explain in detail about regression with categorical independent variables. **(Unit - III, Q.No. 19)**

Rahul Publications

8. Describe briefly about,

   (a) Partition data **(Unit - IV, Q.No. 16)**

   (b) Classification accuracy **(Unit - IV, Q.No.17)**

   (c) Prediction accuracy **(Unit - IV, Q.No. 20)**

(OR)

9. What is cluster analysis ? What are its applications ? **(Unit - IV, Q.No. 10, 12)**

10. Analyze in detail about advantages and disadvantages of simulation techniques. **(Unit - V, Q.No. 3)**

(OR)

11. Explain in detail about

   (a) Risk Analysis **(Unit - V, Q.No. 15)**

   (b) Decision Tree Analysis. **(Unit - V, Q.No. 17)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

MBA III-Semester Examinations

R19

July/August - 2021

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. 12, 14)**

2. Define Business Analytics in detail with its types? **(Unit-I, Q.No. 8)**

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? **(Unit-II, Prob.4)**

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. **(Unit-II, Prob. 1)**

5. Write down the characteristics of correlation coefficient. **(Unit-III, Q.No. 1)**

6. Obtain a regression line associated with the following data by the method of least squares. **(Unit-III, Prob. 3)**

X	1	2	3	4	5
Y	166	184	142	180	338

Rahul Publications

7. (a) What is Classification and Regression Trees (CART)? **(Unit-IV, Q.No. 22)**

   (b) Brief on data exploration and reduction. **(Unit-IV, Q.No. 6)**

8. (a) Analyze in detail about advantages and disadvantages of simulation technique. **(Unit-V, Q.No. 3)**

   (b) Explain the steps involved in Monte-Carlo simulation. **(Unit-V, Q.No. 9)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

MBA III - Semester Examinations

April / May - 2022

R19

DATA ANALYTICS

Time : 3 Hours] **[Max. Marks : 75**

Answer any Five questions

All questions carry equal marks

ANSWERS

1. Define Data Analytics and explain its importance in an organization? **(Unit-I, Q.No. 2)**

2. (a) What is Data Visualization? Explain different tools of data Visualization. **(Unit-I, Q.No. 10, 12)**

   (b) Explain about pivot tables. **(Unit-I, Q.No. 15)**

3. (a) Define discrete probability distribution and solve the below given problem.

   Let X be a discrete random variable with the following PMF. **(Unit-II, Q.No. 6) (Unit-II, Prob. 3)**

$$P_x(k) = \begin{cases} 0.1 & \text{for } k = 0 \\ 0.4 & \text{for } k = 1 \\ 0.3 & \text{for } k = 2 \\ 0.2 & \text{for } k = 3 \\ 0 & \text{otherwise} \end{cases}$$

   (i) Find EX

   (ii) Find Var(X)

   (iii) If $Y = (X - 2)^2$, find EY

   (b) What are measures of Variability. **(Unit-II, Q.No. 4)**

4. (a) Explain about continuous probability distribution. **(Unit-II, Q.No. 9)**

   (b) Explain about Measures of Association. **(Unit-II, Q.No. 5)**

5. (a) What is Two way Anova? Mention merits and demerits **(Unit-III, Q.No. 24)**

   (b) A reputed marketing agency in India has three different training programs for its salesmen. The three programs are Method - A, B, C. To assess the success of the programs. 4 salesmen from each of the programs were sent to the field. Their performances in terms of sales are given in the following table. Test whether there is significant difference among methods and among salesmen.

   (Unit-III, Prob. 7)

Rahul Publications

Salesman	Methods		
	A	B	C
1	4	8	5
2	7	9	8
3	10	5	9
4	6	7	8

6. (a) Explain about multiple correlation. **(Unit-III, Q.No. 9)**

   (b) Explain about method of Least squares. **(Unit-III, Q.No. 17)**

7. (a) Explain the algorithm of KNN? **(Unit-IV, Q.No. 21)**

   (b) We have the data from the questionnaire survey and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not.

   Here are four training samples.

X_1 = Acid Durability (sec)	X_2 = Strength (kg/sqmt)	Y = Classification
8	7	Bad
6	6	Bad
4	5	Good
2	4	Good

   Now the Factory produces a new paper tissue that pass laboratory test with $X_1 = 3$ and $X_2 = 7$. Without another expensive survey can we guess what the classification of this new tissue is?

   (Unit-IV, Prob.1)

8. Explain the process of Monte Carlo Simulation. **(Unit-V, Q.No. 9)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

MBA III - Semester Examinations

March/April - 2023

R19

DATA ANALYTICS

Time : 3 Hours **Max. Marks : 75**

Note :

(i) Question paper consists of Part A, Part B.

(ii) Part A is compulsory, which carries 25 marks. In Part A, Answer all questions.

(iii) In Part B, Answer any one question from each unit. Each question carries 10 marks and may have a, b as sub questions.

PART - A

ANSWERS

1. (a) What is the importance of data analytics in business? **(Unit-I, SQA- 5)**

   (b) What is the nature of Gaussian distribution? How is it, unique? **(Unit-II, Q.No.10)**

   (c) Distinguish between one-way and two-way ANOVA. **(Unit-III, SQA-4)**

   (d) What does data exploration and reduction involve? **(Unit-IV, Q.No.6)**

   (e) Briefly explain the use of decision tree analysis in problems involving a measure of probability, of success or failure. **(Unit-V, Q.No.18)**

Rahul Publications

PART - B

2. (a) Why classification of data is considered hierarchical in nature? **(Unit-I, Q.No.1)**

   (b) Explain the use of bubble chart for comparing stock characteristics. **(Unit-I, Q.No.12)**

OR

3. (a) What do you understand by frequency distributions for categorical data? **(Unit-I, Q.No.14)**

   (b) Explain the process of computing percentiles. **(Unit-I, Q.No.14)**

4. (a) With an example explain multiplication law of probability. **(Unit-II, Q.No.12)**

   (b) Explain the use of scatter chart for determining covariance. **(Unit-III, Q.No. 8)**

OR

5. (a) The students of a class have elected live candidates to represent them on the college management council. **(Unit-II, Prob.6)**

S.No.	Gender	Age in years
1.	Male	18
2.	Male	19
3.	Female	22
4.	Female	20
5.	Male	23

This group decides to elect a spokesperson by randomly drawing a name from a hat. Calculate the probability of the spokesperson being either female or over 21 years.

(b) Brief on data modeling. **(Unit-II, Q.No. 15)**

6. (a) What are assumptions of linear discriminant analysis? **(Unit-III, Q.No.19)**

(b) Briefly explain the process of constructing regression line using the method of least squares. **(Unit-III, Q.No.17)**

OR

7. Calculate the coefficient of correlation between X and Y from the following data.

Marks in English (X)	2	5	4	6	9
Marks in Mathematics (Y)	3	4	4	8	9

(Unit-IV, Prob.1)

8. (a) Explain how discriminant analysis serves the purpose of classifying new data. **(Unit-III, Q.No. 21)**

(b) What is validation data set? How is it useful? **(Unit-IV, Q.No.18)**

OR

9. (a) How do you measure classification performance? **(Unit-IV, Q.No.19)**

(b) What is cause and effect modeling? **(Unit-IV, SQA-11)**

10. (a) Write a small jiote on, stochastic simulation and random numbers. **(Unit-V, Q.No.1)**

(b) What are the advantages and limitations of simulation models? **(Unit-V, Q.No. 3)**

OR

11. Distinguish between solutions derived from simulation models and solutions from analytical models. Highlight some of the problem areas of application, where simulation is considered most preferable. **(Unit-V, Q.No. 4)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

MBA III - Semester Examinations

R22

February - 2024

BUSINESS ANALYTICS

Time : 3 Hours] **[Max. Marks : 60]**

Note : This question paper contains two parts A and B. i) Part - A for 10 marks, ii) Part-B for 50 marks.

- Part-A is a compulsory question which consists of ten sub-questions from all units carrying equal marks.
- Part-B consists of ten questions (numbered from 2 to 11) carrying 10 marks each. From each unit, there are two questions and the student should answer one of them. Hence, the student should answer five questions from Part-B.

PART-A (10 Marks)

ANSWERS

1. (a) Define the term "Data" **(Unit - I, VSQA -1)**

   (b) What is Pivot Table? **(Unit - I, Q.No. 15)**

   (c) What do you mean by "Population" in analytics? **(Unit - II, SQA - 1)**

   (d) Define measure of association in statistics. **(Unit - II, SQA - 12)**

   (e) Define simple Regression. **(Unit - III, Q.No. 2)**

   (f) Define Analysis of Variance (ANOVA). **(Unit - III, VSQA-4)**

   (g) What do you mean by cluster analysis? **(Unit - IV, SQA - 5)**

   (h) What is the primary objective of data mining?

Ans :

The primary goal of data mining is to discover hidden patterns and relationships in the data that can be used to make informed decisions or predictions. This involves exploring the data using various techniques such as clustering, classification, regression analysis, association rule mining, and anomaly detection

   (i) What is decision tree analysis? **(Unit - V, SQA - 8)**

   (j) Define ris analysis in the context of simulation. **(Unit - V, Q.No. 16)**

PART - B (50 Marks)

2. What is the significance of data in today's business environment and examine the application of business analytics in practice? **(Unit - I, Q.No. 8)**

3. What is the importance of data visualization ? Explain about various tools used for data visualization with suitable example. **(Unit - I, Q.No. 10, 12)**

4. (a) What are the measures of dispersion? Explain the purpose of measures of dispersion? **(Unit - II, Q.No. 3)**

   (b) Describe the measures of variability in brief. **(Unit - II, Q.No. 4)**

Rahul Publications

5. (a) Find out the Standard Deviation from the following data :

Age in Years	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of People affected	6	10	32	35	24	14	8

Sol :

Age	(f_i)	Mid value (xi)	d_i	fidi	d^2	fd^2
0 - 10	6	5	– 3	– 18	9	54
10 - 20	10	15	– 2	– 20	4	40
20 - 30	32	25	– 1	– 32	1	32
30 - 40	35	35 A	0	0	0	0
40 - 50	24	45	1	24	1	24
50 - 60	14	55	2	28	4	56
60 - 70	8	65	3	24	9	72
	129			6		278

$$r = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

$$= \sqrt{\frac{278}{129} - \left(\frac{6}{129}\right)^2}$$

$$= \sqrt{2.155 - \frac{36}{16,641}}$$

$$\sqrt{2.155 - 0.002}$$

$$\sqrt{2.1458}$$

$$= 1.465$$

(b) Describe the terms data modeling and distribution fitting. **(Unit - II, Q.No. 17)**

6. (a) Explain the process of constructing regression line using the method of least square **(Unit - III, Q.No. 17)**

(b) How do you build good regression models? **(Unit - III, Q.No. 18)**

7. (a) Calculate Karl Pearson's coefficient of correlation between expenditure on advertising and sales from the data given below:

Advertising expenses ('000 Rs.)	39	65	62	90	82	75	25	98	36	78
Sales (lakh Rs.)	47	53	58	86	62	68	60	91	51	84

Rahul Publications

Sol :

X	Y	$x = x - \bar{x}$	$y = y - \bar{y}$	x^2	y^2	xy
39	47	–26	–19	676	361	494
65	53	0	–13	0	169	0
62	58	–3	–8	9	64	+24
90	86	25	20	625	400	500
82	62	17	– 4	289	16	–68
75	68	10	+2	100	4	+20
25	60	–40	–6	1600	36	+240
98	91	33	+25	1089	625	825
36	51	–29	–15	841	225	435
78	84	13	+18	169	324	234
650	660			5398	2224	2704

$$\bar{x} = \frac{650}{10} = 65 \qquad \bar{y} = \frac{660}{10} = 66$$

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

$$= \frac{2704}{\sqrt{5398 \times 2224}}$$

$$= \frac{2704}{\sqrt{12,005,152}} \quad \frac{2704}{3464.84} = 0.78$$

(b) Brief on linear discriminant analysis importance in analytics. **(Unit - III, Q.No. 20)**

8. (a) Discuss the potential applications and benefits of data mining in various industries.

Ans :

Data mining has a wide range of applications across various industries, offering numerous benefits that can enhance decision-making, efficiency, and customer satisfaction. Here are some key industries and how they utilize data mining:

1. Retail

(i) **Customer Segmentation:** Identifying distinct customer groups based on purchasing behavior.

(ii) **Inventory Management:** Predicting demand to optimize stock levels.

(iii) **Recommendation Systems:** Personalizing product suggestions to increase sales.

2. Health care

(i) **Patient Care Improvement :** Analyzing patient data to identify trends and enhance treatment plans.

(i) **Predictive Analytics** :Forecasting disease outbreaks or patient admissions to improve resource allocation.

(ii) **Fraud Detection** : Identifying unusual billing patterns to prevent health care fraud.

3. Finance

(i) **Credit Scoring :** Assessing the creditworthiness of applicants using historical data.

(ii) **Risk Management :** Analyzing market trends to mitigate financial risks.

(iii) **Fraud Detection :** Monitoring transactions for suspicious activities to prevent fraud.

4. Manufacturing

(i) **Predictive Maintenance :**Analyzing equipment data to predict failures and schedule maintenance.

(ii) **Quality Control** : Identifying defects and improving production processes through data analysis.

(iii) **Supply Chain Optimization :** Enhancing logistics and inventory processes based on demand forecasts.

5. Telecommunications

(i) **Churn Prediction :** Identifying customers likely to switch providers and implementing retention strategies.

(ii) **Network Optimization :** Analyzing usage patterns to enhance network performance and reduce downtime.

(iii) **Fraud Detection :** Monitoring call patterns to detect fraudulent activities.

6. Education

(i) **Student Performance Analysis :** Identifying at-risk students and tailoring interventions.

(ii) **Curriculum Improvement :** Analyzing feedback and performance data to enhance educational programs.

(iii) **Personalized Learning :** Customizing learning experiences based on individual student data.

7. Marketing

(i) **Campaign Optimization :** Evaluating the effectiveness of marketing campaigns and adjusting strategies accordingly.

(ii) **Customer Lifetime Value Prediction :** Estimating the long-term value of customers to inform marketing investments.

(iii) **Sentiment Analysis :** Analyzing social media and feedback to gauge public perception of brands.

8. Travel and Hospitality

(i) **Price Optimization :** Analyzing booking patterns to adjust pricing strategies dynamically.

(ii) **Customer Experience Enhancement :** Personalizing services based on customer preferences and feedback.

(iii) **Demand Forecasting :** Predicting travel trends to optimize resource allocation.

9. Energy

(i) **Consumption Prediction :** Forecasting energy demand to manage supply effectively.

(ii) **Fault Detection :** Analyzing data from sensors to detect and address issues in energy grids.

(iii) **Customer Insights :** Understanding usage patterns to promote energy efficiency programs.

Benefits of Data Mining

(i) **Improved Decision-Making :** Data-driven insights lead to more informed strategic choices.

Rahul Publications

(ii) **Cost Reduction** : Efficient resource allocation and predictive maintenance minimize operational costs.

(iii) **Enhanced Customer Satisfaction** : Personalized services and products increase customer loyalty.

(iv) **Competitive Advantage**

Organizations can identify trends and respond faster than competitors.

(v) **Innovation**

Uncovering new patterns can lead to innovative products and services.

(b) Describe the methods of data exploration. **(Unit - IV, Q.No. 6)**

OR

9. (a) Explain about the K-Nearest Neighbors (KNN) algorithm. **(Unit - IV, Q.No. 21)**

(b) Discuss the logistics regression with an example. **(Unit - IV, Q.No. 23)**

10. Describe the purpose of What If Analysis in the context of decision-making.

Ans :

What-If Analysis is a powerful decision-making tool that allows individuals and organizations to explore the potential outcomes of different scenarios by altering key variables.

Purpose

(i) Predict Outcomes

It enables decision-makers to assess how changes in variables affect outcomes, helping them predict the results of different actions or decisions.

(ii) Test Scenarios

By modifying inputs, users can test various "what-if" scenarios to see how different conditions impact results.

(iii) Identify Risks

It helps identify potential risks and uncertainties associated with different decisions, allowing for better risk management.

(iv) Contingency Planning

Organizations can prepare for various possibilities, ensuring they have strategies in place for unfavourable outcomes.

(v) Optimize Resources

Decision-makers can evaluate how changes in strategy might affect resource allocation and operational efficiency.

(vi) Cost-Benefit Analysis

By simulating different scenarios, organizations can determine the most cost-effective options.

(vii) Data-Driven Insights

What-If Analysis uses quantitative data to guide decisions, reducing reliance on intuition alone.

(viii) Improve Confidence

Decision-makers gain confidence in their choices by understanding potential implications and outcomes.

11. Explain the steps involved in Monte Carlo Simulation. **(Unit - V, Q.No. 9)**

Rahul Publications

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

MBA III - Semester Examinations **R22**

July / August - 2024

BUSINESS ANALYTICS

Time : 3 Hours] **[Max. Marks : 60]**

Note : This question paper contains two parts A and B. i) Part - A for 10 marks, ii) Part-B for 50 marks.

- Part-A is a compulsory question which consists of ten sub-questions from all units carrying equal marks.
- Part-B consists of ten questions (numbered from 2 to 11) carrying 10 marks each. From each unit, there are two questions and the student should answer one of them. Hence, the student should answer five questions from Part-B.

PART-A (10 Marks)

ANSWERS

1. (a) Give a definition of Data Analytics. **(Unit - I, Q.No. 2)**

(b) Elucidate about descriptive statistical measures ?

Ans :

Descriptive statistics helps researchers and analysts to describe the central tendency (mean, median, mode), dispersion (range, variance, and standard deviation), and shape of the distribution of a dataset. It also involves graphical representation of data to aid visualization and understanding.

(c) Define predictive analytics and its uses.

Ans :

Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning. Companies employ predictive analytics to find patterns in this data to identify risks and opportunities

(d) Mean of a Binomial distribution is 24, Standard deviation = 4, what n, p, q respectively are ?

Sol :

$np = 24$

$\sqrt{npq} = 4$

$npq = 16$

$24q = 16$

$q = \frac{16}{24} = \frac{2}{3}$

$p = 1 - q$

$1 - \frac{2}{3} = \frac{1}{3}$

$n \times \frac{1}{3} = 24$

$n = 72$

Rahul Publications

(e) Give the coefficient of correlation being 0.8, what the coefficient of determination will be ?

Ans :

Coefficient of determination = (coefficient of correlation)^2

$= (0.8)^2 = 0.64$

(f) How do you explore data using pivot tables?

Ans :

In the Pivot Table Fields list, you will find all the tables that you imported and the fields in each of them. If the fields are not visible for any table, Click on the arrow next to that table in the Pivot Table Fields list. The fields in that table will be displayed.

(g) Write about distribution fitting and its uses.

Ans :

Distribution fitting is the art of choosing a probability model for an unknown and unknowable population, and calibrating that model using a representative sample from the population. Such a model allows for inferences about the population to be made despite not knowing all of its properties

(h) Enlist the difference between one-way and two - way ANOVA. **(Unit - III, SQA- 4)**

(i) Enumerate about logistics regression. **(Unit - IV, SQA - 10)**

(j) Explain decision tree analysis. **(Unit - V, SQA-8)**

PART-B (50 Marks)

2. Describe about Data, the importance of analytics in data and brief on data visualization tools. **(Unit - I, Q.No. 1, 2, 10)**

3. Enumerate data for business analytics and how business analytics is used in practice, also provide few illustration on Big Data. **(Unit - I, Q.No. 4)**

4. The average starting salary of a college graduate is \$ 19000 according to government's report. The average salary of a random sample of 100 graduates is \$ 18800. The standard error is 800. Is the government's report reliable as the level of significance is 0.05. Find the p-value and test the hypothesis in.

Sol :

(a) $H_0 : \mu = \mu_0 = 19000$ vs. $H_a : \mu \neq \mu_0 = 19000$

$n = 100,\ \bar{x} = 18800,\ s = 800,\ \alpha = 0.05$

Then,

$$|z| = \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| = \left|\frac{18800 - 19000}{800/\sqrt{100}}\right| = |-2.5| = 2.5 > z_{a/2} = z_{0.025} = 1.96$$

Therefore, reject H_0.

(b) P- value = $P(|Z| > |z|) = P(|Z| > 2.5) = 2.\ P(Z > 2.5) = 0.0124 > 0.01$

Therefore, not reject H_0

OR

5. List out random sampling from probability distribution and also brief about continuous Probability sampling with suitable example. **(Unit - II, Q.No. 9, 13)**

6. Find karl pearson's coefficient of correlation from the following data between height of father (x) and height of son (y)

x	64	65	66	67	68	69	70
y	66	67	65	68	70	68	72

Sol :

X	Y	$X = x - \bar{x}$	$Y = y - \bar{y}$	x^2	y^2	xy
64	66	–3	–2	9	4	6
65	67	–2	–1	4	1	2
66	65	–1	–3	1	9	3
67	68	0	0	0	0	0
68	70	+1	+2	1	4	2
69	68	+2	0	4	0	0
70	72	+3	4	9	16	12
469	476			28	34	25

$$\bar{X} = \frac{469}{7} = 67$$

$$\bar{Y} = \frac{476}{7} = 68$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{25}{\sqrt{28 \times 34}} = \frac{25}{\sqrt{952}}$$

$$= \frac{25}{30.85} = 0.810$$

Rahul Publications

7. Describe about Simple and Multiple Regression and Regression by the Method of Least Square with an illustration. **(Unit - III, Q.No. 15,16)**

8. Discuss about Data Exploration and Reduction and the scope of Data Mining. **(Unit - IV, Q.No. 2, 4, 5)**

OR

9. Elucidate in detail about the following:
   (a) Cluster Analysis **(Unit - IV, Q.No. 10)**
   (b) Association Rules **(Unit - IV, Q.No. 13)**
   (c) Supervised Learning **(Unit - IV, Q.No. 15)**

10. Outline in detail about random number regression and brief on advantages, disadvantages of simulation. **(Unit - V, Q.No. 3, 5)**

OR

11. Discuss in detail about Monte Carlo Simulation and What if Analysis ? **(Unit - V, Q.No. 7, 10)**