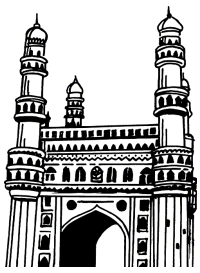
*Rahul's* ✔
*Topper's Voice*

AS PER
CBCS SYLLABUS



# BBA

## *III Year VI Semester*

Latest 2022 Edition

# BASIC BUSINESS ANALYTICS

- Study Manual
- FAQ's and Important Questions
- Short Question & Answers
- Choose the Correct Answer
- Fill in the blanks
- Solved Model Papers

Rs. 159/-


- by -

WELL EXPERIENCED LECTURER



*Rahul Publications*™

Hyderabad. Ph : 66550071, 9391018098



All disputes are subjects to Hyderabad Jurisdiction only

# BBA

*III Year VI Semester*

# BASIC BUSINESS ANALYTICS

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

© *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*

***Price ` . 159/-***

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

VASU BOOK CENTRE

**Shop No. 2, Beside Gokul Chat, Koti, Hyderabad.**

**Maternity Hospital Opp. Lane, Narayan Naik Complex, Koti, Hyderabad.**
**Near Andhra Bank, Subway, Sultan Bazar, Koti, Hyderabad -195.**

# BASIC BUSINESS ANALYTICS

CONTENTS

## STUDY MANUAL


| | |
|---|---|
| **FAQ's & Important Questions** | **III - VI** |
| **Unit - I** | **1 - 24** |
| **Unit - II** | **25 - 76** |
| **Unit - III** | **77 - 160** |
| **Unit - IV** | **161 - 202** |


## SOLVED MODEL PAPERS


| | |
|---|---|
| **Model Paper - I** | **203 - 204** |
| **Model Paper - II** | **205 - 206** |
| **Model Paper - III** | **207 - 208** |

# SYLLABUS

## UNIT - I

### INTRODUCTION TO BUSINESS ANALYTICS :

Definition, Types of Analytics-Descriptive, Predictive and Prescriptive, Business Analytics Applications in Different Areas (BA in Practice), Big Data.

## UNIT - II

### DESCRIPTIVE ANALYTICS 1 :

Types of Data-Population and Sample Data, Quantitative and Categorical Data, Cross-Sectional and Time Series Data, Sources of data, Descriptive Statistics- Measures of Location (central Tendency)-Mean, Median and Mode and relationship between them – Problems.

## UNIT - III

### DESCRIPTIVE ANALYTICS 2 :

Measures of Variability-Range, Variance, Standard deviation, Coefficient of Variation, Percentiles, Quartiles, Analyzing Distributions – Empirical Rule, Identifying Outliers, Box Plots, Measures of Association -Scatter Charts, Covariance, Correlation Coefficient – Problems.

## UNIT - IV

### PREDICTIVE ANALYTICS :

Trend Analysis, Regression Analysis-Least Square Method, Assessing the Fit of Simple Linear Regression, Coefficient of Determination, Introduction to Data Mining- Definition, Methods of Data Mining, Applications of Data Mining.

# Contents

**Topic** **Page No.**


**UNIT - I**

1.1 Introduction To Business Analytics ..... 1

1.1.1 Definition ..... 1

1.2 Types of Analytics ..... 2

1.2.1 Descriptive, Predictive and Prescriptive ..... 2

1.3 Business Analytics Applications in Different Areas (BA in Practice) ..... 9

1.4 Big Data ..... 11

- **Short Questions Answer ..... 20 - 22**
- **Choose the Correct Answer ..... 23 - 23**
- **Fill in the blanks ..... 24 - 24**

**UNIT - II**

2.1 Types of Data ..... 25

2.1.1 Population and Sample Data ..... 26

2.1.2 Quantitative And Categorical Data ..... 42

2.2 Sources of Data ..... 53

2.3 Descriptive Statistics ..... 54

2.3.1 Measures of Location(Central Tendency) ..... 54

2.3.1.1 Mean, Median and Mode and Relationship between them ..... 54

- **Short Questions Answer ..... 69 - 74**
- **Choose the Correct Answer ..... 75 - 75**
- **Fill in the blanks ..... 76 - 76**

**UNIT - III**

3.1 Measures of Variability ..... 77

3.1.1 Range ..... 77

3.1.2 Variance ..... 81

3.1.3 Standard Deviation ..... 89

3.1.4 Coefficient of Variation ..... 93

3.1.5 Percentiles ..... 96

3.1.6 Quartiles ..... 101

**Topic** **Page No.**

3.2 Analysing Distributions ..... 104
3.2.1 Empirical Rule ..... 104
3.2.2 Identifying Outliers ..... 109
3.2.3 Box Plots ..... 115
3.3 Measures of Association ..... 126
3.3.1 Scatter Charts ..... 126
3.3.2 Covariance ..... 134
3.4 Correlation Coefficient ..... 139
- **Short Questions Answer ..... 153 - 157**
- **Choose the Correct Answer ..... 158 - 159**
- **Fill in the blanks ..... 160 - 160**

**UNIT - IV**

4.1 Predictive Analytics ..... 161
4.1.1 Trend Analysis ..... 162
4.2 Regression Analysis ..... 163
4.2.1 Regression by Method of Least Square ..... 165
4.2.2 Assessing the Fit of Simple Linear Regression ..... 168
4.2.3 Coefficient of Determination ..... 175
4.3 Introduction to Data Mining ..... 179
4.3.1 Definition ..... 179
4.3.2 Methods of Data Mining ..... 182
4.4.3 Applications of Data Mining ..... 195
- **Short Questions Answer ..... 197 - 200**
- **Choose the Correct Answer ..... 201 - 201**
- **Fill in the blanks ..... 202 - 202**

# *Frequently Asked & Important Questions*

## UNIT - I

**1. Explain different types of Business Analytical Methods.**

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

Refer Unit-I, Q.No. 2

**2. Explain the different models in business analytics?**

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

Refer Unit-I, Q.No. 5

**3. Describe briefly about the role of business analytics in current business environment.**

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

Refer Unit-I, Q.No. 6

**4. Describe the various stages of Big Data Life Cycle.**

*Ans :* **(Imp.)**

Refer Unit-I, Q.No. 12

**5. What are the technologies available to manage big data?**

*Ans :* **(Imp.)**

Refer Unit-I, Q.No. 14

**6. What are the major types of Big data applications?**

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

Refer Unit-I, Q.No. 16

**7. Explain the role of big data in competing food apps Swiggy and Zomato.**

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

Refer Unit-I, Q.No. 17

## UNIT - II

**1. Explain about population and sample data.**

*Ans :* **(Imp.)**

Refer Unit-II, Q.No. 2

**2. How to calculate variance in Excel ? Explain.**

*Ans :* **(Imp.)**

Refer Unit-II, Q.No. 5

**3. Explain how to calculate population standard deviation with examples.**

*Ans :* **(Imp.)**

Refer Unit-II, Q.No. 7

**4. Explain about quantitative and categorical data with examples.**

*Ans :* **(Imp.)**

Refer Unit-II, Q.No. 8

**5. What is Time Series Data and Cross Sectional Data? Explain with an examples.**

*Ans :* **(Imp.)**

Refer Unit-II, Q.No. 10

**6. Explain the utility of measures of central tendency in understanding performance of business organization.**

*Ans :* **(Dec.-21, Octo.-20, Imp.)**

Refer Unit-II, Q.No. 12

**7. Explain the Relationship between Mean, Median and Mode with examples.**

*Ans :* **(Imp.)**

Refer Unit-II, Q.No. 13

## UNIT - III

**1. What is variance? Explain how to find variance in various modes with an examples.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 4

**2. Explain how to calculate population variance in Excel.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 6

**3. What is standard deviation? Explain with an examples how is Standard Deviation calculated.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 7

**4. Define Quartile? How to find the quartile deviation in statistics? Explain with an examples.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 12

**5. What is the empirical rule? How and where the empirical rule is used ? Explain with an example.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 13

**6. Explain how to identify outliers in statistics.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 15

**7. Explain how box plot can be drawn for the given data.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 16

**8. What is a Scatter Plot? Explain how to draw scatter plot for the given example.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 18

**9. What is correlation coefficient? How to find the Correlation Coefficient.**

*Ans :* **(Imp.)**

Refer Unit-III, Q.No. 22

## UNIT - IV

**1. Explain the Applications of Predictive analysis.**

*Ans :* **(Imp.)**

Refer Unit-IV, Q.No. 3

**2. Explain different types and utility of regression analysis.**

*Ans :* **(Imp.)**

Refer Unit-IV, Q.No. 7

**3. Discuss briefly about least square regression.**

*Ans :* **(Imp.)**

Refer Unit-IV, Q.No. 10

**4. Define Data Mining. Explain the scope of Data Mining.**

*Ans :* **(Imp.)**

Refer Unit-IV, Q.No. 14

**5. Explain the process of data mining.**

*Ans :* **(Imp.)**

Refer Unit-IV, Q.No. 16

**6. Explain the various approaches of Date mining**

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

Refer Unit-IV, Q.No. 18

**7. Explain the various applications of Data Mining.**

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

Refer Unit-IV, Q.No. 27

# UNIT I

## INTRODUCTION TO BUSINESS ANALYTICS :

Definition, Types of Analytics-Descriptive, Predictive and Prescriptive, Business Analytics Applications in Different Areas (BA in Practice), Big Data.

## 1.1 INTRODUCTION TO BUSINESS ANALYTICS

### 1.1.1 Definition

**Q1. What is Business analytics?**

**(OR)**

**Define Business analytics.**

*Ans :*

**Meaning**

Business analytics is the practice of iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by companies committed to data-driven decision-making.

**Definitions**

**(i) According to Schaer (2018),** "allows your business to make predictive analysis rather than reacting to changes in data".

**(ii) According to Gabelli School of Business (2018),** "involves applying models, methods, and tools to data, producing insights that lead to informed business decisions"

**(iii) According to Wells(2008),** "the application of logic and mental processes to find meaning in data"

**(iv) According to Lynda (2018),** "allows us to learn from the past and make better predictions for the future".

Business analytics (BA) refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. Business analytics focuses on developing new insights and understanding of business performance based on data and statistical methods. In contrast, business intelligence traditionally focuses on using a consistent set of metrics to both measure past performance and guide business planning, which is also based on data and statistical methods. Business analytics makes extensive use of statistical analysis, including explanatory and predictive modeling, and fact-based management to drive decision making. It is therefore closely related to management science. Analytics may be used as input for human decisions or may drive fully automated decisions. Business intelligence is querying, reporting, online analytical processing (OLAP), and "alerts."

In other words, querying, reporting, OLAP, and alert tools can answer questions such as what happened, how many, how often, where the problem is, and what actions are needed. Business analytics can answer questions like why is this happening, what if these trends continue, what will happen next (predict), and what is the best outcome that can happen (optimize).

Business analytics. Abbreviated as BA, business analytics is the combination of skills, technologies, applications and processes used by organizations to gain insight in to their business based on data and statistics to drive business planning.

Rahul Publications

## 1.2 Types of Analytics

### 1.2.1 Descriptive, Predictive and Prescriptive

**Q2. Explain different types of Business Analytical Methods.**

**(OR)**

**Describe various categories of Business Analytical Methods.**

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

**1. Descriptive Analytics**

This can be termed as the simplest form of analytics. The mighty size of big data is beyond human comprehension and the first stage, hence involves crunching the data into understandable chunks. The purpose of this analytics type is just to summarize the findings and understand what is going on.

Among some frequently used terms, what people call as advanced analytics or business intelligence is basically usage of descriptive statistics (arithmetic operations, mean, median, max, percentage, etc.) on existing data. It is said that 80% of business analytics mainly involves descriptions based on aggregations of past performance. It is an important step to make raw data understandable to investors, shareholders and managers. This way, it gets easy to identify and address the areas of strengths and weaknesses such that it can help in strategizing.

The two main techniques involved are data aggregation and data mining stating that this method is purely used for understanding the underlying behavior and not to make any estimations. By mining historical data, companies can analyze the consumer behaviors and engagements with their businesses that could be helpful in targeted marketing, service improvement, etc. The tools used in this phase are MS Excel, MATLAB, SPSS, STATA, etc.

**2. Diagnostic Analytics**

Diagnostic Analytics is used to determine why something happened in the past. It is characterized by techniques such as drill-down, data discovery, data mining and correlations. Diagnostic analytics takes a deeper look at data to understand the root causes of the events. It is helpful in determining what factors and events contributed to the outcome. It mostly uses probabilities, likelihoods and the distribution of outcomes for the analysis.

In a time series data of sales, diagnostic analytics would help you understand why the sales have decreased or increased for a specific year or so. However, this type of analytics has a limited ability to give actionable insights. It just provides an understanding of causal relationships and sequences while looking backward.

A few techniques that uses diagnostic analytics include attribute importance, principle components analysis, sensitivity analysis and conjoint analysis. Training algorithms for classification and regression also fall in this type of analytics.

**3. Predictive Analytics**

Predictive Analytics is used to predict future outcomes. However, it is important to note that it cannot predict if an event will occur in the future; it merely forecasts what the probabilities of the occurrence of the event are. A predictive model builds on the preliminary descriptive analytics stage to derive the possibility of the outcomes.

The essence of predictive analytics is to devise models such that the existing data is understood to extrapolate the future occurrence or simply, predict the future data.

Rahul Publications

One of the common applications of predictive analytics is found in sentiment analysis where all the opinions posted on social media are collected and analyzed (existing text data) to predict the person's sentiment on a particular subject as being positive, negative or neutral (future prediction).

Hence, predictive analytics includes building and validation of models that provide accurate predictions. Predictive analytics relies on machine learning algorithms like random forests, SVM, etc. and statistics for learning and testing the data. Usually, companies need trained data scientists and machine learning experts for building these models. The most popular tools for predictive analytics include Python, R, Rapid Miner, etc.

The prediction of future data relies on the existing data as it cannot be obtained otherwise. If the model is properly tuned, it can be used to support complex forecasts in sales and marketing. It goes a step ahead of the standard BI in giving accurate predictions.

**4. Prescriptive Analytics**

The basis of this analytics is predictive analytics, but it goes beyond the three mentioned above to suggest the future solutions. It can suggest all favorable outcomes according to a specified course of action and also suggest various course of actions to get to a particular outcome. Hence, it uses a strong feedback system that constantly learns and updates the relationship between the action and the outcome.

The computations include optimization of some functions that are related to the desired outcome. For example, while calling for a cab online, the application uses GPS to connect you to the correct driver from among a number of drivers found nearby. Hence, it optimizes the distance for faster arrival time. Recommendation engines also use prescriptive analytics.

The other approach includes simulation where all the key performance areas are combined to design the correct solutions. It makes sure whether the key performance metrics are included in the solution. The optimization model will further work on the impact of the previously made forecasts. Because of its power to suggest favorable solutions, prescriptive analytics is the final frontier of advanced analytics or data science, in today's term.

Rahul Publications

**Q3. Differentiate between descriptive analytics and predictive analytics.**

*Ans :*

| S.No | Nature | Descriptive Analysis | Predictive Analysis |
|---|---|---|---|
| 1. | **Data** | It uses historical data and reconfigures it into simple, easy and readable formats. | it also uses historical data but only gaps will be filled in the available data. |
| 2. | **Highlights** | It highlights and describe the nature of business operations. | It highlights and creates data models. |
| 3. | **Techniques** | It involves techniques like data aggregation and data mining. | It involves statistics and other forecasting techniques. |
| 4. | **Results** | It provide accurate result in reports by | It provide inaccurate result in reports. |

**Q4. Describe the evolution of business analytics.**

*Ans :*

The evolution of business analytics started with the introduction of computers in the late 1960's and extended up to 1990's. With the introduction of computers, storing and nasalization of data became much easier and gave rise to the concept of Business Intelligence.

In the year 1958, IBM incorporation introduced a new concept known as 'Business Intelligence', which helps in collecting research information, managing, analyzing and reporting of data.

Business intelligence software is a set of tool that helps the business in retrieving, analyzing and transforming data into essential business requirements.

Statistics is an important part of business which not only summarizes the data but also helps in eliminating the unnecessary data. Statistical method includes basic tools such as description, estimation, interference and exploration. It also includes techniques like forecasting, regression and data mining.

**Q5. Explain the different models in business analytics?**

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

An analytical model is simply a mathematical equation that describes relationships among variables in a historical data set. The equation either estimates or classifies data values. In essence, a model draws a "line" through a set of data points that can be used to predict outcomes. What is a business analysis model?

Simply put, a business analysis model outlines the steps a business takes to complete a specific process, such as ordering a product or on boarding a new hire. Process modeling (or mapping) is key to improving process efficiency, training, and even complying with industry regulations.

Because there are many different kinds of processes, organizations, and functions within a business, BAs employ a variety of visual models to map and analyze data.

The following the different models:

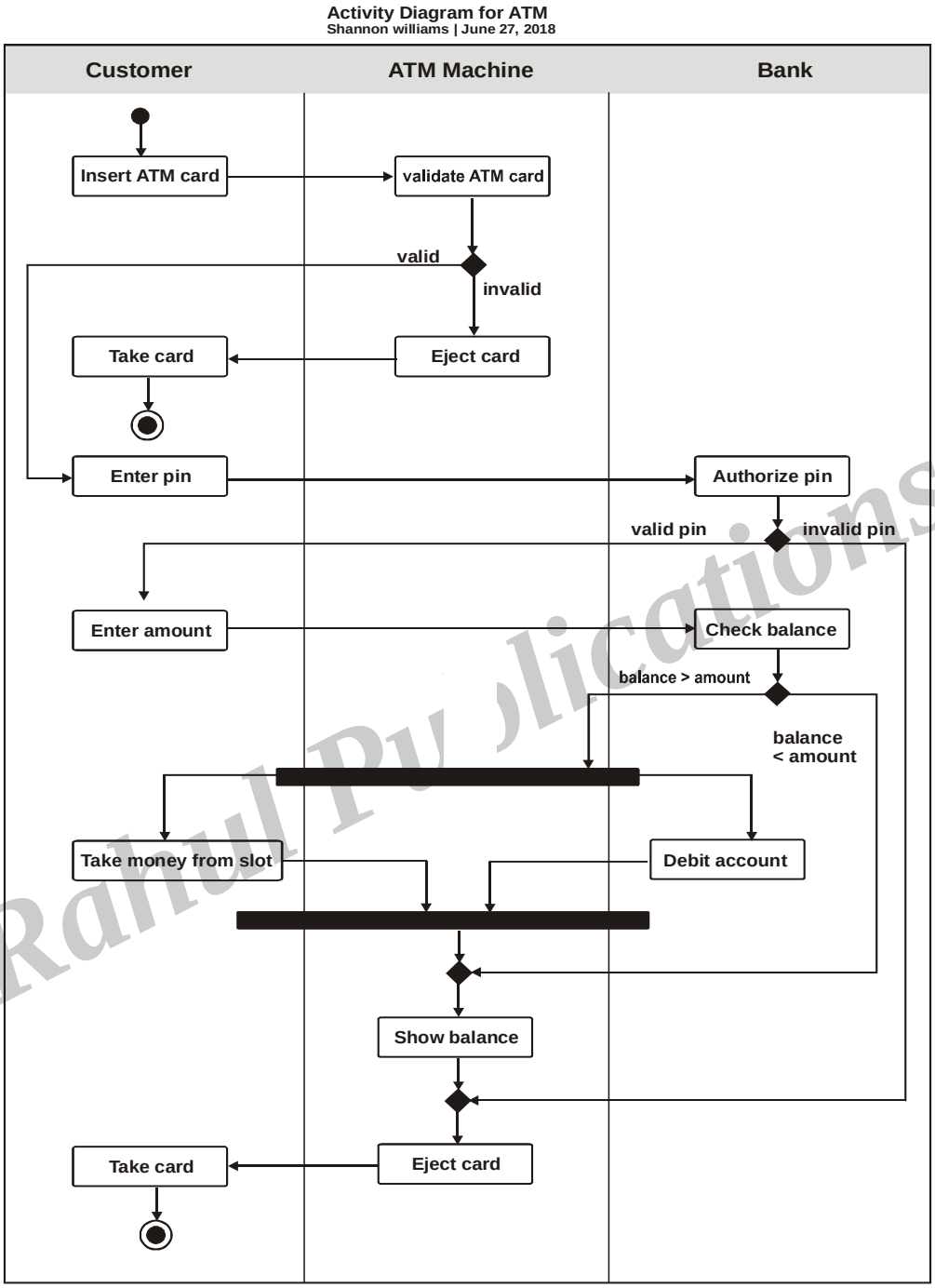
**1. Activity diagrams**

Activity diagrams are a type of UML behavioural diagram that describes what needs to happen in a system. They are particularly useful for communicating process and procedure to stakeholders from both the business and development teams.

A Business analytical might use an activity diagram to map the process of logging in to a website or completing a transaction like withdrawing or depositing money

Rahul Publications

### Activity diagram for ATM



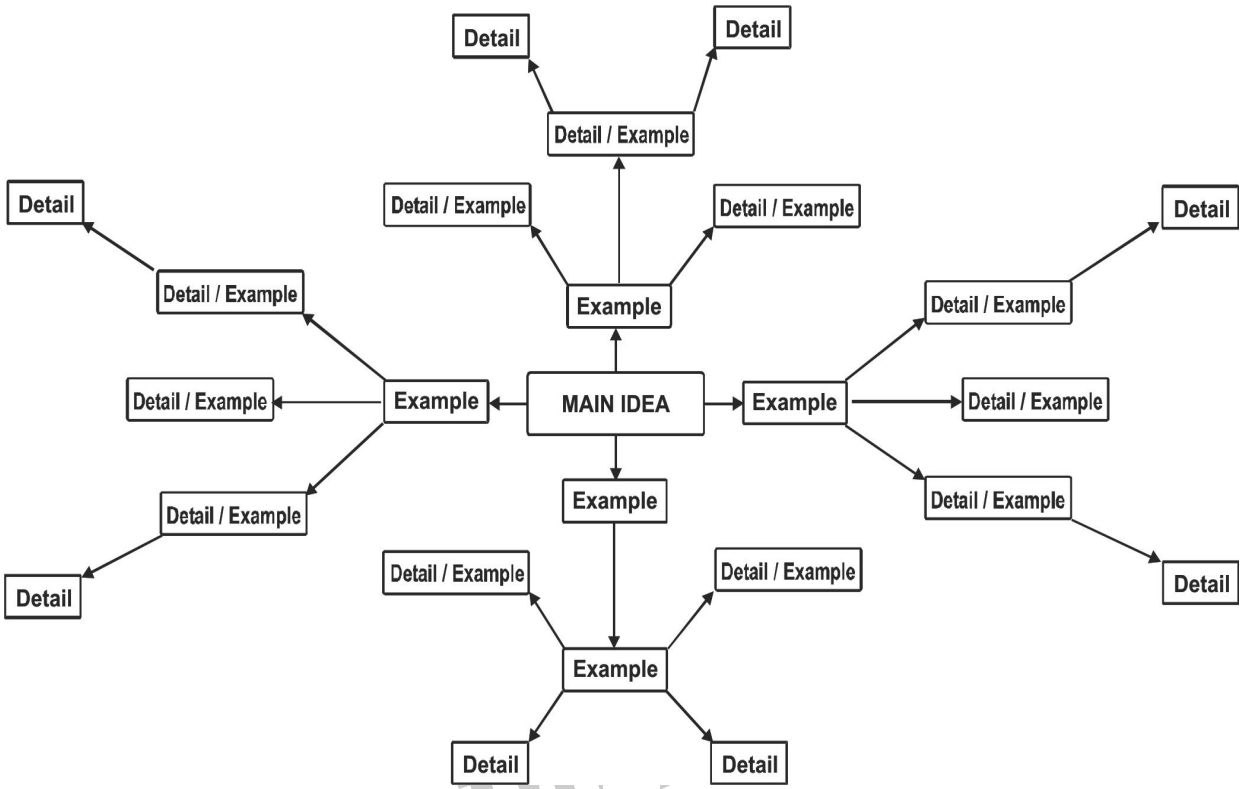

Rahul Publications

## 2. Feature mind maps

Business diagrams aren't just for late-stage analysis or documentation. They are also useful during a project's initial brainstorming phase. Feature mind maps help BAs organize the sometimes messy brainstorm process so that ideas, concerns, and requests are clearly captured and categorized.

This visual ensures initial details and ideas don't fall through the cracks so you can make informed decisions about project direction, goals, and scope down the line.

**Basic mind map**




## 3. Product roadmaps

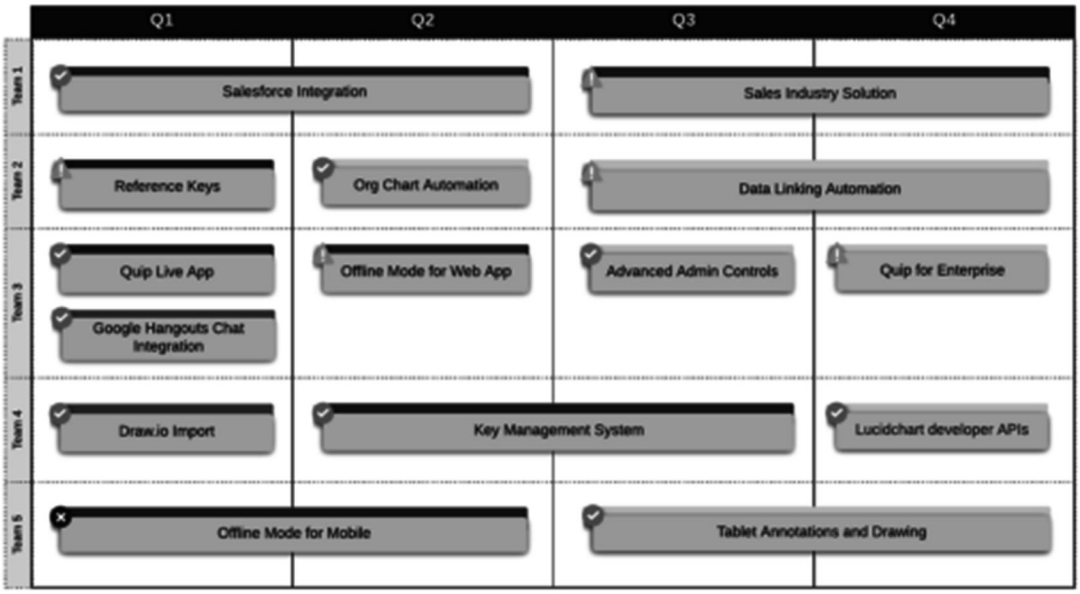
Product (or feature) roadmaps outline the development and launches of a product and its features. They are a focused analysis of a product's evolution, which helps developers and other stakeholders focus on initiatives that add direct value to the user.

The beauty of product roadmaps lies in their flexibility and range of applications. BAs can create different product roadmaps to illustrate different information, including:

- Maintenance and bug fixes
- Feature releases
- High-level strategic product goals

While product roadmaps are commonly used internally by development teams, they are also useful resources for other groups like sales.

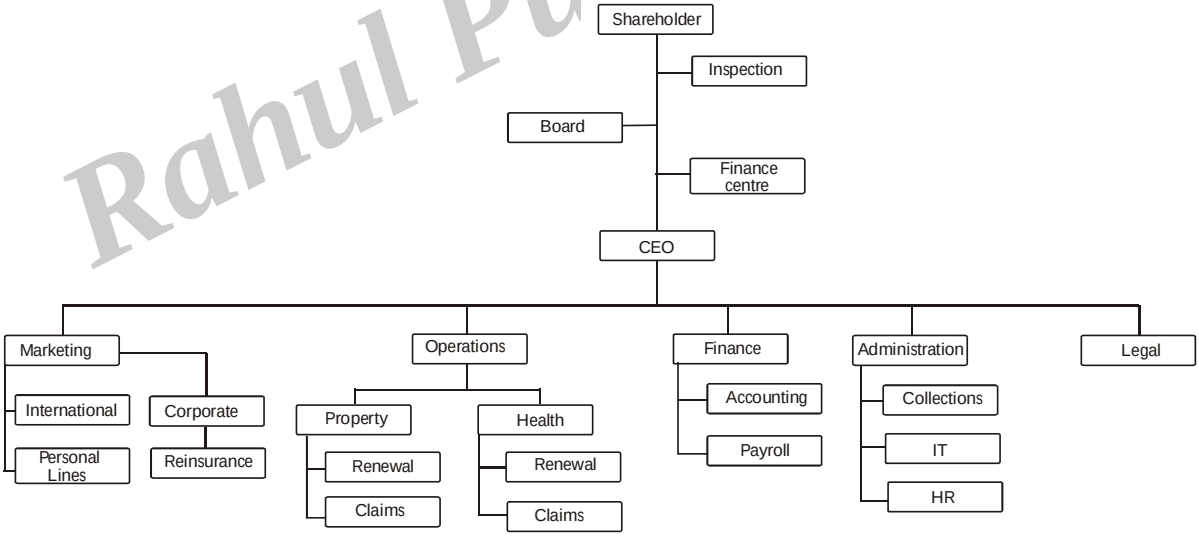
A defined product outline and schedule helps sales stay on the same page as the developers so they can deliver accurate, updated information to their prospects and clients. Because of their versatility and broad applications across teams and organizations, product roadmaps are a core part of an analyst's toolbox.




## 4. Organizational charts

Organizational charts outline the hierarchy of a business or one of its departments or teams. They are especially helpful reference charts for employees to quickly understand how the company is organized and identify key stakeholders and points of contact for projects or queries.

Additionally, organizational charts prove useful for stakeholder analysis and modeling new groupings and teams following organizational shifts.




## 5. SWOT analysis

The SWOT analysis is a fundamental tool in a Business analytics. SWOT stands for strengths, weaknesses, opportunities, and threats. A SWOT analysis evaluates a business's strengths and weaknesses and identifies any opportunities or threats to that business.

SWOT analysis helps stakeholders make strategic decisions regarding their business. The goal is to capitalize on strengths and opportunities while reducing the impact of internal or external threats and weaknesses.

Rahul Publications

From a visual modeling perspective, SWOT analysis is fairly straight forward. A typical model will have four boxes or quadrants-one for each category-with bulleted lists outlining the respective results.

**SWOT Analysis**

| STRENGTHS | WEAKNESSES |
|---|---|
| • Outlet one<br>• Outlet two<br>• Outlet three<br>• Outlet four | • Outlet one<br>• Outlet two<br>• Outlet three<br>• Outlet four |
| **OPPORTUNITIES** | **THREATS** |
| • Outlet one<br>• Outlet two<br>• Outlet three<br>• Outlet four | • Outlet one<br>• Outlet two<br>• Outlet three<br>• Outlet four |

Rahul Publications

**6. User interface wireframe**

Another essential business diagram is the UI wireframe. Software development teams use wireframes (also called mockups or prototypes) to visually outline and design a layout for a specific screen. In other words, wireframes are the blueprints for a website or software program. They help stakeholders assess navigational needs and experience for a successful practical application.

Wireframes range from low-fidelity to high-fidelity prototypes. Low-fidelity wireframes are the most basic outlines, showing only the bare-bones layout of the screen. High-fidelity wireframes are typically rendered in the later planning stages and will include specific UI elements (e.g., buttons, drop-down bars, text fields, etc.) and represent how the final implementation should look on the screen.

Website Design Wireframe (Click on image to modify online)

**7. Process flow diagram**

A process flow diagram (PFD) is typically used in chemical and process engineering to identify the basic flow of plant processes, but it can also be used in other fields to help stakeholders understand how their organization operates.

**A PFD is best used to:**

- Document a process.
- Study a process to make changes or improvements.
- Improve understanding and communication between stakeholders.

These diagrams focus on broad, high-level systems rather than annotating minor process details.

**8. PESTLE analysis**

A PESTLE analysis often goes hand-in-hand with a SWOT analysis. PESTLE evaluates external factors that could impact business performance. This acronym stands for six elements affecting business: political, economic, technological, environmental, legal, and sociological.

PESTLE analysis assesses the possible factors within each category, as well as their potential impact, duration of effect, type of impact (i.e., negative or positive), and level of importance.

This type of business analysis helps stakeholders manage risk, strategically plan and review business goals and performance, and potentially gain an advantage over competitors.

**9. Entity-relationship diagram**

An entity-relationship diagram (ER diagram) illustrates how entities (e.g., people, objects, or concepts) relate to one another in a system. For example, a logical ER diagram visually shows how the terms in an organization's business glossary relate to one another.

**ER diagrams comprise three main parts:**

- Entities
- Relationships
- Attributes

Attributes apply to the entities, describing further details about the concept. Relationships are where the key insights from ER diagrams arise. In a visual model, the relationships between entities are illustrated either numerically or via crow's foot notation.

These diagrams are most commonly used to model database structures in software engineering and business information systems and are particularly valuable tools for Business analytics in those fields.

## 1.3 Business Analytics Applications in Different Areas (BA in Practice)

**Q6. Describe briefly about the role of business analytics in current business environment.**

**(OR)**

**Explain the applications of business analytics in business.**

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

Organizations use analytic techniques in their businesses to attain competitive advantage and solve various problems. Business analytics offers various simple tools like graphs and reports as well as sophisticated tools like datamining, simulation and optimization. Organizations use business analytics by following prescriptive, predictive and descriptive analytic methods based upon the degree of complexity and competitive advantage. The prescriptive analytic method consists of optimization, decision analysis and simulation. Predictive analytic method consists of predictive modeling, forecasting and data mining. Descriptive analytic method consists of descriptive statistics, data visualization, data query and standard reporting. Prescriptive and predictive analytics are often called advanced analytics. Thus, business analytics is used by multiple organization including government organizations, some of them are, Google, Amazon.com, Netflix, IBM, the Internal Revenue Services, UPS, General Electric and Procter and Gamble.

Some of the types of analytics practiced by various organizations are 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 constructing optimal portfolios of investments and allocating assets. Addition to this, simulation is also used for ascertaining risk in the financial sector.

**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 rise 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.

**3. Human Resource (HR) Analytics**

HR function utilizes analytics to ensure that

Rahul Publications

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.

**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.

**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.

**6. Analytics for Government and Non-Profit Organizations**

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

**7. Sports Analytics**

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

**8. Web Analytics**

The analysis of internet activity including visits of users to social media sites like LinkedIn 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 to determine statistically the reasons for differences in the sales and website traffic.

Rahul Publications

**Q7. Explain the various challenges in business analytics.**

*Ans :*

- **Executive Ownership:** Business Analytics requires buy-in from senior leadership and a clear corporate strategy for integrating predictive models.
- **IT Involvement:** Technology infrastructure and tools must be able to handle the data and Business Analytics processes.
- **Available Production Data vs. Cleansed Modeling Data :** Watch for technology infrastructure that restrict available data for historical modeling, and know the difference

between historical data for model development and real-time data in production.

- **Project Management Office (PMO)** : The correct project management structure must be in place in order to implement predictive models and adopt an agile approach.
- **End user Involvement and Buy-In :** End users should be involved in adopting Business Analytics and have a stake in the predictive model.
- **Change Management :** Organizations should be prepared for the changes that Business Analytics bring to current business and technology operations.
- **Explain ability vs. the "Perfect Lift"**: Balance building precise statistical models with being able to explain the model and how it will produce results.

## 1.4 BIG DATA

**Q8. Define date. Explain types of data.**

*Ans :*

A collection of facts from which conclusions may be drawn.

Following are the types of data,

**1. Structured Data**

Structured data is the data that can be stored in an organized format in rows and columns, so that it can be effectively retrieved and processed by an application. Examples: Data Base Management System (DB ) is used for saving the structured data.

**2. Semi-structured Data**

Semi-structured data is the data which does not accommodates to a data model but has a structure. Semi-structured data is not used easily by a computer program.

**Examples:** Emails, XML, Markup languages like HTML etc.

**3. Unstructured Data**

Unstructured data is the data that cannot be stored in systematic format in rows and columns, due to which it becomes complicated for an application to query and retrieve the data.

**Q9. What is Big Data? State the characteristics of Big Data.**

*Ans :*

**Meaning**

The term Big-data can be used to describe huge volumes of data both in structured and unstructured format taken from webtraffic, e-mail messages and content of social media such as tweets and status messages. Interestingly, the big data moves beyond the processing ability to capture, store and analyse typical database systems. As a consequence a large quantity of data is generated which moves fast and fail to fit into the structures of database. The big data is measured in the range of petabyte and in exabyte.

The different aspects of big-data can be characterized as,

1. Volume
2. Velocity
3. Variety.

**1. Volume**

Multiple sources are responsible for the accumulation of data with in the organization. These sources could be business transactions, social media and information from sensor or machine-to-machine data. Unlike previous technologies, with the use of new technology such as Hadoop, it can be stored easily.

**2. Velocity**

The exceptional speeds of the data streams is controlled in a timely manner. For example- RFID tags, sensors and smart metering can be controlled using torrents of data.

**3. Variety**

This ensures that big-data has all types of data such as structures, numeric data, unstructured text documents, email, video, audio and financial transactions.

Rahul Publications

**Q10. Explain the evolution of Big Data.**

*Ans :*

The evolution of big data is discussed below,

(i) 1970s and before

(ii) 1980s and 1990s

(iii) 2000s and beyond.

**(i) 1970s and Before**

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

**(ii) 1980s and 1990s**

In 1980s and 1990s the evolution of relational data bases took place. The relational data utilization is complex and thus this era comprises of data intensive applications.

**(iii) 2000s and Beyond**

The World Wide Web (WWW) and the Internet of Things (10T) have an aggression of structured, unstructured and multimedia data. The data driven is complex and unstructured.

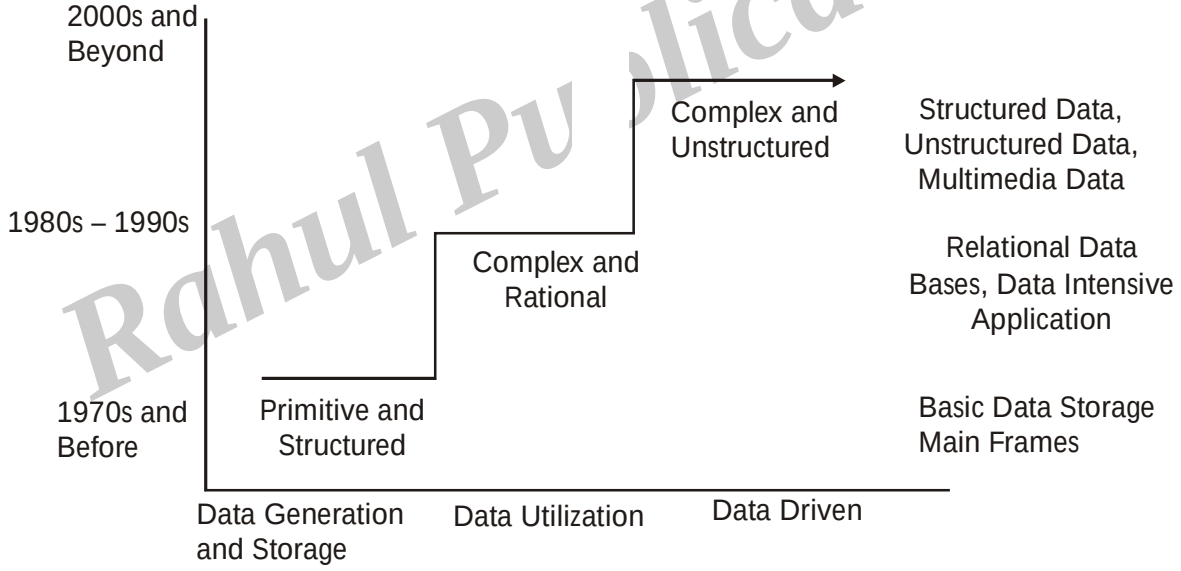



**Fig.: The Evolution of Big Data**

**Q11, What are the advantages of Big Data?**

**(OR)**

**State the advantages of Big Data.**

*Ans :*

➢ One of the biggest advantages of Big Data is Predictive Analysis. Big Data Analytics tools can predict outcomes accurately, thereby allowing businesses and organizations to make better decisions, while simultaneously optimizing their operational efficiencies and reducing risks.

- By harnessing data from social media platforms using Big Data analytics tools, businesses around the world are streamlining their digital marketing strategies to enhance the overall consumer experience. Big Data provides insights into the customer pain points and allows companies to improve upon their products and services.
- Being accurate, Big Data combines relevant data from multiple sources to produce highly actionable insights. Almost 43% of companies lack the necessary tools to filter out irrelevant data, which eventually costs them millions of dollars to hash out useful data from the bulk. Big Data tools can help reduce this, saving you both time and money.
- Big Data Analytics could help companies generate more sales leads which would naturally mean a boost in revenue. Businesses are using Big Data Analytics tools to understand how well their products/services are doing in the market and how the customers are responding to them. Thus, they can understand better where to invest their time and money.
- With Big Data insights, you can always stay a step ahead of your competitors. You can screen the market to know what kind of promotions and offers your rivals are providing, and then you can come up with better offers for your customers. Also, Big Data insights allow you to learn customer behavior to understand the customer trends and provide a highly 'personalized' experience to them.

**Q12. Explain the life cycle of Big Data.**

**(OR)**

**Describe the various stages of Big Data Life Cycle.**

*Ans :* **(Imp.)**

In today's big data context, the previous approaches are either incomplete or suboptimal. For example, the SEMMA methodology disregards completely data collection and pre-processing of different data sources. These stages normally constitute most of the work in a successful big data project.

A Big Data Analytics Cycle can be described by the following stages:

1. Business Problem Definition
2. Research
3. Human Resources Assessment
4. Data Acquisition
5. Data Mugging
6. Data Storage
7. Exploratory Data Analysis
8. Data Preparation for Modeling and Assessment
9. Modeling
10. Implementation

**1. Business Problem Definition**

This is a point common in traditional BI and big data analytics lifecycle. Normally, it is a non-trivial stage of a big data project to define the problem and evaluate correctly how much potential gain it may have for an organization. It seems obvious to mention this, but it has to be evaluated what are the expected gains and costs of the project.

**2. Research**

Analyze what other companies have done in the same situation. This involves looking for solutions that are reasonable for your company, even though it involves adapting other solutions to the resources and requirements that your company has. In this stage, a methodology for the future stages should be defined.

**3. Human Resources Assessment**

Once the problem is defined, it is reasonable to continue analyzing if the current staff is able to complete the project successfully. Traditional BI teams might not be capable to deliver an optimal solution to all the stages.

Rahul Publications

So, it should be considered before starting the project if there is a need to outsource a part of the project (or) hire more people.

**4. Data Acquisition**

This section is key in a big data life cycle; it defines which type of profiles would be needed to deliver the resultant data product. It is a non-trivial step of the process; it normally involves gathering unstructured data from different sources. To give an example, it could involve writing a crawler to retrieve reviews from a website. This involves dealing with text, perhaps in different languages normally requiring a significant amount of time to be completed.

**5. Data Mugging**

Once the data is retrieved, for example, from the web, it needs to be stored in an easy-to-use format. To continue with the reviews examples, let's assume the data is retrieved from different sites where each has a different display of the data.

Suppose one data source gives reviews in terms of rating in stars. Therefore, it is possible to read this as a mapping for the response variable $y \in \{1, 2, 3, 4, 5\}$. Another data source gives reviews using two arrows system, one for up voting and the other for down voting. This would imply a response variable of the form $y \in \{\text{positive, negative}\}$.

In order to combine both the data sources, a decision has to be made in order to make these two response representations equivalent. This can involve converting the first data source response representation to the second form, considering one star as negative and five stars as positive. This process often requires a large time allocation to be delivered with good quality.

**6. Data Storage**

Once the data is processed, it sometimes needs to be stored in a database. Big data technologies offer plenty of alternatives regarding this point. The most common alternative is using the Hadoop File System for storage that provides users a limited version of SQL, known as HIVE Query Language. This allows most analytics task to be done in similar ways as would be done in traditional BI data warehouses, from the user perspective. Other storage options to be considered are Mongo DB, Redis and SPARK.

This stage of the cycle is related to the human resources knowledge in terms of their abilities to implement different architectures. Modified versions of traditional data warehouses are still being used in large-scale applications. For example, Teradata and IBM offer SQL databases that can handle terabytes of data; open source solutions such as postgreSQL and MySQL are still being used for large-scale applications.

Even though there are differences in how the different storages work in the background, from the client side, most solutions provide a SQL API. Hence, having a good understanding of SQL is still a key skill to have for big data analytics.

This stage a priori seems to be the most important topic; in practice, this is not true. It is not even an essential stage. It is possible to implement a big data solution that would be working with real-time data. So, in this case, we only need to gather data to develop the model and then implement it in real time. So, there would not be a need to formally store the data at all.

**7. Exploratory Data Analysis**

Once the data has been cleaned and stored in a way that insights can be retrieved from it, the data exploration phase is mandatory. The objective of this stage is to understand the data. This is normally done with statistical techniques and also plotting the data. This is a good stage to evaluate whether the problem definition makes sense or is feasible.

**8. Data Preparation for Modeling and Assessment**

This stage involves reshaping the cleaned data retrieved previously and using statistical pre-processing for missing values imputation,

Rahul Publications

outlier detection, normalization, feature extraction and feature selection.

**9. Modelling**

The prior stage should have produced several data sets for training and testing, e.g., a predictive model. This stage involves trying different models and looking forward to solving the business problem at hand. In practice, it is normally desired that the model would give some insight into the business. Finally, the best model or combination of models is selected evaluating its performance on a left-out data set.

**10. Implementation**

In this stage, the data product developed is implemented in the data pipeline of the company. This involves setting up a validation scheme while the data product is working in order to track its performance. For example, in case of implementing a predictive model, this stage would involve applying the model to new data and once the response is available, evaluate the model.

**Q13. Explain the uses of Big Data.**

*Ans :*

The people who are using Big Data know better that, what Big Data is. Let's look at some such industries:

**(i) Healthcare**

Big Data has already started to create a huge difference in the healthcare sector. With the help of predictive analytics, medical professionals and HCPs are now able to provide personalized healthcare services to individual patients. Apart from that, fitness wearables, telemedicine, remote monitoring all powered by Big Data and AI – are helping change lives for the better.

**(ii) Academia**

Big Data is also helping enhance education today. Education is no more limited to the physical bounds of the classroom there are numerous online educational courses to learn from. Academic institutions are investing in digital courses powered by Big Data technologies to aid the all-round development of budding learners.

**(iii) Banking**

The banking sector relies on Big Data for fraud detection. Big Data tools can efficiently detect fraudulent acts in real-time such as misuse of credit/debit cards, archival of inspection tracks, faulty alteration in customer stats, etc.

**(iv) Manufacturing**

According to TCS 2013 Global Trend Study, the most significant benefit of Big Data in manufacturing is improving the supply strategies and product quality. In the manufacturing sector, Big Data helps create a transparent infrastructure, thereby predicting uncertainties and incompetencies that can affect the business adversely.

**(v) IT**

One of the largest users of Big Data, IT companies around the world are using Big Data to optimize their functioning, enhance employee productivity and minimize risks in business operations. By combining Big Data technologies with ML and AI, the IT sector is continually powering innovation to find solutions even for the most complex of problems.

**Q14. What are the technologies available to manage big data?**

*Ans :* **(Imp.)**

The most popular technologies for collecting, storing, processing and analyzing Big-data are as follows,

1. MapReduce framework, including Hadoop distributed file system (HDFS)
2. NOSQL (Not only SQL) data stores
3. MPP (Massively Parallel Processing) databases
4. In-memory database processing.

Rahul Publications

**1. MapReduce Framework, Including Hadoop Distributed File System (HDFS)**

The Hadoop project includes both distributed file system called HDFS (Hadoop Distributed File System) modeled on Google File Systems (GFS) and distributed processing framework by implementing MapReduce concepts.

Inorder to acquire a broader understanding of the technology in Hadoop, it is necessary to accentuate MapReduce programming model. The name of the model is taken from map and reduces functions. It carries out parallel processing on large sets of data using huge number of computer nodes connected through network interconnect in the form of cluster. It has four characteristics such as parallelism, fault tolerance, scalability and data locality.

**2. NOSQL (Not only SQL) Data Stores**

Unlike traditional database systems, big data requirements corresponding to data storage and retrieval, data are different. An unstructured large data usually demands high-speed insertion, deletion and retrieval, where as if the data is structured it is given minimum importance. Thus, it is necessary to focus upon the capabilities to store and retrieve large amount of data.

As a consequence, need for highly efficient storage and retrieval has entailed on additional class of database systems called NOSQL (Not only SQL). The NOSQL make use of various approaches for managing the unstructured data or non-relational data. Such systems are also referred to as key-value stores which focuses upon large scaling "on-demand" data model flexibility and easy application development and deployment.

Some of the common attributes of NOSQL database system implementations are distributed and fault tolerant architecture, schema-free design and ACID properties like atomicity, consistency, isolation and durability.

**3. MPP (Massively Parallel Processing) Databases**

The MPP databases system is an essential elements of the Big data technology. The standard size of the MPP database system is from few terabytes to few petabytes of structured data.

According to the research made, three main infrastructure models of databases can be deduced. They are, shared everything, shared disk and shared nothing. We observe that high volumes of small transactional queries such as on-line transaction processing (OLTP) system can be architectured in general propose databases. The loads in OLTP demands fast access quick updates to small set of data records. The processing of the quarry is done in local sector of disks using less number of parallel processes. The architecture of OLTP database system can be-shared everything or shared disk architectures. In former architecture a single large server a symmetric multi processors (SMP) along with many processors memory is utilized. Apart from this it also incorporates disk scalability and performance limitation. On the other hand, the latter one is a system which contains one or many compute nodes linked to storage area network or either to shared storage. Interestingly, the shared disk architecture can be applied on OLTP as every single node takes the subset of query and process them alone. During the process, the consistency of the system is not compromised.

**4. In-memory Database Processing**

The in-memory database offers many trends and technologies. Here compute and memory architectures changes consistently. Apart from this, multi-core architectures and main-memory also changes.

**Q15. Explain the various sources of Big Data.**

*Ans :*

Data sources are broadly categorized into three types. They are as follows,

Rahul Publications

**1. People-to-People Communications**

At present electronic networks are widely used by people and corporations. With the development of multimedia, communication has become easier and effective. Majority of the people are communicating with mobile phones, e-mails and internet websites as they eradicate the time and distance. For example, people are using high resolution Cameras in mobile phones to capture photos and videos and it is also easy to share them instantly with others. Telecom and internet service providers are the intermediaries who are storing all communications. In the present scenario, social media is the important means of human-human communications.

**(i) Social Media :** Social media has become a part of current modem world. There are different types of social media like Facebook. LinkedIn, Flicker, Youtube, Twitter, Skype, Snap Chat, WhatsApp etc. These media is used for different purposes. Facebook and Youtube is used to share videos, messages and pictures instantly. Flicker is used to share photo albums. Twitter is used to communicate short asynchronous messages, Skype and WhatsApp is used to make a audio or video call, text messages, etc.

All these streams are part of big data. They helps to understand patterns of communication and importance of different types of conversations.

**2. People-to-Machine Communications**

Sensors and web are two types of machines where people use them for communication. Siri and Cortana are the new man-machine communications which were considered as personal assistants. They understand the human requests in natural language and also try to fulfill them. Addition to this, smart devices like fitbit and smart watch deals with people personal data like weight, food and exercise data, blood pressure, sleep hours etc. They read, store and analyze the people's personal data in order to answer their queries. However, these medias are useful for different purposes and have stunning effects. People-to-machine communication can be simply done with the help of web or by accessing World Wide Web (www).

**(i) Web Access:** The World Wide Web (www) has become a part of all human and machine activities. The web users are increasing every time and even the requests for web pages are in billions in number. For accessing a web page a request is to be generated. The web page provider tracks the identity of the requesting device and user. This information is in small pieces of computer code and data which are known as cookies. It normally tracks the webpage received, date/time of access to analyze and acquire new opportunities for marketing.

An application which is used to monitor streaming web access logs is a web log analyzer. It examines the website performance and minimize the errors.

**3. Machine-to-Machine (M2M) Communications**

Machine-to-Machine (M2M) communications is also called as Internet of Things (IOT). In order to communicate with each other or with some master machines trillion devices are connected to the internet. The owners of those machines will access and control the data.

---

**Q16. What are the major types of Big data applications?**

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

**1. Monitoring and Tracking Applications**

Monitoring and tracking applications are the first and basic applications which helps for the improvement in efficiency of the business of various industries. Following are the few specific applications of monitoring and tracking applications,

**(i) Public Health Monitoring:** It is performed by the health care centres by

Rahul Publications

maintaining the interoperability and data standards. The advance big data analytics will enable secondary use of health data.

(ii) **Consumer Sentiment Monitoring:** Majority of the consumer goods companies are using social media instead of traditional media for advertisement of their products. Social media streams includes, Facebook posts, Twitter tweets and blog posts are filtered and analyzed in accordance with certain keywords, demographics and regions. The information regarding such analysis is forwarded to the marketing professionals at the moment when the product is new to the market.

(iii) **Asset Tracking:** Asset tracking is done by a tiny RFID chip which tracks the performance of the asset or product. It enables to track with the help of RFID tags by arranging RF readers. These sensors are majorly used by Airplanes where it tracks every part of the plane.

(iv) **Supply Chain Monitoring:** Retailers and suppliers tracks the status and location of all containers on ships around the globe using RFID tags.

**2. Analysis and Insight Applications**

Analysis and insight applications are considered as the next level of big data applications. These are structured and analyzed to provide insights and patterns which makes a better business. The various applications of analysis and insight includes.

(i) **Predictive Policing :** The concept of predictive policing is invented by the Los Angeles Policy Department (LAPD). The LAPD combinely analyzed with UC Berkely researchers about the data base of 13 million crimes that are recorded over 80 years. It identifies the hotspots of crime and predicts the future crime.

(ii) **Winning Political Elections:** The big data had played a vital role in the 2008 elections of United States. The US president Barack Obama is the first person to utilize big data in an effective way. The data gather about million s of people including supporters helped him in building an effective nation.

**3. New Product Development**

These are considered as completely new concepts of applications and does not exists before. It comprises of transformative potential for generating new modes of revenue for business. Following are some of the applications related to them,

(i) **Flexible Auto Insurance:** In order to calculate the risk of accidents on the basis of travel patterns, an auto insurance company can use the GPS data from cars. This GPS can track the performance of a car with the help of car sensor data.

(ii) **Location - Based Promotion:** Based on the location data obtained through Global Positioning System (GPS), the retailers and third party advertisers targets the customers with various promotions and coupons. They are delivered through mobile apps, sms and email.

**Q17. Explain the role of big data in competing food apps Swiggy and Zomato.**

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

As the online food ordering trend is becoming more and more prominent in India, the food delivery platforms like Swiggy and Zomato are growing their user base at an exponential rate.

**1. Swiggy**

According to a report, the number of user interactions on Swiggy has grown

Rahul Publications

exponentially from 2 billion in October 2017 to a massive 40 billion in January 2019. To keep up with the massive growth the company looks up to Artificial intelligence as a solution to many of the problems. Head of the Engineering and Data Science Team at Swiggy, Dale Vaz says "AI is critical for us to sustain our growth,".

Artificial intelligence helps Swiggy distinguish the dishes from images classifying them as vegan or non-vegan dishes. Natural Language Processing can greatly help the platform in serving wider geography without having to consider linguistic boundaries that enables search using colloquial terms which customers could use to obtain accurate results.

**2. Zomato**

As seen by Swiggy as its arch-rival in the food delivery market, Zomato doesn't seem to be backing down too. Recently the company had raised Rs 284 crore from a US investor Glade Brook Capital Partners as part of its strategy to acquire more market share from its rivals. Last month Zomato made a claim of achieving a 28 million monthly order run rate as of December compared to the 21 million in October which also helps the company in projecting future order volume. The platforms Gold Subscription package also claims to have worked out for the company, bringing on board 7 lakh members and over 6,000 restaurant partners, up from 6 lakh members and 4,000 restaurants.

It was in the late December that Zomato acquired Lucknow-based startup TechEagle Innovations looking forward to establishing a drone-based delivery network in India.

Zomato's Founder and CEO Deepinder Goyal said in a press release -"We believe that robots powering the last-mile delivery is an inevitable part of the future and hence is going to be a significant area of investment for us."

Rahul Publications

# Short Question and Answers

**1. Define business analytics.**

*Ans :*

**Meaning**

Business analytics is the practice of iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by companies committed to data-driven decision-making.

**Definitions**

**(i) According to Schaer (2018),** "allows your business to make predictive analysis rather than reacting to changes in data".

**(ii) According to Gabelli School of Business (2018),** "involves applying models, methods, and tools to data, producing insights that lead to informed business decisions"

**(iii) According to Wells(2008),** "the application of logic and mental processes to find meaning in data"

**(iv) According to Lynda (2018),** "allows us to learn from the past and make better predictions for the future".

**2. Descriptive Analytics.**

*Ans :*

This can be termed as the simplest form of analytics. The mighty size of big data is beyond human comprehension and the first stage, hence involves crunching the data into understandable chunks. The purpose of this analytics type is just to summarize the findings and understand what is going on.

Among some frequently used terms, what people call as advanced analytics or business intelligence is basically usage of descriptive statistics (arithmetic operations, mean, median, max, percentage, etc.) on existing data. It is said that 80% of business analytics mainly involves descriptions based on aggregations of past performance. It is an important step to make raw data understandable to investors, shareholders and managers. This way, it gets easy to identify and address the areas of strengths and weaknesses such that it can help in strategizing.

**3. Diagnostic Analytics.**

*Ans :*

Diagnostic Analytics is used to determine why something happened in the past. It is characterized by techniques such as drill-down, data discovery, data mining and correlations. Diagnostic analytics takes a deeper look at data to understand the root causes of the events. It is helpful in determining what factors and events contributed to the outcome. It mostly uses probabilities, likelihoods and the distribution of outcomes for the analysis.

In a time series data of sales, diagnostic analytics would help you understand why the sales have decreased or increased for a specific year or so. However, this type of analytics has a limited ability to give actionable insights. It just provides an understanding of causal relationships and sequences while looking backward.

A few techniques that uses diagnostic analytics include attribute importance, principle components analysis, sensitivity analysis and conjoint analysis. Training algorithms for classification and regression also fall in this type of analytics.

**4. Predictive Analytics.**

*Ans :*

Predictive Analytics is used to predict future outcomes. However, it is important to note that it cannot predict if an event will occur in the future; it merely forecasts what the probabilities of the occurrence of the event are. A predictive model builds on the preliminary descriptive analytics stage to derive the possibility of the outcomes.

The essence of predictive analytics is to devise models such that the existing data is understood to extrapolate the future occurrence or simply, predict the future data. One of the common

Rahul Publications

applications of predictive analytics is found in sentiment analysis where all the opinions posted on social media are collected and analyzed (existing text data) to predict the person's sentiment on a particular subject as being positive, negative or neutral (future prediction).

**5. Prescriptive Analytics**

*Ans :*

The basis of this analytics is predictive analytics, but it goes beyond the three mentioned above to suggest the future solutions. It can suggest all favorable outcomes according to a specified course of action and also suggest various course of actions to get to a particular outcome. Hence, it uses a strong feedback system that constantly learns and updates the relationship between the action and the outcome.

The computations include optimization of some functions that are related to the desired outcome. For example, while calling for a cab online, the application uses GPS to connect you to the correct driver from among a number of drivers found nearby. Hence, it optimizes the distance for faster arrival time. Recommendation engines also use prescriptive analytics.

**6. What is Big Data?**

*Ans :*

**Meaning**

The term Big-data can be used to describe huge volumes of data both in structured and unstructured format taken from webtraffic, e-mail messages and content of social media such as tweets and status messages. Interestingly, the big data moves beyond the processing ability to capture, store and analyse typical database systems. As a consequence a large quantity of data is generated which moves fast and fail to fit into the structures of database. The big data is measured in the range of petabyte and in exabyte.

The different aspects of big-data can be characterized as,

1. Volume
2. Velocity
3. Variety.

**7. State the advantages of Big Data.**

*Ans :*

- One of the biggest advantages of Big Data is Predictive Analysis. Big Data Analytics tools can predict outcomes accurately, thereby allowing businesses and organizations to make better decisions, while simultaneously optimizing their operational efficiencies and reducing risks.
- By harnessing data from social media platforms using Big Data analytics tools, businesses around the world are streamlining their digital marketing strategies to enhance the overall consumer experience. Big Data provides insights into the customer pain points and allows companies to improve upon their products and services.
- Being accurate, Big Data combines relevant data from multiple sources to produce highly actionable insights. Almost 43% of companies lack the necessary tools to filter out irrelevant data, which eventually costs them millions of dollars to hash out useful data from the bulk. Big Data tools can help reduce this, saving you both time and money.
- Big Data Analytics could help companies generate more sales leads which would naturally mean a boost in revenue. Businesses are using Big Data Analytics tools to understand how well their products/ services are doing in the market and how the customers are responding to them. Thus, they can understand better where to invest their time and money.

Rahul Publications

**8. Explain different types of data.**

*Ans :*

Following are the types of data,

**1. Structured Data**

Structured data is the data that can be stored in an organized format in rows and columns, so that it can be effectively retrieved and processed by an application. Examples: Data Base Management System (DB ) is used for saving the structured data.

**2. Semi-structured Data**

Semi-structured data is the data which does not accommodates to a data model but has a structure. Semi-structured data is not used easily by a computer program.

**Examples:** Emails, XML, Markup languages like HTML etc.

**3. Unstructured Data**

Unstructured data is the data that cannot be stored in systematic format in rows and columns, due to which it becomes complicated for an application to query and retrieve the data.

**9. What are the differences between descriptive analytics?**

*Ans :*

| S.No | Nature | Descriptive Analysis | Predictive Analysis |
|---|---|---|---|
| 1. | **Data** | It uses historical data and reconfigures it into simple, easy and readable formats. | it also uses historical data but only gaps will be filled in the available data. |
| 2. | **Highlights** | It highlights and describe the nature of business operations. | It highlights and creates data models. |
| 3. | **Techniques** | It involves techniques like data aggregation and data mining. | It involves statistics and other forecasting techniques. |
| 4. | **Results** | It provide accurate result in reports by | It provide inaccurate result in reports. |

**10. What is Web Analytics?**

*Ans :*

The analysis of internet activity including visits of users to social media sites like LinkedIn 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 to determine statistically the reasons for differences in the sales and website traffic.

Rahul Publications

# Choose the Correct Answers

1. Facebook Tackles Big Data with _____ based on Hadoop. [ a ]
   (a) Project Prism (b) Prism
   (c) Project Data (d) Project Bid
2. All of the following accurately describe Hadoop, EXCEPT: [ b ]
   (a) Open Source (b) Real-time
   (c) Java-based (d) Distributed computing approach
3. What are the main components _____ of Big Data? [ d ]
   (a) Map Reduce (b) HDFS
   (c) YARN (d) All of these
4. _____ has the world's largest Hadoop cluster. [ c ]
   (a) Apple (b) Datamatics
   (c) Facebook (d) None of the mentioned
5. According to analysts, for what can traditional IT systems provide a foundation when they are integrated with Big Data technologies like Hadoop? [ a ]
   (a) Big Data management and data mining
   (b) Data warehousing and business intelligence
   (c) Management of Hadoop clusters
   (d) Collecting and storing unstructured data
6. What are the five V's of Big Data? [ d ]
   (a) Volume (b) Velocity
   (c) Variety (d) All the above
7. What are the different features of Big Data Analytics? [ d ]
   (a) Open Source (b) Scalability
   (c) Data Recovery (d) All the above
8. _____ Data refers to the data that lacks any specific form [ b ]
   (a) Structured data (b) Unstructured data
   (c) Both (d) None of the above
9. _____ is the last stage is Big data life cycle. [ a ]
   (a) Implementation (b) Datastorage
   (c) Data Mugging (d) Research
10. _____ analyze what other companies have done in the same situations. [ d ]
   (a) Implementation (b) Datastorage
   (c) Data Mugging (d) Research

Rahul Publications

# Fill in the Blanks

1. _____ refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.
2. _____ tools range from spreadsheets with statistical functions to complex data mining and predictive modeling applications.
3. _____ analytics it is really valuable, but largely not used.
4. An _____ model is simply a mathematical equation that describes relationships among variables in a historical data set.
5. _____ analytics can be useful in the sales cycle,
6. _____ (or feature) roadmaps outline the development and launches of a product and its features.
7. _____ outline the hierarchy of a business or one of its departments or teams.
8. PFD stands for _____.
9. _____ can be measured by quality of transactions, events and amount of history.
10. _____ models Improve Operational Efficiencies Increase Revenues Achieve Competitive Differentiation Reduce risks and costs

**ANSWERS**

1. Business analytics (BA)
2. Analytics
3. Prescriptive
4. Analytical
5. Descriptive
6. Product
7. Organizational charts
8. Process flow Diagram
9 Data Volume
10. Big data

Rahul Publications

UNIT II

**DESCRIPTIVE ANALYTICS 1 :**

Types of Data-Population and Sample Data, Quantitative and Categorical Data, Cross-Sectional and Time Series Data, Sources of data, Descriptive Statistics- Measures of Location (central Tendency)-Mean, Median and Mode and relationship between them – Problems.

## 2.1 Types of Data

**Q1. Write briefly about types of data in statistics.**

*Ans ;*

There are different types of data in Statistics, that are collected, analysed, interpreted and presented. The data are the individual pieces of factual information recorded, and it is used for the purpose of the analysis process. The two processes of data analysis are interpretation and presentation. Statistics are the result of data analysis.

The data is classified into following categories:

**1. Population**

It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a parameter. For example, All people living in India indicates the population of India.

There are different types of population. They are:

- Finite Population
- Infinite Population
- Existent Population
- Hypothetical Population

**Example:**

The population may be "ALL people living in the US."

**2. Sample**

It includes one or more observations that are drawn from the population and the measurable characteristic of a sample is a statistic. Sampling is the process of selecting the sample from the population. For example, some people living in India is the sample of the population.

Basically, there are two types of sampling. They are:

- Probability sampling
- Non-probability sampling

**Example:**

The sample may be "SOME people living in the US."

**3. Quantitative Data**

Quantitative: Has numerical values for which arithmetic operations (e.g., addition or averaging) make sense.

**Examples:**

age, height, # of AP classes, SAT score.

**Categorical Data**

Places an individual into one of several groups or categories.

**Examples:**

eye color, race, gender. May have numerical values assigned: 1 = White, 2 = Hispanic, 3 = Asian, etc. Other numeric categorical variables include baseball jersey number or zip code.

**5. Cross-sectional Data**

Cross-section data is collected in a single time period and is characterized by individual units - people, companies, countries, etc. Some examples include:

Rahul Publications

- Student grades at the end of the current semester;
- Household data of the previous year - expenditure on food, unemployment, income, etc.
- Car data - average speed, horsepower, color, etc.

**6. Time Series Data**

Data collected at a number of specific points in time is called time series data. Such examples include stock prices, interest rates, exchange rates as well as product prices, GDP, etc. Time series data can be observed at many different frequencies (hourly, daily, weekly, monthly, quarterly, anually, etc.)

### 2.1.1 Population and Sample Data

**Q2. Explain about population and sample data.**

*Ans :* **(Imp.)**

**I) Population**

It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a parameter. For example, All people living in India indicates the population of India.

There are different types of population. They are:

i) Finite Population

ii) Infinite Population

iii) Existent Population

iv) Hypothetical Population

**(i) Finite Population**

The finite population is also known as a countable population in which the population can be counted. In other words, it is defined as the population of all the individuals or objects that are finite. For statistical analysis, the finite population is more advantageous than the infinite population. Examples of finite populations are employees of a company, potential consumer in a market.

**(ii) Infinite Population**

The infinite population is also known as an uncountable population in which the counting of units in the population is not possible. Example of an infinite population is the number of germs in the patient's body is uncountable.

**(iii) Existent Population**

The existing population is defined as the population of concrete individuals. In other words, the population whose unit is available in solid form is known as existent population. Examples are books, students etc.

**(iv) Hypothetical Population**

The population in which whose unit is not available in solid form is known as the hypothetical population. A population consists of sets of observations, objects etc that are all something in common. In some situations, the populations are only hypothetical. Examples are an outcome of rolling the dice, the outcome of tossing a coin.

**II) Sample**

It includes one or more observations that are drawn from the population and the measurable characteristic of a sample is a statistic. Sampling is the process of selecting the sample from the population. For example, some people living in India is the sample of the population.

Basically, there are two types of sampling. They are:

1. Probability sampling
2. Non-probability sampling

**1. Probability Sampling**

In probability sampling, the population units cannot be selected at the discretion of the researcher. This can be dealt with following certain procedures which will ensure that every unit of the population consists of one fixed probability being included in the sample. Such a method is also called random sampling. Some of the techniques used for probability sampling are:

- Simple random sampling
- Cluster sampling

Rahul Publications

- Stratified Sampling
- Disproportionate sampling
- Proportionate sampling
- Optimum allocation stratified sampling
- Multi-stage sampling

**2. Non Probability Sampling**

In non-probability sampling, the population units can be selected at the discretion of the researcher. Those samples will use the human judgements for selecting units and has no theoretical basis for estimating the characteristics of the population. Some of the techniques used for non-probability sampling are

- Quota sampling
- Judgement sampling
- Purposive sampling

**Population and Sample Examples**

- All the people who have the ID proofs is the population and a group of people who only have voter id with them is the sample.
- All the students in the class are population whereas the top 10 students in the class are the sample.
- All the members of the parliament is population and the female candidates present there is the sample.

**Population and Sample Formulas**

We will demonstrate here the formulas for mean absolute deviation (MAD), variance and standard deviation based on population and given sample. Suppose n denotes the size of the population and n-1 denotes the sample size, then the formulas for mean absolute deviation, variance and standard deviation are given by;

$$\text{Population MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\text{Population MAD} = \frac{1}{n-1}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\text{Population Variance} = (\sigma x)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Sample Variance} = (Sx)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Population Standard Deviation} = \sigma x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{Sample Standard Deviation} = Sx = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Rahul Publications

**Q3. Explain how to calculate Mean Absolute Deviation of Population.**

*Ans :*

**Mean Absolute Deviation**

The Mean Absolute Deviation (MAD) of a set of datais the average distance between each data value and the mean.

The steps to find the MAD include:

1. find the mean (average)
2. find the difference between each data value and the mean
3. take the absolute value of each difference
4. find the mean (average) of these differences

We will look at two different approaches to find the population MAD.

$$\text{Population MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Data Set: {92, 83, 88, 94, 91, 85, 89, 90}

**Find the Mean Absolute Deviation.**

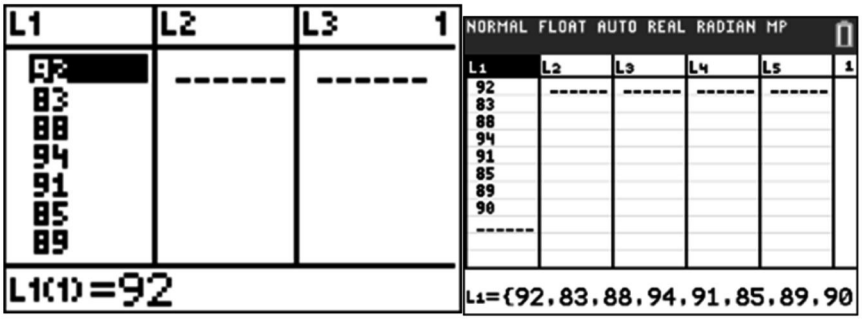
**Method 1:**

(doing one step at a time with the lists = easy)

Enter the data:

**Step 1:**

Enter the data into a list (L1)




**Calculate the differences:**

**Step 2:**

Tell L2 to calculate the values from L1 minus the mean of L1.

Arrow up onto L2.

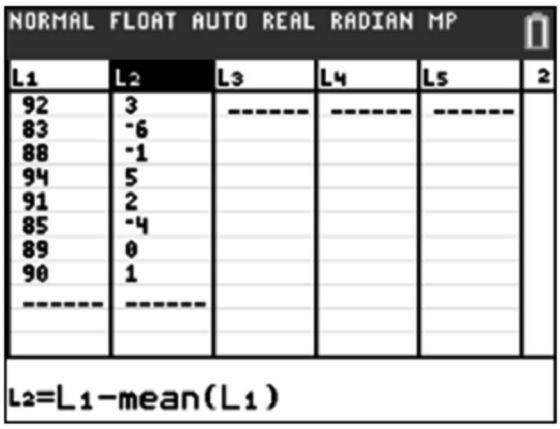
Enter L1 - mean(L1) (you can find "mean" in the Catalog)

[mean( is also found in 2nd LIST, MATH, #3 mean( ]

Hit ENTER.

Notice that if you add the sum of the positive differences and the negative differences, you will get zero.

Rahul Publications




Take the absolute value of the differences:

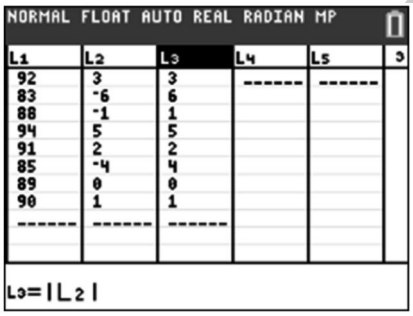
**Step 3:**

Arrow up onto L3.

Enter abs(L2) (you can find "abs" in the Catalog)

Hit ENTER.

(This step could have been combined with Step 2.)




Find the mean (average) of the absolute values:

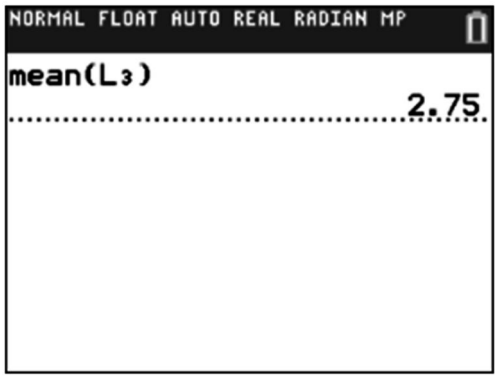
**Step 3:**

Go to the HOME screen. (2nd QUIT)

Enter mean(L3).

Hit ENTER.

This value is the Mean Absolute Deviation (MAD).




Rahul Publications

Since the MAD is small, it implies that the mean of 89 is indicative of the other values within the data set.

**Method 2:**

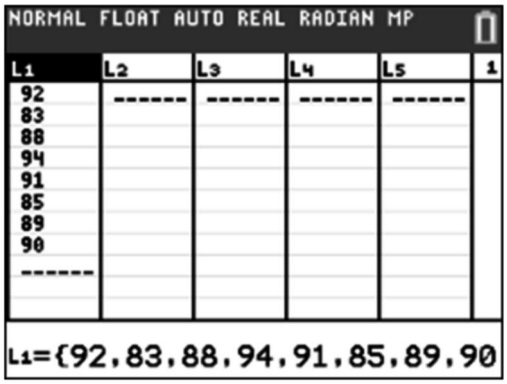
(doing most work in one statement - harder to remember the locations of the functions)

Enter the data:

**Step 1:**

Enter the data into a list (L1)

(See Basic Commands for entering data.)




Have the calculator compute the mean and the count of entries.

**Step 2:**

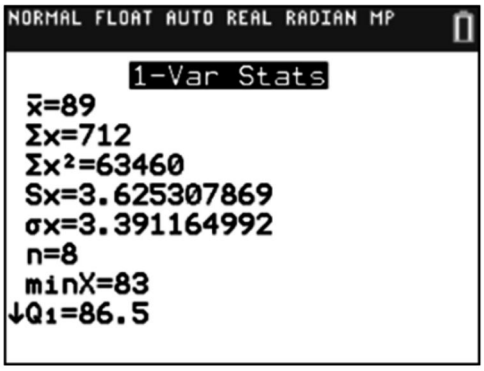
Hit STAT '! CALC

Choose #1 1-Var Stats

On the home screen: #1 1-Var Stats L1

Hit ENTER.

The calculator now has stored the value for the mean, , and it has the count of the number of entries in the list, n.

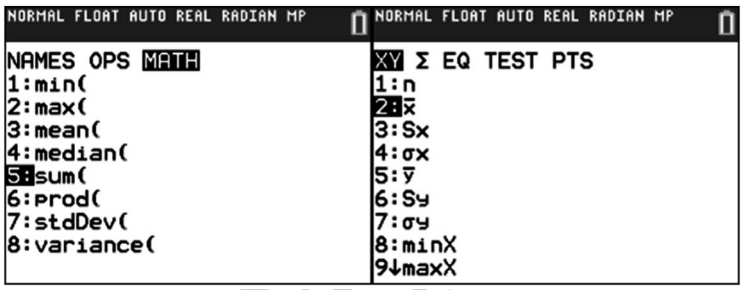



Compute the answer using one formula entry:

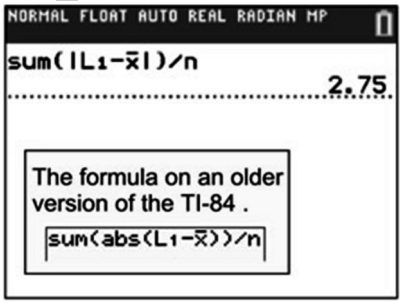
**Step 3:**

- Go to the HOME screen.
- left parenthesis
- 2nd LIST '! MATH #5 sum(
- from Catalog, abs(
- Enter L1 (2nd and the key for "1")
- Subtraction symbol
- VARS, #5 Statistics, #2
- Division symbol
- VARS, #5 Statistics, #1 n

  Hit ENTER.




2nd LIST '! MATH #5 sum(VARS, #5 Statistics, #2




Hit ENTER

**Note:**

When adding the entries in a list, be sure to use the sum( function which appears under LIST - MATH. Do not use the sigma notation " from ALPHA-WINDOW, as it does not understand how to access numbers from a list.

**Q4. What is population variance? Explain how to calculate population variance with examples.**

*Ans :*

Population variance is a measure of the spread of population data. Hence, population variance can be defined as the average of the distances from each data point in a particular population to the mean squared, and it indicates how data points are spread out in the population. Population variance is an

important measure of dispersion used in statistics. Statisticians calculate variance to determine how individual numbers in a data set relate to each other.

Population variance can be calculated by using the following formula:

$$\sigma_2 = \frac{\Sigma_{i=1}^{n}(x_i - \mu)^2}{N}$$

where

- $\sigma^2$ is population variance,
- $x_1, x_2, x_3, \ldots x_n$ are the observations
- N is the number of observations,
- $\mu$ is the mean of the data set

**Step by Step Calculation of Population Variance**

The formula for population variance can be calculated by using the following five simple steps:

1. **Step 1:** Calculate the mean ($\mu$) of the given data. In order to calculate the mean, add all the observations and then divide that by the number of observations (N).
2. **Make a table:** Please note that constructing a table is not compulsory, but presenting it in a tabular format would make the calculations easier. In the first column, write each observation ($x_1, x_2, x_3, \ldots x_n$).
3. **In the second column:** Write the deviation of each observation from the mean ($x_i - \mu$).
4. **In the third column:** Write the square of each observation from the mean $(x_i - \mu)^2$. In other words, square each of the numbers obtained in column 2.
5. **Subsequently:** We need to add the numbers obtained in the third column. Find the sum of the squared deviations and divide the sum so obtained by the number of observations (N). This will help us to obtain which is the population variance.

**Examples**

**Example #1:**

Calculate the population variance from the following 5 observations: 50, 55, 45, 60, 40.

**Solution:**

Use the following data for the calculation of population variance.

| | A | B | C |
|---|---|---|---|
| 1 | No.of Items (N) | Observations(x) | |
| 2 | 1 | 50 | |
| 3 | 2 | 55 | |
| 4 | 3 | 45 | |
| 5 | 4 | 60 | |
| 6 | 5 | 40 | |
| 7 | | | |

There are a total of 5 observations. Hence, N=5.

$\mu = (50+55+45+60+40)/5 = 250/5 = 50$

So, the Calculation of population variance ó² can be done as follows-

$$\sigma^2 = \frac{\Sigma_{i=1}^{n}(x_i - \mu)^2}{N}$$

=(E2+E3+E4+E5+E6)/COUNT(A2:A6)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | No.of Items (N) | Observations(x) | μ | x- μ | (x- μ)² |
| 2 | 1 | 50 | | 0 | 0 |
| 3 | 2 | 55 | | 5 | 25 |
| 4 | 3 | 45 | 50 | -5 | 25 |
| 5 | 4 | 60 | | 10 | 100 |
| 6 | 5 | 40 | | -10 | 100 |
| 7 | | | | | |
| 8 | σ² | =(E2+E3+E4+E5+E6) /COUNT(A2:A6) | | | |
| 9 | | | | | |
| 10 | | | | | |

Using Population Variance Formula

$\sigma^2$ = 250/5

Population Variance $\sigma^2$ will be-

B8 =(E2+E3+E4+E5+E6)/COUNT(A2:A6)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | No.of Items (N) | Observations(x) | μ | x- μ | (x- μ)² | |
| 2 | 1 | 50 | | 0 | 0 | |
| 3 | 2 | 55 | | 5 | 25 | |
| 4 | 3 | 45 | 50 | -5 | 25 | |
| 5 | 4 | 60 | | 10 | 100 | |
| 6 | 5 | 40 | | -10 | 100 | |
| 7 | | | | | | |
| 8 | σ² | 50 | | | | |
| 9 | | | | | | |

**Population Variance ($\sigma^2$) = 50**

The population variance is 50.

**Example #2**

XYZ Ltd. is a small firm and consists of only 6 employees. The CEO believes that there should not be high dispersion in the salaries of these employees. For this purpose, he wants you to calculate the variance of these salaries. The salaries of these employees are as under. Calculate the population variance of the salaries for the CEO.

**Solution:**

Use the following data for the calculation of population variance.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | No.of Items (N) | Employee Name | Salary (in $) | |
| 2 | 1 | Chris Sawyer | 30 | |
| 3 | 2 | Tom Hawkins | 27 | |
| 4 | 3 | Bella Smith | 20 | |
| 5 | 4 | George Obama | 40 | |
| 6 | 5 | Marcus Smith | 32 | |
| 7 | 6 | Nancy Lewis | 31 | |
| 8 | | | | |

There are a total of 6 observations. Hence, N = 6.

$= (30+27+20+40+32+31)/6 \;\; = 180/6 \;\; = \$\ 30$

So, the Calculation of population variance ó² can be done as follows-

$$\sigma^2 = \frac{\Sigma_{i=1}^{n}(x_i - \mu)^2}{N}$$

=(F2+F3+F4+F5+F6+F7)/COUNT(A2:A7)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | No.of Items (N) | Employee Name | Salary (in $) | μ | x- μ | (x- μ)² |
| 2 | 1 | Chris Sawyer | 30 | 30 | 0 | 0 |
| 3 | 2 | Tom Hawkins | 27 | | -3 | 9 |
| 4 | 3 | Bella Smith | 20 | | -10 | 100 |
| 5 | 4 | George Obama | 40 | | 10 | 100 |
| 6 | 5 | Marcus Smith | 32 | | 2 | 4 |
| 7 | 6 | Nancy Lewis | 31 | | 1 | 1 |
| 8 | | | | | | |
| 9 | $\sigma^2$ | =(F2+F3+F4+F5+F6+F7)/COUNT(A2:A7) | | | | |
| 10 | | | | | | |

Using Population Variance Formula

$\sigma^2 = 214/6$

Population Variance $\sigma^2$ will be-

B9 =(F2+F3+F4+F5+F6+F7)/COUNT(A2:A7)

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | No.of Items (N) | Employee Name | Salary (in $) | μ | x- μ | (x- μ)² | |
| 2 | 1 | Chris Sawyer | 30 | 30 | 0 | 0 | |
| 3 | 2 | Tom Hawkins | 27 | | -3 | 9 | |
| 4 | 3 | Bella Smith | 20 | | -10 | 100 | |
| 5 | 4 | George Obama | 40 | | 10 | 100 | |
| 6 | 5 | Marcus Smith | 32 | | 2 | 4 | |
| 7 | 6 | Nancy Lewis | 31 | | 1 | 1 | |
| 8 | | | | | | | |
| 9 | $\sigma^2$ | 35.67 | | | | | |
| 10 | | | | | | | |

**Population Variance ($\sigma^2$) = 35.67**

The population variance of the salaries is 35.67.

**Q5. How to calculate variance in Excel ? Explain.**

*Ans :* **(Imp.)**

A sample is a set of data extracted from the entire population. And the variance calculated from a sample is called sample variance.

For example, if you want to know how people's heights vary, it would be technically unfeasible for you to measure every person on the earth. The solution is to take a sample of the population, say 1,000 people, and estimate the heights of the whole population based on that sample.

Sample variance is calculated with this formula:

$$\text{Sample variance} = \frac{\Sigma(x-\bar{x})^2}{(n-1)}$$

Where,

- $\bar{x}$ is the mean (simple average) of the sample values.
- n is the sample size, i.e. the number of values in the sample.

There are 3 functions to find sample variance in Excel: VAR, VAR.S and VARA.

VAR function in Excel

It is the oldest Excel function to estimate variance based on a sample. The VAR function is available in all versions of Excel 2000 to 2019.

**VAR(number1, [number2], ...)**

**Note:**

In Excel 2010, the VAR function was replaced with VAR.S that provides improved accuracy. Although VAR is still available for backward compatibility, it is recommended to use VAR.S in the current versions of Excel.

VAR.S function in Excel

It is the modern counterpart of the Excel VAR function. Use the VAR.S function to find sample variance in Excel 2010 and later.

**VAR.S(number1, [number2], ...)**

VARA function in Excel

The Excel VARA function returns a sample variance based on a set of numbers, text, and logical values as shown in this table.

**VARA(value1, [value2], ...)**

Sample variance formula in Excel

When working with a numeric set of data you can use any of the above functions to calculate sample variance in Excel.

As an example, let's find the variance of a sample consisting of 6 items (B2:B7). For this, you can use one of the below formulas:

= VAR(B2:B7)

= VAR.S(B2:B7)

= VARA(B2:B7)

As shown in the screenshot, all the formulas return the same result (rounded to 2 decimal places):

Rahul Publications

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Student | Score | | Sample variance | | | |
| 2 | Daniela | 85 | | VAR | 34.97 | =VAR(B2:B7) | |
| 3 | Tommy | 79 | | VAR.S | 34.97 | =VAR.S(B2:B7) | |
| 4 | Edward | 90 | | VARA | 34.97 | =VARA(B2:B7) | |
| 5 | Julia | 88 | | | | | |
| 6 | Timothy | 88 | | | | | |
| 7 | Peter | 75 | | | | | |

To check the result, let's do var calculation manually:

1. Find the mean by using the AVERAGE function:

   =AVERAGE(B2:B7)

   The average goes to any empty cell, say B8.

2. Subtract the average from each number in the sample:

   =B2-$B$8

   The differences go to column C, beginning in C2.

3. Square each difference and put the results to column D, beginning in D2:

   =C2 ^ 2

4. Add up the squared differences and divide the result by the number of items in the sample minus 1:

   =SUM(D2:D7)/(6-1)

As you can see, the result of our manual var calculation is exactly the same as the number returned by Excel's built-in functions:

If your data set contains the Boolean and/or text values, the VARA function will return a different result. The reason is that VAR and VAR.S ignore any values other than numbers in references, while VARA evaluates text values as zeros, TRUE as 1, and FALSE as 0. So, please carefully choose the variance function for your calculations depending on whether you want to process or ignore text and logicals.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Student | Score | | Sample variance | | | |
| 2 | Daniela | 85 | | VAR | 39.30 | =VAR(B2:B7) | |
| 3 | Tommy | 79 | | VAR.S | 39.30 | =VAR.S(B2:B7) | |
| 4 | Edward | 90 | | VARA | 1190.70 | =VARA(B2:B7) | |
| 5 | Julia | 88 | | | | | |
| 6 | Timothy | N/A | | | | | |
| 7 | Peter | 75 | | | | | |

Rahul Publications

**Q6. How to calculate population variance in Excel? Excel.**

*Ans :*

Population is all members of a given group, i.e. all observations in the field of study. Population variance describes how data points in the entire population are spread out.

The population variance can be found with this formula:

Population variance = $\frac{\Sigma(x-\bar{x})^2}{n}$

Where,

- x is the mean of the population.
- n is the population size, i.e. the total number of values in the population.

There are 3 functions to calculate population variance in Excel: VARP, VAR.P and VARPA.

**VARP function in Excel**

The Excel VARP function returns the variance of a population based on the entire set of numbers. It is available in all versions of Excel 2000 to 2019.

**VARP(number1, [number2], ...)**

**Note:** In Excel 2010, VARP was replaced with VAR.P but is still kept for backward compatibility. It is recommended to use VAR.P in the current versions of Excel because there is no guarantee that the VARP function will be available in future versions of Excel.

**VAR.P function in Excel**

It is an improved version of the VARP function available in Excel 2010 and later.

**VAR.P(number1, [number2], ...)**

**VARPA function in Excel**

The VARPA function calculates the variance of a population based on the entire set of numbers, text, and logical values. It is available in all version of Excel 2000 through 2019.

**VARA(value1, [value2], ...)**

**Population variance formula in Excel**

In the sample var calculation example, we found a variance of 5 exam scores assuming those scores were a selection from a bigger group of students. If you collect data on all the students in the group, that data will represent the entire population, and you will calculate a population variance by using the above functions.

Let's say, we have the exam scores of a group of 10 students (B2:B11). The scores constitute the entire population, so we will do variance with these formulas:

=VARP(B2:B11)

=VAR.P(B2:B11)

=VARPA(B2:B11)

Rahul Publications

And all the formulas will return the identical result:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Student | Score | | Population variance | | | |
| 2 | Daniela | 85 | | VARP | 36.41 | =VARP(B2:B11) | |
| 3 | Tommy | 79 | | VAR.P | 36.41 | =VAR.P(B2:B11) | |
| 4 | Edward | 90 | | VARPA | 36.41 | =VARPA(B2:B11) | |
| 5 | Julia | 88 | | | | | |
| 6 | Timothy | 88 | | | | | |
| 7 | Peter | 75 | | | | | |
| 8 | Neal | 92 | | | | | |
| 9 | Sally | 74 | | | | | |
| 10 | Mike | 83 | | | | | |
| 11 | Adam | 89 | | | | | |

To make sure Excel has done the variance right, you can check it with the manual var calculation formula shown in the screenshot below:

C2 | =B2-$B$12 | =C2^2

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Student | Score | Dif | Dif$^2$ | | Population variance | | | |
| 2 | Daniela | 85 | 0.70 | 0.49 | | VARP | 36.41 | =VARP(B2:B11) | |
| 3 | Tommy | 79 | -5.30 | 28.09 | | VAR.P | 36.41 | =VAR.P(B2:B11) | |
| 4 | Edward | 90 | 5.70 | 32.49 | | VARPA | 36.41 | =VARPA(B2:B11) | |
| 5 | Julia | 88 | 3.70 | 13.69 | | | | | |
| 6 | Timothy | 88 | 3.70 | 13.69 | | Manual | 36.41 | =SUM(D2:D11)/10 | |
| 7 | Peter | 75 | -9.30 | 86.49 | | | | | |
| 8 | Neal | 92 | 7.70 | 59.29 | | | | | |
| 9 | Sally | 74 | -10.30 | 106.09 | | | | | |
| 10 | Mike | 83 | -1.30 | 1.69 | | | | | |
| 11 | Adam | 89 | 4.70 | 22.09 | | | | | |
| 12 | Average | 84.30 | =AVERAGE(B2:B11) | | | | | | |

If some of the students did not take the exam and have N/A instead of a score number, the VARPA function will return a different result. The reason is that VARPA evaluates text values as zeros while VARP and VAR.P ignore text and logical values in references. Please see VAR.P vs. VARPA for full details.

**Q7. Explain how to calculate population standard deviation with examples.**

*Ans :* **(Imp.)**

**Population and sample standard deviation**

Standard deviation measures the spread of a data distribution. It measures the typical distance between each data point and the mean.

The formula we use for standard deviation depends on whether the data is being considered a population of its own, or the data is a sample representing a larger population.

- If the data is being considered a population on its own, we divide by the number of data points, N.
- If the data is a sample from a larger population, we divide by one fewer than the number of data points in the sample, n-1n, minus, 1.

**Population standard deviation:**

Population standard deviation :

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Sample standard deviation :

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Here's how to calculate population standard deviation:

**Step 1:**

Calculate the mean of the data—this is \muìmu in the formula.

**Step 2:**

Subtract the mean from each data point. These differences are called deviations. Data points below the mean will have negative deviations, and data points above the mean will have positive deviations.

**Step 3:**

Square each deviation to make it positive.

**Step 4:**

Add the squared deviations together.

**Step 5:**

Divide the sum by the number of data points in the population. The result is called the variance.

**Step 6:**

Take the square root of the variance to get the standard deviation.

**Example: Population standard deviation**

Four friends were comparing their scores on a recent essay.

**Calculate the standard deviation of their scores:**

6, 2, 3, 1

**Step 1:**

Find the mean.

$$\mu = \frac{6+2+3+1}{4} = \frac{12}{4} = 3$$

The mean is 3 points.

Rahul Publications

**Step 2:**

Subtract the mean from each score.

| Score: $x_i$ | Deviation: $(x_i - \mu)$ |
|---|---|
| 6 | $6 - 3 = 3$ |
| 2 | $2 - 3 = -1$ |
| 3 | $3 - 3 = 0$ |
| 1 | $1 - 3 = -2$ |

**Step 3:**

Square each deviation.

| Score: $x_i$ | Deviation: $(x_i - \mu)$ | Squared deviation: $(x_i - \mu)^2$ |
|---|---|---|
| 6 | $6 - 3 = 3$ | $(3)^2 = 9$ |
| 2 | $2 - 3 = -1$ | $(-1)^2 = 1$ |
| 3 | $3 - 3 = 0$ | $(0)^2 = 0$ |
| 1 | $1 - 3 = -2$ | $(-2)^2 = 4$ |

**Step 4:**

Add the squared deviations.

$9 + 1 + 0 + 4 = 14$

**Step 5:**

Divide the sum by the number of scores.

$\frac{14}{4} = 3.5$

**Step 6:**

Take the square root of the result from Step 5.

$\sqrt{3.5} \approx 1.87$

Calculating standard deviation of a sample and population

Depending on the nature of your data, use one of the following formulas:

- To calculate standard deviation based on the entire population, i.e. the full list of values (B2:B50 in this example), use the STDEV.P function:

  = STDEV.P(B2:B50)

- To find standard deviation based on a sample that constitutes a part, or subset, of the population (B2:B10 in this example), use the STDEV.S function:

  = STDEV.S(B2:B10)

As you can see in the screenshot below, the formulas return slightly different numbers (the smaller a sample, the bigger a difference):

Rahul Publications

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Name | Score | | Population SD | 15.78 | =STDEV.P(B2:B50) | |
| 2 | Neal | 85 | | Sample SD | 14.44 | =STDEV.S(B2:B10) | |
| 3 | Alex | 76 | | | | | |
| 4 | Dora | 73 | | | | | |
| 5 | Tim | 80 | | | | | |
| 6 | Laura | 72 | | | | | |
| 7 | Kate | 84 | | | | | |
| 8 | Peter | 79 | | | | | |
| 9 | Isabella | 120 | | | | | |
| 10 | Lucas | 85 | | | | | |
| 11 | Ethan | 73 | | | | | |
| 12 | Olivia | 77 | | | | | |
| 13 | Ava | 83 | | | | | |
| 14 | Mason | 108 | | | | | |

In Excel 2007 and lower, you'd use STDEVP and STDEV functions instead:

- To get population standard deviation:

  = STDEVP(B2:B50)

- To calculate sample standard deviation:

  = STDEV(B2:B10)

Calculating standard deviation for text representations of numbers

When discussing different functions to calculate standard deviation in Excel, we sometimes mentioned "text representations of numbers" and you might be curious to know what that actually means.

In this context, "text representations of numbers" are simply numbers formatted as text. How can such numbers appear in your worksheets? Most often, they are exported from external sources. Or, returned by so-called Text functions that are designed to manipulate text strings, e.g. TEXT, MID, RIGHT, LEFT, etc. Some of those functions can work with numbers too, but their output is always text, even if it looks much like a number.

=RIGHT(A2,LEN(A2)-SEARCH("-",A2,1))

The problem is that using an Excel standard deviation formula on the extracted numbers returns either #DIV/0! or 0 like shown in the screenshot below:

=RIGHT(A2,LEN(A2)-SEARCH("-",A2,1))

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Code | Qty. | | |
| 2 | Jeans-105 | 105 | | |
| 3 | Dress-79 | 79 | | |
| 4 | Blouse-98 | 98 | | |
| 5 | Shorts-82 | 82 | | |
| 6 | | | | |
| 7 | STDEV.S | #DIV/0! | =STDEV.S($B$2:$B$4) | |
| 8 | STDEVA | 0 | =STDEVA($B$2:$B$4) | |

Rahul Publications

Why such weird results? As mentioned above, the output of the RIGHT function is always a text string. But neither STDEV.S nor STDEVA can handle numbers formatted as text in references (the former simply ignores them while the latter counts as zeros). To get the standard deviation of such "text-numbers", you need to supply them directly to the list of arguments, which can be done by embedding all RIGHT functions into your STDEV.S or STDEVA formula:

=STDEV.S(RIGHT(A2,LEN(A2)-SEARCH("-",A2,1)), RIGHT(A3,LEN(A3)-SEARCH("-",A3,1)), RIGHT(A4,LEN(A4)-SEARCH("-",A4,1)), RIGHT(A5,LEN(A5)-SEARCH("-",A5,1)))

=STDEVA(RIGHT(A2,LEN(A2)-SEARCH("-",A2,1)), RIGHT(A3,LEN(A3)-SEARCH("-",A3,1)), RIGHT(A4,LEN(A4)-SEARCH("-",A4,1)), RIGHT(A5,LEN(A5)-SEARCH("-",A5,1)))

=STDEV.S(RIGHT(A2,LEN(A2)-SEARCH("-",A2,1)), RIGHT(A3,LEN(A3)-SEARCH("-",A3,1)), RIGHT(A4,LEN(A4)-SEARCH("-",A4,1)), RIGHT(A5,LEN(A5)-SEARCH("-",A5,1)))

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Code | Qty. | | | | | | |
| 2 | Jeans-105 | 105 | | | | | | |
| 3 | Dress-79 | 79 | | | | | | |
| 4 | Blouse-98 | 98 | | | | | | |
| 5 | Shorts-82 | 82 | | | | | | |
| 6 | | | | | | | | |
| 7 | STDEV.S | 12.51666 | | | | | | |
| 8 | STDEVA | 12.51666 | | | | | | |

The formulas are a bit cumbersome, but that might be a working solution for a small sample. For a bigger one, not to mention the entire population, it is definitely not an option. In this case, a more elegant solution would be having the VALUE function convert "text-numbers" to numbers that any standard deviation formula can understand (please notice the right-aligned numbers in the screenshot below as opposed to the left-aligned text strings on the screenshot above):

=VALUE(RIGHT(A2,LEN(A2)-SEARCH("-",A2,1)))

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Code | Qty. | | |
| 2 | Jeans-105 | 105 | | |
| 3 | Dress-79 | 79 | | |
| 4 | Blouse-98 | 98 | | |
| 5 | Shorts-82 | 82 | | |
| 6 | | | | |
| 7 | STDEV.S | 12.51666 | =STDEV.S(B2:B5) | |
| 8 | STDEVA | 12.51666 | =STDEVA(B2:B5) | |

### 2.1.2 Quantitative And Categorical Data

**Q8. Explain about quantitative and categorical data with examples.**

*Ans :* **(Imp.)**

Categorical data is the statistical data comprising categorical variables of data that are converted into categories. One of the examples is a grouped data. More precisely, categorical data could be derived from qualitative data analysis that are countable, or from quantitative data analysis grouped within given

intervals. These data are summarised in the form of a probability table. However, when we consider data analysis, it is referred to use the term "categorical data", which is applied to data sets. Also, it is to be noted that, while containing some categorical variables, the data set may also contain non-categorical variables.

### Categorical or Qualitative Data

The categorical data consists of categorical variables which represent the characteristics such as a person's gender, hometown etc. Categorical measurements are expressed in terms of natural language descriptions, but not in terms of numbers. Sometimes categorical data can take numerical values, but those numbers do not have mathematical meaning. Some of the examples of the categorical data are as follows:

- Birthdate
- Favourite sport
- School Postcode
- Travel method to school etc.

When you observe the above example, birthdate and postcode contain numbers. Even though it contains numerals, it is considered as categorical data. The easy way to determine whether the given data is categorical or numerical data is to calculate the average. If you are able to calculate the average, then it is considered to be a numerical data. If you cannot calculate the average, then it is considered to be a categorical data. Like the example mentioned above, the average of birthdate and the postal code has no meaning, so it is taken as categorical data.

### Types of Categorical Data

In general, categorical data has values and observations which can be sorted into categories or groups. The best way to represent these data is bar graphs and pie charts. Categorical data are further classified into two types namely,

- Nominal Data
- Ordinal Data

### Nominal Data

Nominal data is a type of data that is used to label the variables without providing any numerical value. It is also known as the nominal scale. Nominal data cannot be ordered and measured. But sometimes nominal data can be qualitative and quantitative. Some of the few common examples of nominal data are letters, words, symbols, gender etc.

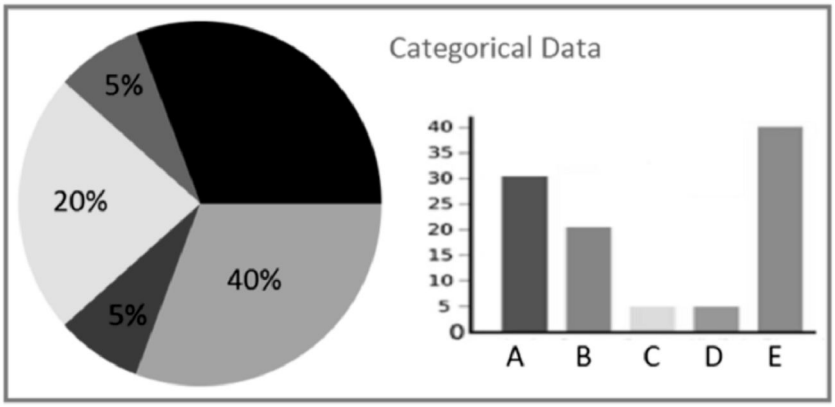
These data are analysed with the help of the grouping method. The variables are grouped together into categories and the percentage or frequency can be calculated. It can be presented visually using the pie chart.

### Ordinal Data

Ordinal data is a type of data that follows a natural order. The notable features of ordinal data are that the difference between data values cannot be determined. It is commonly encountered in surveys, questionnaires, finance and economics.

The data can be analysed using visualisation tools. It is commonly represented using a bar chart. Sometimes the data may be represented using tables in which each row in the table indicates the distinct category.

Rahul Publications




**Categorical Variables**

In statistics, a categorical variable is a variable that contains limited, and usually a fixed number of possible values. They take values which are normally names or labels. Examples are:

- The colour of a wall, like red, blue, pink, gree, etc.,
- Gender of people, like male, female and transgender
- Blood group of a person: A, B, O, AB, etc.,

These variables are used to assign each individual or another unit of observation to a particular group or nominal category based on some qualitative property. Generally, each of the potential values of a categorical variable is said to be as a level. The probability distribution linked with a random categorical variable is known as categorical distribution.

**Categorical and Numerical Data**

- Categorical data or Qualitative data consist of categorical values or variables, where the data are represented in labelled or given a name. Such as the breed of a dog, colour of the car, and so on
- Numerical data or Quantitative data comprising numbers or numerical values to represent the data, such as height, weight, age of a person

**Example:**

The FAA monitors airlines for safety and customer service. For each flight, the carrier must report the type of aircraft, flight number, number of passengers, and whether or not the flights departed and arrived on schedule.

What variables are reported for each flight, and are they quantitative or categorical?

| Answers: Variables | Quantitative or | Categorical |
|---|---|---|
| (1) Type of aircraft | Quantitative | **Categorical** |
| (2) Flight Number | Quantitative | **Categorical** |
| (3) Number of Passengers | **Quantitative** | Categorical |
| (4) Arrived/Departed on Schedule | Quantitative | **Categorical** |

**Presenting Categorical Data Graphically**

Categorical, or qualitative, data are pieces of information that allow us to classify the objects under investigation into various categories. We usually begin working with categorical data by summarizing the data into a frequency table.

Rahul Publications

**Example 1**

An insurance company determines vehicle insurance premiums based on known risk factors. If a person is considered a higher risk, their premiums will be higher. One potential factor is the color of your car. The insurance company believes that people with some color cars are more likely to get in accidents. To research this, they examine police reports for recent total-loss collisions. The data is summarized in the frequency table below.

| Color | Frequency |
|---|---|
| Blue | 25 |
| Green | 52 |
| Red | 41 |
| White | 36 |
| Black | 39 |
| Grey | 23 |

Sometimes we need an even more intuitive way of displaying data. This is where charts and graphs come in. There are many, many ways of displaying data graphically, but we will concentrate on one very useful type of graph called a bar graph. In this section we will work with bar graphs that display categorical data; the next section will be devoted to bar graphs that display quantitative data.

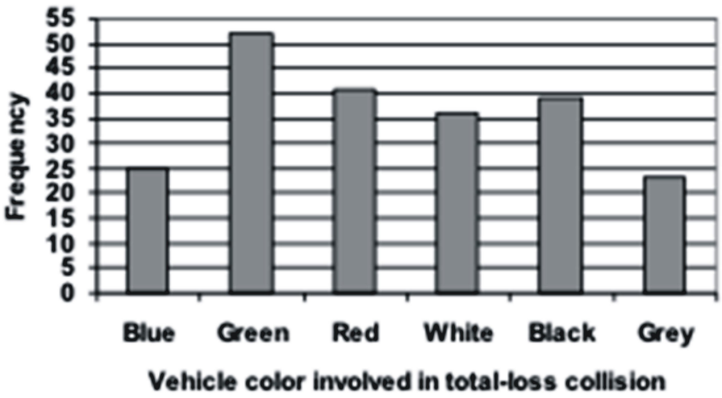
**BAR GRAPH**

A bar graph is a graph that displays a bar for each category with the length of each bar indicating the frequency of that category.

To construct a bar graph, we need to draw a vertical axis and a horizontal axis. The vertical direction will have a scale and measure the frequency of each category; the horizontal axis has no scale in this instance. The construction of a bar chart is most easily described by use of an example.

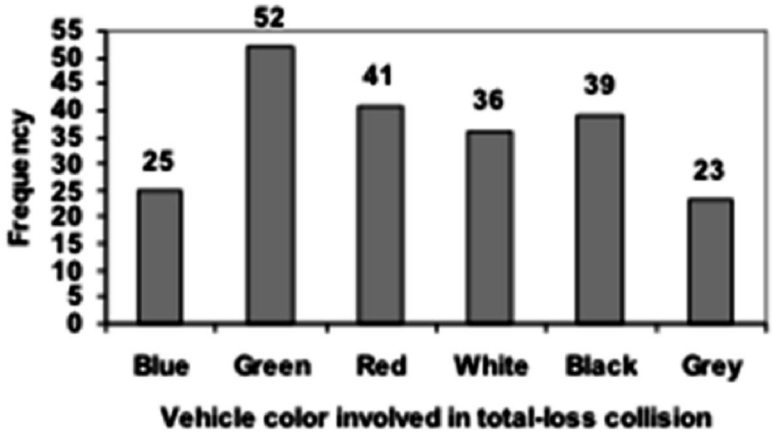
**Example 2**

Using our car data from above, note the highest frequency is 52, so our vertical axis needs to go from 0 to 52, but we might as well use 0 to 55, so that we can put a hash mark every 5 units:




Notice that the height of each bar is determined by the frequency of the corresponding color. The horizontal gridlines are a nice touch, but not necessary. In practice, you will find it useful to draw bar graphs using graph paper, so the gridlines will already be in place, or using technology. Instead of gridlines, we might also list the frequencies at the top of each bar, like this:

Rahul Publications




In this case, our chart might benefit from being reordered from largest to smallest frequency values. This arrangement can make it easier to compare similar values in the chart, even without gridlines. When we arrange the categories in decreasing frequency order like this, it is called a Pareto chart.

**Example:**

Radio Station Formats (Distribution of a categorical variable) The radio audience rating service Arbitron places the country's 13,838 radio stations into categories that describe the kinds of programs they broadcast. Here are two different tables showing the distribution of station formats:

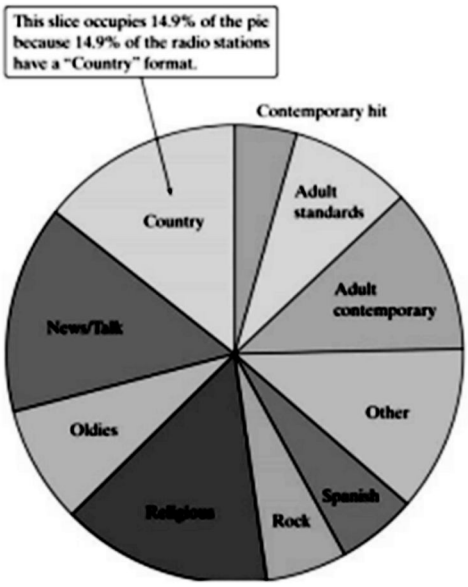
| Frequency table | |
|---|---|
| **Format** | **Count of stations** |
| Adult contemporary | 1,556 |
| Adult standards | 1,196 |
| Contemporary hit | 569 |
| Country | 2,066 |
| News/Talk/Information | 2,179 |
| Oldies | 1,060 |
| Religious | 2,014 |
| Rock | 869 |
| Spanish language | 750 |
| Other formats | 1,579 |
| Total | 13,838 |

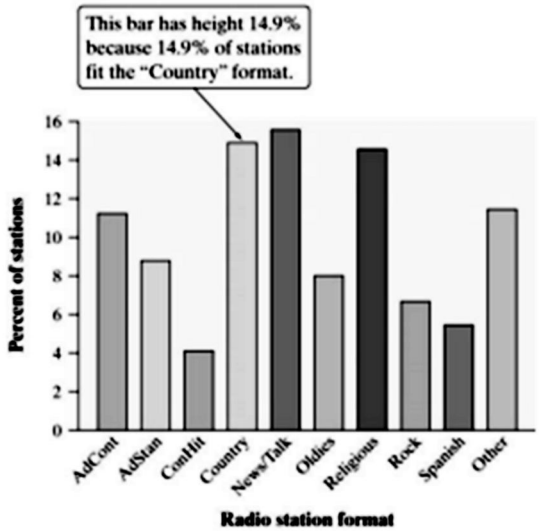
| Relative frequency table | |
|---|---|
| **Format** | **Percent of stations** |
| Adult contemporary | 11.2 |
| Adult standards | 8.6 |
| Contemporary hit | 4.1 |
| Country | 14.9 |
| News/Talk/Information | 15.7 |
| Oldies | 7.7 |
| Religious | 14.6 |
| Rock | 6.3 |
| Spanish language | 5.4 |
| Other formats | 11.4 |
| Total | 99.9 |

The on the left, which we call a frequency table, displays the counts (frequencies) of stations in each format category. On the right, we see a relative frequency table of the data that shows the percents (relative frequencies) of stations in each format category.

**Displaying categorical data**

Frequency tables can be difficult to read. Sometimes is easier to analyze a distribution by displaying it with a bar graph or pie chart.



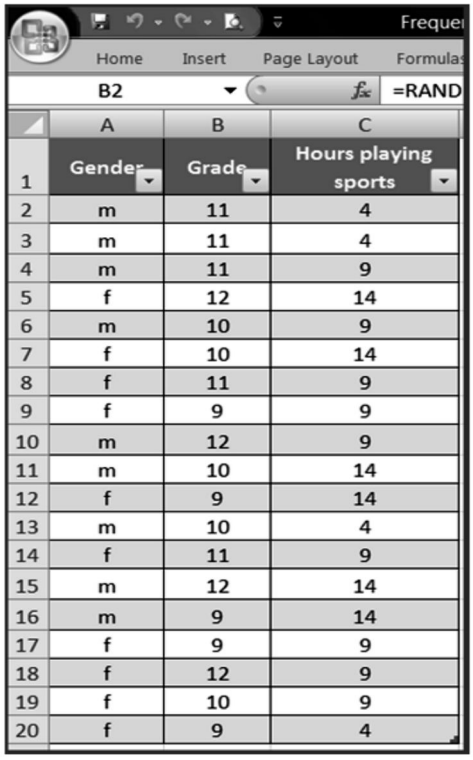

**Q9. How to organize the categorical data on excel?**

*Ans :*

**Organizing and Graphing Categorical Data on Excel**

**Make a Frequency Table**

Excel will count how many of each response you have in a given column and you can easily organize the data into a frequency table.



Home Insert Page Layout Formulas

B2 =RAND

| | A | B | C |
|---|---|---|---|
| 1 | Gender | Grade | Hours playing sports |
| 2 | m | 11 | 4 |
| 3 | m | 11 | 4 |
| 4 | m | 11 | 9 |
| 5 | f | 12 | 14 |
| 6 | m | 10 | 9 |
| 7 | f | 10 | 14 |
| 8 | f | 11 | 9 |
| 9 | f | 9 | 9 |
| 10 | m | 12 | 9 |
| 11 | m | 10 | 14 |
| 12 | f | 9 | 14 |
| 13 | m | 10 | 4 |
| 14 | f | 11 | 9 |
| 15 | m | 12 | 14 |
| 16 | m | 9 | 14 |
| 17 | f | 9 | 9 |
| 18 | f | 12 | 9 |
| 19 | f | 10 | 9 |
| 20 | f | 9 | 4 |

Rahul Publications

**To make a frequency table of simple numeric data (no intervals)**

1. Use the COUNTIF formula.
2. The range of data is the column you are counting.
3. The criteria is the number you are counting, in this case grade.

| Grade | Frequency |
|---|---|
| 9 | =COUNTIF(B2:B20,9) |
| 10 | =COUNTIF(B2:B20,10) |
| 11 | =COUNTIF(B2:B20,11) |
| 12 | =COUNTIF(B2:B20,12) |

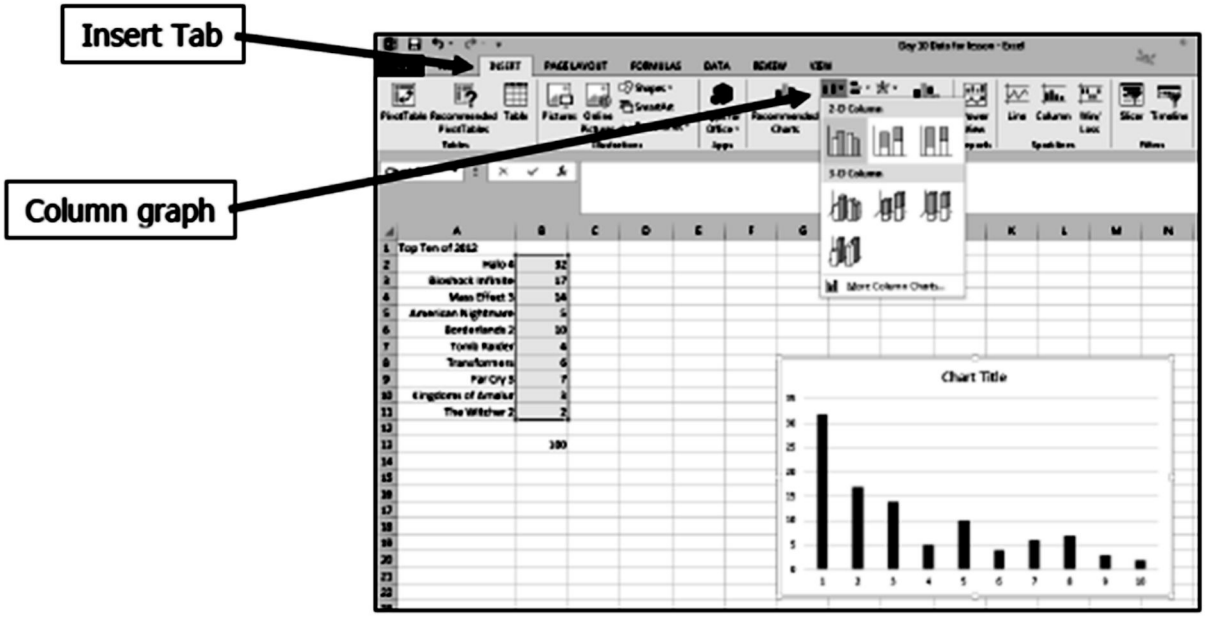
**To make a frequency table for text data (counting words)**

1. Use the COUNTIF formula.
2. The range of data is the column you are counting.
3. Put the word you are counting in quotations.

| Gender | Frequency |
|---|---|
| Male | =COUNTIF(A2:A20,"m") |
| Female | =COUNTIF(A2:A20,"f") |

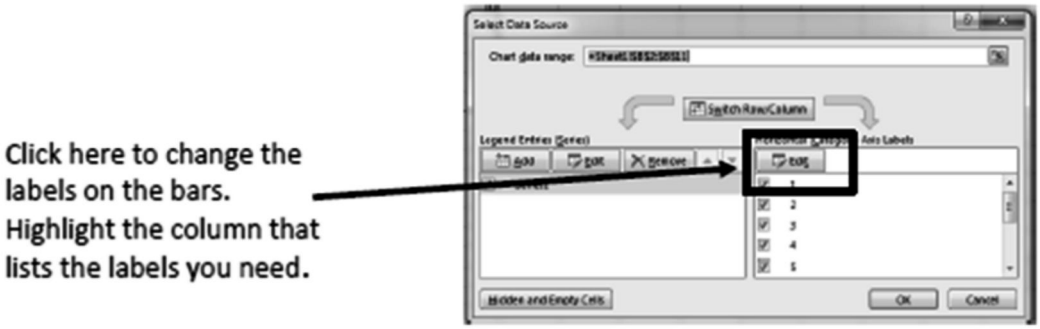
**Make a Bar Graph**

Highlight the column that includes the data you would like in your bar graph. On the Insert Tab click on the bar or column graph depending on your preference.



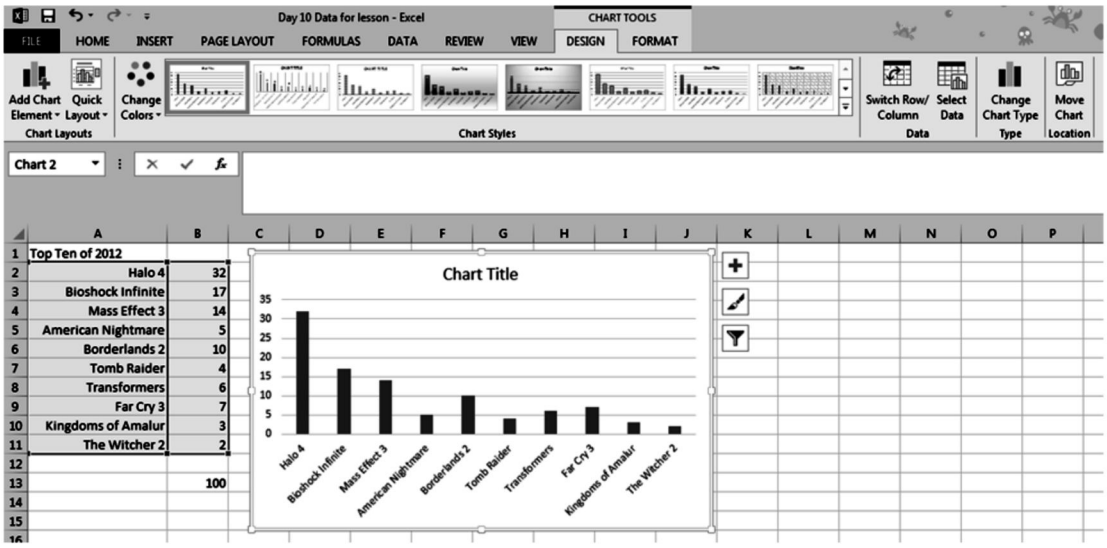

A bar graph will now appear, however the layout and labels may not be correct. These can be adjusted using the Design Tab in the Chart Tools Tab when a chart is selected.

Choose Select Data. You will get a popup that looks like this:




### Adjusting Your Graph

By using the Chart Tools – Design, adjust the features of the graph to customize it for your needs.



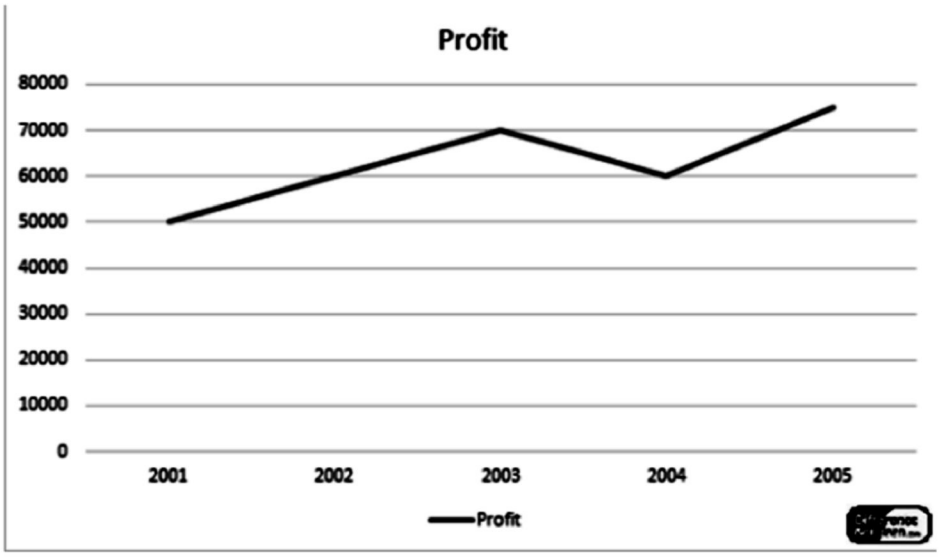

## 2.1.3 Cross-sectional and Time Series Data

**Q10. What is Time Series Data and Cross Sectional Data? Explain with an examples.**

*Ans :* **(Imp.)**

Time series data focuses on observations of a single individual at different times usually at uniform intervals. It is the data of the same variable over a period of time such as months, quarters, years etc. The time series data takes the form of Xt. The t represents the time. Below is an example of the profit of an organization over a period of 5 years' time. Profit is the variable that changes each year.

| Year | Profit |
|---|---|
| 2001 | 50000 |
| 2002 | 60000 |
| 2003 | 70000 |
| 2004 | 60000 |
| 2005 | 75000 |




Usually, time series data is useful in business applications. Time measurement can be months, quarters or years but it can also be any time interval. Generally, the time has uniform intervals.

### Cross Sectional Data

In cross sectional data, there are several variables at the same point in time. Data set with maximum temperature, humidity, wind speed of few cities on a single day is an example of a cross sectional data.

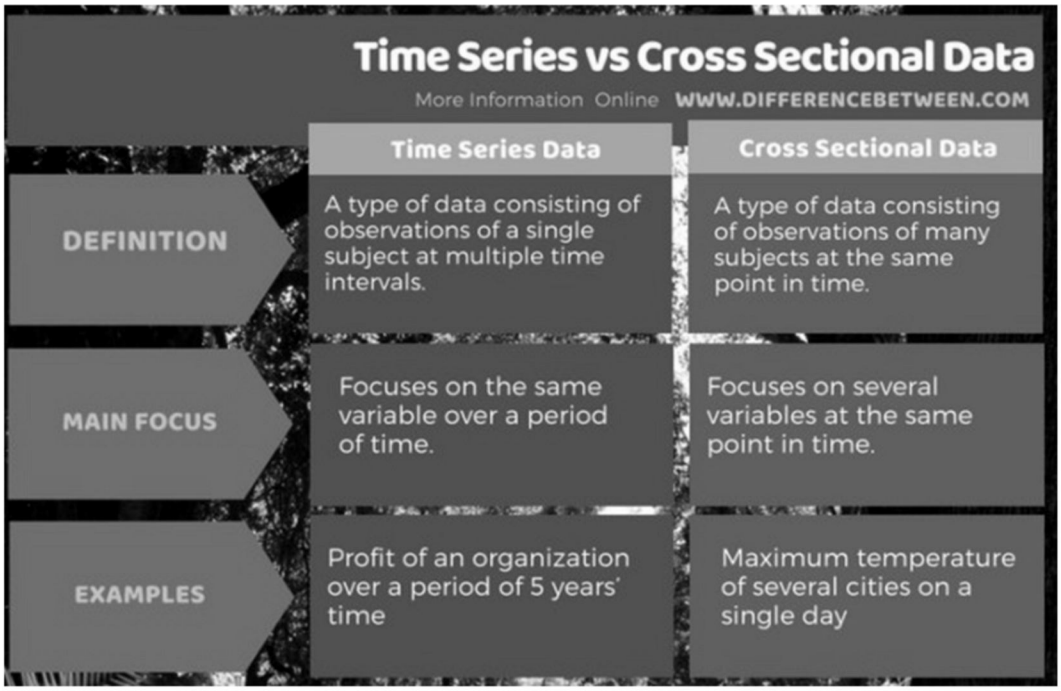
| City | Maximum Temperature | Humidity | Wind Speed |
|---|---|---|---|
| City A | 29 | 60% | 20mph |
| City B | 27 | 65% | 26mph |
| City C | 30 | 60% | 21mph |

Another example is the sales revenue, sales volume, number of customers and expenses of an organization in the past month. Cross sectional data takes the form of Xi. Expanding the data from several months will convert the cross sectional data to time series data.

### Difference Between Time Series and Cross Sectional Data

Time series data consist of observations of a single subject at multiple time intervals. Cross sectional data consist of observations of many subjects at the same point in time. Time series data focuses on the same variable over a period of time. On the other hand, cross sectional data focuses on several variables at the same point in time. This is the main difference between time series and cross sectional data.

Profit of an organization over a period of 5 years' time is an example for a time series data while maximum temperature of several cities on a single day is an example for a cross sectional data.




**Example1.**

The last period's forecast was 70 and demand was 60. What is the simple exponential smoothing forecast with alpha of 0.4 for the next period.

**Solution:**

$Y_{t-1} = 70$

$S_{t-1} = 60$

Alpha = 0.4

Substituting the values we get

0.4*60 + 0.6*70 = 24 + 42 = 66

**Example2.**

If the demand is 100 during October 2016, 200 in November 2016, 300 in December 2016, 400 in January 2017. What is the 3-month simple moving average for February 2017?

**Solution:**

$X' = (x_{t-3} + x_{t-2} + x_{t-1}) / 3$

(200+300+400)/ 3 = 900/3 =300

**Example 3.**

Suppose you are given a time series dataset which has only 4 columns (id, Time, X, Target).

| Id | Time | X | Target |
|---|---|---|---|
| 1 | 1 | 100 | 10 |
| 2 | 2 | 200 | 20 |
| 3 | 3 | 300 | 30 |
| 1 | 4 | 400 | 40 |
| 2 | 5 | 500 | 50 |
| 3 | 6 | 600 | 60 |
| 1 | 7 | 500 | 50 |
| 2 | 8 | 400 | 40 |
| 3 | 9 | 500 | 30 |
| 4 | 10 | 700 | 20 |

What would be the rolling mean of feature X if you are given the window size 2?

Note: X column represents rolling mean.

Solution:

$X' = x_{t-2} + x_{t-1} /2$

| Quater | Time | X' | Target |
|---|---|---|---|
| 1 | 1 | NaN | 10 |
| 2 | 2 | NaN | 20 |
| 3 | 3 | 150 | 30 |
| 1 | 4 | 250 | 40 |
| 2 | 5 | 350 | 50 |
| 3 | 6 | 450 | 60 |
| 1 | 7 | 550 | 50 |
| 2 | 8 | 550 | 40 |
| 3 | 9 | 450 | 30 |
| 4 | 10 | 450 | 20 |

Based on the above formula: (100 + 200) /2 = 150; (200 + 300)/2 = 250 and so on.

**Example 4.**

Consider the following set of data:

{23.32 32.33 32.88 28.98 33.16 26.33 29.88 32.69 18.98 21.23 26.66 29.89}

**What is the lag-one sample autocorrelation of the time series?**

**Solution:**

$\hat{p}_1 = PT_{t=2}(x_{t-1} - x)(x_t - x)\ PT_{t=1}(xt - x)^2$

$= (23.32 - x)(32.33 - x) + (32.33 - x)(32.88 - x) + \cdots\ PT_{t=1}(x_t - x)^2$

$= 0.130394786$

Where $\bar{x}$ is the mean of the series which is 28.0275

## 2.2 Sources of Data

**Q11. Write about various sources of data and how to collect them.**

*Ans :*

### Sources of Data

The sources of data can be classified into two types: statistical and non-statistical. Statistical sources refer to data that is gathered for some official purposes, incorporate censuses, and officially administered surveys. Non-statistical sources refer to the collection of data for other administrative purposes or for the private sector.

The following are the two sources of data:

**1. Internal sources**

- When data is collected from reports and records of the organisation itself, they are known as the internal sources.
- For example, a company publishes its annual report' on profit and loss, total sales, loans, wages, etc.

**2. External sources**

- When data is collected from sources outside the organisation, they are known as the external sources. For example, if a tour and travel company obtains information on Karnataka tourism from Karnataka Transport Corporation, it would be known as an external source of data.

### Types of Data

**(a) Primary data**

- Primary data means first-hand information collected by an investigator.
- It is collected for the first time.
- It is original and more reliable.
- For example, the population census conducted by the government of India after every ten years is primary data.

**(b) Secondary data**

- Secondary data refers to second-hand information.
- It is not originally collected and rather obtained from already published or unpublished sources.
- For example, the address of a person taken from the telephone directory or the phone number of a company taken from Just Dial are secondary data.

Students can also refer to Meaning and Sources of Secondary Data

### Methods of Collecting Primary Data

**(a) Collection of Primary Data**

Collection of Primary Data can be done through various methods, which are:

- **Direct Personal Investigation:** In this method, surveyors or investigators collect the data themselves. This method is suitable for small projects where the required data needs to be reliable and excessive effort is not mandatory.

Rahul Publications

- **Collection with the Help of Investigators**: In this method, a single or a group of correspondents collects the data for the surveyor. These correspondents are trained investigators who are employed for this course of action. This type of data collecting method is useful for a large population.
- **Collection Assisted by Questionnaires**: When the amount of data that is required to be collected is significantly large, questionnaires are used to make the data collecting process easier. Questionnaires are nothing but a set of questions that, when answered, provide the required data. Surveyors can also mail questionnaires to the respondents for added convenience.

**(b) Collection of Secondary Data**

The collection of secondary data is much easier than collecting primary data. Secondary data is available on various sources, both published and unpublished.

However, the investigator of this kind of data must ensure that the data is reliable, suitable for analysis, whether bias is involved during sampling of the said data, etc.

## 2.3 Descriptive Statistics

### 2.3.1 Measures of Location(Central Tendency)

#### 2.3.1.1 Mean, Median and Mode and Relationship between them

**Q12. Write briefly about various measures of tendency with examples.**

**(OR)**

**Explain Mean, Median and Mode with an examples.**

**(OR)**

**Discuss the measure of central tendency.**

**(OR)**

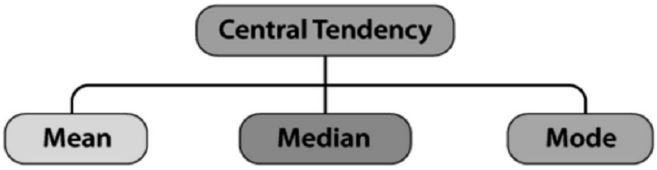
**Explain the utility of measures of central tendency in understanding performance of business organization.**

*Ans :* **(Dec.-21, Octo.-20, Imp.)**

The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset. It aims to provide an accurate description of the entire data in the distribution.

**Measures of Central Tendency**

The central tendency of the dataset can be found out using the three important measures namely mean, median and mode.




**I) Mean**

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the arithmetic mean. Some other measures of mean used to find the central tendency are as follows:

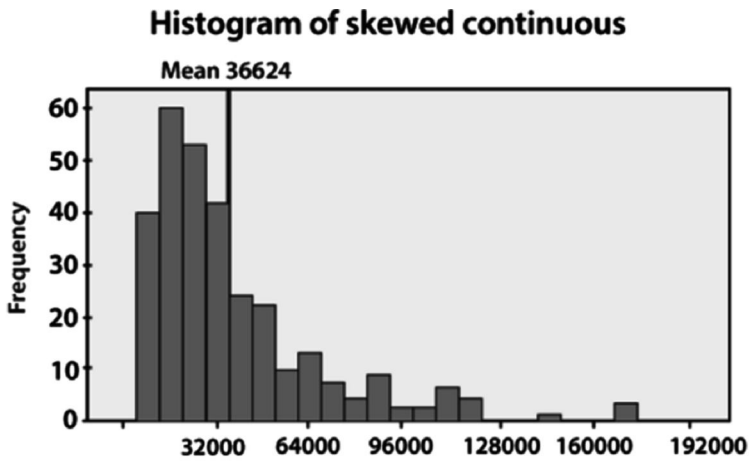


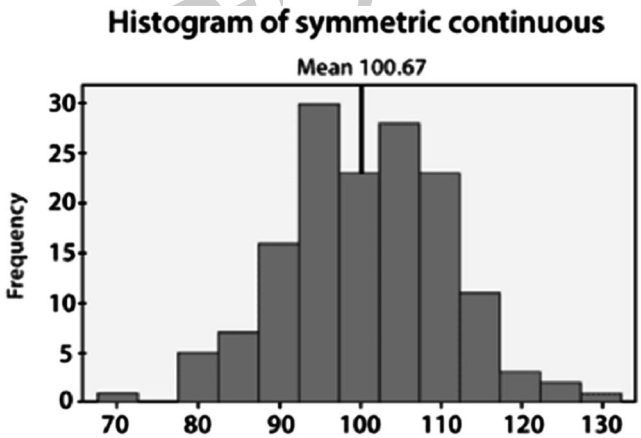
Rahul Publications

- Geometric Mean
- Harmonic Mean
- Weighted Mean

It is observed that if all the values in the dataset are the same, then all geometric, arithmetic and harmonic mean values are the same. If there is variability in the data, then the mean value differs. Calculating the mean value is completely easy. The formula to calculate the mean value is given as

The histogram given below shows that the mean value of symmetric continuous data and the skewed continuous data.




In symmetric data distribution, the mean value is located accurately at the centre. But in the skewed continuous data distribution, the extreme values in the extended tail pull the mean value away from the centre. So it is recommended that the mean can be used for the symmetric distributions.

Mean is the most commonly used measure of central tendency. It actually represents the average of the given collection of data. It is applicable for both continuous and discrete data.

It is equal to the sum of all the values in the collection of data divided by the total number of values.

Suppose we have n values in a set of data namely as

$x_1, x_2, x_3 \ldots\ldots\ldots\ldots\ldots\ldots x_n$

then the mean of data is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n}$$

It can also be denoted as:

$$\bar{x} = \frac{\sum_{t=1}^{n} x_t}{n}$$

For grouped data, we can calculate the mean using three different methods of formula.

| Direct method | Assumed mean method | Step deviation method |
|---|---|---|
| Mean $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$ Here, $\sum f_i$ = Sum of all frequencies | Mean $(\bar{x}) = a + \frac{\sum f_i d_i}{\sum f_i}$ Here, a = Assumed mean $d_i = x_i - a$ $\sum f_i$ = Sum of all frequencies | Mean $(\bar{x}) = a + h\frac{\sum f_i u_i}{\sum f_i}$ Here, a = Assumed mean $u_i = (x_i - a)/h$ h = Class size $\sum f_i$ = Sum of all frequencies |

**Example 1:**

Find the Median of 14, 63 and 55

**Solution:**

Put them in ascending order: 14, 55, 63

The middle number is 55, so the median is 55.

**Example 2:**

Find the median of the following:

4, 17, 77, 25, 22, 23, 92, 82, 40, 24, 14, 12, 67, 23, 29

**Solution:**

When we put those numbers in the order we have:

4, 12, 14, 17, 22, 23, 23, 24, 25, 29, 40, 67, 77, 82, 92,

There are fifteen numbers. Our middle is the eighth number:

The median value of this set of numbers is 24.

**Example 3:**

Rahul's family drove through 7 states on summer vacation. The prices of Gasoline differ from state to state. Calculate the median of gasoline cost.

1.79, 1.61, 2.09, 1.84, 1.96, 2.11, 1.75

**Solution:**

By organizing the data from smallest to greatest, we get:

1.61, 1.75, 1.79, 1.84 , 1.96, 2.09, 2.11

**II) Median**

Generally median represents the mid-value of the given set of data when arranged in a particular order.

Median: Given that the data collection is arranged in ascending or descending order, the following method is applied:

- If number of values or observations in the given data is odd, then the median is given by
- If in the given data set, the number of values or observations is even then the median is given by the average of observation.

$$\left(\frac{n}{2}\right)^{th} \text{ and } \left(\frac{n}{2}+1\right)^{th}$$

Median for grouped data can be calculated using the formula,

**Example 1:**

Find the Median of 14, 63 and 55

**Solution:**

Put them in ascending order: 14, 55, 63

The middle number is 55, so the median is 55.

**Example 2:**

Find the median of the following:

4, 17, 77, 25, 22, 23, 92, 82, 40, 24, 14, 12, 67, 23, 29

**Solution:**

When we put those numbers in the order we have:

4, 12, 14, 17, 22, 23, 23, 24, 25, 29, 40, 67, 77, 82, 92,

There are fifteen numbers. Our middle is the eighth number:

The median value of this set of numbers is 24.

**Example 3:**

Rahul's family drove through 7 states on summer vacation. The prices of Gasoline differ from state to state. Calculate the median of gasoline cost.

1.79, 1.61, 2.09, 1.84, 1.96, 2.11, 1.75

**Solution:**

By organizing the data from smallest to greatest, we get:

1.61, 1.75, 1.79, 1.84 , 1.96, 2.09, 2.11

Hence, the median of the gasoline cost is 1.84. There are three states with greater gasoline costs and 3 with smaller prices.

Rahul Publications

### III) Mode

The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.

Consider the following data set which represents the marks obtained by different students in a subject.

| Name | Anmol | Kushagra | Garima | Ashwini | Geetika | Shakshi |
|---|---|---|---|---|---|---|
| Marks Obtained (out of 100) | 73 | 80 | 73 | 70 | 73 | 65 |

The maximum frequency observation is 73 ( as three students scored 73 marks), so the mode of the given data collection is 73.

We can calculate the mode for grouped data using the below formula:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

**Mode Formula For Grouped Data**

In the case of grouped frequency distribution, calculation of mode just by looking into the frequency is not possible. To determine the mode of data in such cases we calculate the modal class. Mode lies inside the modal class. The mode of data is given by the formula:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where,

l = lower limit of the modal class

h = size of the class interval

$f_1$ = frequency of the modal class

$f_0$ = frequency of the class preceding the modal class

$f_2$ = frequency of the class succeeding the modal class

Let us take an example to understand this clearly.

**Example 1:**

Find the mode of the given data set: 3, 3, 6, 9, 15, 15, 15, 27, 27, 37, 48.

**Solution:**

In the following list of numbers,

3, 3, 6, 9, 15, 15, 15, 27, 27, 37, 48

15 is the mode since it is appearing more number of times in the set compared to other numbers.

**Example 2:**

Find the mode of 4, 4, 4, 9, 15, 15, 15, 27, 37, 48 data set.

**Solution:**

Given: 4, 4, 4, 9, 15, 15, 15, 27, 37, 48 is the data set.

Rahul Publications

As we know, a data set or set of values can have more than one mode if more than one value occurs with equal frequency and number of time compared to the other values in the set.

Hence, here both the number 4 and 15 are modes of the set.

**Example 3:**

Find the mode of 3, 6, 9, 16, 27, 37, 48.

**Solution:**

If no value or number in a data set appears more than once, then the set has no mode.

Hence, for set 3, 6, 9, 16, 27, 37, 48, there is no mode available.

**Example 4:**

In a class of 30 students marks obtained by students in mathematics out of 50 is tabulated as below. Calculate the mode of data given.

| Marks Obtained | Number of Student |
|---|---|
| 10 – 20 | 5 |
| 20 – 30 | 12 |
| 30 – 40 | 8 |
| 40 – 50 | 5 |

**Solution:**

The maximum class frequency is 12 and the class interval corresponding to this frequency is 20 – 30. Thus, the modal class is 20 – 30.

Lower limit of the modal class (l) = 20

Size of the class interval (h) = 10

Frequency of the modal class ($f_1$) = 12

Frequency of the class preceding the modal class ($f_0$) = 5

Frequency of the class succeeding the modal class ($f_2$) = 8

Substituting these values in the formula we get;

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h = 20 + \left(\frac{12-5}{2\times 12 - 5 - 8}\right) \times 10 = 26.364$$

**Q13. Explain the Relationship between Mean, Median and Mode with examples.**

*Ans :* **(Imp.)**

**Relationship between Mean, Median and Mode**

In case of a moderately skewed distribution, the difference between mean and mode is almost equal to three times the difference between the mean and median. Thus, the empirical mean median mode relation is given as:

**Mean – Mode = 3 (Mean – Median)**

**(OR)**

**Mode = 3 Median – 2 Mean**

Either of these two ways of equations can be used as per the convenience since by expanding the first representation we get the second one as shown below:

Mean – Mode = 3 (Mean – Median)

Mean – Mode = 3 Mean – 3 Median

By rearranging the terms,

Mode = Mean – 3 Mean + 3 Median

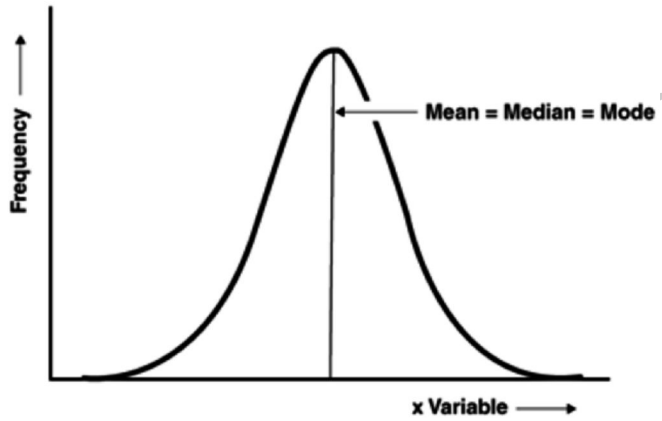
Mode = 3 Median – 2 Mean

However, we can define the relation between mean, median and mode for different types of distributions as explained below:

Mean Median Mode Relation With Frequency Distribution

- **Frequency Distribution with Symmetrical Frequency Curve**

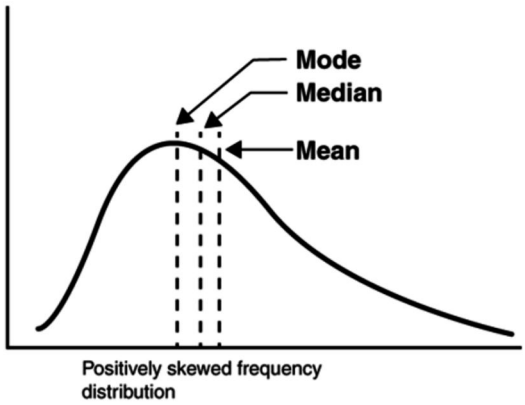
  If a frequency distribution graph has a symmetrical frequency curve, then mean, median and mode will be equal.




**Mean = Median = Mode**

- **For Positively Skewed Frequency Distribution**

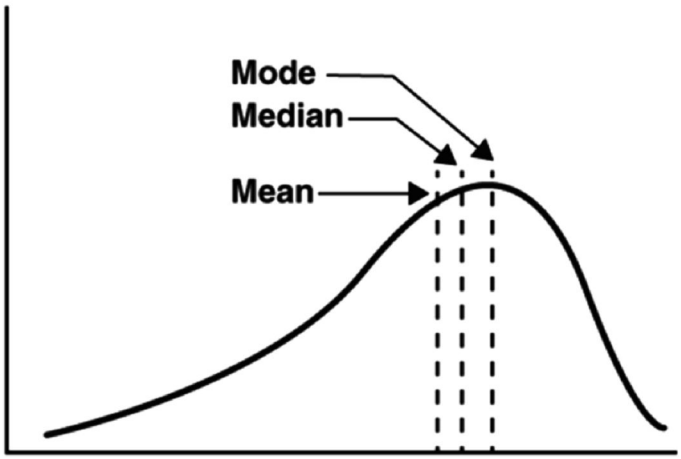
  In case of a positively skewed frequency distribution, the mean is always greater than median and the median is always greater than the mode.



**Mean > Median > Mode**

- **For Negatively Skewed Frequency Distribution**

In case of a negatively skewed frequency distribution, the mean is always lesser than median and the median is always lesser than the mode.




Negatively skewed frequency distribution

**Mean < Median < Mode**

**Examples Using the Mean, Median and Mode Relationship**

**Example 1:**

In a moderately skewed distribution, the median is 20 and the mean is 22.5. Using these values, find the approximate value of the mode.

**Solution:**

Given,

Mean = 22.5

Median = 20

Mode = x

Now, using the relationship between mean mode and median we get,

(Mean – Mode) = 3 (Mean – Median)

So,

22.5 – x = 3 (22.5 – 20)

22.5 – x = 7.5

∴ x = 15

So, Mode = 15.

Rahul Publications

**Q14. Explain how to calculate mean, median and mode in Excel.**

*Ans :* **(Imp.)**

**Calculation of mean in Excel**

Arithmetic mean, also referred to as average, is probably the measure you are most familiar with. The mean is calculated by adding up a group of numbers and then dividing the sum by the count of those numbers.

For example, to calculate the mean of numbers {1, 2, 2, 3, 4, 6}, you add them up, and then divide the sum by 6, which yields 3: (1+2+2+3+4+6)/6=3.

In Microsoft Excel, the mean can be calculated by using one of the following functions:

- **AVERAGE** - returns an average of numbers.
- **AVERAGEA** - returns an average of cells with any data (numbers, Boolean and text values).
- **AVERAGEIF** - finds an average of numbers based on a single criterion.
- **AVERAGEIFS** - finds an average of numbers based on multiple criteria.

For the in-depth tutorials, please follow the above links. To get a conceptual idea of how these functions work, consider the following example.

In a sales report (please see the screenshot below), supposing you want to get the average of values in cells C2:C8. For this, use this simple formula:

=AVERAGE(C2:C8)

To get the average of only "Banana" sales, use an AVERAGEIF formula:

=AVERAGEIF(A2:A8, "Banana", C2:C8)

To calculate the mean based on 2 conditions, say, the average of "Banana" sales with the status "Delivered", use AVERAGEIFS:

=AVERAGEIFS(C2:C8,A2:A8, "Banana", B2:B8, "Delivered")

You can also enter your conditions in separate cells, and reference those cells in your formulas, like this:

| | 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 | | | | | |

Rahul Publications

### Calculation of median in Excel

Median is the middle value in a group of numbers, which are arranged in ascending or descending order, i.e. half the numbers are greater than the median and half the numbers are less than the median. For example, the median of the data set {1, 2, 2, 3, 4, 6, 9} is 3.

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 3 | 4 | 6 | 9 |

This works fine when there are an odd number of values in the group. But what if you have an even number of values? In this case, the median is the arithmetic mean (average) of the two middle values. For example, the median of {1, 2, 2, 3, 4, 6} is 2.5. To calculate it, you take the 3rd and 4th values in the data set and average them to get a median of 2.5.

| | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 2 | 3 | 4 | 6 |

In Microsoft Excel, a median is calculated by using the MEDIAN function. For example, to get the median of all amounts in our sales report, use this formula:

**=MEDIAN(C2:C8)**

To make the example more illustrative, I've sorted the numbers in column C in ascending order (though it is not actually required for the Excel Median formula to work):

| | 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) | |

In contrast to average, Microsoft Excel does not provide any special function to calculate median with one or more conditions. However, you can "emulate" the functionality of MEDIANIF and MEDIANIFS by using a combination of two or more functions like shown in these examples:

### Calculation of mode in Excel

Mode is the most frequently occurring value in the dataset. While the mean and median require some calculations, a mode value can be found simply by counting the number of times each value occurs.

For example, the mode of the set of values {1, 2, 2, 3, 4, 6} is 2. In Microsoft Excel, you can calculate a mode by using the function of the same name, the MODE function. For our sample data set, the formula goes as follows:

**=MODE(C2:C8)**

| | 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) | |

In situations when there are two or more modes in your data set, the Excel MODE function will return the lowest mode.

**Q15. For the given distribution, find the Mean.**

| Xi | Fi |
|---|---|
| 1 | 3 |
| 2 | 5 |
| 3 | 8 |
| 4 | 4 |
| ΣFi = 20 | |

*Ans :*

Mean = $(1\times3+2\times5+3\times8+4\times4)/20 = 2.65$

**Q16. Find the median of the set = { 2,4,4,3,8,67,23}**

*Ans :*

As we can see the list is not arranged in any order. The sorted list in ascending order = {2,3,4,4,8,23,67}. The list contains 7 terms, thus $4^{th}$ term of the list will be the median, so the median is 4.

If the list contains 'n' terms (n is an odd number), the median will be the (n+1)/2 term.

In case the list consists of even number of terms, the median will be the average of $n^{th}$ and $(n+1)^{th}$ term.

**Q17. Find the median of the set = {11,22,33,55,66,99}**

*Ans :*

As we can see the list is already in ascending order and the list contains 6 terms, hence the average of the third and fourth term will be the median.

Median = (33+55)/2 = 44.

Rahul Publications

**Q18. Find the median of a series of all the even terms from 4 to 296.**

*Ans :*

The given sequence is 4,6,8,10,12,14….296. As we can see, the given sequence is an Arithmetic progression ( An arithmetic progression is a sequence of terms where any two consecutive terms differ by a constant difference). To find out the median, we need to know the number of terms. We will use the $n^{th}$ term of an arithmetic progression formula ($a_n = a_1 + (n-1)d$) to calculate the number of terms. Then, depending on whether n is odd or even we can find out the media. The entire process will take up a lot of time.

Remember: If the given sequence is Arithmetic sequence, then

Median = First term+Last term/2 = Mean

For this sequence, Median = 4+296/2 = 150.

**Q19. Find the mode of the Set = {1,3,3,6,9}**

*Ans :*

In the sequence, the value '1' occurs maximum number of times, hence the mode is 1.

Remember: There can be more than one mode in a series. For example, in the set = {2,4,4,6,8,9,9}, both 4 and 9 are the Modes as their frequency of occurrence is more than other values.

**Q20. It is given that in a moderately skewed distribution, median = 10 and mean = 12. Using these values, find the approximate value of the mode.**

*Ans :*

We know that the relation between mean, median, and mode in a moderately skewed distribution is 3 median = mode + 2 mean. Let us take mode to be 'x'. We have been given that the median = 10 and mean = 12. Now, using the relationship between mean, mode, and median we get,

$3 \times 10 = x + 2 \times 12$

$30 = x + 24$

$x = 30-24$

$x = 6$

Therefore, the value of mode is 6.

**Q21. Find the possible range of median of a positively skewed distribution, if the values of mean and mode are 30 and 20 respectively.**

*Ans :*

For a positively skewed frequency distribution, the empirical relation between mean, median, and mode is mean > median > mode. On the basis of this, the range of the median if the mean is 30 and mode is 20 is 30 > median > 20. It means that the median will be greater than 20 and less than 30.

If the heights of 5 persons are 140 cm, 150 cm, 152 cm, 158 cm and 161cm respectively, find the mean height.

It is given that

heights of 5 persons are -140cm, 150cm, 152cm, 158cm and 161cm.

Rahul Publications

$$\therefore \quad \text{Mean height} = \frac{\text{Sum of heights}}{\text{Total No of persons}}$$

$$= \frac{140+150+152+158+161}{5}$$

$$= \frac{761}{5}$$

$$= 152.2$$

**22. Calculate the mean for the following distribution:**

| x : | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|
| f : | 4 | 8 | 14 | 11 | 3 |

*Sol :*

| x | f | fx |
|---|---|---|
| 5 | 4 | 20 |
| 6 | 8 | 48 |
| 7 | 14 | 98 |
| 8 | 11 | 88 |
| 9 | 3 | 27 |
| | $N = 40$ | $\sum fx$ 284 |

$$\therefore \quad \text{Mean } \bar{x} = \frac{\sum fx}{N}$$

$$= \frac{281}{40}$$

$$= 7.025$$

**23. Find the median of the following data**

**25, 34, 31, 23, 22, 26, 35, 29, 20, 32**

*Sol :*

Given number are 25, 34, 31, 23, 22, 26. 35, 29, 20. 32

Arranging in increasing order

20, 22, 23, 25, 26, 29, 31, 32, 34, 35

10 = n (even)

$$\therefore \quad \text{Median} = \frac{\frac{n^{th}}{2}\text{value} + \left(\frac{n^{th}}{2}+1\right)\text{value}}{2}$$

Rahul Publications

$$= \frac{\frac{10^{th}}{2}\text{value} + \left(\frac{10^{th}}{2}+1\right)\text{value}}{2}$$

$$= \frac{5^{th}\text{value} + 6^{th}\text{value}}{2}$$

$$= \frac{26+29}{2} = \frac{55}{2}$$

**24. If the difference of mode and median of a data is 24, then find the difference of median and mean.**

*Sol :*

Given that the difference of mode and median of a data is 24. That is,

Mode - Median = 24

$\Rightarrow$ Mode = Median + 24

We have to find the difference between median and mean

We know that

Mode = 3 × Median – 2 × Mean

$\Rightarrow$ Median + 24 = 3 × Median – 2 × Mean

$\Rightarrow$ 24 = 3 × Median – Median – 2 × Mean

$\Rightarrow$ 24 = 2 × Median – 2 × Mean

$\Rightarrow$ 2 × Median – 2 × Mean = 24

$\Rightarrow$ 2(Median – Mean) = 24

$\Rightarrow$ Median – Mean = $\frac{24}{2}$

$\Rightarrow$ Median - Mean = 12

**25. The arithmetic mean and mode of a data are 24 and 12 respectively, then find the median of the data.**

*Sol :*

Given that the arithmetic mean and mode of a data are 24 and 12 respectively.

That is,

Rahul Publications

Mean = 24

Mode = 12

We have to find median

We know that

Mode = 3 × Median – 2 × Mean

$\Rightarrow$ 12 = 3 × Median – 2 × 24

$\Rightarrow$ 3 × Median = 12 + (2 × 24)

$\Rightarrow$ 3 × Median = 12 + 48

$\Rightarrow$ 3 × Median = 60

$\Rightarrow$ Median = $\frac{60}{3}$

$\Rightarrow$ Median = 20

Rahul Publications

# Short Question and Answers

**1. Population.**

*Ans :*

It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a parameter. For example, All people living in India indicates the population of India.

There are different types of population. They are:

- Finite Population
- Infinite Population
- Existent Population
- Hypothetical Population

**2. Sample.**

*Ans :*

It includes one or more observations that are drawn from the population and the measurable characteristic of a sample is a statistic. Sampling is the process of selecting the sample from the population. For example, some people living in India is the sample of the population.

Basically, there are two types of sampling. They are:

i) Probability sampling

ii) Non-probability sampling

**(i) Probability Sampling**

In probability sampling, the population units cannot be selected at the discretion of the researcher. This can be dealt with following certain procedures which will ensure that every unit of the population consists of one fixed probability being included in the sample. Such a method is also called random sampling. Some of the techniques used for probability sampling are:

- Simple random sampling
- Cluster sampling
- Stratified Sampling
- Disproportionate sampling
- Proportionate sampling
- Optimum allocation stratified sampling
- Multi-stage sampling

**(ii) Non Probability Sampling**

In non-probability sampling, the population units can be selected at the discretion of the researcher. Those samples will use the human judgements for selecting units and has no theoretical basis for estimating the characteristics of the population. Some of the techniques used for non-probability sampling are

- Quota sampling
- Judgement sampling
- Purposive sampling

**3. What is population variance?**

*Ans :*

Population variance is a measure of the spread of population data. Hence, population variance can be defined as the average of the distances from each data point in a particular population to the mean squared, and it indicates how data points are spread out in the population. Population variance is an important measure of dispersion used in statistics. Statisticians calculate variance to determine how individual numbers in a data set relate to each other.

Population variance can be calculated by using the following formula:

$$\sigma_2 = \frac{\Sigma_{i=1}^{n}(x_i - \mu)^2}{N}$$

where

- $\sigma^2$ is population variance,
- $x_1, x_2, x_3, \ldots x_n$ are the observations
- N is the number of observations,
- $\mu$ is the mean of the data set

Rahul Publications

**4. How to calculate variance in Excel?**

*Ans :*

A sample is a set of data extracted from the entire population. And the variance calculated from a sample is called sample variance.

For example, if you want to know how people's heights vary, it would be technically unfeasible for you to measure every person on the earth. The solution is to take a sample of the population, say 1,000 people, and estimate the heights of the whole population based on that sample.

Sample variance is calculated with this formula:

$$\text{Sample variance} = \frac{\Sigma(x-\bar{x})^2}{(n-1)}$$

Where,

- $\bar{x}$ is the mean (simple average) of the sample values.
- n is the sample size, i.e. the number of values in the sample.

There are 3 functions to find sample variance in Excel: VAR, VAR.S and VARA.

VAR function in Excel

It is the oldest Excel function to estimate variance based on a sample. The VAR function is available in all versions of Excel 2000 to 2019.

**5. Categorical Data**

*Ans :*

The categorical data consists of categorical variables which represent the characteristics such as a person's gender, hometown etc. Categorical measurements are expressed in terms of natural language descriptions, but not in terms of numbers. Sometimes categorical data can take numerical values, but those numbers do not have mathematical meaning. Some of the examples of the categorical data are as follows:

- Birthdate
- Favourite sport
- School Postcode
- Travel method to school etc.

When you observe the above example, birthdate and postcode contain numbers. Even though it contains numerals, it is considered as categorical data. The easy way to determine whether the given data is categorical or numerical data is to calculate the average. If you are able to calculate the average, then it is considered to be a numerical data. If you cannot calculate the average, then it is considered to be a categorical data. Like the example mentioned above, the average of birthdate and the postal code has no meaning, so it is taken as categorical data.

**Types of Categorical Data**

In general, categorical data has values and observations which can be sorted into categories or groups. The best way to represent these data is bar graphs and pie charts. Categorical data are further classified into two types namely,

- Nominal Data
- Ordinal Data

**6. Time Series Data.**

*Ans :*

Time series data focuses on observations of a single individual at different times usually at uniform intervals. It is the data of the same variable over a period of time such as months, quarters, years etc. The time series data takes the form of Xt. The t represents the time. Below is an example of the profit of an organization over a period of 5 years' time. Profit is the variable that changes each year.

Rahul Publications

**7. Cross Sectional Data.**

*Ans :*

In cross sectional data, there are several variables at the same point in time. Data set with maximum temperature, humidity, wind speed of few cities on a single day is an example of a cross sectional data.

| City | Maximum Temperature | Humidity | Wind Speed |
|---|---|---|---|
| City A | 29 | 60% | 20mph |
| City B | 27 | 65% | 26mph |
| City C | 30 | 60% | 21mph |

Another example is the sales revenue, sales volume, number of customers and expenses of an organization in the past month. Cross sectional data takes the form of Xi. Expanding the data from several months will convert the cross sectional data to time series data.

**8. Difference Between Time Series and Cross Sectional Data**

*Ans :*

Time series data consist of observations of a single subject at multiple time intervals. Cross sectional data consist of observations of many subjects at the same point in time. Time series data focuses on the same variable over a period of time. On the other hand, cross sectional data focuses on several variables at the same point in time. This is the main difference between time series and cross sectional data.

Profit of an organization over a period of 5 years' time is an example for a time series data while maximum temperature of several cities on a single day is an example for a cross sectional data.

**9. Sources of Data**

*Ans :*

The sources of data can be classified into two types: statistical and non-statistical. Statistical sources refer to data that is gathered for some official purposes, incorporate censuses, and officially administered surveys. Non-statistical sources refer to the collection of data for other administrative purposes or for the private sector.

The following are the two sources of data:

**(i) Internal sources**

- When data is collected from reports and records of the organisation itself, they are known as the internal sources.
- For example, a company publishes its annual report' on profit and loss, total sales, loans, wages, etc.

**(ii) External sources**

- When data is collected from sources outside the organisation, they are known as the external sources. For example, if a tour and travel company obtains information on Karnataka tourism from Karnataka Transport Corporation, it would be known as an external source of data.

Rahul Publications

**10. Types of Data**

*Ans :*

**(a) Primary data**

- Primary data means first-hand information collected by an investigator.
- It is collected for the first time.
- It is original and more reliable.
- For example, the population census conducted by the government of India after every ten years is primary data.

**(b) Secondary data**

- Secondary data refers to second-hand information.
- It is not originally collected and rather obtained from already published or unpublished sources.
- For example, the address of a person taken from the telephone directory or the phone number of a company taken from Just Dial are secondary data.

Students can also refer to Meaning and Sources of Secondary Data

**11. Mean**

*Ans :*

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the arithmetic mean. Some other measures of mean used to find the central tendency are as follows:

- Geometric Mean
- Harmonic Mean
- Weighted Mean

It is observed that if all the values in the dataset are the same, then all geometric, arithmetic and harmonic mean values are the same. If there is variability in the data, then the mean value differs. Calculating the mean value is completely easy. The formula to calculate the mean value is given as

The histogram given below shows that the mean value of symmetric continuous data and the skewed continuous data.

**12. Median.**

*Ans :*

Generally median represents the mid-value of the given set of data when arranged in a particular order.

Median: Given that the data collection is arranged in ascending or descending order, the following method is applied:

Rahul Publications

- If number of values or observations in the given data is odd, then the median is given by
- If in the given data set, the number of values or observations is even then the median is given by the average of observation.

$$\left(\frac{n}{2}\right)^{th} \text{ and } \left(\frac{n}{2}+1\right)^{th}$$

Median for grouped data can be calculated using the formula.

**13. Mode**

*Ans :*

The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.

Consider the following data set which represents the marks obtained by different students in a subject.

| Name | Anmol | Kushagra | Garima | Ashwini | Geetika | Shakshi |
|---|---|---|---|---|---|---|
| Marks Obtained (out of 100) | 73 | 80 | 73 | 70 | 73 | 65 |

The maximum frequency observation is 73 ( as three students scored 73 marks), so the mode of the given data collection is 73.

We can calculate the mode for grouped data using the below formula:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

**Mode Formula For Grouped Data**

In the case of grouped frequency distribution, calculation of mode just by looking into the frequency is not possible. To determine the mode of data in such cases we calculate the modal class. Mode lies inside the modal class. The mode of data is given by the formula:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where,

l = lower limit of the modal class

h = size of the class interval

$f_1$ = frequency of the modal class

$f_0$ = frequency of the class preceding the modal class

$f_2$ = frequency of the class succeeding the modal class

Let us take an example to understand this clearly.

Rahul Publications

**14. Relationship between Mean, Median and Mode.**

*Ans :*

In case of a moderately skewed distribution, the difference between mean and mode is almost equal to three times the difference between the mean and median. Thus, the empirical mean median mode relation is given as:

**Mean – Mode = 3 (Mean – Median)**

**(OR)**

**Mode = 3 Median – 2 Mean**

Either of these two ways of equations can be used as per the convenience since by expanding the first representation we get the second one as shown below:

Mean – Mode = 3 (Mean – Median)

Mean – Mode = 3 Mean – 3 Median

By rearranging the terms,

Mode = Mean – 3 Mean + 3 Median

Mode = 3 Median – 2 Mean

Rahul Publications

# Choose the Correct Answers

1. An ideal measure of central tendency is [ a ]
   (a) Arithmetic mean (b) Moving average
   (c) Median (d) Harmonic Mean
2. Mathematical average is called [ a ]
   (a) Arithmetic mean (b) Geometric mean
   (c) Mode (d) None of these
3. Sum of deviations of the items is zero from [ a ]
   (a) Mean (b) Median
   (c) Mode (d) Geometric mean
4. A _____ is a characteristic that takes different values at different times, places or situations. [ d ]
   (a) Attributes (b) Data
   (c) Statistics (d) Variable
5. Measure of central tendency _____. [ d ]
   (a) Mean (b) Mode
   (c) Median (d) All
6. Which of the following measures of central tendency will always change if a single value in the data changes? [ a ]
   (a) Mean (b) Median
   (c) Mode (d) All of these
7. If a positively skewed distribution has a median of 50, which of the following statement is true? [ c ]
   (a) Mean is greater than 50 (b) Mean is less than 50
   (c) Both (a) and (c) (d) Mode is greater than 50
8. If the variance of a dataset is correctly computed with the formula using (n - 1) in the denominator, which of the following option is true? [ c ]
   (a) Dataset is a sample
   (b) Dataset is a population
   (c) Dataset could be either a sample or a population
   (d) Dataset is from a census
9. The difference between the highest and the lowest value of the observations in a data is called: [ b ]
   (a) Mean (b) Range
   (c) Total frequency (d) Sum of observation
10. The range of the data: 6,14,20,16,6,5,4,18,25,15, and 5 is [ b ]
   (a) 4 (b) 21
   (c) 25 (d) 20

Rahul Publications

# Fill in the Blanks

1. _____ is a branch of mathematics that deals with collecting, interpreting, organization and interpretation of data.
2. _____ statistics is used to summarize data and make sense out of the raw data collected during the research.
3. _____ It is the midpoint of a distribution of data.
4. _____ stands for business integer.
5. _____ reports using Ms.Excel can display in a number of ways.
6. _____ is usually performed on categorical data that can be divided into mutually exclusive groups.
7. _____ Charts are used to graphically summarize data and explore complicated data.
8. KPI stands for _____.
9. _____ Represent the information in Rows and Columns.
10. _____ allow managers to monitor the contribution of the various departments in the organization.

**ANSWERS**

1. Statistics
2. Descriptive
3. Median
4. BI
5. Data analysis
6. Cross tabulation
7. Pivot
8. Key perofrmance indicators
9. Tables
10. Dashboards

Rahul Publications

UNIT III

**DESCRIPTIVE ANALYTICS 2 :**

Measures of Variability-Range, Variance, Standard deviation, Coefficient of Variation, Percentiles, Quartiles, Analyzing Distributions – Empirical Rule, Identifying Outliers, Box Plots, Measures of Association -Scatter Charts, Covariance, Correlation Coefficient – Problems.

## 3.1 Measures of Variability

**Q1. What is variability?**

*Ans :*

**Meaning**

Variability describes how far apart data points lie from each other and from the center of a distribution. Along with measures of central tendency, measures of variability give you descriptive statistics that summarize your data.

Variability is also referred to as spread, scatter or dispersion. It is most commonly measured with the following:

- **Range:** the difference between the highest and lowest values.
- **Interquartile range:** the range of the middle half of a distribution.
- **Standard deviation:** average distance from the mean.
- **Variance:** average of squared distances from the mean.

### 3.1.1 Range

**Q2. What is Range? How to find Range in Statistics?**

*Ans :*

The range in statistics for a given data set is the difference between the highest and lowest values. For example, if the given data set is {2,5,8,10,3}, then the range will be 10 – 2 = 8.

Thus, the range could also be defined as the difference between the highest observation and lowest observation. The obtained result is called the range of observation. The range in statistics represents the spread of observations.

**Formula**

The formula of the range in statistics, can simply be given by the difference between highest and lowest value.

Range = Highest Value – Lowest Value

OR

Range = Highest observation – Lowest observation

OR

Range = Maximum value – Minimum Value

**Finding Range:**

To find the range in statistics, we need to arrange the given values or set of data or set of observations in ascending order. That means, firstly write the observations from lowest to highest value. Now, we need to use the formula to find the range of observations.

**Example 1:**

Find the range of given observations: 32, 41, 28, 54, 35, 26, 23, 33, 38, 40.

*Sol :*

Let us first arrange the given values in ascending order.

23, 26, 28, 32, 33, 35, 38, 40, 41, 54

Since, 23 is the lowest value and 54 is the highest value, therefore, the range of the observations will be;

Range (X) = Max (X) – Min (X)

= 54 – 23

= 31

Rahul Publications

**Example 2:**

Following are the marks of students in Mathematics: 50, 53, 50, 51, 48, 93, 90, 92, 91, 90. Find the range of the marks.

*Sol :*

Arrange the following marks in ascending order, we get;

48, 50, 50, 51, 53, 90, 90, 91, 92, 93

Thus, the range of marks will be:

Range = Maximum marks – Minimum marks

Range = 93 – 48 = 45

Thus, 45 is the required range.

**Arithmetic Mean and Range in Statistics**

In statistics, groups of data are commonly represented by arithmetic mean. Sometimes, arithmetic mean is also referred to as average or just 'mean'.

Basically, mean is the central value of given data. To find the arithmetic mean of the data set, we have to add all the values in the set and then divide the resulting value by the total number of values.

Arithmetic mean = (Sum of all observations)/(Total number of observations)

**Example 1:**

Find the mean of the data set: 32, 41, 28, 54, 35, 26, 23, 33, 38, 40.

*Sol :*

To find the mean, we have to add all the given values first.

Sum of observations = 32 + 41 + 28 + 54 + 35 + 26 + 23 + 33 + 38 + 40 = 350

Total number of observations = 10

Therefore, the mean of observations is:

Mean = (Sum of all observations)/(Total number of observations)

Mean = 350/10 = 35

Hence, 35 is the required arithmetic mean.

**Example 2:**

Following are the marks of students in Mathematics: 50, 53, 50, 51, 48, 93, 90, 92, 91, 90. Find the mean of the marks.

*Sol :*

Given, the marks of the students are:

50, 53, 50, 51, 48, 93, 90, 92, 91, 90

Mean = (Sum of all observations)/(Total number of observations)

Thus,

Sum of observations = 50 + 53 + 50 + 51 + 48 + 93 + 90 + 92 + 91 + 90 = 708

Rahul Publications

Total observations = 10

Therefore,

Arithmetic mean = 708/10 = 70.8

Hence, 70.8 is the required mean.

**Example 3:**

If the data set has observations as: 4, 6, 7, 5, 3, 5, 4, 5, 2, 6, 2, 5, 1, 9, 6, 5, 8, 4, 6, 7. Then find:

(a) The maximum value?

(b) The minimum value?

(c) Range of data set

*Sol :*

Let us arrange the given values from lowest to highest (increasing order).

1, 2, 2, 3, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 9.

Clearly from the above arrangement we can conclude that;

(a) The maximum value is 9.

(b) The minimum value is 1

(c) Range = 9 – 1 = 8

**Example 4:**

What is the arithmetic mean of 4, 6, 7, 5, 3, 5, 4, 5, 2, 6, 2, 5, 1, 9, 6, 5, 8, 4, 6, 7?

*Sol :*

To find the arithmetic mean, we have to use the below formula:

Arithmetic mean = (Sum of all observations)/(Total number of observations)

Sum of all observation

= 1 + 2 + 2 + 3 + 4 + 4 + 4 + 5 + 5 + 5 + 5 + 5 + 6 + 6 + 6 + 6 + 7 + 7 + 8 + 9 = 100

Total Number of Observation = 20

Arithmetic mean = (100/20) = 5

**Q3. Explain how to Calculate Range in Excel.**

*Ans :*

If you have a list of sorted values, you just have to subtract the first value from the last value (assuming that the sorting is in the ascending order).

Excel has the functions to find out the maximum and the minimum value from a range (the MAX and the MIN function).

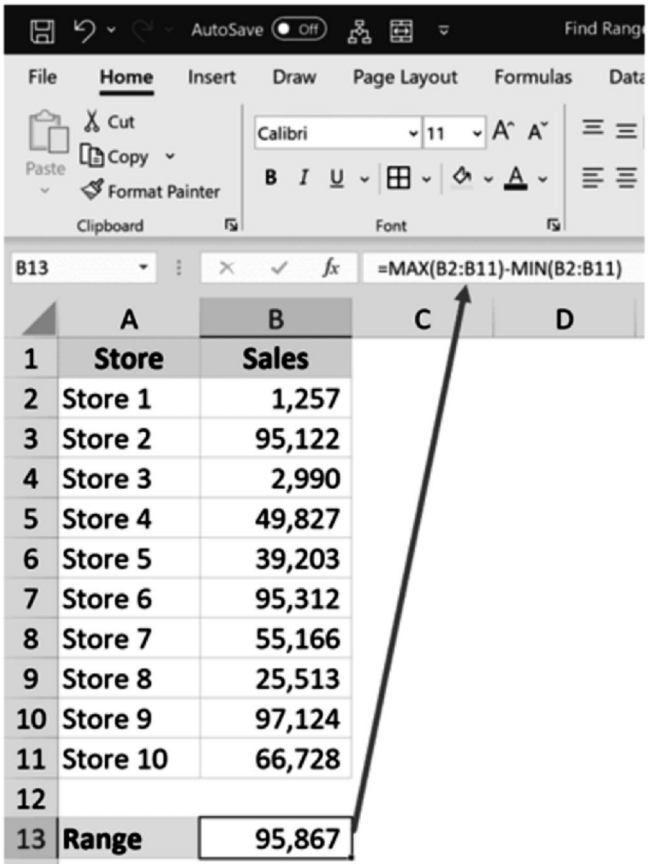
Suppose you have a data set as shown below, and you want to calculate the range for the data in column B.

Rahul Publications

| | A | B |
|---|---|---|
| 1 | Store | Sales |
| 2 | Store 1 | 1,257 |
| 3 | Store 2 | 95,122 |
| 4 | Store 3 | 2,990 |
| 5 | Store 4 | 49,827 |
| 6 | Store 5 | 39,203 |
| 7 | Store 6 | 95,312 |
| 8 | Store 7 | 55,166 |
| 9 | Store 8 | 25,513 |
| 10 | Store 9 | 97,124 |
| 11 | Store 10 | 66,728 |

below is the formula to calculate the range for this data set:

=MAX(B2:B11)-MIN(B2:B11)



### 3.1.2 Variance

**Q4. What is variance? Explain how to find variance in various modes with an examples.**

*Ans :* **(Imp.)**

Variance is a measure of how data points differ from the mean. According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value.

Variance means to find the expected difference of deviation from actual value. Therefore, variance depends on the standard deviation of the given data set.

The more the value of variance, the data is more scattered from its mean and if the value of variance is low or minimum, then it is less scattered from mean. Therefore, it is called a measure of spread of data from mean.

**Population variance formula**

Use the population form of the equation when you have values for all members of the group of interest. In this case, you are not using the sample to estimate the population. Instead, you have measured all people or items and need the variance for that specific group. For example, if you have measured test scores for all class members and need to know the value for that class, use the population variance formula.

The formula for the variance of an entire population is

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{n}$$

In the population variance formula:

- $\sigma^2$ is the population variance.
- $X_i$ is the $i^{th}$ data point.
- $\mu$ is the population mean.
- n is the number of observations.

To find the variance, take a data point, subtract the population mean, and square that difference. Repeat this process for all data points. Then, sum all of those squared values and divide by the number of observations. Hence, it's the average squared difference.

**Sample variance formula**

Use the sample variance formula when you're using a sample to estimate the value for a population. For example, if you have taken a random sample of statistics students, recorded their test scores, and need to use the sample as an estimate for the population of statistics students, use the sample variance formula.

The population formula tends to underestimate variability when you use it with a sample. The sample formula below corrects for that bias.

$$s^2 = \frac{\sum (X_i - \bar{x})^2}{n-1}$$

In the sample variance formula:

Rahul Publications

- $s^2$ is the sample variance.
- $X_i$ is the $i^{th}$ data point.
- x is the sample mean.
- n – 1 is the degrees of freedom.

The calculation process for samples is very similar to the population method. However, you're working with a sample instead of a population, and you're dividing by n–1. This denominator counteracts a bias where samples tend to underestimate the population value.

**Variance and Standard Deviation**

Standard deviation is the positive square root of the variance. The symbols ó and s are used correspondingly to represent population and sample standard deviations.

Standard Deviation is a measure of how spread out the data is. Its formula is simple; it is the square root of the variance for that data set. It's represented by the Greek symbol sigma (ó).

How to Calculate Variance

Variance can be calculated easily by following the steps given below:

- Find the mean of the given data set. Calculate the average of a given set of values
- Now subtract the mean from each value and square them
- Find the average of these squared values, that will result in variance

Say if $x_1, x_2, x_3, x_4, \ldots, x_n$ are the given values.

Therefore, the mean of all these values is:

$\bar{x} = (x_1 + x_2 + x_3 + \ldots + x_n)/n$

Now subtract the mean value from each value of the given data set and square them.

$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, (x_3 - \bar{x})^2, \ldots\ldots, (x_n - \bar{x})^2$

Find the average of the above values to get the variance.

$Var(X) = [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots\ldots + (x_n - \bar{x})^2]/n$

Hence, the variance is calculated.

**Example1:**

**Let's say the heights (in mm) are 610, 450, 160, 420, 310 find variance**

*Sol :*

Mean and Variance is interrelated. The first step is finding the mean which is done as follows,

Mean = ( 610 + 450 + 160 + 420 + 310)/ 5 = 390

So the mean average is 390 mm.

To calculate the Variance, compute the difference of each from the mean, square it and find then find the average once again.

So for this particular case the variance is :

$= (220^2 + 60^2 + (-230)^2 + 30^2 + (-80)^2)/5$

$= (48400 + 3600 + 52900 + 900 + 6400)/5$

Variance = 22440

**Example 2:**

**Find the variance of the numbers 3, 8, 6, 10, 12, 9, 11, 10, 12, 7.**

*Sol :*

Given,

3, 8, 6, 10, 12, 9, 11, 10, 12, 7

**Step 1:**

Compute the mean of the 10 values given.

Mean = (3+8+6+10+12+9+11+10+12+7) / 10 = 88 / 10 = 8.8

**Step 2:**

Make a table with three columns, one for the X values, the second for the deviations and the third for squared deviations. As the data is not given as sample data so we use the formula for population variance. Thus, the mean is denoted by $\mu$.

| X | $X - \mu$ | $(X - \mu)^2$ |
|---|---|---|
| 3 | -5.8 | 33.64 |
| 8 | -0.8 | 0.64 |
| 6 | -2.8 | 7.84 |
| 10 | 1.2 | 1.44 |
| 12 | 3.2 | 10.24 |
| 9 | 0.2 | 0.04 |
| 11 | 2.2 | 4.84 |
| 10 | 1.2 | 1.44 |
| 12 | 3.2 | 10.24 |
| 7 | -1.8 | 3.24 |
| Total | 0 | 73.6 |

Rahul Publications

**Step 3:**

$\sigma^2$ = 73.6 / 10

= 7.36

**Example:**

Here's an example of how to calculate the variance using the sample formula. The dataset has 17 observations in the table below. The numbers in parentheses correspond to table columns.

To calculate the statistic, take each data value (1) and subtract the mean (2) to calculate the difference (3), and then square the difference (4).

At the bottom of the worksheet, I sum the squared values, and divide it by 17 – 1 = 16 because we're finding the sample value.

The variance for this dataset is 201.

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Data Point | Mean | Difference | Squared Difference |
| 11 | 32 | -21 | 441 |
| 16 | 32 | -16 | 256 |
| 19 | 32 | -13 | 169 |
| 20 | 32 | -12 | 144 |
| 21 | 32 | -11 | 121 |
| 22 | 32 | -10 | 100 |
| 25 | 32 | -7 | 49 |
| 26 | 32 | -6 | 36 |
| 29 | 32 | -3 | 9 |
| 33 | 32 | 1 | 1 |
| 34 | 32 | 2 | 4 |
| 38 | 32 | 6 | 36 |
| 39 | 32 | 7 | 49 |
| 46 | 32 | 14 | 196 |
| 52 | 32 | 20 | 400 |
| 55 | 32 | 23 | 529 |
| 58 | 32 | 26 | 676 |

| | |
|---|---|
| Sum | 3216 |
| Divide by 16 | 201 |
| | Variance |

**Q5. Explain how to calculate sample variance in Excel.**

*Ans :*

A sample is a set of data extracted from the entire population. And the variance calculated from a sample is called sample variance.

**For example,** if you want to know how people's heights vary, it would be technically unfeasible for you to measure every person on the earth. The solution is to take a sample of the population, say 1,000 people, and estimate the heights of the whole population based on that sample.

Sample variance is calculated with this formula:

$$\text{Sample variance} = \frac{\sum(x-\bar{x})^2}{(n-1)}$$

Where,

- $\bar{x}$ is the mean (simple average) of the sample values.
- n is the sample size, i.e. the number of values in the sample.

There are 3 functions to find sample variance in Excel: VAR, VAR.S and VARA.

VAR function in Excel

It is the oldest Excel function to estimate variance based on a sample. The VAR function is available in all versions of Excel 2000 to 2019.

**VAR(number1, [number2], ...)**

**Note:** In Excel 2010, the VAR function was replaced with VAR.S that provides improved accuracy. Although VAR is still available for backward compatibility, it is recommended to use VAR.S in the current versions of Excel.

**VAR.S function in Excel**

It is the modern counterpart of the Excel VAR function. Use the VAR.S function to find sample variance in Excel 2010 and later.

**VAR.S(number1, [number2], ...)**

VARA function in Excel

The Excel VARA function returns a sample variance based on a set of numbers, text, and logical values as shown in this table.

**VARA(value1, [value2], ...)**

**Sample variance formula in Excel**

When working with a numeric set of data you can use any of the above functions to calculate sample variance in Excel.

As an example, let's find the variance of a sample consisting of 6 items (B2:B7). For this, you can use one of the below formulas:

=VAR(B2:B7)

=VARS(B2:B7)

=VARA(B2:B7)

As shown in the screenshot, all the formulas return the same result (rounded to 2 decimal places):

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Student | Score | | Sample variance | | | |
| 2 | Daniela | 85 | | VAR | 34.97 | =VAR(B2:B7) | |
| 3 | Tommy | 79 | | VAR.S | 34.97 | =VAR.S(B2:B7) | |
| 4 | Edward | 90 | | VARA | 34.97 | =VARA(B2:B7) | |
| 5 | Julia | 88 | | | | | |
| 6 | Timothy | 88 | | | | | |
| 7 | Peter | 75 | | | | | |

To check the result, let's do var calculation manually:

1. Find the mean by using the AVERAGE function:

   **=AVERAGE(B2:B7)**

   The average goes to any empty cell, say B8.

2. Subtract the average from each number in the sample:

   **=B2-$B$8**

The differences go to column C, beginning in C2.

3. Square each difference and put the results to column D, beginning in D2:

**=C2 ^ 2**

4. Add up the squared differences and divide the result by the number of items in the sample minus 1:

**=SUM(D2:D7)/(6-1)**

As you can see, the result of our manual var calculation is exactly the same as the number returned by Excel's built-in functions:

C2 | =B2-$B$8 | =C2^2

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Student | Score | Dif | Dif² | | Sample variance | | | |
| 2 | Daniela | 85 | 0.83 | 0.69 | | VAR | 34.97 | =VAR(B2:B7) | |
| 3 | Tommy | 79 | -5.17 | 26.69 | | VAR.S | 34.97 | =VAR.S(B2:B7) | |
| 4 | Edward | 90 | 5.83 | 34.03 | | VARA | 34.97 | =VARA(B2:B7) | |
| 5 | Julia | 88 | 3.83 | 14.69 | | | | | |
| 6 | Timothy | 88 | 3.83 | 14.69 | | Manual | 34.97 | =SUM(D2:D7)/5 | |
| 7 | Peter | 75 | -9.17 | 84.03 | | | | | |
| 8 | Average | 84.17 | =AVERAGE(B2:B7) | | | | | | |

If your data set contains the Boolean and/or text values, the VARA function will return a different result. The reason is that VAR and VAR.S ignore any values other than numbers in references, while VARA evaluates text values as zeros, TRUE as 1, and FALSE as 0. So, please carefully choose the variance function for your calculations depending on whether you want to process or ignore text and logicals.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Student | Score | | Sample variance | | | |
| 2 | Daniela | 85 | | VAR | 39.30 | =VAR(B2:B7) | |
| 3 | Tommy | 79 | | VAR.S | 39.30 | =VAR.S(B2:B7) | |
| 4 | Edward | 90 | | VARA | 1190.70 | =VARA(B2:B7) | |
| 5 | Julia | 88 | | | | | |
| 6 | Timothy | N/A | | | | | |
| 7 | Peter | 75 | | | | | |

**Q6. Explain how to calculate population variance in Excel.**

*Ans :* **(Imp.)**

Population is all members of a given group, i.e. all observations in the field of study. Population variance describes how data points in the entire population are spread out.

The population variance can be found with this formula:

Population variance = $\frac{\sum(x-\bar{x})^2}{n}$

Where,

- $\bar{x}$ is the mean of the population.
- n is the population size, i.e. the total number of values in the population.

There are 3 functions to calculate population variance in Excel: VARP, VAR.P and VARPA.

VARP function in Excel

The Excel VARP function returns the variance of a population based on the entire set of numbers. It is available in all versions of Excel 2000 to 2019.

**VARP(number1, [number2], ...)**

**Note.** In Excel 2010, VARP was replaced with VAR.P but is still kept for backward compatibility. It is recommended to use VAR.P in the current versions of Excel because there is no guarantee that the VARP function will be available in future versions of Excel.

VAR.P function in Excel

It is an improved version of the VARP function available in Excel 2010 and later.

**VAR.P(number1, [number2], ...)**

VARPA function in Excel

The VARPA function calculates the variance of a population based on the entire set of numbers, text, and logical values. It is available in all version of Excel 2000 through 2019.

**VARA(value1, [value2], ...)**

**Population variance formula in Excel**

In the sample var calculation example, we found a variance of 5 exam scores assuming those scores were a selection from a bigger group of students. If you collect data on all the students in the group, that data will represent the entire population, and you will calculate a population variance by using the above functions.

Let's say, we have the exam scores of a group of 10 students (B2:B11). The scores constitute the entire population, so we will do variance with these formulas:

=VARP(B2:B11)

=VAR.P(B2:B11)

=VARPA(B2:B11)

And all the formulas will return the identical result:

Rahul Publications

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Student | Score | | Population variance | | | |
| 2 | Daniela | 85 | | VARP | 36.41 | =VARP(B2:B11) | |
| 3 | Tommy | 79 | | VAR.P | 36.41 | =VAR.P(B2:B11) | |
| 4 | Edward | 90 | | VARPA | 36.41 | =VARPA(B2:B11) | |
| 5 | Julia | 88 | | | | | |
| 6 | Timothy | 88 | | | | | |
| 7 | Peter | 75 | | | | | |
| 8 | Neal | 92 | | | | | |
| 9 | Sally | 74 | | | | | |
| 10 | Mike | 83 | | | | | |
| 11 | Adam | 89 | | | | | |

To make sure Excel has done the variance right, you can check it with the manual var calculation formula shown in the screenshot below:

C2 : =B2-$B$12 =C2^2

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Student | Score | Dif | Dif$^2$ | | Population variance | | | |
| 2 | Daniela | 85 | 0.70 | 0.49 | | VARP | 36.41 | =VARP(B2:B11) | |
| 3 | Tommy | 79 | -5.30 | 28.09 | | VAR.P | 36.41 | =VAR.P(B2:B11) | |
| 4 | Edward | 90 | 5.70 | 32.49 | | VARPA | 36.41 | =VARPA(B2:B11) | |
| 5 | Julia | 88 | 3.70 | 13.69 | | | | | |
| 6 | Timothy | 88 | 3.70 | 13.69 | | Manual | 36.41 | =SUM(D2:D11)/10 | |
| 7 | Peter | 75 | -9.30 | 86.49 | | | | | |
| 8 | Neal | 92 | 7.70 | 59.29 | | | | | |
| 9 | Sally | 74 | -10.30 | 106.09 | | | | | |
| 10 | Mike | 83 | -1.30 | 1.69 | | | | | |
| 11 | Adam | 89 | 4.70 | 22.09 | | | | | |
| 12 | Average | 84.30 | =AVERAGE(B2:B11) | | | | | | |

If some of the students did not take the exam and have N/A instead of a score number, the VARPA function will return a different result. The reason is that VARPA evaluates text values as zeros while VARP and VAR.P ignore text and logical values in references. Please see VAR.P vs. VARPA for full details.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Student | Score | | Population variance | | | |
| 2 | Daniela | 85 | | VARP | 43.50 | =VARP(B2:B11) | |
| 3 | Tommy | 79 | | VAR.P | 43.50 | =VAR.P(B2:B11) | |
| 4 | Edward | 90 | | VARPA | 1163.76 | =VARPA(B2:B11) | |
| 5 | Julia | 88 | | | | | |
| 6 | Timothy | N/A | | | | | |
| 7 | Peter | 75 | | | | | |
| 8 | Neal | 92 | | | | | |
| 9 | Sally | 74 | | | | | |
| 10 | Mike | N/A | | | | | |
| 11 | Adam | 89 | | | | | |

### 3.1.3 Standard Deviation

**Q7. What is standard deviation? Explain with an examples how is Standard Deviation calculated.**

*Ans :* **(Imp.)**

Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a "typical" deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set. Like the variance, if the data points are close to the mean, there is a small variation whereas the data points are highly spread out from the mean, then it has a high variance. Standard deviation calculates the extent to which the values differ from the average. Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore a change in even one value affects the value of standard deviation. It is independent of origin but not of scale. It is also useful in certain advanced statistical problems.

**Formula**

The population standard deviation formula is given as:

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

Here,

$\sigma$ = Population standard deviation

N = Number of observations in population

$X_i$ = ith observation in the population

$\mu$ = Population mean

Similarly, the sample standard deviation formula is:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)}$$

Here,

s = Sample standard deviation

n = Number of observations in sample

xi = ith observation in the sample

$\bar{x}$ = Sample mean

**Calculating Standard deviation**

The formula for standard deviation makes use of three variables. The first variable is the value of each point within a data set, with a sum-number indicating each additional variable (x, $x_1$, $x_2$, $x_3$, etc). The mean is applied to the values of the variable M and the number of data that is assigned to the variable n. Variance is the average of the values of squared differences from the arithmetic mean.

To calculate the mean value, the values of the data elements have to be added together and the total is divided by the number of data entities that were involved.

Standard deviation, denoted by the symbol ó, describes the square root of the mean of the squares of all the values of a series derived from the arithmetic mean which is also called the root-mean-square deviation. 0 is the smallest value of standard deviation since it cannot be negative. When the elements in a series are more isolated from the mean, then the standard deviation is also large.

The statistical tool of standard deviation is the measures of dispersion that computes the erraticism of the dispersion among the data. For instance, mean, median and mode are the measures of central tendency. Therefore, these are considered to be the central first order averages. The measures of dispersion that are mentioned directly over are averages of deviations that result from the average values, therefore these are called second-order averages.

**Example :**

Let's calculate the standard deviation for the number of gold coins on a ship run by pirates.

There are a total of 100 pirates on the ship. Statistically, it means that the population is 100. We use the standard deviation equation for the entire population if we know a number of gold coins every pirate has.

Statistically, let's consider a sample of 5 and here you can use the standard deviation equation for this sample population.

This means we have a sample size of 5 and in this case, we use the standard deviation equation for the sample of a population.

Consider the number of gold coins 5 pirates have; 4, 2, 5, 8, 6.

**Mean:**

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

$$= \frac{x_1 + x_2 + x_3 + x_4 ..... + x_n}{n}$$

$= (4 + 2 + 5 + 6 + 8)/5$

$= 5$

$x_n - \bar{x}$

for every value of the sample:

$x_1 - \bar{x} = 4 - 5 = -1$

$x_2 - x = 2 - 5 = -3$

$x_3 - x = 5 - 5 = 0$

$x_4 - x = 8 - 5 = 3$

$x_5 - \bar{x} = 6 - 5 = 1$

$\sum (x_n - \bar{x})^2$

$= [x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_5 - \bar{x})$

$= (-l + (-3)^2 + 0^2 + 3^2 + l^2$

$= 20$

**Standard deviation:**

$$S.D = \sqrt{\frac{\sum (x_n - x)^2}{n-1}}$$

$$= \sqrt{\frac{20}{4}}$$

$$= \sqrt{5}$$

$$= 2.236$$

**Standard deviation of Grouped Data**

In case of grouped data or grouped frequency distribution, the standard deviation can be found by considering the frequency of data values. This can be understood with the help of an example.

**Example 1:**

Calculate the mean, variance and standard deviation for the following data:

| Class Interval | 0 – 10 | 10 – 20 | 20 – 30 | 30 – 40 | 40 – 50 | 50 – 60 |
|---|---|---|---|---|---|---|
| Frequency | 27 | 10 | 7 | 5 | 4 | 2 |

Rahul Publications

*Sol :*

| Class Interval | Frequency (f) | Mid Value ($x_i$) | $fx_i$ | $fx_i^2$ |
|---|---|---|---|---|
| 0 – 10 | 27 | 5 | 135 | 675 |
| 10 – 20 | 10 | 15 | 150 | 2250 |
| 20 – 30 | 7 | 25 | 175 | 4375 |
| 30 – 40 | 5 | 35 | 175 | 6125 |
| 40 – 50 | 4 | 45 | 180 | 8100 |
| 50 – 60 | 2 | 55 | 110 | 6050 |
| | $\sum f = 55$ | | $\sum fx_i = 925$ | $\sum fx_i^2 = 27575$ |

$N = \sum f = 55$

Mean $= \left(\sum fx_i\right)/N = 925/55 = 16.818$

Variance $= 1/(N-1)\ [\Sigma fx_i^2 - 1/N(\Sigma fx_i)^2]$

$= 1/(55-1)\ [27575 - (1/55)\ (925)^2]$

$= (1/54)\ [27575 - 15556.8182]$

$= 222.559$

Standard deviation $= \sqrt{\text{variance}} = \sqrt{222.559} = 14.918$

**Example 2 :**

**Let's calculate the standard deviation for the data given below:**

| | | | | | |
|---|---|---|---|---|---|
| xi | 6 | 8 | 10 | 12 | 14 |
| fi | 2 | 3 | 4 | 5 | 4 |

Calculate mean($\bar{x}$): $(6+8+10+12+14)/5 = 10$

| $x_i$ | $f_i$ | $f_ix_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $ff_i(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 6 | 2 | 12 | –4 | 16 | 32 |
| 8 | 3 | 24 | –2 | 4 | 12 |
| 10 | 4 | 40 | 0 | 0 | 0 |
| 12 | 5 | 60 | 2 | 4 | 20 |
| 14 | 4 | 56 | 4 | 16 | 64 |
| | 18 | 192 | | | 128 |

$N = 18, \sum fixi = 192, \sum fi(x_i - \bar{x})^2 = 128$

Calculate variance: $\sigma^2 = 1/N \sum i = 1Nfi(X_i - \bar{x})^2$

$= 1/18 \times 128 = 7.1$

Calculate SD: $\sigma$ = "Variance = $\sqrt{7.1} = 2.66$

### 3.1.4 Coefficient of Variation

**Q8. Explain the concept of coefficient of variation with examples.**

*Ans :*

The coefficient of variation is a type of measure of dispersion. A measure of dispersion is a quantity that is used to gauge the extent of variability of data. Thus, the coefficient of variation is used to measure the dispersion of data from the average or the mean value. CV is the abbreviated form of the coefficient of variation.

The formula for coefficient of variation is given below:

$$\text{coefficient of variation} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100\%$$

As per sample and population data type, the formula for standard deviation may vary.

$$\text{Sample Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

$$\text{Population Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$$

Where,

$x_i$ = Terms given in the data

$\bar{x}$ = Mean

n = Total number of terms.

**Example 1:**

A researcher is comparing two multiple-choice tests with different conditions. In the first test, a typical multiple-choice test is administered. In the second test, alternative choices (i.e. incorrect answers) are randomly assigned to test takers. The results from the two tests are:

| | Regular Test | Randomized Answer |
|---|---|---|
| **Mean** | 59.9 | 44.8 |
| **SD** | 10.2 | 12.7 |

Trying to compare the two test results is challenging. Comparing standard deviations doesn't really work, because the means are also different. Calculation using the formula CV=(SD/Mean)*100 helps to make sense of the data:

| | Regular Test | Randomized Answer |
|---|---|---|
| **Mean** | 59.9 | 44.8 |
| **SD** | 10.2 | 12.7 |
| **CV** | 17.03 | 28.35 |

Looking at the standard deviations 10.2 and 12.7, you might think that the tests have similar results. However, when you adjust for the difference in the means, the results have more significance:

Regular test: CV = 17.03

Randomized answers: CV = 28.35

**Example 2:**

Find the coefficient of variation of the following sample set of numbers.

{1, 5, 6, 8, 10, 40, 65, 88}.

*Sol :*

Given sample set: {1, 5, 6, 8, 10, 40, 65, 88}.

Sample mean = (1 + 5 + 6 + 8 + 10 + 40 + 65 + 88)/8 = 223/8 = 27.875

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (1 - 27.875)^2 + (5 - 27.875)^2 + (6 - 27.875)^2 + (8 - 27.875) + (10 - 27.875)^2 + (40 - 27.875)^2 + (65 - 27.875)^2 + (88 - 27.875)^2$$

$$= 27.875$$

**Variance:**

$$\frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1} = \frac{7578.875}{7} = 1082.696$$

Standard deviation:

$$\sigma = \sqrt{\sum_{i=1}^{n}\frac{(x_i - x)^2}{n-1}} = \sqrt{1082.696} = 32.904$$

Coefficient of variation = 32.901/27.875 = 1.180

**Q9. Explain how to Calculate the Coefficient of Variation in Excel.**

*Ans :*

**Step 1:**

Calculate the standard deviation of all the points in the series. It can be calculated using any of the below given three functions based upon our requirements:

**(i) STDEV():** Used for calculating the standard deviation of a general series.

**(ii) STDEV.P():** Used for calculating the standard deviation of a population.

**(iii) STDEV.S():** Used for calculating the standard deviation of a sample.

Rahul Publications

B23 =STDEV(A2:A21)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Age (in Years) | | | | |
| 2 | 35 | | | | |
| 3 | 25 | | | | |
| 4 | 43 | | | | |
| 5 | 78 | | | | |
| 6 | 22 | | | | |
| 7 | 73 | | | | |
| 8 | 53 | | | | |
| 9 | 24 | | | | |
| 10 | 18 | | | | |
| 11 | 41 | | | | |
| 12 | 19 | | | | |
| 13 | 22 | | | | |
| 14 | 47 | | | | |
| 15 | 56 | | | | |
| 16 | 64 | | | | |
| 17 | 41 | | | | |
| 18 | 38 | | | | |
| 19 | 55 | | | | |
| 20 | 51 | | | | |
| 21 | 47 | | | | |
| 22 | | | | | |
| 23 | Standard Deviation | 17.67662745 | | | |
| 24 | Mean | | | | |
| 25 | Coefficent of Variation (CV) | | | | |
| 26 | | | | | |

**Step 2:**

Calculate the arithmetic mean of all the data points in the series using the AVERAGE() function.

B24 =AVERAGE(A2:A21)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Age (in Years) | | | | |
| 2 | 35 | | | | |
| 3 | 25 | | | | |
| 4 | 43 | | | | |
| 5 | 78 | | | | |
| 6 | 22 | | | | |
| 7 | 73 | | | | |
| 8 | 53 | | | | |
| 9 | 24 | | | | |
| 10 | 18 | | | | |
| 11 | 41 | | | | |
| 12 | 19 | | | | |
| 13 | 22 | | | | |
| 14 | 47 | | | | |
| 15 | 56 | | | | |
| 16 | 64 | | | | |
| 17 | 41 | | | | |
| 18 | 38 | | | | |
| 19 | 55 | | | | |
| 20 | 51 | | | | |
| 21 | 47 | | | | |
| 22 | | | | | |
| 23 | Standard Deviation | 17.67662745 | | | |
| 24 | Mean | 42.6 | | | |
| 25 | Coefficent of Variation (CV) | | | | |
| 26 | | | | | |

Rahul Publications

**Step 3:**

Calculate the ratio by dividing the standard deviation value by the mean value.

B25 | fx =B23/B24

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Age (in Years) | | | | |
| 2 | 35 | | | | |
| 3 | 25 | | | | |
| 4 | 43 | | | | |
| 5 | 78 | | | | |
| 6 | 22 | | | | |
| 7 | 73 | | | | |
| 8 | 53 | | | | |
| 9 | 24 | | | | |
| 10 | 18 | | | | |
| 11 | 41 | | | | |
| 12 | 19 | | | | |
| 13 | 22 | | | | |
| 14 | 47 | | | | |
| 15 | 56 | | | | |
| 16 | 64 | | | | |
| 17 | 41 | | | | |
| 18 | 38 | | | | |
| 19 | 55 | | | | |
| 20 | 51 | | | | |
| 21 | 47 | | | | |
| 22 | | | | | |
| 23 | Standard Deviation | 17.67662745 | | | |
| 24 | Mean | 42.6 | | | |
| 25 | Coefficent of Variation (CV) | 0.414944306 | | | |
| 26 | | | | | |

### 3.1.5 Percentiles

**Q10. How to find percentiles in statistics? Explain with an example.**

*Ans :*

Percentile is defined as the value below which a given percentage falls under. For example, in a group of 20 children, Ben is the 4th tallest and 80% of the children are shorter than you. Hence, it means that Ben is at the 80th percentile. It is most commonly used in competitive exams such as SAT, LSAT, etc.

The Percentile Formula is given as,

Percentile = (Number of Values Below "x" / Total Number of Values) × 100

Another formula to find the percentile is given by:

$$P = (n/N) \times 100$$

P = (nth percentile/100) × Total number of values in the list

Here,

n = Ordinal rank of the given value or value below the number

N = Number of values in the data set

P = Percentile

Rank = Percentile/100

Ordinal rank for Percentile value = Rank × Total number of values in the list

**Example 1:**

The scores for student are 40, 45, 49, 53, 61, 65, 71, 79, 85, 91. What is the percentile for score 71?

*Sol :*

Given,

No. of.scores below 71 = 6

Total no. of.scores = 10

The formula for percentile is given as,

Percentile = (Number of Values Below "x" / Total Number of Values) × 100

Percentile of 71

= (6/10) × 100

= 0.6 × 100 = 60

**Example 2:**

Consider the list {50, 45, 60, 25, 30}. Find the 5th, 30th, 40th, 50th and 100th percentiles of the list given.

*Sol :*

Given list – 50, 45, 60, 25, 30

Ordered list – 25, 30, 45, 50, 60

N = 5

| Percentile (P) | Ordinal rank | Percentile value |
|---|---|---|
| 5th | (5/100) × 5 = [0.25] = 1 | 1st number in the ordered list = 25 |
| 30th | (30/100) × 5 = [1.5] = 2 | 2nd number in the ordered list = 30 |
| 40th | (40/100) × 5 = 2 | 2nd number in the ordered list = 30 |
| 50th | (50/100) × 5 = [2.5] = 3 | 3rd number in the ordered list = 45 |
| 100th | (100/100) × 5 = 5 | 5th number in the ordered list = 60 |

**Q11. Explain, how to find percentile in excel.**

*Ans :*

There are three variations of the percentile function available in Excel. If you're using Excel 2010 or versions after that, you will have access to all these three functions.

- **PERCENTILE** – this is old function that is now kept for backward compatibility purposes. You can use this, but it's best to use the new ones (if you have those in your version of Excel). The result of this function is a value between 0 and 1
- **PERCENTILE.INC** – this is the new formula (which works exactly like the PERCENTILE function). In most cases, this is the function you would need to use. The result of this function is a value between 0 and 1
- **PERCENTILE.EXC** – this also works like the PERCENTILE.INC function with one difference – the result of this function will be a value between 0 and 1, but excludes K values between 0 to 1/(N+1) as well as N/(N+1) to 1 (where N is the size of the sample)

Below is the syntax of the PERCENTILE.INC function in Excel:

=PERCENTILE.INC(array,k)

where:

1. array is the range of cells where you have the values for which you want to find out the K-th percentile
2. k is the value between 0 and 1, and gives you the k-th percentile value. For example, if you want to calculate 90th percentile value, this would be 0.9 or 90%, and for 50th percntile value, this would be 0.5 or 50%

The syntax remains the same for the PERCENTILE and PERCENTILE.EXC functions.

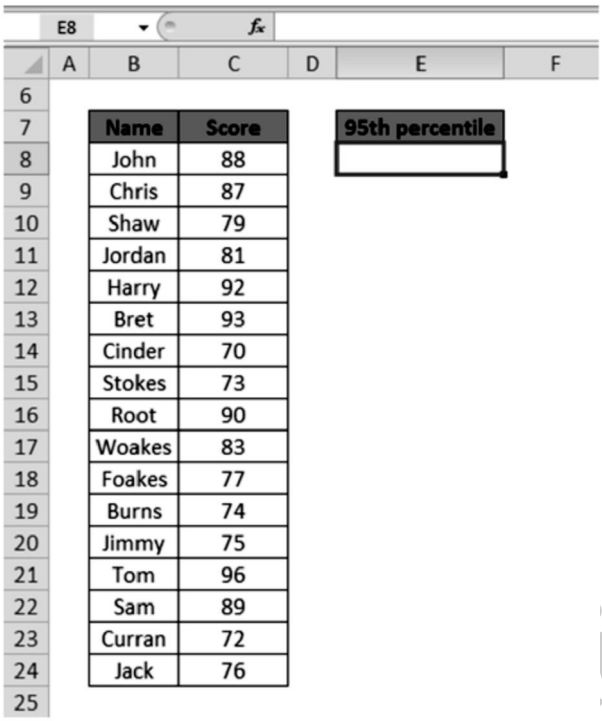
PERCENTILE in Excel – Example #1

PERCENTILE Function To Calculate 95th Percentile In Entrance Exam

In the below-mentioned table, it contains student name in the column B (B8 to B24) &Their score in column C (C8 to C24) I need to find out the score for the 95th percentile

| Name | Score |
|---|---|
| John | 88 |
| Chris | 87 |
| Shaw | 79 |
| Jordan | 81 |
| Harry | 92 |
| Bret | 93 |
| Cinder | 70 |
| Stokes | 73 |
| Root | 90 |
| Woakes | 83 |
| Foakes | 77 |
| Burns | 74 |
| Jimmy | 75 |
| Tom | 96 |
| Sam | 89 |
| Curran | 72 |
| Jack | 76 |

Let's apply the PERCENTILE function in cell "E8". Select the cell "E8". where the PERCENTILE function needs to be applied.

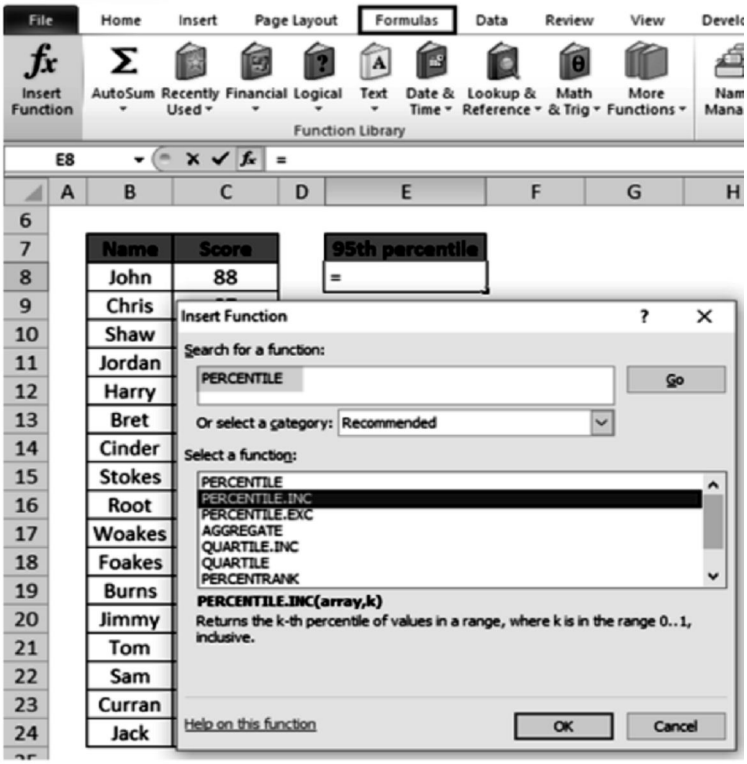


| Name | Score |
|---|---|
| John | 88 |
| Chris | 87 |
| Shaw | 79 |
| Jordan | 81 |
| Harry | 92 |
| Bret | 93 |
| Cinder | 70 |
| Stokes | 73 |
| Root | 90 |
| Woakes | 83 |
| Foakes | 77 |
| Burns | 74 |
| Jimmy | 75 |
| Tom | 96 |
| Sam | 89 |
| Curran | 72 |
| Jack | 76 |

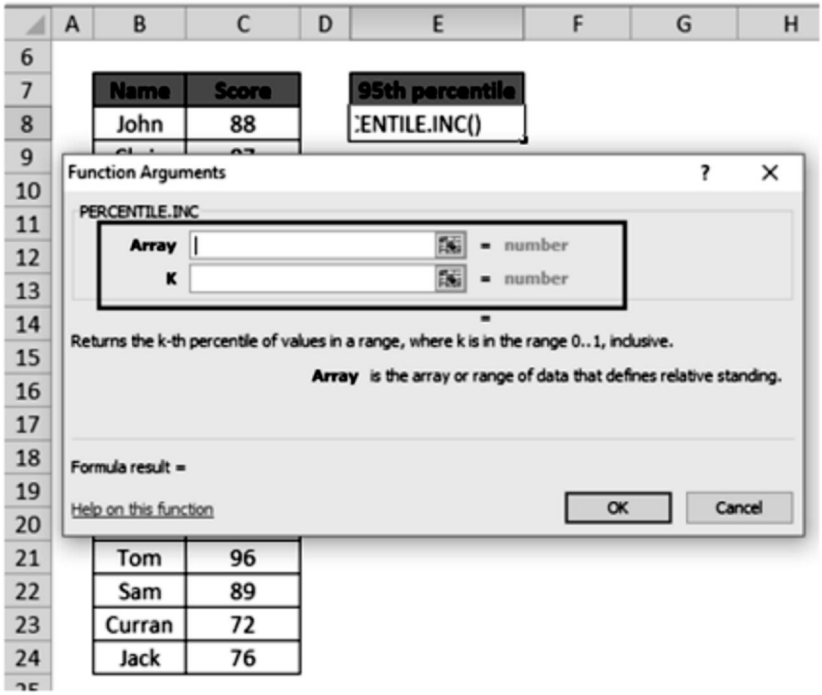
95th percentile

Click the insert function button (fx) under the formula toolbar, the dialog box will appear, type the keyword "PERCENTILE" in the search for a function box, the PERCENTILE function will appear in select a function box. Three options appear in select a function box, i.e. PERCENTILE, PERCENTILE.EXC and PERCENTILE.INC function.

Double click on PERCENTILE.INC.



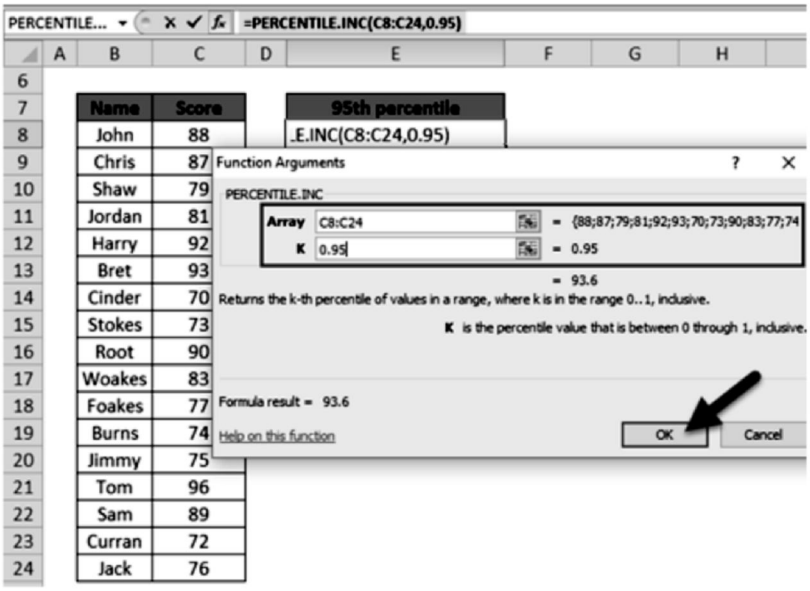
A dialog box appears where arguments for PERCENTILE.INC function needs to be filled or entered, i.e. =PERCENTILE (array, k)




i.e. **=PERCENTILE.INC(C8:C24,0.95)** Here, the score data is present in the range (C8 to C24) for which we need to apply PERCENTILE.INC function

To enter Array argument, click inside cell C8 and you'll see the cell selected, then Select the cells till C24. So that column range will get selected, i.e. C8:C24

K, it is the percentile value we are looking for. Here I need to find out the value for the 95th percentile so that I will use "0.95" as the percentile value. (K is any percentage expressed as a decimal, i.e. 0.95 for 95%) Click ok.



After entering both the arguments. **=PERCENTILE.INC (C8:C24,0.95),** i.e. returns the score for the 95th percentile, i.e. 93.6, as a result in the cell E8. 95th percentile falls between 93 & 96. Excel has interpolated between 93 to 96 score to produce the result 93.6.

E8 =PERCENTILE.INC(C8:C24,0.95)

| Name | Score |
|---|---|
| John | 88 |
| Chris | 87 |
| Shaw | 79 |
| Jordan | 81 |
| Harry | 92 |
| Bret | 93 |
| Cinder | 70 |
| Stokes | 73 |
| Root | 90 |
| Woakes | 83 |
| Foakes | 77 |
| Burns | 74 |
| Jimmy | 75 |
| Tom | 96 |
| Sam | 89 |
| Curran | 72 |
| Jack | 76 |

| 95th percentile |
|---|
| 93.6 |

### 3.1.6 Quartiles

**Q12. Define Quartile? How to find the quartile deviation in statistics? Explain with an examples.**

*Ans :* **(Imp.)**

Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by $Q_1$, $Q_2$ and $Q_3$, respectively. $Q_2$ is nothing but the median, since it indicates the position of the item in the list and thus, is a positional average. To find quartiles of a group of data, we have to arrange the data in ascending order.

**Quartiles Formula**

Suppose, $Q_3$ is the upper quartile is the median of the upper half of the data set. Whereas, $Q_1$ is the lower quartile and median of the lower half of the data set. $Q_2$ is the median. Consider, we have n number of items in a data set. Then the quartiles are given by;

$Q_1 = [(n+1)/4]$th item

$Q_2 = [(n+1)/2]$th item

$Q_3 = [3(n+1)/4]$th item

Hence, the formula for quartile can be given by;

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f}(l_2 - l_1)$$

Where, Qr is the $r^{th}$ quartile

$l_1$ is the lower limit

$l_2$ is the upper limit

f is the frequency

c is the cumulative frequency of the class preceding the quartile class.

### Quartiles in Statistics

Similar to the median which divides the data into half so that 50% of the estimation lies below the median and 50% lies above it, the quartile splits the data into quarters so that 25% of the estimation are less than the lower quartile, 50% of estimation are less than the mean, and 75% of estimation are less than the upper quartile. Usually, the data is ordered from smallest to largest:

- First quartile: 25% from smallest to largest of numbers
- Second quartile: between 25.1% and 50% (till median)
- Third quartile: 51% to 75% (above the median)
- Fourth quartile: 25% of largest numbers

### Quartile Deviation

You have learned about standard deviation in statistics. Quartile deviation is defined as half of the distance between the third and the first quartile. It is also called Semi Interquartile range. If $Q_1$ is the first quartile and $Q_3$ is the third quartile, then the formula for deviation is given by;

**Quartile deviation = $(Q_3 - Q_1)/2$**

### Interquartile Range

The interquartile range (IQR) is the difference between the upper and lower quartile of a given data set and is also called a midspread. It is a measure of statistical distribution, which is equal to the difference between the upper and lower quartiles. Also, it is a calculation of variation while dividing a data set into quartiles. If $Q_1$ is the first quartile and $Q_3$ is the third quartile, then the IQR formula is given by;

$IQR = Q_3 - Q_1$

Rahul Publications

**Example 1:**

Find the quartiles of the following data: 4, 6, 7, 8, 10, 23, 34.

***Sol :***

Here the numbers are arranged in the ascending order and number of items, n = 7

Lower quartile, $Q_1$ = [(n+1)/4] th item

$Q_1$= 7+1/4 = 2nd item = 6

Median, $Q_2$ = [(n+1)/2]th item

$Q_2$= 7+1/2 item = 4th item = 8

Upper Quartile, $Q_3$ = [3(n+1)/4]th item

$Q_3$ = 3(7+1)/4 item = 6th item = 23

**Example 2:**

Find the Quartiles of the following age:-

23, 13, 37, 16, 26, 35, 26, 35

*Sol :*

First, we need to arrange the numbers in increasing order.

Therefore, 13, 16, 23, 26, 26, 35, 35, 37

Number of items, n = 8

Lower quartile, $Q_1$ = [(n+1)/4] th item

$Q_1$ = 8+1/4 = 9/4 = 2.25th term

From the quartile formula we can write;

$Q_1$ = 2nd term + 0.25(3rd term-2nd term)

$Q_1$= 16+0.25(23-26) = 15.25

Similarly,

Median, $Q_2$ = [(n+1)/2]th item

$Q_2$ = 8+1/2 = 9/2 = 4.5

$Q_2$ = 4th term+0.5 (5th term-4th term)

$Q_2$= 26+0.5(26 – 26) = 26

And,

Upper Quartile, $Q_3$ = [3(n+1)/4]th item

$Q_3$ = 3(8+1)/4 = 6.75th term

$Q_3$ = 6th term + 0.75(7th term-6th term)

$Q_3$ = 35+0.75(35-35) = 35

Rahul Publications

## 3.2 Analysing Distributions

### 3.2.1 Empirical Rule

**Q13. What is the empirical rule? How and where the empirical rule is used ? Explain with an example.**

*Ans :* **(Imp.)**

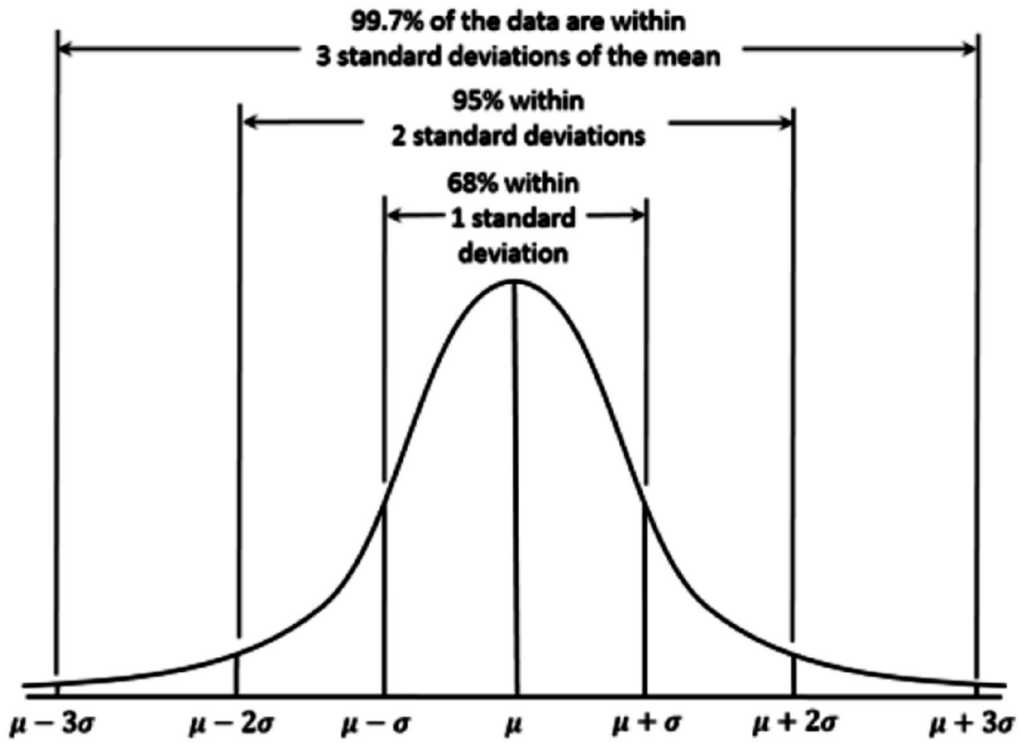
The empirical rule is a statistical rule (also called the three-sigma rule or the 68-95-99.7 rule) which states that, for normally distributed data, almost all of the data will fall within three standard deviations either side of the mean.

More specifically, you'll find:

- **68% of data within 1 standard deviation**
- **95% of data within 2 standard deviations**
- **99.7% of data within 3 standard deviations**

Standard deviation is a measure of spread; it tells how much the data varies from the average, i.e., how diverse the dataset is. The smaller value, the more narrow the range of data is.

Normal distribution is a distribution that is symmetric about the mean, with data near the mean are more frequent in occurrence than data far from the mean. In graphical form, normal distributions appear as a bell shaped curve, as you can see below:




The empirical rule - formula

The algorithm below explains how to use the empirical rule:

**1. Calculate the mean of your values:**

$\mu = (\Sigma x_i) / n$

- $\Sigma$ - sum
- $x_i$ - each individual value from your data
- n - the number of samples

**2. Calculate the standard deviation:**

$$\sigma = \sqrt{\left(\sum x_i - \mu^2\right) / (n-1))}$$

**3. Apply the empirical rule formula:**

- 68% of data falls within 1 standard deviation from the mean - that means between $\mu - \sigma$ and $\mu + \sigma$.
- 95% of data falls within 2 standard deviations from the mean - between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- 99.7% of data falls within 3 standard deviations from the mean - between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Enter the mean and standard deviation into the empirical rule calculator, and it will output the intervals for you.

**An example of how to use the empirical rule**

Intelligence quotient (IQ) scores are normally distributed with the mean of 100 and the standard deviation equal to 15. Let's have a look at the maths behind the 68 95 99 rule calculator:

1. Mean: $\mu = 100$
2. Standard deviation: $\sigma = 15$
3. Empirical rule formula:

   $\mu - \sigma = 100 - 15 = 85$

   $\mu + \sigma = 100 + 15 = 115$

   68% of people have an IQ between 85 and 115.

   $\mu - 2\sigma = 100 - 2*15 = 70$

   $\mu + 2\sigma = 100 + 2*15 = 130$

   95% of people have an IQ between 70 and 130.

   $\mu - 3\sigma = 100 - 3*15 = 55$

   $\mu + 3\sigma = 100 + 3*15 = 145$

   99.7% of people have an IQ between 55 and 145.

For quicker and easier calculations, input the mean and standard deviation into this empirical rule calculator, and watch as it does the rest for you.

**Uses**

The rule is widely used in empirical research, such as when calculating the probability of a certain piece of data occurring, or for forecasting outcomes when not all data is available. It gives insight into the characteristics of a population without the need to test everyone and helps to determine whether a given data set is normally distributed. It is also used to find outliers – results that differ significantly from others - which may be the result of experimental errors.

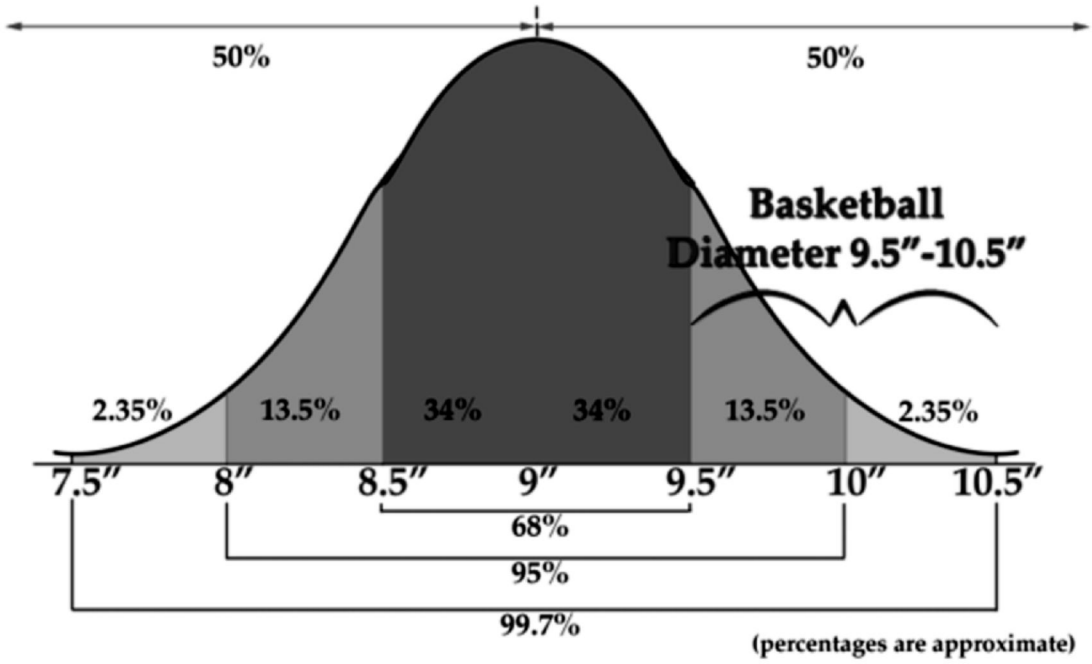
Rahul Publications

**Example**

If the diameter of a basketball is normally distributed, with a mean ($\mu$) of 93 , and a standard deviation ($\sigma$) of 0.53 , what is the probability that a randomly chosen basketball will have a diameter between 9.53 and 10.53?

*Sol :*

Since the $\sigma$ = 0.53 and the $\mu$ = 93 , we are evaluating the probability that a randomly chosen ball will have a diameter between 1 and 3 standard deviations above the mean. The graphic below shows the portion of the normal distribution included between 1 and 3 SDs:




The percentage of the data spanning the 2nd and 3rd SDs is 13.5% + 2.35% = 15.85%

The probability that a randomly chosen basketball will have a diameter between 9.5 and 10.5 inches is 15.85%.

**Q14. Explain how to apply the Empirical Rule in Excel.**

*Ans :* **(Imp.)**

Suppose we have a normally-distributed dataset with a mean of 7 and a standard deviation of 2.2. The following screenshot shows how to apply the Empirical Rule to this dataset in Excel to find which values 68% of the data falls between, which values 95% of the data falls between, and which values 99.7% of the data falls between:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | | mean | 7 | | | | |
| 3 | | standard deviation | 2.2 | | | | |
| 4 | | | | | | | |
| 5 | | | | | | | |
| 6 | | 68% of data falls between: | 4.8 | 9.2 | | =$C$2 - $C$3 | =$C$2 + $C$3 |
| 7 | | 95% of data falls between: | 2.6 | 11.4 | | =$C$2 - 2*$C$3 | =$C$2 + 2*$C$3 |
| 8 | | 99.7% of data falls between: | 0.4 | 13.6 | | =$C$2 - 3*$C$3 | =$C$2 + 3*$C$3 |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | | | | | | | |
| 16 | | | | | | | |
| 17 | | | | | | | |
| 18 | | | | | | | |
| 19 | | | | | | | |

From this output, we can see:

- 68% of the data falls between 4.8 and 9.2
- 95% of the data falls between 2.6 and 11.4
- 99.7% of the data falls between 0.4 and 13.6

The cells in columns F and G show the formulas that were used to find these values.

To apply the Empirical Rule to a different dataset, we simply need to change the mean and standard deviation in cells C2 and C3. For example, here is how to apply the Empirical Rule to a dataset with a mean of 40 and a standard deviation of 3.75:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | | mean | 40 | | | | |
| 3 | | standard deviation | 3.75 | | | | |
| 4 | | | | | | | |
| 5 | | | | | | | |
| 6 | | 68% of data falls between: | 36.25 | 43.75 | | =$C$2 - $C$3 | =$C$2 + $C$3 |
| 7 | | 95% of data falls between: | 32.5 | 47.5 | | =$C$2 - 2*$C$3 | =$C$2 + 2*$C$3 |
| 8 | | 99.7% of data falls between: | 28.75 | 51.25 | | =$C$2 - 3*$C$3 | =$C$2 + 3*$C$3 |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | | | | | | | |
| 16 | | | | | | | |
| 17 | | | | | | | |
| 18 | | | | | | | |
| 19 | | | | | | | |

From this output, we can see:

- 68% of the data falls between 36.25 and 43.75
- 95% of the data falls between 32.5 and 47.5
- 99.7% of the data falls between 28.75 and 51.25

And here is one more example of how to apply the Empirical Rule to a dataset with a mean of 100 and a standard deviation of 5:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | | mean | 100 | | | | |
| 3 | | standard deviation | 5 | | | | |
| 4 | | | | | | | |
| 5 | | | | | | | |
| 6 | | 68% of data falls between: | 95 | 105 | | =$C$2 - $C$3 | =$C$2 + $C$3 |
| 7 | | 95% of data falls between: | 90 | 110 | | =$C$2 - 2*$C$3 | =$C$2 + 2*$C$3 |
| 8 | | 99.7% of data falls between: | 85 | 115 | | =$C$2 - 3*$C$3 | =$C$2 + 3*$C$3 |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | | | | | | | |
| 16 | | | | | | | |
| 17 | | | | | | | |
| 18 | | | | | | | |
| 19 | | | | | | | |

From this output, we can see:

- 68% of the data falls between 95 and 105
- 95% of the data falls between 90 and 110
- 99.7% of the data falls between 85 and 115

Finding What Percentage of Data Falls Between Certain Values

Another question you might have is: What percentage of data falls between certain values?

For example, suppose you have a normally-distributed dataset with a mean of 100, a standard deviation of 5, and you want to know what percentage of the data falls between the values 99 and 105.

In Excel, we can easily answer this question by using the function = NORM.DIST(), which takes the following arguments:

NORM.DIST(x, mean, standard_dev, cumulative)

where,

- x is the value we're interested in
- mean is the mean of the distribution

- standard_dev is the standard deviation of the distribution
- cumulative takes a value of "TRUE" (returns the CDF) or "FALSE" (returns the PDF) – we'll use "TRUE" to get the value of the cumulative distribution function.
- The following screenshot shows how to use the NORM.DIST() function to find the percentage of the data that falls between the values 99 and 105 for a distribution that has a mean of 100 and a standard deviation of 5:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | | | | | | | |
| 3 | | | | | | | |
| 4 | | | | | | | |
| 5 | | =NORM.DIST(105, 100, 5, TRUE) - NORM.DIST(99, 100, 5, TRUE) | | | | | |
| 6 | | 42.1% | | | | | |
| 7 | | | | | | | |
| 8 | | | | | | | |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | | | | | | | |

- We see that 42.1% of the data falls between the values 105 and 99 for this distribution.

Rahul

### 3.2.2 Identifying Outliers

**Q15. Explain how to identify outliers in statistics.**

*Ans :* **(Imp.)**

Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

Unfortunately, there are no strict statistical rules for definitively identifying outliers. Finding outliers depends on subject-area knowledge and an understanding of the data collection process. While there is no solid mathematical definition, there are guidelines and statistical tests you can use to find outlier candidates

**Outliers and Their Impact**

Outliers are a simple concept—they are values that are notably different from other data points, and they can cause problems in statistical procedures.

To demonstrate how much a single outlier can affect the results, let's examine the properties of an example dataset. It contains 15 height measurements of human males. One of those values is an outlier. The table below shows the mean height and standard deviation with and without the outlier.

| With Outlier | Without Outlier | Difference |
|---|---|---|
| 2.4m (7' 10.5") | 1.8m (5' 10.8") | 0.6m (~2 feet) |
| 2.3m (7' 6") | 0.14m (5.5 inches) | 2.16m (~7 feet) |

From the table, it's easy to see how a single outlier can distort reality. A single value changes the mean height by 0.6m (2 feet) and the standard deviation by a whopping 2.16m (7 feet)! Hypothesis tests that use the mean with the outlier are off the mark.

Before performing statistical analyses, you should identify potential outliers. That's the subject of this post. In the next post, we'll move on to figuring out what to do with them.

There are a variety of ways to find outliers

**1. Sorting Your Datasheet to Find Outliers**

Sorting your datasheet is a simple but effective way to highlight unusual values. Simply sort your data sheet for each variable and then look for unusually high or low values.

For example, I've sorted the example dataset in ascending order, as shown below. The highest value is clearly different than the others. While this approach doesn't quantify the outlier's degree of unusualness, I like it because, at a glance, you'll find the unusually high or low values.

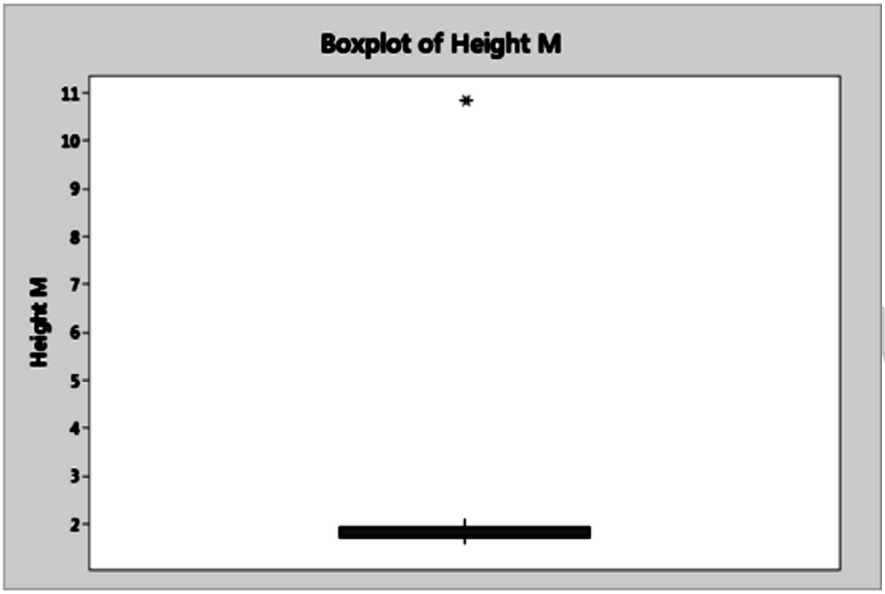
| Height M |
|---|
| 1.5895 |
| 1.6508 |
| 1.7131 |
| 1.7136 |
| 1.7212 |
| 1.7296 |
| 1.7343 |
| 1.7663 |
| 1.8018 |
| 1.8394 |
| 1.8869 |
| 1.9357 |
| 1.9482 |
| 2.1038 |
| 10.8135 |

Rahul Publications

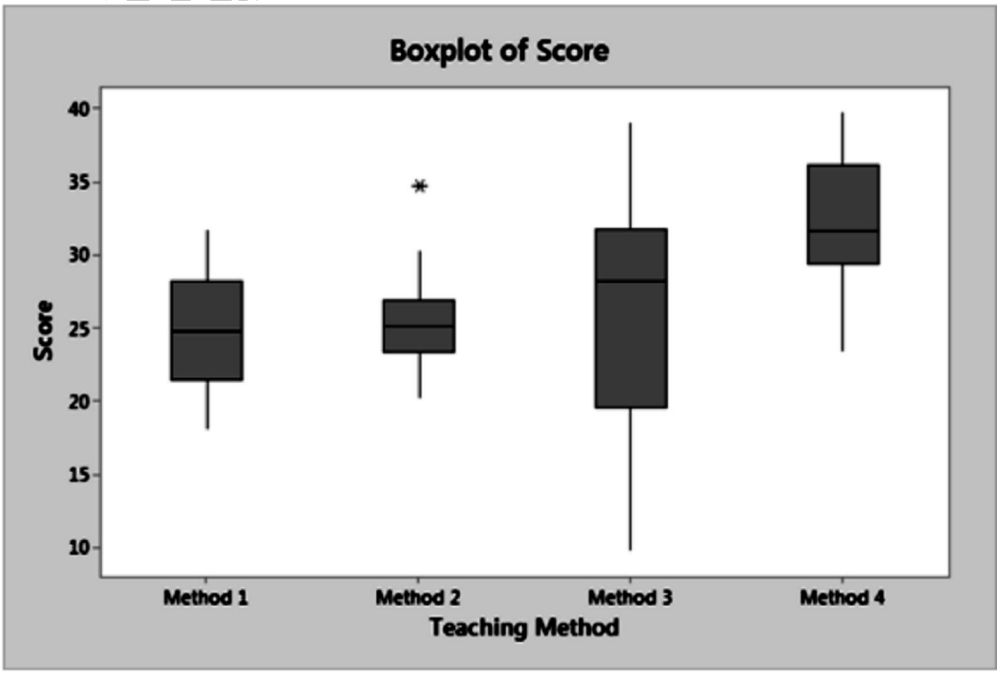
## 1. Graphing Your Data to Identify Outliers

Boxplots, histograms, and scatterplots can highlight outliers.

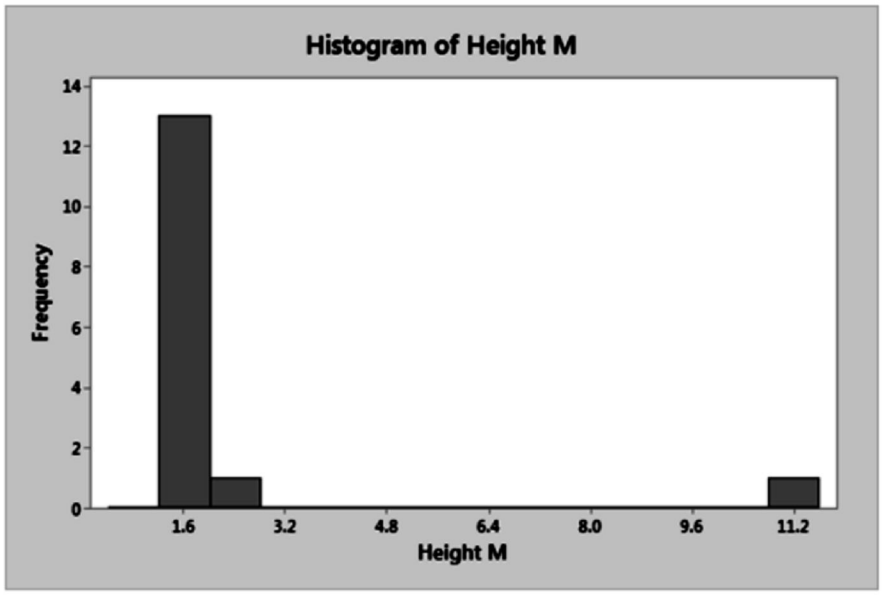
Boxplots display asterisks or other symbols on the graph to indicate explicitly when datasets contain outliers. These graphs use the interquartile method with fences to find outliers, which I explain later. The boxplot below displays our example dataset. It's clear that the outlier is quite different than the typical data value.




You can also use boxplots to find outliers when you have groups in your data. The boxplot below shows a different dataset that has an outlier in the Method 2 group.

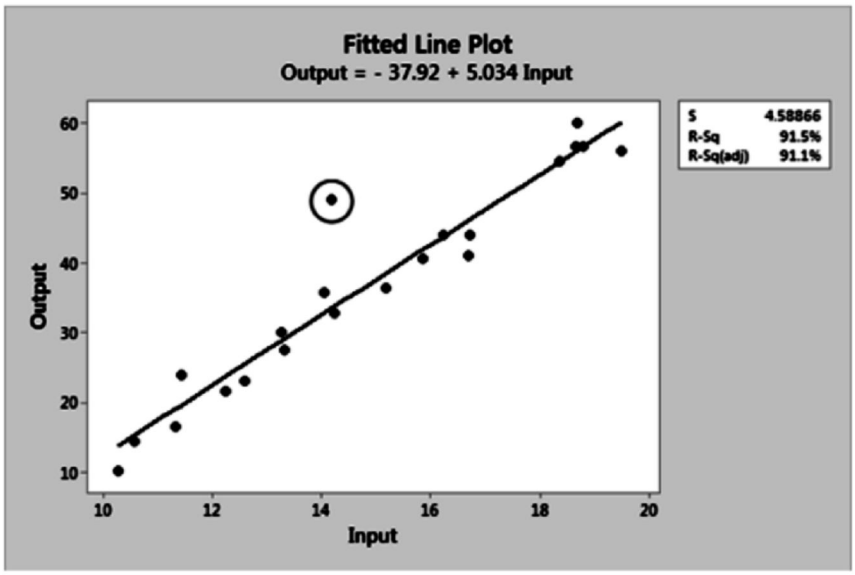


Histograms also emphasize the existence of outliers. Look for isolated bars, as shown below. Our outlier is the bar far to the right. The graph crams the legitimate data points on the far left.




Most of the outliers I discuss in this post are univariate outliers. We look at a data distribution for a single variable and find values that fall outside the distribution. However, you can use a scatterplot to detect outliers in a multivariate setting.

In the graph below, we're looking at two variables, Input and Output. The scatterplot with regression line shows how most of the points follow the fitted line for the model. However, the circled point does not fit the model well.




Interestingly, the Input value (~14) for this observation isn't unusual at all because the other Input values range from 10 through 20 on the X-axis. Also, notice how the Output value (~50) is similarly within the range of values on the Y-axis (10 – 60). Neither the Input nor the Output values themselves are unusual in this dataset. Instead, it's an outlier because it doesn't fit the model.

This type of outlier can be a problem in regression analysis. Given the multifaceted nature of multivariate regression, there are numerous types of outliers in that realm. In my ebook about regression analysis, I detail various methods and tests for identifying outliers in a multivariate context.

For the rest of this post, we'll focus on univariate outliers.

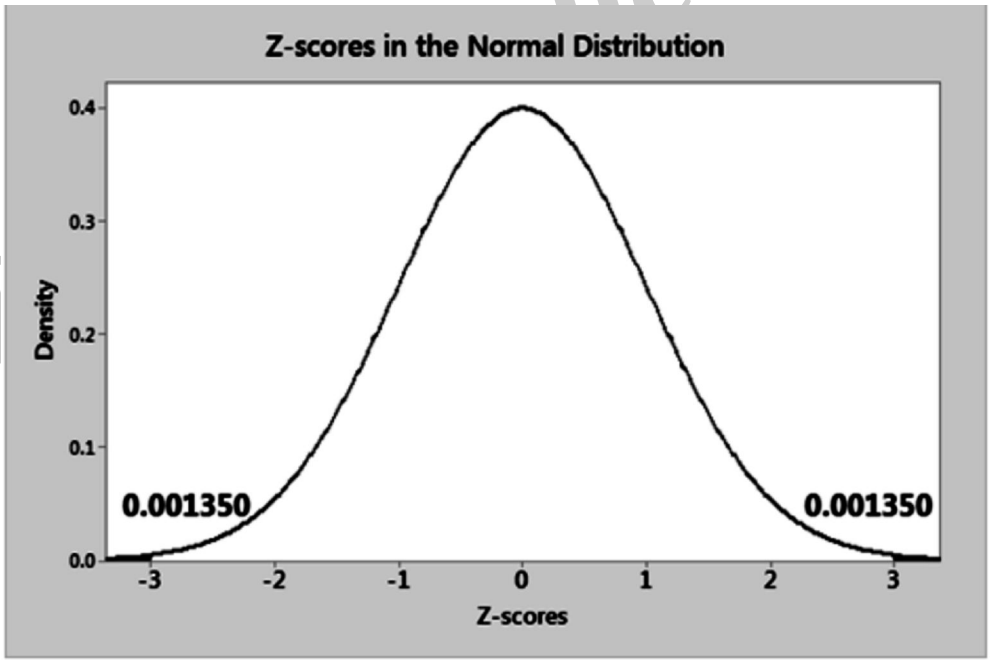
### 3. Using Z-scores to Detect Outliers

Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean.

To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation. Mathematically, the formula for that process is the following:

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

The further away an observation's Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers are Z-scores of +/–3 or further from zero. The probability distribution below displays the distribution of Z-scores in a standard normal distribution. Z-scores beyond +/–3 are so extreme you can barely see the shading under the curve.




In a population that follows the normal distribution, Z-score values more extreme than +/- 3 have a probability of 0.0027 (2 * 0.00135), which is about 1 in 370 observations. However, if your data don't follow the normal distribution, this approach might not be accurate.

### Z-scores and Our Example Dataset

In our example dataset below, I display the values in the example dataset along with the Z-scores. This approach identifies the same observation as being an outlier.

| Height M | Z-score |
|---|---|
| 1.5895 | -0.34603 |
| 1.6508 | -0.31975 |
| 1.7131 | -0.29301 |
| 1.7136 | -0.29283 |
| 1.7212 | -0.28954 |
| 1.7296 | -0.28595 |
| 1.7343 | -0.28394 |
| 1.7663 | -0.27020 |
| 1.8018 | -0.25501 |
| 1.8394 | -0.23888 |
| 1.8869 | -0.21852 |
| 1.9357 | -0.19757 |
| 1.9482 | -0.19223 |
| 2.1038 | -0.12551 |
| 10.8135 | 3.60910 |

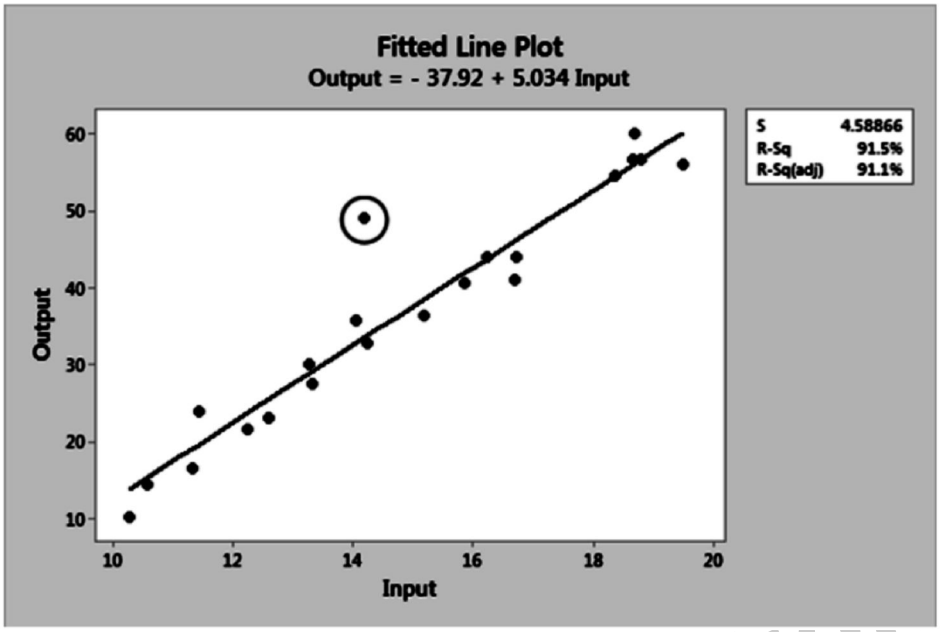
Note that Z-scores can be misleading with small datasets because the maximum Z-score is limited to $(n - 1) / \sqrt{n}$.*

Indeed, our Z-score of ~3.6 is right near the maximum value for a sample size of 15. Sample sizes of 10 or fewer observations cannot have Z-scores that exceed a cutoff value of +/–3.

Also, note that the outlier's presence throws off the Z-scores because it inflates the mean and standard deviation as we saw earlier. Notice how all the Z-scores are negative except the outlier's value. If we calculated Z-scores without the outlier, they'd be different! Be aware that if your dataset contains outliers, Z-values are biased such that they appear to be less extreme (i.e., closer to zero).

Most of the outliers I discuss in this post are univariate outliers. We look at a data distribution for a single variable and find values that fall outside the distribution. However, you can use a scatterplot to detect outliers in a multivariate setting.

In the graph below, we're looking at two variables, Input and Output. The scatterplot with regression line shows how most of the points follow the fitted line for the model. However, the circled point does not fit the model well.




Interestingly, the Input value (~14) for this observation isn't unusual at all because the other Input values range from 10 through 20 on the X-axis. Also, notice how the Output value (~50) is similarly within the range of values on the Y-axis (10 – 60). Neither the Input nor the Output values themselves are unusual in this dataset. Instead, it's an outlier because it doesn't fit the model.

This type of outlier can be a problem in regression analysis. Given the multifaceted nature of multivariate regression, there are numerous types of outliers in that realm.

### 3.2.3 Box Plots

**Q16. Explain how box plot can be drawn for the given data.**

*Ans :* **(Imp.)**

The method to summarize a set of data that is measured using an interval scale is called a box and whisker plot. These are maximum used for data analysis. We use these types of graphs or graphical representation to know:

- Distribution Shape
- Central Value of it
- Variability of it

A box plot is a chart that shows data from a five-number summary including one of the measures of central tendency. It does not show the distribution in particular as much as a stem and leaf plot or histogram does. But it is primarily used to indicate a distribution is skewed or not and if there are potential unusual observations (also called outliers) present in the data set. Boxplots are also very beneficial when large numbers of data sets are involved or compared.

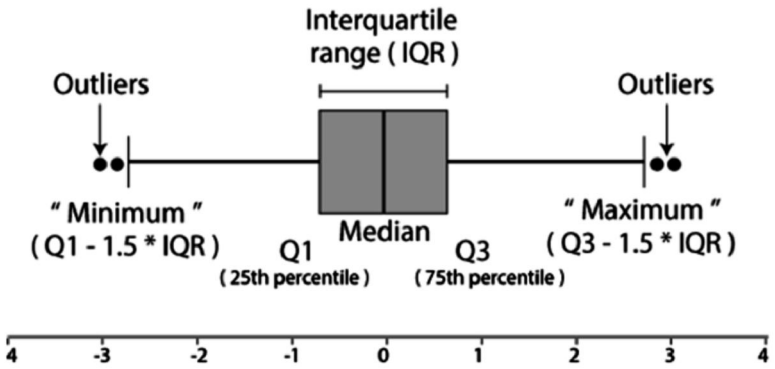
In simple words, we can define the box plot in terms of descriptive statistics related concepts. That means box or whiskers plot is a method used for depicting groups of numerical data through their quartiles graphically. These may also have some lines extending from the boxes or whiskers which indicates the variability outside the lower and upper quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers can be indicated as individual points.

It helps to find out how much the data values vary or spread out with the help of graphs. As we need more information than just knowing the measures of central tendency, this is where the box plot helps. This also takes less space. It is also a type of pictorial representation of data.

Since, the centre, spread and overall range are immediately apparent, using these boxplots the distributions can be compared easily.

**Parts of Box Plots**

Check the image below which shows the minimum, maximum, first quartile, third quartile, median and outliers.




**Different parts of boxplot**

**Minimum:**

The minimum value in the given dataset

**First Quartile (Q1):**

The first quartile is the median of the lower half of the data set.

**Median:**

The median is the middle value of the dataset, which divides the given dataset into two equal parts. The median is considered as the second quartile.

**Third Quartile (Q3):**

The third quartile is the median of the upper half of the data.

**Maximum:**

The maximum value in the given dataset.

Apart from these five terms, the other terms used in the box plot are:

**Interquartile Range (IQR):**

The difference between the third quartile and first quartile is known as the interquartile range. (i.e.) IQR = Q3-Q1

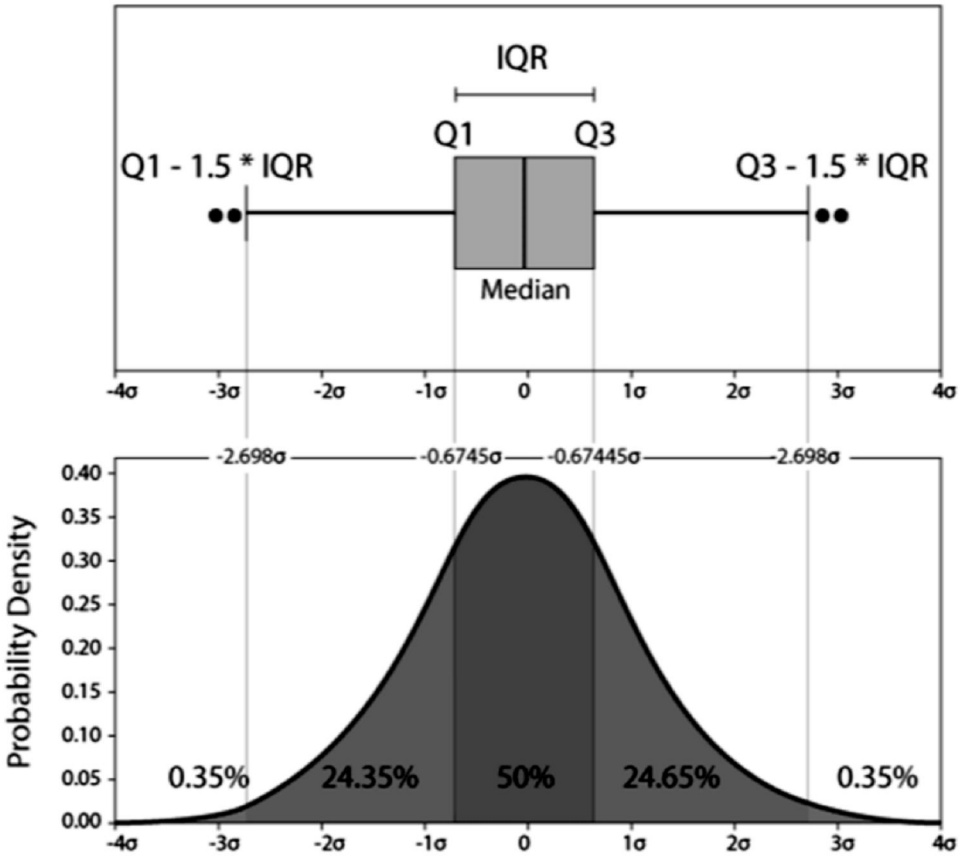
**Outlier:**

The data that falls on the far left or right side of the ordered data is tested to be the outliers. Generally, the outliers fall more than the specified distance from the first and third quartile.

(i.e.) Outliers are greater than Q3+(1.5 . IQR) or less than Q1-(1.5 .IQR).

**Boxplot Distribution**

The box plot distribution will explain how tightly the data is grouped, how the data is skewed, and also about the symmetry of data.




**Boxplot on a normal distribution**

**Positively Skewed:**

If the distance from the median to the maximum is greater than the distance from the median to the minimum, then the box plot is positively skewed.

**Negatively Skewed:**

If the distance from the median to minimum is greater than the distance from the median to the maximum, then the box plot is negatively skewed.

**Symmetric:**

The box plot is said to be symmetric if the median is equidistant from the maximum and minimum values.

**Box Plot Chart**

In a box and whisker plot:

- the ends of the box are the upper and lower quartiles so that the box crosses the interquartile range
- a vertical line inside the box marks the median
- the two lines outside the box are the whiskers extending to the highest and lowest observations.

**Example 1:**

Find the maximum, minimum, median, first quartile, third quartile for the given data set: 23, 42, 12, 10, 15, 14, 9.

***Sol :***

Given: 23, 42, 12, 10, 15, 14, 9.

Arrange the given dataset in ascending order.

9, 10, 12, 14, 15, 23, 42

Hence,

Minimum = 9

Maximum = 42

Median = 14

First Quartile = 10 (Middle value of 9, 10, 12 is 10)

Third Quartile = 23 (Middle value of 15, 23, 42 is 23).

**Example 2 :**

Finding the five-number summary

A sample of 10 boxes of raisins has these weights (in grams):

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

Make a box plot of the data.

**Step 1:**

Order the data from smallest to largest.

Our data is already in order.

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

**Step 2:**

Find the median.

The median is the mean of the middle two numbers:

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

(30 + 34) / 2 = 32

The median is 32.

**Step 3:**

Find the quartiles.

The first quartile is the median of the data points to the left of the median.

25, 28, 29, 29, 30

Q1 =29

The third quartile is the median of the data points to the right of the median.

34, 35, 35, 37, 38

Q3 =35

Rahul Publications

**Step 4:**

Complete the five-number summary by finding the min and the max.

The min is the smallest data point, which is 25.

The max is the largest data point, which is 38.

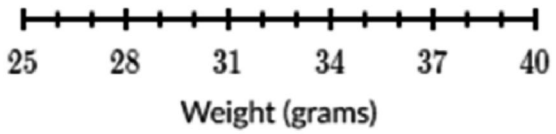
The five-number summary is 25, 29, 32, 35, 38.

**Making a box plot**

Let's make a box plot for the same dataset from above.

**Step 1:**

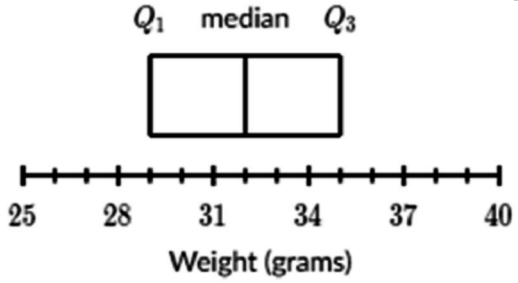
Scale and label an axis that fits the five-number summary.




**Step 2:**

Draw a box from Q1 t to Q3 , with a vertical line through the median.

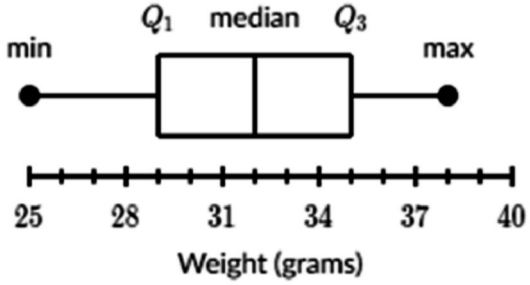
Recall that Q1 =29, the median is 32, and Q3 =35.



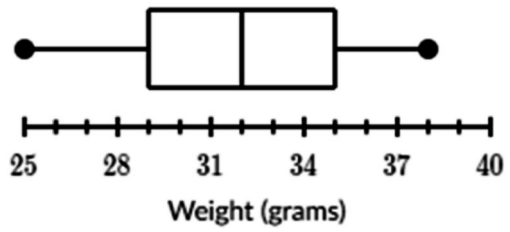

**Step 3:**

Draw a whisker from Q1, to the min and from Q3 to the max.

Recall that the min is 25 and the max is 38.




We don't need the labels on the final product:



**Q17. Explain, how to draw a box plot in excel.**

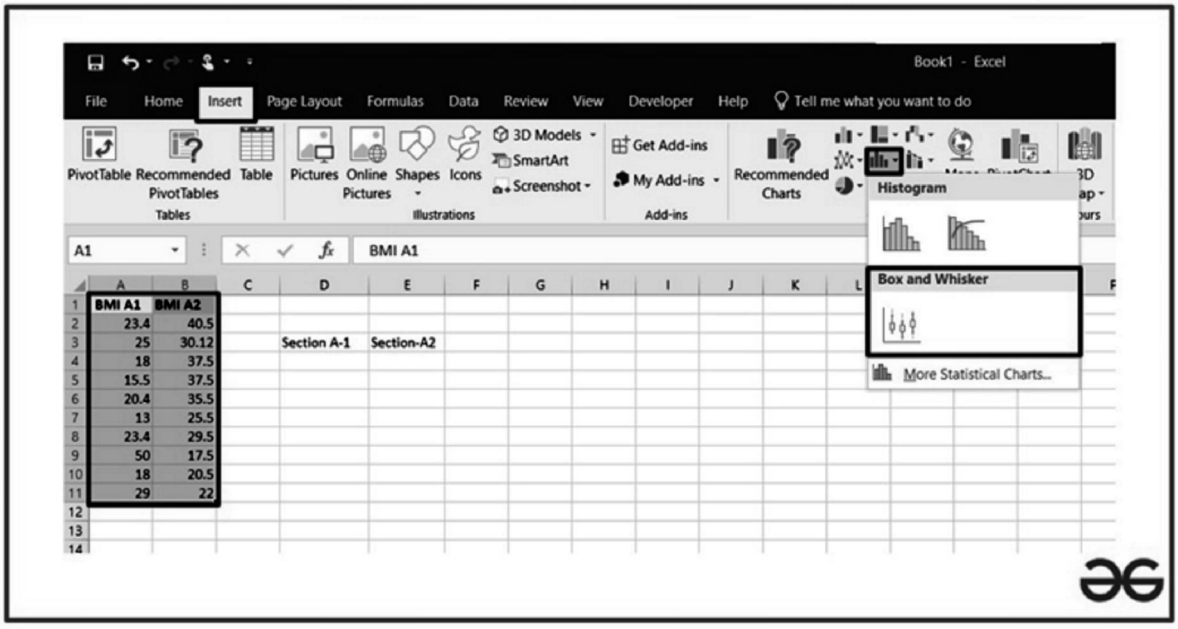
*Ans :*

Consider the BMI of ten students from section A-1 and that of section A-2. BMI stands for Body Mass Index which is an important parameter to judge the body fat and health of a person on the basis of height and weight of a person.

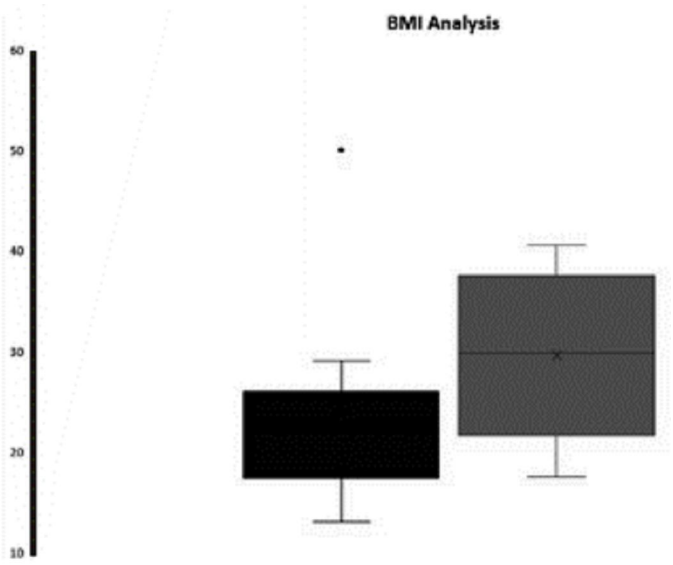
| BMI A1 | BMI A2 |
|---|---|
| 23.4 | 40.5 |
| 25 | 30.12 |
| 18 | 37.5 |
| 15.5 | 37.5 |
| 20.4 | 35.5 |
| 13 | 25.5 |
| 23.4 | 29.5 |
| 50 | 17.5 |
| 18 | 20.5 |
| 29 | 22 |

The steps to create a box plot :

1. Insert the data in the cells as shown above.
2. Select the data and go to the Insert tab at the top of the Excel window.
3. Now click on the Statistical Chart menu. A drop-down will occur.
4. Now select Box and Whisker chart.




Rahul Publications

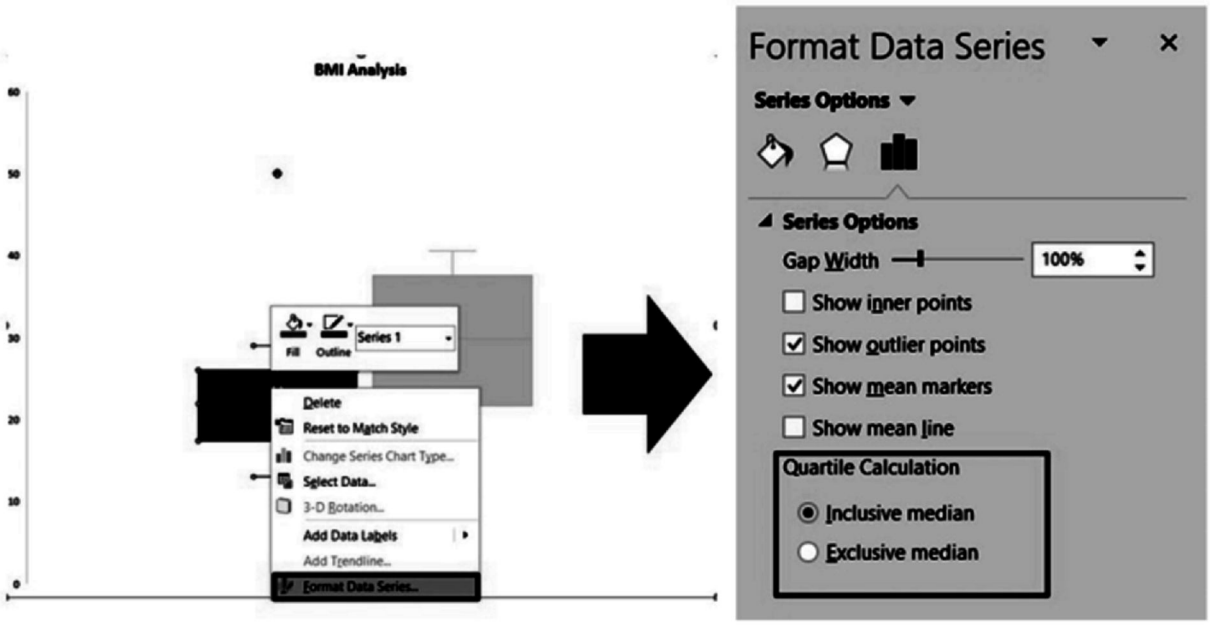



*Box Plot*

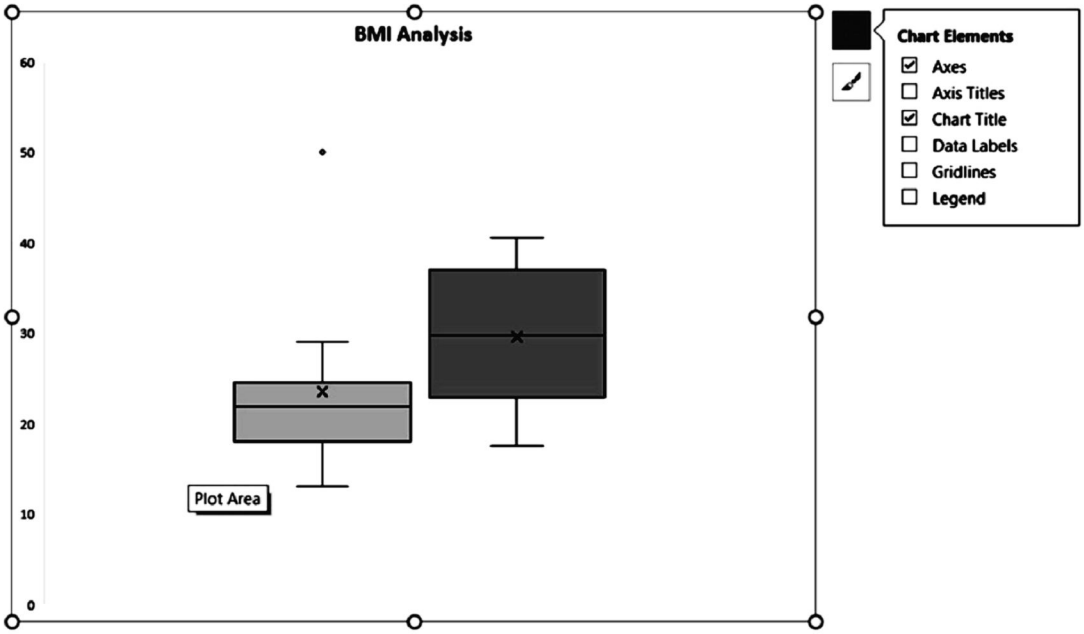
**Box Plot**

The box plot by default will be exclusive of the mean value. In order to make it inclusive of mean :

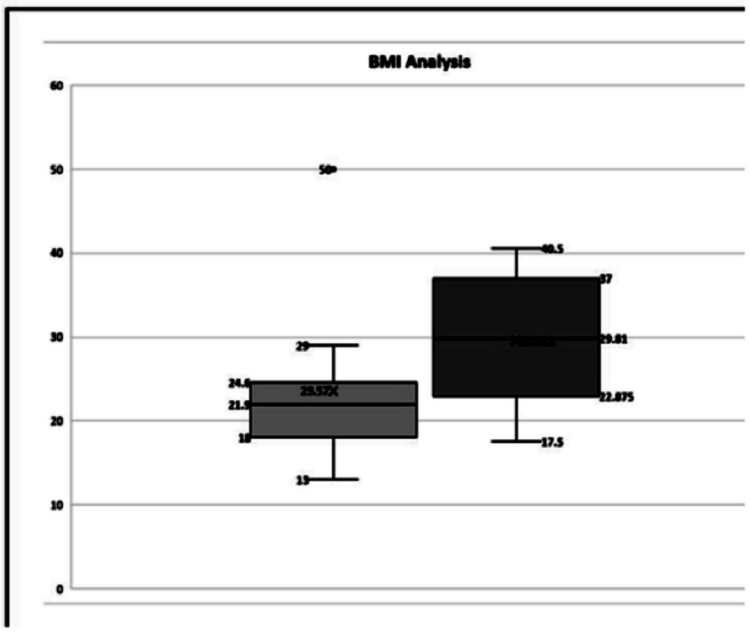
1. Select the box plot.
2. Right-Click and select Format Data Series.
3. In the Format Data Series dialog box check "Inclusive Mean" in Quartile Calculation.




To format a box plot use the + symbol in the top right corner of the chart as shown below :




Check the Data Labels option to add data labels in the box plots and make the plot more insightful.




You can examine the data labels values using the following section where we are going to discuss how to calculate these parameters using Excel formulas.

Formula to calculate parameters associated with the box plot:

In order to calculate the different quartile values use the formula :

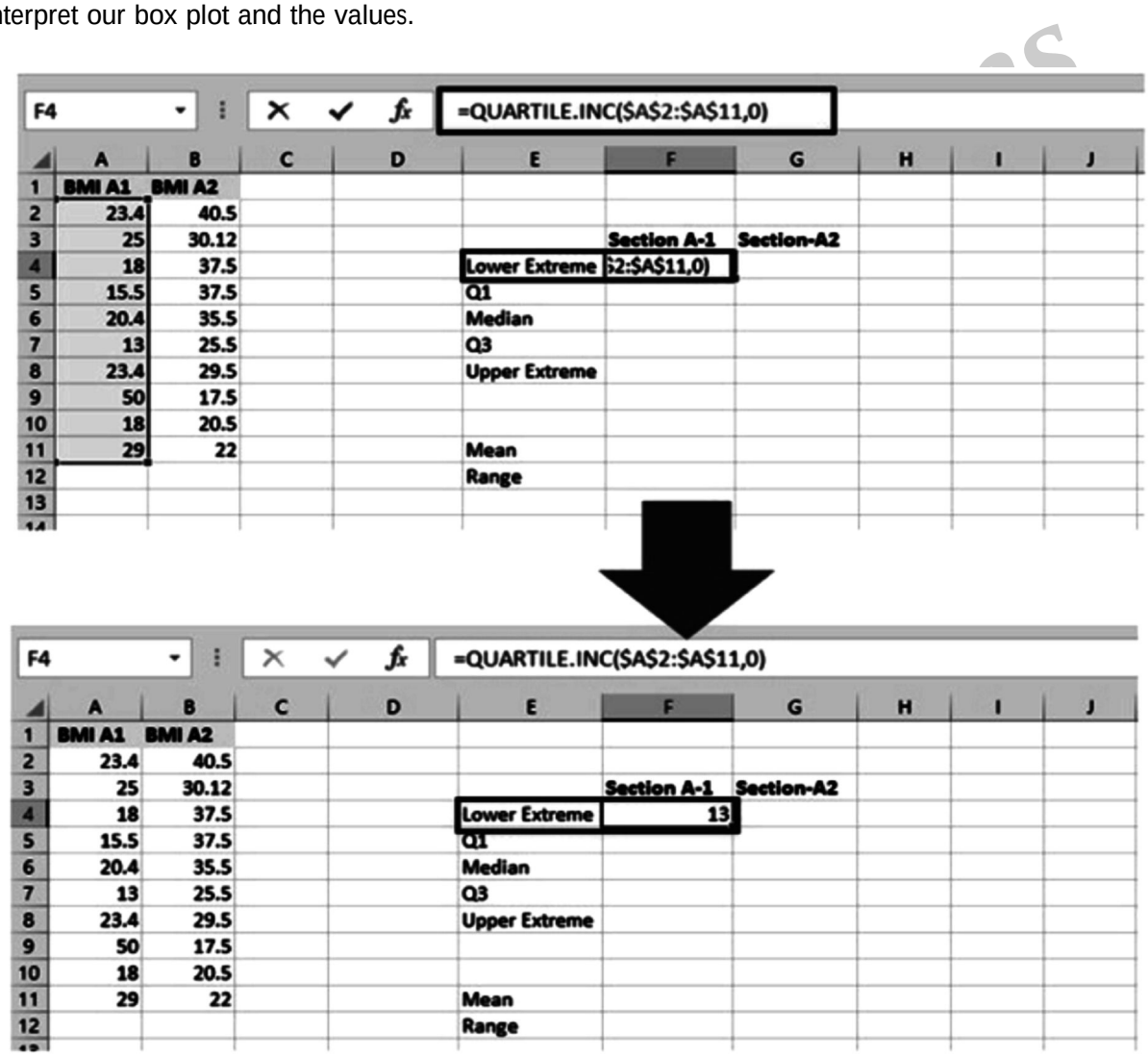
= QUARTILE.INC(Cell_Range, integer)

Here,

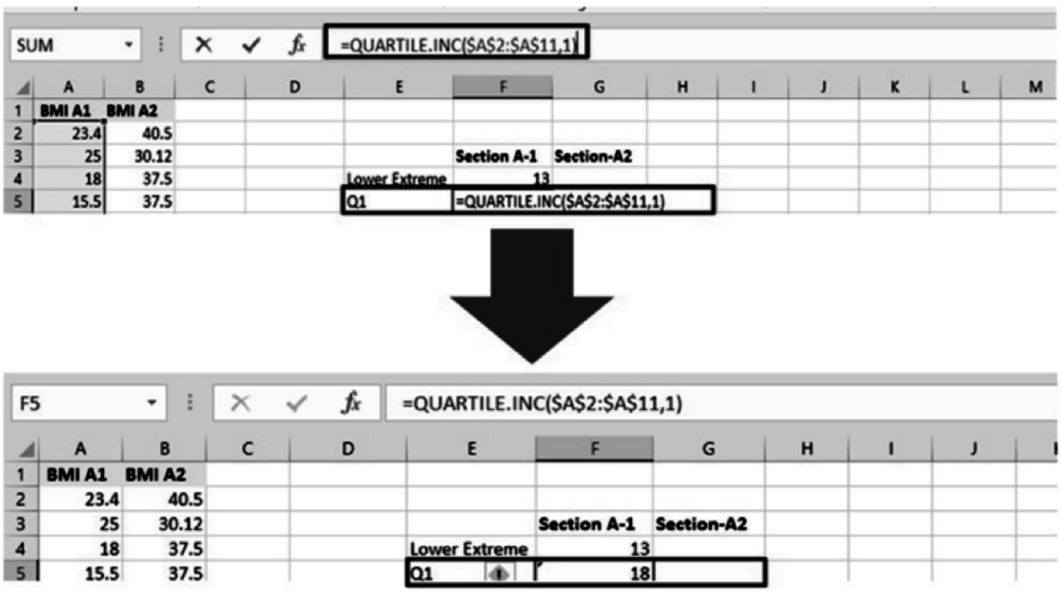
- Cell range: Range of cells. In our case, it is A2 to A11 for section A-1 and B2 to B11 for section A – 2
- integer : [0,4]

| **Quartile Values** | **Formula** |
|---|---|
| Lower Extreme | = QUARTILE.INC(Cell_Range, 0) |
| Q1 | = QUARTILE.INC(Cell_Range, 1) |
| Median | = QUARTILE.INC(Cell_Range, 2) |
| Q3 | = QUARTILE.INC(Cell_Range, 3) |
| Upper Extreme | = QUARTILE.INC(Cell_Range, 4) |

Make a helper table in Excel to calculate the above formulas. The helper table can be used to interpret our box plot and the values.




Lower Limit Calculation



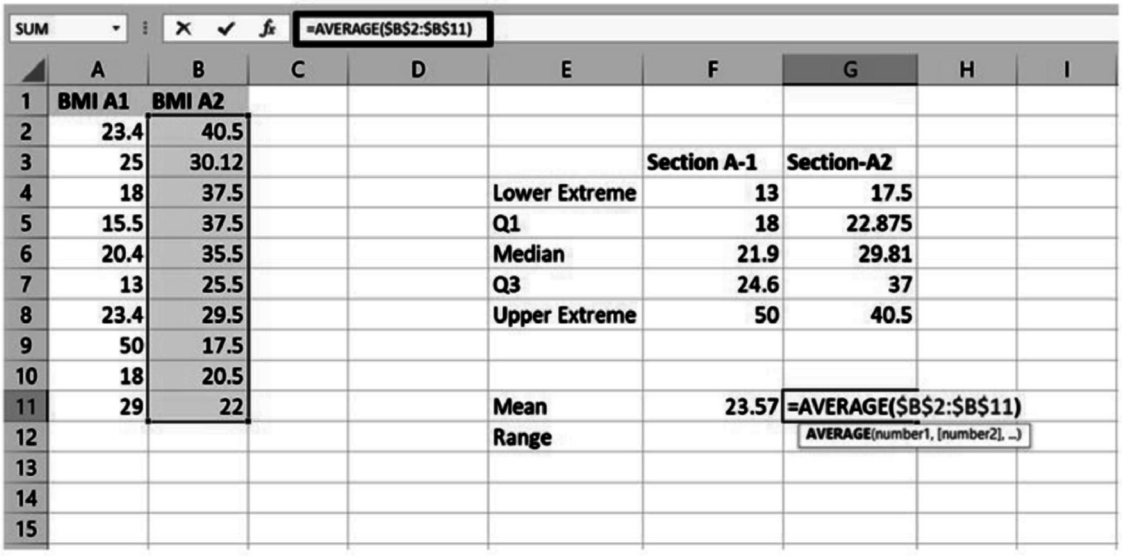

Quartile 1 calculation

Similarly, you can calculate all the other parameters for both sections. The final table will look like this:

| | Section A-1 | Section-A2 |
|---|---|---|
| Lower Extreme | 13 | 17.5 |
| Q1 | 18 | 22.875 |
| Median | 21.9 | 29.81 |
| Q3 | 24.6 | 37 |
| Upper Extreme | 50 | 40.5 |

Some other important parameters in a box plot are (1) Mean (2) Range. The formulas are :

= AVERAGE(Cell_Range) = (Upper Extreme - Lower Extreme)



F4 =F8-F4

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BMI A1 | BMI A2 | | | | | | | | | |
| 2 | 23.4 | 40.5 | | | | | | | | | |
| 3 | 25 | 30.12 | | | | Section A-1 | Section-A2 | | | | |
| 4 | 18 | 37.5 | | | Lower Extreme | 13 | 17.5 | | | | |
| 5 | 15.5 | 37.5 | | | Q1 | 18 | 22.875 | | | | |
| 6 | 20.4 | 35.5 | | | Median | 21.9 | 29.81 | | | | |
| 7 | 13 | 25.5 | | | Q3 | 24.6 | 37 | | | | |
| 8 | 23.4 | 29.5 | | | Upper Extreme | 50 | 40.5 | | | | |
| 9 | 50 | 17.5 | | | | | | | | | |
| 10 | 18 | 20.5 | | | | | | | | | |
| 11 | 29 | 22 | | | Mean | 23.57 | 29.612 | | | | |
| 12 | | | | | Range | =F8-F4 | | | | | |
| 13 | | | | | | | | | | | |

| | Section A-1 | Section-A2 |
|---|---|---|
| Lower Extreme | 13 | 17.5 |
| Q1 | 18 | 22.875 |
| Median | 21.9 | 29.81 |
| Q3 | 24.6 | 37 |
| Upper Extreme | 50 | 40.5 |
| Mean | 23.57 | 29.612 |
| Range | 37 | 23 |

**Helper Table**

Another important parameter in a box plot is an outlier which depends on the value of Interquartile Range (IQR). The formula for IQR is :

IQR = Quartile_3 - Quartile_1

In our example, the value of IQR is 6.6 which you can calculate from the helper table. Now, a point is an outlier if the value is :

below (Quartile_1 - IQR*1.5) and

above (Quartile_3 + IQR*1.5)

In the given example for section A-1 we have an outlier at value 50 which is the maximum value of BMI. After calculation the value will be :

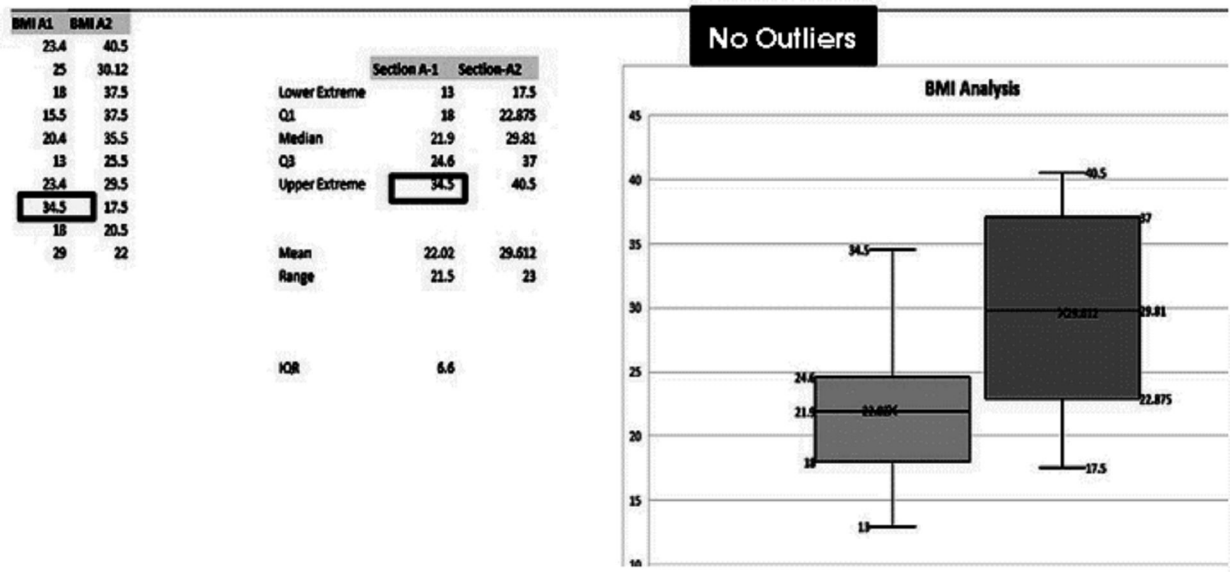
IQR * 1.5 = 9.9

Q3 + IQR * 1.5 = 34.5

Since, 50 > 34.5 so it is in the outlier of the box plot.

Similarly, you can calculate the above parameters for the second box plot, and you can observe that all the five parameters are within the range and hence there are no outliers.

In order to remove the outlier in Box plot-1, you have to modify the maximum value from 50 to any value less or equal to 34.5.




## 3.3 Measures of Association

### 3.3.1 Scatter Charts

**Q18. What is a Scatter Plot? Explain how to draw scatter plot for the given example.**

*Ans :* **(Imp.)**

A scatter plot is a means to represent data in a graphical format. A simple scatter plot makes use of the Coordinate axes to plot the points, based on their values. The following scatter plot excel data for age (of the child in years) and height (of the child in feet) can be represented as a scatter plot.

| Age of the Child | Height |
|---|---|
| 3 | 2.3 |
| 4 | 2.7 |
| 5 | 3.1 |
| 6 | 3.6 |
| 7 | 3.8 |
| 8 | 4 |
| 9 | 4.3 |
| 10 | 4.5 |

**Scatter Plot Application**

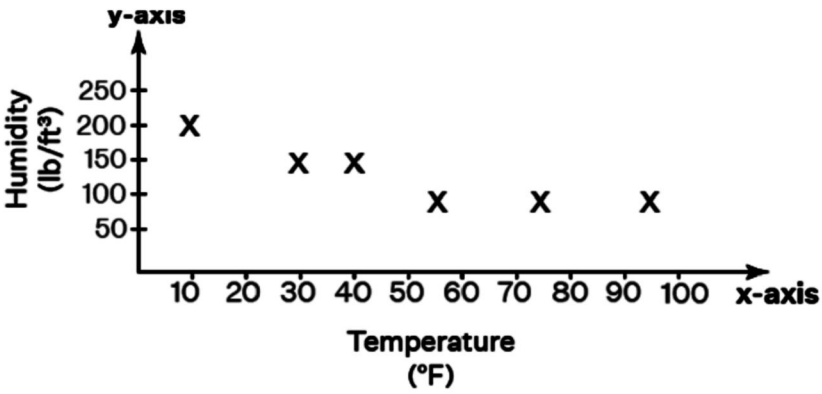
Here let us look at real-life application represented by scatter plot.

Example: Days of the week and the sales

**How to Construct a Scatter Plot?**

There are three simple steps to plot a scatter plot.

- **STEP I:** Identify the x-axis and y-axis for the scatter plot.
- **STEP II:** Define the scale for each of the axes.
- **STEP III:** Plot the points based on their values.



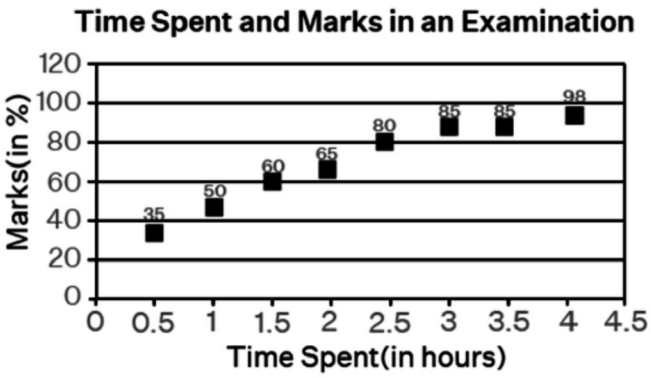
### Types of Scatter Plot

A scatter plot helps find the relationship between two variables. This relationship is referred to as a correlation. Based on the correlation, scatter plots can be classified as follows.

- Scatter Plot for Positive Correlation
- Scatter Plot for Negative Correlation
- Scatter Plot for Null Correlation

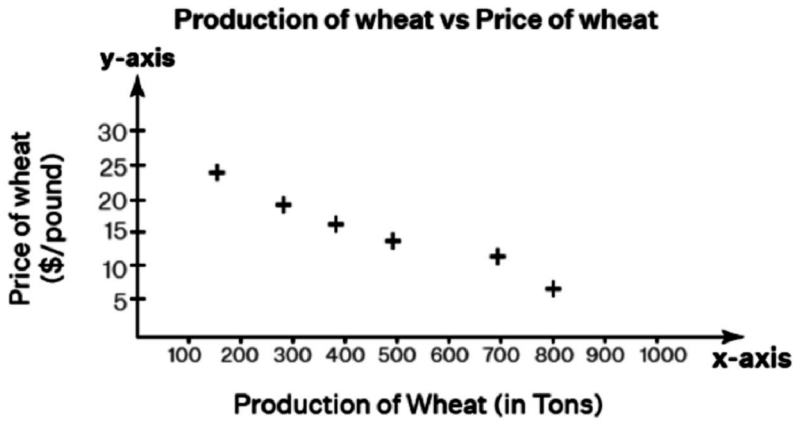
- **Scatter Plot for Positive Correlation**

A scatter plot with increasing values of both variables can be said to have a positive correlation. The scatter plot for the relationship between the time spent studying for an examination and the marks scored can be referred to as having a positive correlation.



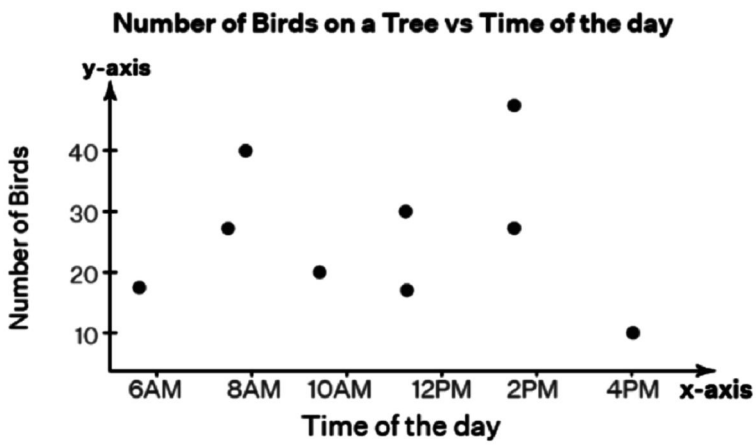

- **Scatter Plot for Negative Correlation**

A scatter plot with an increasing value of one variable and a decreasing value for another variable can be said to have a negative correlation. Observe the below image of negative scatter plot depicting the amount of production of wheat against the respective price of wheat.




- **Scatter Plot for Null Correlation**

A scatter plot with no clear increasing or decreasing trend in the values of the variables is said to have no correlation. Here the points are distributed randomly across the graph. For example, the data for the number of birds on a tree at different times of the day does not show any correlation. Observe the below scatter plot showing the number of birds on a tree versus time of the day.




**Example 1:**

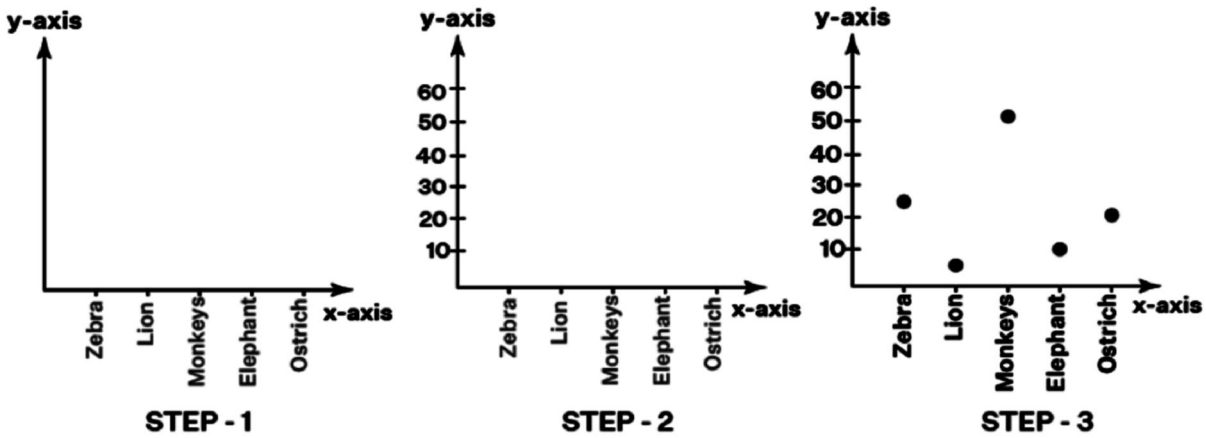
Laurell had visited a zoo recently and had collected the following data. How can Laurell use a scatter plot to represent this data?

| Type of Animal | Number of Animals in the Zoo |
|---|---|
| Zebra | 25 |
| Lions | 5 |
| Monkeys | 50 |
| Elephants | 10 |
| Ostriches | 20 |

*Sol :*

The aim is to present the above data in a scatter plot.

- **Step 1:** Mark the points on the x-axis and write the names of the animals beside each of the markings.
- **Step 2:** Marks the points as 10, 20, 30, 40, 50, 60 on the y-axis to represent the number of animals.
- **Step 3:** Identify the animals marked on the x-axis and mark a point above is based on the number given in the table. Refer to the y-axis to measure and mark the points.




Therefore the points representing the number of animals have been plotted on the scatter plot.

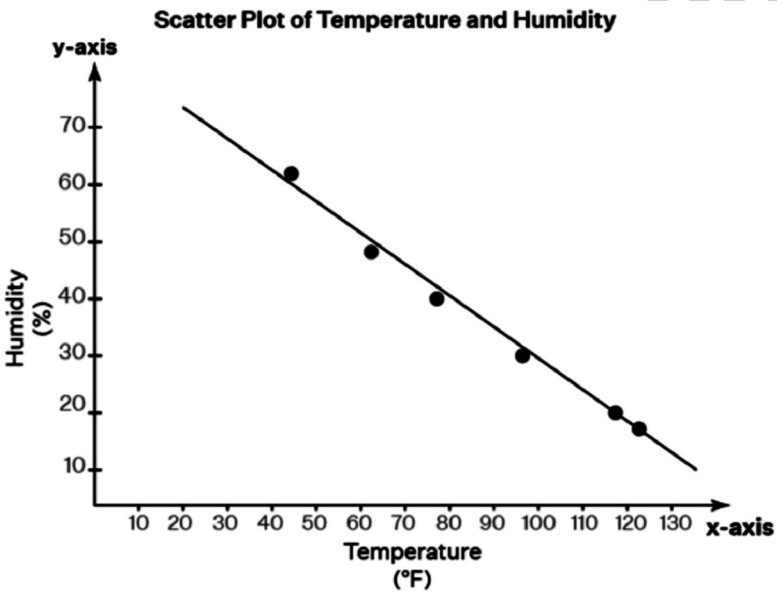
**Example 2:**

The meteorological department has collected the following data about the temperature and humidity in their town. Refer to the table given below and indicate the method to find the humidity at a temperature of 60 degrees Fahrenheit.

| Temperature (Degree Fahrenheit) | Humidity(%) |
|---|---|
| 45 | 60 |
| 62 | 48 |
| 77 | 40 |
| 97 | 30 |
| 118 | 20 |
| 122 | 18 |

*Sol :*

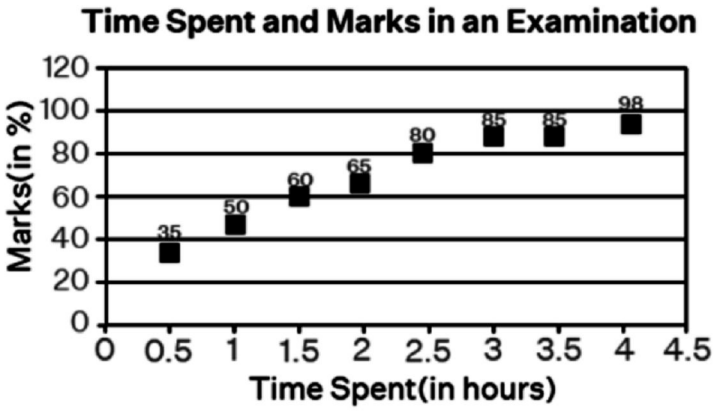
The collected data of the temperature and humidity can be presented in the form of a scatter plot.




- Temperature is marked on the x-axis and humidity is on the y-axis.
- To calculate the humidity at a temperature of 60 degrees Fahrenheit, we need to first draw a line of best fit.
- A line of "Best Fit" is a straight line drawn to pass through most of these data points.
- Now draw a vertical line from the mark of 60 degrees Fahrenheit on the x-axis, so that it cuts the line of "Best Fit".
- At the point where this line cuts the line of "Best Fit", the corresponding marking on the y-axis represents the humidity at 60 degrees Fahrenheit.
- Therefore, the humidity at a temperature of 60 degrees Fahrenheit is 50%.

**Example 3:**

In a school, a teacher has prepared a scatter plot on her computer to show the marks of 8 students and the time spent in preparation for the examination. How can we help the teacher find the outlier?




*Sol :*

The data in the scatter plot shows a positive correlation; the marks increase with an increase in time spent on preparation. But the data point referring to the student who has to spend 2.5 hours of time for preparation and has secured 40% of marks is distinct from the correlation and can thus be identified as an outlier.

Therefore, the data point of the student with 40% marks and time of 2.5 hours is the outlier.

**Q19. Explain how to make a scatter plot in Excel.**

*Ans :* **(Imp.)**

The following steps to make a scatter plot.

**Step 1: Organize your data**

Ensure that your data is in the correct format. Since scatter graphs are meant to show how two numeric values are related to each other, they should both be displayed in two separate columns.

The first column will usually be plotted on the X-axis and the second column on the Y-axis. The independent variable usually falls on the X-axis and the dependent variable on the Y-axis.

**Step 2: Select the relevant data**

Highlight the columns containing both sets of variables. If the columns are non-contiguous, hold down the Ctrl key between selections. Only select the columns with the two sets of data that are being examined for a cause/effect relationship.

**Step 3: Select the desired type of scatter plot**

From the Insert tab, go to the Charts group and click the Scatter graph symbol.

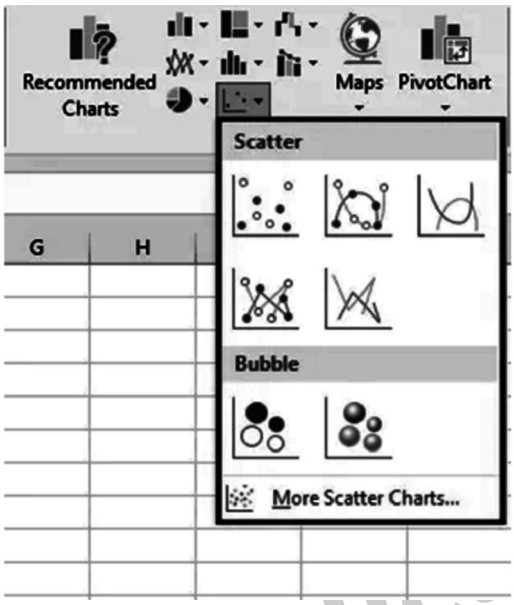
**Types of scatter plots**

Several types of scatter plots are available from the Insert Charts menu. These include:

- 'Classic' scatter chart (solely with data points)
- Scatter with smooth lines and markers

Rahul Publications

- Scatter with smooth lines
- Scatter with straight lines and markers
- Scatter with straight lines




Scatter charts with lines are best used when you have few data points. Otherwise, the plot area can begin to look quite cluttered.

### Multiple XY pairs

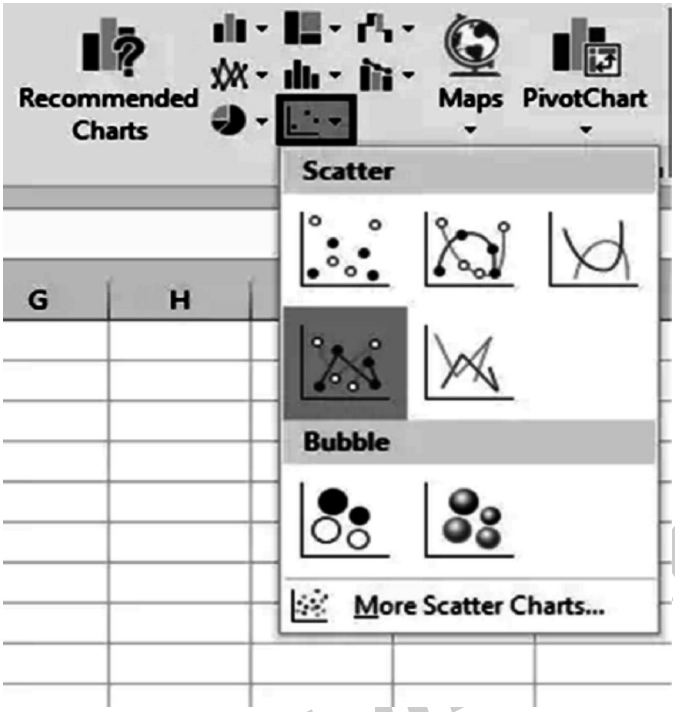
Indeed you can, and in fact, this can be done in more ways than one. Below is perhaps the simplest and most straightforward method:

1. Create two separate data sets. Organize them as previously shown, whereby for each data set the dependent variable should be to the right of the independent variable, as seen below.

| | A | B | C |
|---|---|---|---|
| 1 | | Rainfall (mm) | Pollen Count |
| 2 | | 179 | 79 |
| 3 | | 136 | 83 |
| 4 | | 116 | 79 |
| 5 | Series 1 | 75 | 52 |
| 6 | | 62 | 51 |
| 7 | | 53 | 56 |
| 8 | | 36 | 41 |
| 9 | | 43 | 33 |
| 10 | | 31 | 21 |
| 11 | | | |
| 12 | | Daily Temp. (°F) | Pollen Count |
| 13 | | 37 | 21 |
| 14 | | 40 | 33 |
| 15 | | 43 | 41 |
| 16 | Series 2 | 49 | 56 |
| 17 | | 55 | 55 |
| 18 | | 60 | 52 |
| 19 | | 71 | 79 |
| 20 | | 72 | 83 |
| 21 | | 65 | 67 |

Rahul Publications

2. Create a scatter plot from the first data set by highlighting the data and using the Insert > Chart > Scatter sequence.




In the above image, the Scatter with straight lines and markers was selected, but of course, any one will do. The scatter plot for your first series will be placed on the worksheet.

3. Select the chart.
4. Go to the Design tab and click Select Data.
5. In the Select Data Source dialog box, below Legend Entries (Series), click Add.
6. Another dialog box, Edit Series, will appear. From here, you will enter the details pertaining to the second data set, including the series name, range of X-axis values, and range of Y-axis values.
7. Click OK, and this will take you back to the Select Data Source dialog box.
8. Click OK, and you will see the second set of XY values plotted on the scatter diagram.

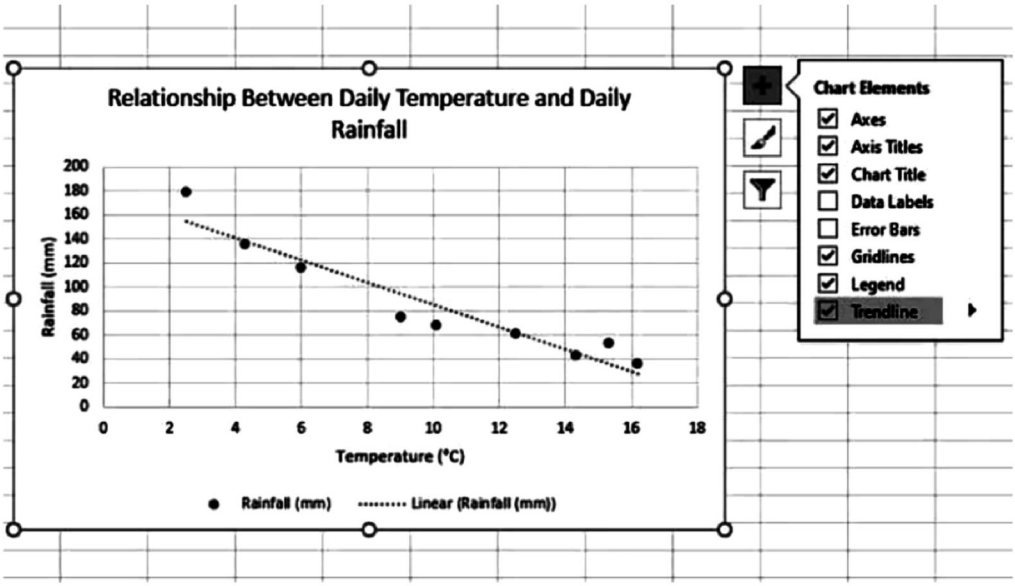
**Customize a scatter plot**

Steps to customize a scatter plot are similar to doing so for most other charts. You can customize your graph by changing, adding, or removing things like:

- Data labels
- A chart title
- Axis titles
- A trendline

Do this by selecting the graph, then clicking on the green plus (+) symbol at the upper right corner for the Chart Elements shortcut. All the elements available for your chart type will be shown, with an expanding arrow to the right of each one, offering additional options.

For example, a Trendline is particularly useful in identifying patterns and can be added by simply checking the Trendline checkbox from the Chart Elements shortcut.




Scatter graphs may be one of the most useful Excel charts you're not yet using. Important points to remember are:

- If dates or a timeline are important, then you may need to think about a line graph or maybe even a column chart.
- If you think that two variables have a correlation and you want to highlight or determine that relationship, a scatter diagram is your best bet.
- Independent variables are usually shown on the horizontal axis and dependent variables on the vertical axis.
- A trendline can help you to establish the type of relationship if it isn't immediately apparent.

### 3.3.2 Covariance

**Q20. What is covariance? How to calculate covariance? Explain with an example.**

*Ans :*

Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

**Types of Covariance**

Covariance can have both positive and negative values. Based on this, it has two types:

1. Positive Covariance
2. Negative Covariance

**1. Positive Covariance**

If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

**2. Negative Covariance**

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

**Covariance Formula**

Covariance formula is a statistical formula, used to evaluate the relationship between two variables. It is one of the statistical measurements to know the relationship between the variance between the two variables. Let us say X and Y are any two variables, whose relationship has to be calculated. Thus the covariance of these two variables is denoted by Cov(X,Y). The formula is given below for both population covariance and sample covariance.

Population covariance

$$\text{Cov}(x, y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample covariance

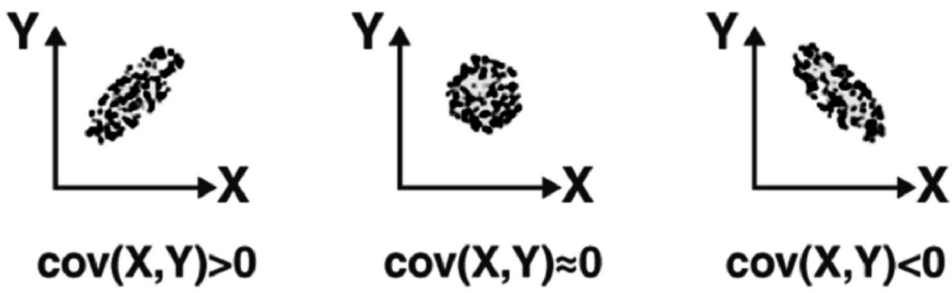
$$\text{Cov}(x, y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Where,

- xi = data value of x
- yi = data value of y
- x = mean of x
- 3 = mean of y
- N = number of data values.

**Covariance of X and Y**

Below figure shows the covariance of X and Y.




Rahul Publications

If cov(X, Y) is greater than zero, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.

If cov(X, Y) is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.

If cov(X, Y) is zero, then we can say that there is no relation between two variables.

**Correlation Coefficient Formula**

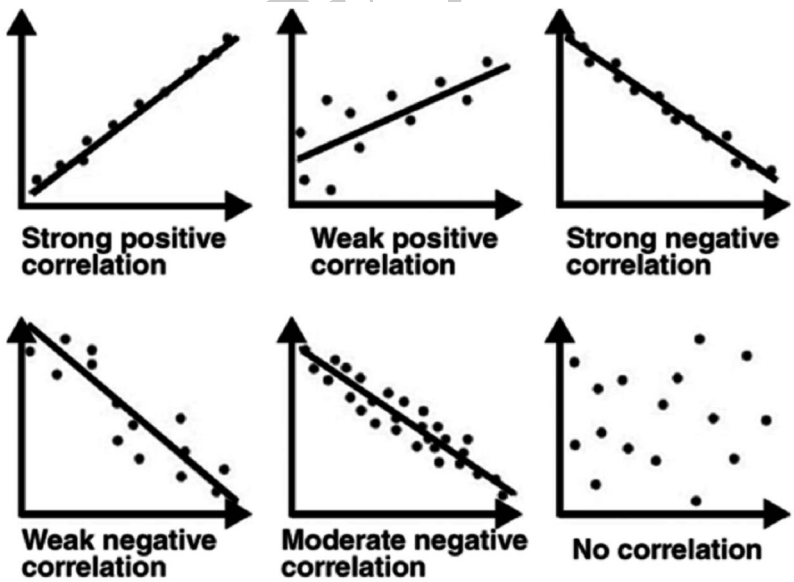
We have already discussed covariance, which is the evaluation of changes between any two variables. Correlation estimates the depth of the relationship between variables. It is the estimated measure of covariance and is dimensionless. In other words, the correlation coefficient is a constant value always and does not have any units. The relationship between the correlation coefficient and covariance is given by;

**Correlation,ñ(X,Y) = Cov(X,Y)/$\sigma_x$ $\sigma_y$**

Where,

- $\rho(X,Y)$ = correlation between the variables X and Y
- Cov(X,Y) = covariance between the variables X and Y
- $\sigma X$ = standard deviation of the X variable
- $\sigma Y$ = standard deviation of the Y variable

Based on the value of correlation coefficient, we can estimate the type of correlation between the given two variables. Also, the graphical representation of correlation among two variables is given in the below figure.




**Example 1:**

Calculate the coefficient of covariance for the following data:

| | | | | | | |
|---|---|---|---|---|---|---|
| X | 2 | 8 | 18 | 20 | 28 | 30 |
| Y | 5 | 12 | 18 | 23 | 45 | 50 |

*Sol :*

Number of observations = 6

Mean of X = 17.67

Mean of Y = 25.5

Cov(X, Y) = (116) [(2 – 17.67)(5 – 25.5) + (8 – 17.67)(12 – 25.5) + (18 – 17.67)(18 – 25.5) + (20 – 17.67)(23 – 25.5) + (28 – 17.67)(45 – 25.5) + (30 – 17.67)(50 – 25.5)]

= 157.83

**Example 2 :**

John is an investor. His portfolio primarily tracks the performance of the S&P 500 and John wants to add the stock of ABC Corp. Before adding the stock to his portfolio, he wants to assess the directional relationship between the stock and the S&P 500.

*Sol :*

John does not want to increase the unsystematic risk of his portfolio. Thus, he is not interested in owning securities in the portfolio that tend to move in the same direction.

John can calculate the covariance between the stock of ABC Corp. and S&P 500 by following the steps below:

**1. Obtain the data**

First, John obtains the figures for both ABC Corp. stock and the S&P 500. The prices obtained are summarized in the table below:

| | S & P 500 | ABC Corp. |
|---|---|---|
| 2013 | 1,692 | 68 |
| 2014 | 1,978 | 102 |
| 2015 | 1,884 | 110 |
| 2016 | 2,151 | 112 |
| 2017 | 2,519 | 154 |

**2. Calculate the mean (average) prices for each asset**

$$\text{Mean (S \& P 500)} = \frac{1,692 + 1,978 + 1,884 + 2,151 + 2,519}{5} = 2,044.80$$

$$\text{Mean (ABC Corp.)} = \frac{68 + 102 + 110 + 112 + 154}{5} = 109.20$$

**3. For each security, find the difference between each value and mean price**

| | | | Step 3 | | Step 4 |
|---|---|---|---|---|---|
| | S&P 500 | ABC Corp. | a | b | a x b |
| 2013 | 1,692 | 68 | -352.80 | -41.20 | 14,535.36 |
| 2014 | 1,978 | 102 | -66.80 | -7.20 | 480.96 |
| 2015 | 1,884 | 110 | -160.80 | 0.80 | -128.64 |
| 2016 | 2,151 | 112 | 106.20 | 2.80 | 297.36 |
| 2017 | 2,519 | 154 | 474.20 | 44.80 | 21,244.16 |
| Mean | 2,044.80 | 109.20 | Sum | | 36,429.20 |

Rahul Publications

**4. Multiply the results obtained in the previous step.**

**5. Using the number calculated in step 4, find the covariance.**

$$\text{Cov(S \& P 500, ABC Corp.)} = \frac{36,429.20}{5-1} = 9,107.30$$

In such a case, the positive covariance indicates that the price of the stock and the S&P 500 tend to move in the same direction.

**Q21. Explain how to calculate covariance in Excel?**

*Ans :*

In this Excel tutorial, you will learn how to calculate covariance in Excel.

We can detect the existence of a correlation relationship by examining the covariance.

Covariance is the average of the products of the deviations of each data point pair.

Use covariance to define the relationship between two datasets.

- A positive value of covariance occurs when both examined characteristics are moving in the same direction.
- Negative covariance occurs when the increase in the value of one feature tends to decrease the value of the other.
- It is also possible that the covariance is zero. This means that the variables are not correlated with each other.

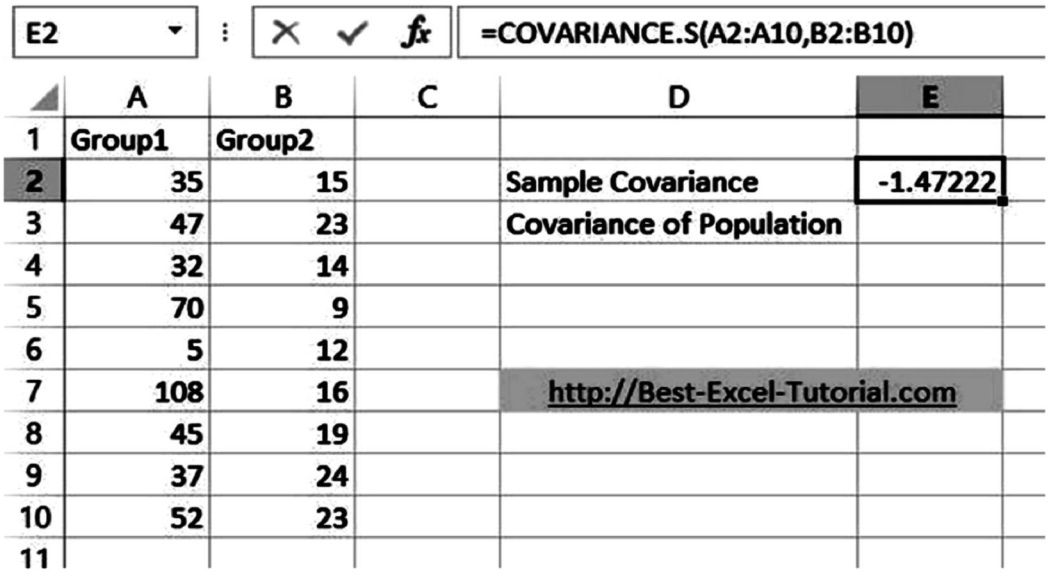
**Sample Covariance**

To calculate sample covariance use the COVARIANCE.S Excel function.

**COVARIANCE.S syntax**

=COVARIANCE.S(array1,array2)

In the given example sample covariance formula is =COVARIANCE.S(A2:A10,B2:B10)




Rahul Publications

**Population Covariance**

To calculate population covariance use the COVARIANCE.P Excel function.

**COVARIANCE.P syntax**

=COVARIANCE.S(array1,array2)

In the given example sample covariance formula is =COVARIANCE.P(A2:A10,B2:B10)

E3 | =COVARIANCE.P(A2:A10,B2:B10)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Group1 | Group2 | | | |
| 2 | 35 | 15 | | Sample Covariance | -1.47222 |
| 3 | 47 | 23 | | Covariance of Population | -1.30864 |
| 4 | 32 | 14 | | | |
| 5 | 70 | 9 | | | |
| 6 | 5 | 12 | | | |
| 7 | 108 | 16 | | http://Best-Excel-Tutorial.com | |
| 8 | 45 | 19 | | | |
| 9 | 37 | 24 | | | |
| 10 | 52 | 23 | | | |
| 11 | | | | | |

In most cases, you will use sample covariance. Use covariance of population function only when it is specifically said that it is population covariance

## 3.4 Correlation Coefficient

**Q22. What is correlation coefficient? How to find the Correlation Coefficient.**

*Ans :* **(Imp.)**

Correlation Coefficient is a statistical concept, which helps in establishing a relation between predicted and actual values obtained in a statistical experiment. The calculated value of the correlation coefficient explains the exactness between the predicted and actual values.

Correlation Coefficient value always lies between -1 to +1. If correlation coefficient value is positive, then there is a similar and identical relation between the two variables. Else it indicates the dissimilarity between the two variables.

The covariance of two variables divided by the product of their standard deviations gives Pearson's correlation coefficient. It is usually represented by ρ (rho).

$$\rho(X,Y) = \text{cov}(X,Y) / \sigma X.\sigma Y.$$

Here cov is the covariance. óX is the standard deviation of X and óY is the standard deviation of Y. The given equation for correlation coefficient can be expressed in terms of means and expectations.

$$\rho(X, Y) = E\,\frac{(X-\mu_x)(Y-\mu_y)}{\sigma x.\sigma y}$$

μx and μy are mean of x and mean of y respectively. E is the expectation.

### Assumptions of Karl Pearson's Correlation Coefficient

The assumptions and requirements for calculating Pearson's correlation coefficient are as follows:

1. The data set which is to be correlated should approximate to the normal distribution. If the data is normally distributed, then the data points tend to lie closer to the mean.

2. The word homoscedastic is a greek originated meaning 'able to disperse'. Homoscedasticity means 'equal variances'. For all the values of the independent variable, the error term is the same. Suppose the error term is smaller for a certain set of values of independent variable and larger for another set of values, then homoscedasticity is violated. It can be checked visually through a scatter plot. The data is said to be homoscedastic if the points lie equally on both sides of the line of best fit.

3. When the data follows a linear relationship, it is said to be linearity. If the data points are in the form of a straight line on the scatter plot, then the data satisfies the condition of linearity.

4. The variables which can take any value in an interval are continuous variables. The data set must contain continuous variables to compute the Pearson correlation coefficient. If one of the data sets is ordinal, then Spearman's rank correlation is an appropriate measure.

5. The data points must be in pairs which are termed as paired observations. There exists a dependent variable for every observation of the independent variable.

6. There must be no outliers in the data. If the outliers are present, then they can skew the correlation coefficient and make it inappropriate. A point is considered to be an outlier if it is beyond +3.29 or –3.29 standard deviations away. They can be easily determined visually from a scatter plot.

### Pearson Correlation Coefficient Formula

The linear correlation coefficient defines the degree of relation between two variables and is denoted by "r". It is also called as Cross correlation coefficient as it predicts the relation between two quantities. Now let us proceed to a statistical way of calculating the correlation coefficient.

If x & y are the two variables of discussion, then the correlation coefficient can be calculated using the formula

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

Here,

n = Number of values or elements

$\Sigma x$ = Sum of 1st values list

$\Sigma y$ = Sum of 2nd values list

$\Sigma xy$ = Sum of the product of 1st and 2nd values

$\Sigma x^2$ = Sum of squares of 1st values

$\Sigma y^2$ = Sum of squares of 2nd values

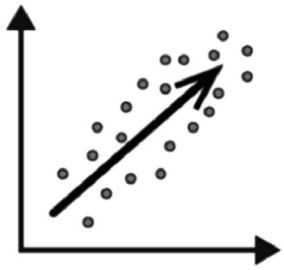
Rahul Publications

### finding Correlation Coefficient

Correlation is used almost everywhere in statistics. Correction illustrates the relationship between two or more variables. It is expressed in the form of a number that is known as correlation coefficient. There are mainly two types of correlations:

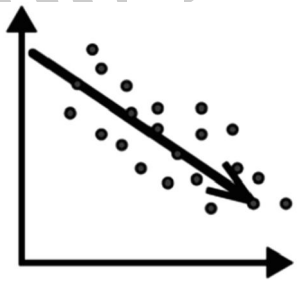
- Positive Correlation
- Negative Correlation

- **Positive Correlation :** The value of one variable increases linearly with increase in another variable. This indicates a similar relation between both the variables. So its correlation coefficient would be positive or 1 in this case.



Positive correlation

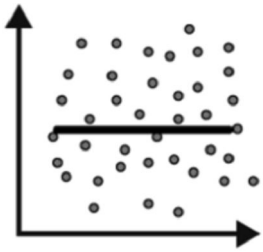
- **Negative Correlation:** When there is a decrease in values of one variable with increase in values of other variable. In that case, correlation coefficient would be negative.



Negative correlation

### Zero Correlation or No Correlation

There is one more situation when there is no specific relation between two variables.



No correlation

Rahul Publications

**Q23. What are the peroperties of correlation coefficient?**

*Ans :*

**Properties**

Correlation coefficient is all about establishing relationships between two variables. Some properties of correlation coefficient are as follows:

1) Correlation coefficient remains in the same measurement as in which the two variables are.
2) The sign which correlations of coefficient have will always be the same as the variance.
3) The numerical value of correlation of coefficient will be in between -1 to + 1. It is known as real number value.
4) The negative value of coefficient suggests that the correlation is strong and negative. And if 'r' goes on approaching toward -1 then it means that the relationship is going towards the negative side.

   When 'r' approaches to the side of + 1 then it means the relationship is strong and positive. By this we can say that if +1 is the result of the correlation then the relationship is in a positive state.
5) The weak correlation is signaled when the coefficient of correlation approaches to zero. When 'r' is near about zero then we can deduce that the relationship is weak.
6) Correlation coefficient can be very dicey because we cannot say that the participants are truthful or not.

   The coefficient of correlation is not affected when we interchange the two variables.
7) Coefficient of correlation is a pure number without effect of any units on it. It also not get affected when we add the same number to all the values of one variable. We can multiply all the variables by the same positive number. It does not affect the correlation coefficient. As we discussed, 'r 'is not affected by any unit because 'r' is a scale invariant.
8) We use correlation for measuring the association but that does not mean we are talking about causation. By this, we simply mean that when we are correlating the two variables then it might be the possibility that the third variable may be influencing them.

Rahul Publications

**PROBLEMS**

**1. Calculate the Correlation coefficient of given data:**

| | | | | | |
|---|---|---|---|---|---|
| **x** | **50** | **51** | **52** | **53** | **54** |
| **y** | **3.1** | **3.2** | **3.3** | **3.4** | **3.5** |

*Sol :*

Here n = 5

| | | | | | |
|---|---|---|---|---|---|
| x | 50 | 51 | 52 | 53 | 54 |
| y | 3.1 | 3.2 | 3.3 | 3.4 | 3.5 |
| xy | 155 | 163.2 | 171.6 | 180.2 | 189 |
| $x^2$ | 2500 | 2601 | 2704 | 2809 | 2916 |
| $y^2$ | 9.61 | 10.24 | 10.89 | 11.56 | 12.25 |

sum x = 260

sum y = 16.5

sumxy = 859

sum $x^2$ = 13530

sum $y^2$ = 54.55

By substituting all the values in formula, we get r = 1. This shows a positive correlation coefficient.

**2. Calculate the Correlation coefficient of given data:**

| | | | | | |
|---|---|---|---|---|---|
| **x** | **12** | **15** | **18** | **21** | **27** |
| **y** | **2** | **4** | **6** | **8** | **12** |

*Sol :*

Here n = 5

| | | | | | |
|---|---|---|---|---|---|
| x | 12 | 15 | 18 | 21 | 27 |
| y | 2 | 4 | 6 | 8 | 12 |
| xy | 24 | 60 | 94 | 168 | 324 |
| $x^2$ | 144 | 225 | 324 | 441 | 729 |
| $y^2$ | 4 | 16 | 36 | 64 | 144 |

sum x = 93

sum y = 32

sumxy = 670

sum $x^2$ = 1863

sum $y^2$ = 264

Now, putting all the values in below formula

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

We have, r = 0.84

**3. Explain, how to find Correlation Coefficient in Excel.**

*Ans :*

Finding the Correlation Coefficient in Excel:

**1. Using CORREL function**

In Excel to find the correlation coefficient use the formula :

=CORREL(array1,array2)

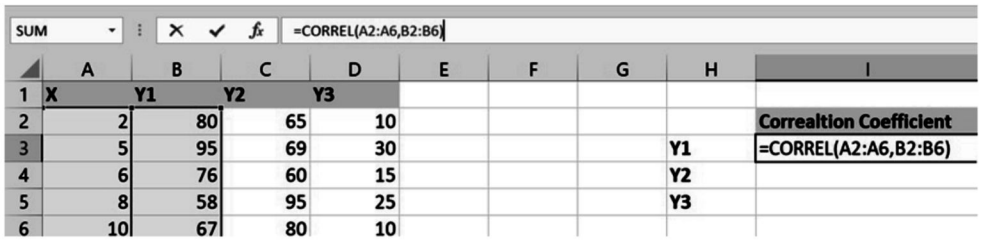
array1 : array of variable x

array2: array of variable y

Rahul Publications

To insert array1 and array2 just select the cell range for both.

1. Let's find the correlation coefficient for the variables and X and Y1.




Correlation coefficient of x and y1

**array1 :** Set of values of X. The cell range is from A2 to A6.

**array2 :** Set of values of Y1. The cell range is from B2 to B6.

Similarly, you can find the correlation coefficients for (X , Y2) and (X , Y3) using the Excel formula. Finally, the correlation coefficients are as follows :

| | Correlation Coefficient |
|---|---|
| XY1 | -0.627317836 |
| XY2 | 0.633224784 |
| XY3 | 0.01814885 |

From the above table we can infer that :

X and Y1 have negative correlation coefficient.

X and Y2 have positive correlation coefficient.

X and Y3 are not correlated as the correlation coefficient is almost zero.

**Example:**

Now, let's proceed to the further two methods using a new data set. Consider the following data set :

| X | Y1 | Y2 | Y3 |
|---|---|---|---|
| 2 | 80 | 45 | 10 |
| 5 | 95 | 79 | 30 |
| 6 | 66 | 94 | 15 |
| 8 | 58 | 90 | 25 |
| 10 | 47 | 98 | 10 |

Using Data Analysis

We can also analyze the given dataset and calculate the correlation coefficient: To do so follow the below steps:

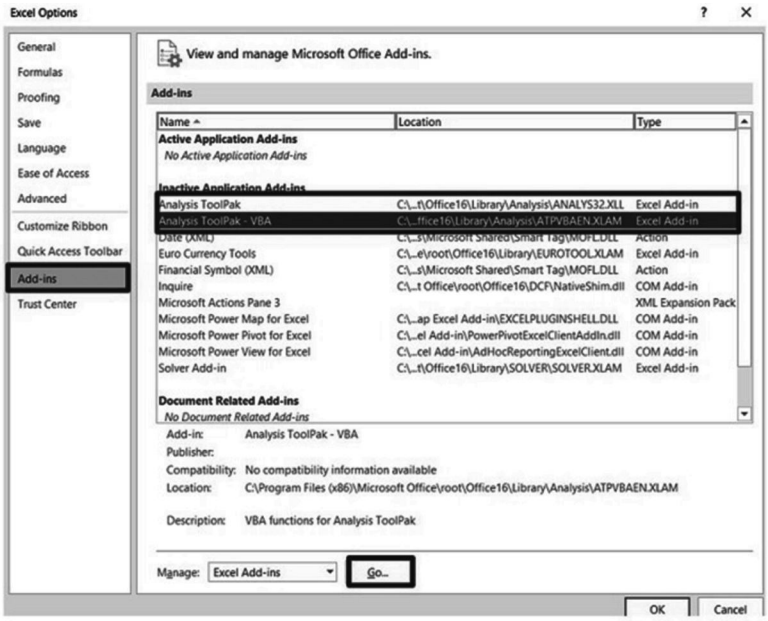
**Step 1:**

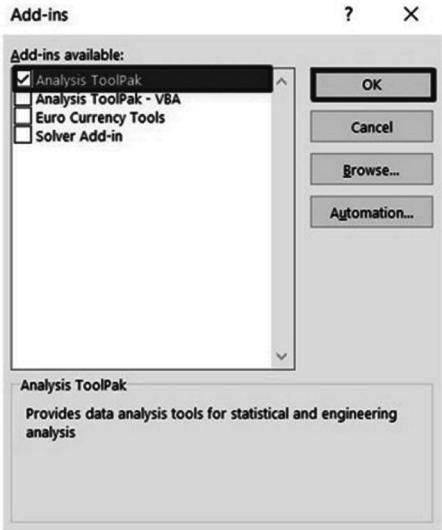
First you need to enable Data Analysis ToolPak in Excel. To enable :

1. Go to File tab in the top left corner of the Excel window and choose Options.
2. The Excel Options dialog box opens. Now go to the Add-Ins option and in the Manage select Excel Add-ins from the drop down.

Rahul Publications

3. Click on Go button.
4. The Add-ins dialog box opens. In this check the option Analysis ToolPak.
5. Click OK!



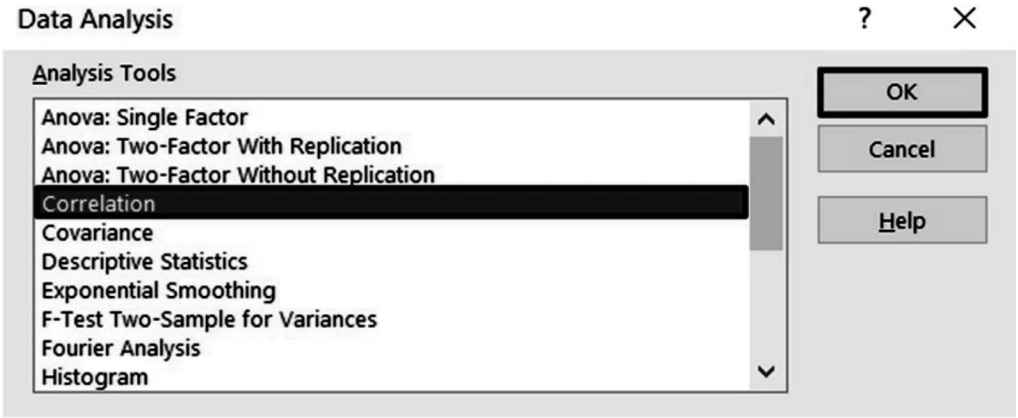
**Data Analysis tab added**

**Step 2:**

Now click on Data followed by Data Analysis. A dialog box will appear.

**Step 3:**

In the dialog box select Correlation from the list of options. Click OK!



**Step 4:**

The Correlation menu will appear.

**Step 5:**

In this menu first provide the Input Range. The input range is the cell range of X and Y1 columns as highlighted in the picture below.

**Step 6:**

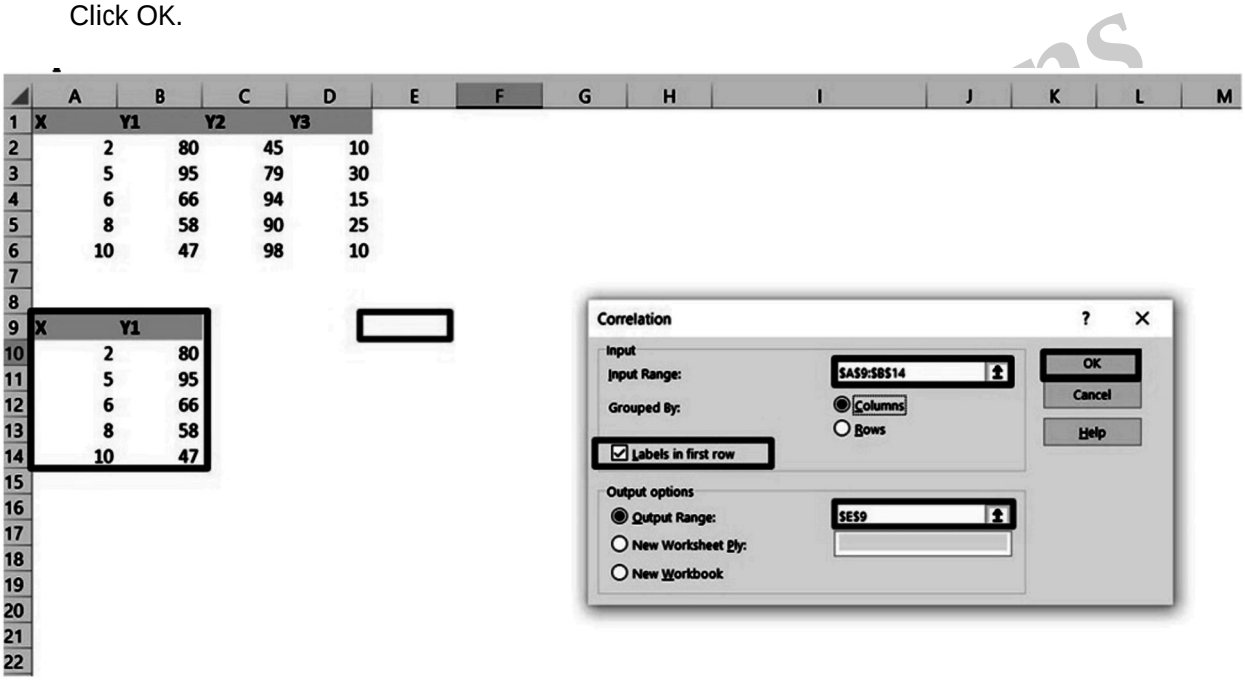
Also, supply the Output Range as the cell number where you want to display the result. By default, the output will appear in the new Excel sheet in case if you don't provide any Output Range.

**Step 7:**

Check the Labels in first-row option if you have labels in the dataset. In our case column 1 has label X and column 2 has label Y1.

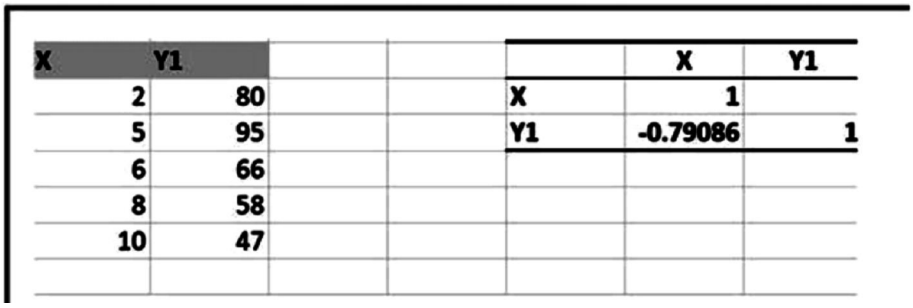
**Step 8:**

Click OK.




**Step 9:**

The Data Analysis table is now ready. Here, you can see the correlation coefficient between X and Y1 in the analysis table.



| X | Y1 |
|---|---|
| 2 | 80 |
| 5 | 95 |
| 6 | 66 |
| 8 | 58 |
| 10 | 47 |

| | X | Y1 |
|---|---|---|
| X | 1 | |
| Y1 | -0.79086 | 1 |

Similarly, you can find correlation coefficients of XY2 and that of XY3. Finally, all the correlation coefficients are :

| | X | Y1 |
|---|---|---|
| X | 1 | |
| Y1 | -0.79086 | 1 |

| | X | Y2 |
|---|---|---|
| X | 1 | |
| Y2 | 0.89109 | 1 |

| | X | Y3 |
|---|---|---|
| X | 1 | |
| Y3 | 0.01815 | 1 |

### Using PEARSON Function

It is exactly similar to the CORREL function which we have discussed in the above section. The syntax for PEARSON function is :

=PEARSON(array1,array2)

**array1 :** array of variable x

**array2:** array of variable y

To insert array1 and array2 just select the cell range for both.

Let's find the correlation coefficient for X and Y1 in the data set of Example 2 using PEARSON function.

Publications

SUM =PEARSON(A2:A6,B2:B6)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | X | Y1 | Y2 | Y3 | | | | | Correlation Coefficient |
| 2 | 2 | 80 | 45 | 10 | | | | XY1 | =PEARSON(A2:A6,B2:B6) |
| 3 | 5 | 95 | 79 | 30 | | | | | |
| 4 | 6 | 66 | 94 | 15 | | | | | |
| 5 | 8 | 58 | 90 | 25 | | | | | |
| 6 | 10 | 47 | 98 | 10 | | | | | |
| 7 | | | | | | | | | |

The formula will return the correlation coefficient of X and Y1. Similarly, you can do for others.

The final correlation coefficients are :

| | Correlation Coefficient |
|---|---|
| XY1 | -0.790857344 |
| XY2 | 0.891091446 |
| XY3 | 0.01814885 |

## PROBLEMS

**4. Find covariance for following data set x = {2,5,6,8,9}, y = {4,3,7,5,6}**

*Sol :*

Given data sets x = {2,5,6,8,9}, y = {4,3,7,5,6} and N = 5

Mean(x) = (2 + 5 + 6 + 8 + 9) / 5

= 30 / 5

= 6

Mean(y) = (4 + 3 +7 + 5 + 6) / 5

= 25 / 5

= 5

Sample covariance Cov(x,y) = $\Sigma(x_i - x) \times (y_i - y)/ (N - 1)$

= [(2 – 6)(4 – 5) + (5 – 6)(3 – 5) + (6 – 6)(7 – 5) + (8 – 6)(5 – 5) + (9 – 6)(6 – 5)] / 5 – 1

= 4 + 2 + 0 + 0 + 3 / 4

= 9 / 4

= 2.25

Population covariance Cov(x,y) = $\Sigma(x_i - x) \times (y_i - y)/ (N)$

= [(2 – 6)(4 – 5) + (5 – 6)(3 – 5) + (6 – 6)(7 – 5) + (8 – 6)(5 – 5) + (9 – 6)(6 – 5)] / 5

= 4 + 2 + 0 + 0 + 3 /

= 9 / 5

= 1.8

The sample covariance is 2.25 and the population covariance is 1.8.

**5. Using the covariance formula, find covariance for following data set x = {5,6,8,11,4,6}, y = {1,4,3,7,9,12}.**

*Sol :*

Given data sets x = {5,6,8,11,4,6}, y = {1,4,3,7,9,12} and N = 6

Mean(x) = (5 + 6 + 8 + 11 + 4 + 6) / 6

= 40 / 6

= 6.67

Mean(y) = (1 + 4 + 3 + 7 + 9 + 12) / 6

= 36 / 6

= 6

Using sample covariance formula Cov(x,y) = $\Sigma(x_i - x) \times (y_i - y)/ (N - 1)$

= – 0.4

Population covariance Cov(x,y) = $\Sigma(x_i - x) \times (y_i - y)/ (N)$

= – 0.33

The sample covariance is – 0.4 and the population covariance is – 0.33.

Rahul Publications

**6. Find covariance for following data set x = {13,15,17,18,19}, y = {10,11, 12,14,16} using the covariance formula.**

*Sol :*

Given data sets x = {13,15,17,18,19}, y = {10,11,12,14,16} and N = 5

Mean(x) = (13 + 15 + 17 + 18 + 19) / 5

= 82 / 5

= 16.4

Mean(y) = (10 + 11 +12 + 14 + 16) / 5

= 63 / 5

= 12.6

Sample covariance Cov(x,y) = $\Sigma(x_i - x) \times (y_i - y)/ (N - 1)$

= [(16.4 – 13)(12.6 – 10) + (16.4 – 15)(12.6 – 11) + (16.4 – 17)(12.6 – 12) + (16.4 –18) (12.6 – 14) + (16.4 – 19)(12.6 – 16)] / 5 – 1

= (8.84 + 2.24 – 0.36 + 2.24 + 8.84) / 4

= 21.8 / 4

= 5.45

Population covariance Cov(x,y) = $\Sigma(x_i - x) \times (y_i - y)/ (N)$

= [(16.4 – 13)(12.6 – 10) + (16.4 – 15)(12.6 – 11) + (16.4 – 17)(12.6 – 12) + (16.4 – 18) (12.6 – 14) + (16.4 – 19)(12.6 – 16)] / 5

= (8.84 + 2.24 – 0.36 + 2.24 + 8.84) / 5

= 21.8 / 5

= 4.36

The sample covariance is 5.45 and the population covariance is 4.36.

**7. Compute Pearsons coefficient of correlation between advertisement cost and sales as per the data given below:**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Advertisement Cost in 1000's** | **39** | **65** | **62** | **90** | **82** | **75** | **25** | **98** | **36** | **78** |
| **Sales in lakhs** | **47** | **53** | **58** | **86** | **62** | **68** | **60** | **91** | **51** | **84** |

*Sol :*

Ho: The correlation coefficient r is not significant

H1: The correlation coefficient r is significant.

Level of significance 5%

From the data

n = 10

Rahul Publications

$\sum x = 650 \sum y = 660 \sum xy = 45604 \sum x^2 = 47648 \sum y^2 = 4578$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}}$$

$$= \frac{45604 - \frac{(650)(660)}{10}}{\sqrt{47648 - \frac{(650)^2}{10}}\sqrt{45784 - \frac{(660)^2}{10}}}$$

$$= \frac{45604 - 42900}{(73.47)(47.1)} = 0.7804$$

Correlation coefficient is positively correlated.

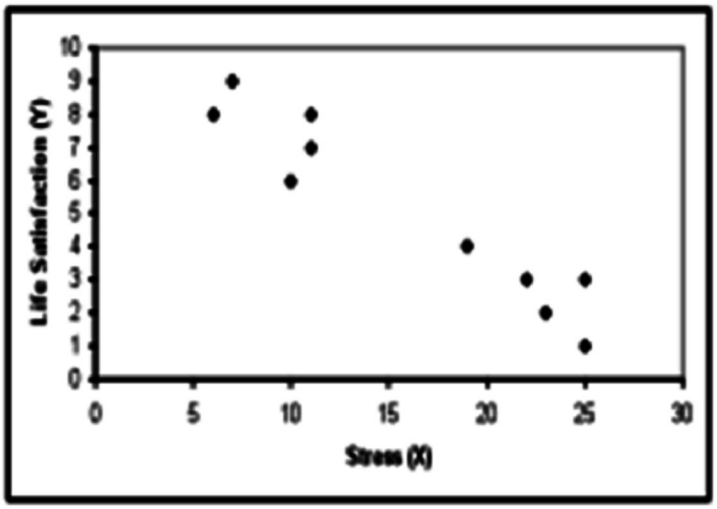
**8. Janice and Paul did a study on feelings of stress and life satisfaction. Participants completed a measure on how stressed they were feeling (on a 1 to 30 scale) and a measure of how satisfied they felt with their lives (measured on a 1 to 10 scale). The table below indicates the participants' scores. Using this data, answer the following questions:**

| Participant | Stress score (X) | Life Satisfaction (Y) |
|---|---|---|
| 1 | 11 | 7 |
| 2 | 25 | 1 |
| 3 | 19 | 4 |
| 4 | 7 | 9 |
| 5 | 23 | 2 |
| 6 | 6 | 8 |
| 7 | 11 | 8 |
| 8 | 22 | 3 |
| 9 | 25 | 3 |
| 10 | 10 | 6 |
| Σ | 159 | 51 |
| Mean | 15.9 | 5.1 |
| SD | 7.23 | 2.70 |

Rahul Publications

*Sol :*

(a) On a scrap paper, try to draw a rough scatterplot of the data, just to get an idea of what these look like. You didn't have to include this, but essentially a scatterplot would look something like this:




There are multiple ways to calculate a correlation coefficient r (that is, a standardized indicator of the relation between two variables). We can calculate the covariation between two variables (X and Y), and then adjust this by the standard deviation and sample size.

$$r = \frac{\text{cov}(x,y)}{S_x S_y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{(N)S_x S_y}$$

Alternatively, we can calculate r from the Z scores :

$$r = \frac{\sum(Z_x Z_y)}{N}$$

$$r = \frac{\text{cov}(x,y)}{S_x S_y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{(N)S_x S_y}$$

Alternatively, we can calculate r from the Z scores :

$$r = \frac{\sum(Z_x Z_y)}{N}$$

(b) Using either method, calculate the correlation (r) between stress and life satisfaction. This was probably the most time intensive element of this assignment. We can use either formula – I've done both here as an example, with the covariance method in green and the Z score method in red.

| # | X | $X-M_x$ | $Z_x$ | Y | $Y - M_v$ | Zy | (X-M,)( $Y-M_v$) | ZxZy |
|---|---|---|---|---|---|---|---|---|
| 1 | 11 | –4.9 | –0.68 | 7 | 1.9 | 0.70 | –9.31 | –0.48 |
| 2 | 25 | 9.1 | 1.26 | 1 | –4.1 | –1.52 | –37.31 | –1.91 |
| 3 | 19 | 3.1 | 0.43 | 4 | –1.1 | –0.41 | –3.41 | –0.17 |
| 4 | 7 | –8.9 | –1.23 | 9 | 3.9 | 1.44 | –34.71 | –1.78 |
| 5 | 23 | 7.1 | 0.98 | 2 | –3.1 | –1.15 | –22.01 | –1.13 |
| 6 | 6 | –9.9 | –1.37 | 8 | 2.9 | 1.07 | –28.71 | –1.47 |
| 7 | 11 | –4.9 | –0.68 | 8 | 2.9 | 1.07 | –14.21 | –0.73 |
| 8 | 22 | 6.1 | 0.84 | 3 | –2.1 | –0.78 | –12.81 | –0.66 |
| 9 | 25 | 9.1 | 1.26 | 3 | –2.1 | –0.78 | –19.11 | –0.98 |
| 10 | 10 | –5.9 | –0.82 | 6 | 0.9 | 0.33 | –5.31 | –0.27 |
| Σ | 159 | 0 | 0.00 | 51 | 0.00 | 0.00 | –186.90 | –9.57 |
| Mean | 15.9 | | | 5.1 | | | | |
| ξ | 7.23 | | | 2.70 | | | | |

The we add these to the formulas.

Using the covariance : $r = \frac{(-186.90)}{(10 * 7.23 * 2.70)} = -.957$

Using the covariance : $r = \frac{-9.57}{10} = -.957$

Round to 2 decimal points, so r = .96

Rahul Publications

# Short Question and Answers

**1. What is variability?**

*Ans :*

**Meaning**

Variability describes how far apart data points lie from each other and from the center of a distribution. Along with measures of central tendency, measures of variability give you descriptive statistics that summarize your data.

Variability is also referred to as spread, scatter or dispersion. It is most commonly measured with the following:

- **Range:** the difference between the highest and lowest values
- **Interquartile range:** the range of the middle half of a distribution
- **Standard deviation:** average distance from the mean
- **Variance:** average of squared distances from the mean

**2. What is Range?**

*Ans :*

The range in statistics for a given data set is the difference between the highest and lowest values. For example, if the given data set is {2,5,8,10,3}, then the range will be 10 – 2 = 8.

Thus, the range could also be defined as the difference between the highest observation and lowest observation. The obtained result is called the range of observation. The range in statistics represents the spread of observations.

**Formula**

The formula of the range in statistics, can simply be given by the difference between highest and lowest value.

Range = Highest Value – Lowest Value

OR

Range = Highest observation – Lowest observation

OR

Range = Maximum value – Minimum Value

**3. What is variance?**

*Ans :*

Variance is a measure of how data points differ from the mean. According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value.

Variance means to find the expected difference of deviation from actual value. Therefore, variance depends on the standard deviation of the given data set.

The more the value of variance, the data is more scattered from its mean and if the value of variance is low or minimum, then it is less scattered from mean. Therefore, it is called a measure of spread of data from mean.

**Population variance formula**

Use the population form of the equation when you have values for all members of the group of interest. In this case, you are not using the sample to estimate the population. Instead, you have measured all people or items and need the variance for that specific group. For example, if you have measured test scores for all class members and need to know the value for that class, use the population variance formula.

The formula for the variance of an entire population is the following:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{n}$$

In the population variance formula:

- $\sigma^2$ is the population variance.
- $X_i$ is the $i^{th}$ data point.
- $\mu$ is the population mean.
- n is the number of observations.

To find the variance, take a data point, subtract the population mean, and square that difference. Repeat this process for all data points. Then, sum all of those squared values and divide by the number of observations. Hence, it's the average squared difference.

Rahul Publications

**4. What is standard deviation?**

*Ans :*

Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a "typical" deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set. Like the variance, if the data points are close to the mean, there is a small variation whereas the data points are highly spread out from the mean, then it has a high variance. Standard deviation calculates the extent to which the values differ from the average. Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore a change in even one value affects the value of standard deviation. It is independent of origin but not of scale. It is also useful in certain advanced statistical problems.

**Formula**

The population standard deviation formula is given as:

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

Here,

$\sigma$ = Population standard deviation

N = Number of observations in population

$X_i$ = ith observation in the population

$\mu$ = Population mean

Similarly, the sample standard deviation formula is:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)}$$

Here,

s = Sample standard deviation

n = Number of observations in sample

xi = ith observation in the sample

$\bar{x}$ = Sample mean

**5. Coefficient of Variation.**

*Ans :*

The coefficient of variation is a type of measure of dispersion. A measure of dispersion is a quantity that is used to gauge the extent of variability of data. Thus, the coefficient of variation is used to measure the dispersion of data from the average or the mean value. CV is the abbreviated form of the coefficient of variation.

The formula for coefficient of variation is given below:

$$\text{Coefficient of variation} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100\%$$

As per sample and population data type, the formula for standard deviation may vary.

Sample Standard Deviation = $\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$

Population Standard Deviation = $\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$

Where,

$x_i$ = Terms given in the data

$\bar{x}$ = Mean

n = Total number of terms.

**6. Percentiles**

*Ans :*

Percentile is defined as the value below which a given percentage falls under. For example, in a group of 20 children, Ben is the 4th tallest and 80% of the children are shorter than you. Hence, it means that Ben is at the 80th percentile. It is most commonly used in competitive exams such as SAT, LSAT, etc.

The Percentile Formula is given as,

Percentile = (Number of Values Below "x" / Total Number of Values) × 100

Another formula to find the percentile is given by:

P = (n/N) × 100

P = (nth percentile/100) × Total number of values in the list

Here,

n = Ordinal rank of the given value or value below the number

N = Number of values in the data set

P = Percentile

Rank = Percentile/100

Ordinal rank for Percentile value = Rank × Total number of values in the list

**7. Quartiles.**

*Ans :*

Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by $Q_1$, $Q_2$ and $Q_3$, respectively. $Q_2$ is nothing but the median, since it indicates the position of the item in the list and thus, is a positional average. To find quartiles of a group of data, we have to arrange the data in ascending order.

**Quartiles Formula**

Suppose, $Q_3$ is the upper quartile is the median of the upper half of the data set. Whereas, $Q_1$ is the lower quartile and median of the lower half of the data set. $Q_2$ is the median. Consider, we have n number of items in a data set. Then the quartiles are given by;

$Q_1 = [(n+1)/4]$th item

$Q_2 = [(n+1)/2]$th item

$Q_3 = [3(n+1)/4]$th item

Hence, the formula for quartile can be given by;

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f}(l_2 - l_1)$$

Where, Qr is the $r^{th}$ quartile

$l_1$ is the lower limit

$l_2$ is the upper limit

f is the frequency

c is the cumulative frequency of the class preceding the quartile class.

**8. Empirical Rule.**

*Ans :*

The empirical rule is a statistical rule (also called the three-sigma rule or the 68-95-99.7 rule) which states that, for normally distributed data, almost all of the data will fall within three standard deviations either side of the mean.

More specifically, you'll find:

- **68% of data within 1 standard deviation**
- **95% of data within 2 standard deviations**
- **99.7% of data within 3 standard deviations**

Standard deviation is a measure of spread; it tells how much the data varies from the average, i.e., how diverse the dataset is. The smaller value, the more narrow the range of data is.

Normal distribution is a distribution that is symmetric about the mean, with data near the mean are more frequent in occurrence than data far from the mean.

**9. Box Plots.**

*Ans :*

The method to summarize a set of data that is measured using an interval scale is called a box and whisker plot. These are maximum used for data analysis. We use these types of graphs or graphical representation to know:

- Distribution Shape
- Central Value of it
- Variability of it

A box plot is a chart that shows data from a five-number summary including one of the measures of central tendency. It does not show the distribution in particular as much as a stem and leaf plot or histogram does. But it is primarily used to indicate a distribution is skewed or not and if there are potential unusual observations (also called outliers) present in the data set. Boxplots are also very beneficial when large numbers of data sets are involved or compared.

In simple words, we can define the box plot in terms of descriptive statistics related concepts. That means box or whiskers plot is a method used for depicting groups of numerical data through their quartiles graphically. These may also have some lines extending from the boxes or whiskers which indicates the variability outside the lower and upper quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers can be indicated as individual points.

**10. Covariance.**

*Ans :*

Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

**Types of Covariance**

Covariance can have both positive and negative values. Based on this, it has two types:

1. Positive Covariance
2. Negative Covariance

Rahul Publications

**1. Positive Covariance**

If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

**2. Negative Covariance**

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

**Covariance Formula**

Covariance formula is a statistical formula, used to evaluate the relationship between two variables. It is one of the statistical measurements to know the relationship between the variance between the two variables. Let us say X and Y are any two variables, whose relationship has to be calculated. Thus the covariance of these two variables is denoted by Cov(X,Y). The formula is given below for both population covariance and sample covariance.

Population covariance

$$Cov(x, y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample covariance

$$Cov(x, y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Where,

- xi = data value of x
- yi = data value of y
- x = mean of x
- 3 = mean of y
- N = number of data values.

**11. What is correlation coefficient?**

*Ans :*

Correlation Coefficient is a statistical concept, which helps in establishing a relation between predicted and actual values obtained in a statistical experiment. The calculated value of the correlation coefficient explains the exactness between the predicted and actual values.

Correlation Coefficient value always lies between -1 to +1. If correlation coefficient value is positive, then there is a similar and identical relation between the two variables. Else it indicates the dissimilarity between the two variables.

The covariance of two variables divided by the product of their standard deviations gives Pearson's correlation coefficient. It is usually represented by ρ (rho).

$\rho$ (X,Y) = cov (X,Y) / $\sigma$X.$\sigma$Y.

Here cov is the covariance. óX is the standard deviation of X and óY is the standard deviation of Y. The given equation for correlation coefficient can be expressed in terms of means and expectations.

$$\rho(X, Y) = E\ \frac{(X - \mu_x)(Y - \mu_y)}{\sigma x.\sigma y}$$

$\mu$x and $\mu$y are mean of x and mean of y respectively. E is the expectation.

Rahul Publications

# Choose the Correct Answers

1. Which of the following are methods under measures of dispersion? [ d ]

   (a) Standard deviation (b) Mean deviation

   (c) Range (d) All of the above

2. Which of the following are characteristics of a good measure of dispersion? [ d ]

   (a) It should be easy to calculate

   (b) It should be based on all the observations within a series

   (c) It should not be affected by the fluctuations within the sampling

   (d) All of the above

3. The coefficient of variation is a percentage expression for _____. [ a ]

   (a) Standard deviation (b) Quartile deviation

   (c) Mean deviation (d) None of the above

4. While calculating the standard deviation, the deviations are only taken from _____. [ d ]

   (a) The mode value of a series (b) The median value of a series

   (c) The quartile value of a series (d) The mean value of a series

5. The numerical value of a standard deviation can never be_____. [ a ]

   (a) Negative (b) Zero

   (c) Larger than the variance (d) None of the above

6. The average of squared deviations from the arithmetic mean is known as _____. [ c ]

   (a) Quartile deviation (b) Standard deviation

   (c) Variance (d) None of the above

7. Which of the following cannot be calculated for open-ended distributions? [ b ]

   (a) Standard deviation (b) Mean deviation

   (c) Range (d) None of the above

8. The standard deviation of a set of 90 observations is 105. If the value of each observation is decreased by 9, then the new standard deviation of these observations would be _____. [ c ]

   (a) 96 (b) 100

   (c) 105 (d) None of the above

Rahul Publications

9. The average daily wage of 100 workers in a shipyard was Rs. 200, with a standard deviation of 40. Now, if each worker gets an increment of 20% in their wages, how will it affect the mean wage? [ d ]

(a) The mean wage will remain unchanged

(b) The mean wage will be increased by 20%

(c) The mean wage will be Rs. 240

(d) Both b and c

10. If you secure 97 percentile in an examination, it means that your position is below _____of the total candidates who had appeared in the exam. [ b ]

(a) 97 percent (b) 3 percent

(c) 90 percent (d) None of the above

11. Which of the following measures of dispersion can attain a negative value? [ b ]

(a) Mean deviation (b) Range

(c) Standard deviation (d) None of the above

12. The range represents _____. [ a ]

(a) The lowest number

(b) The highest number

(c) The middle number

(d) The difference between the lowest and highest number

13. The square of standard deviation is _____ . [ c ]

(a) Square deviation (b) Mean square deviation

(c) Variance (d) None of the above

14. An example of the application of range in a real-world scenario would be _____. [ d ]

(a) Fluctuation in share prices (b) Weather forecasts

(c) Quality control (d) All of the above

15. The scatter within a distribution that is high on each side indicates _____. [ b ]

(a) High uniformity of data (b) Outliers of data

(c) Low uniformity of data (d) None of the above

Rahul Publications

# Fill in the Blanks

1. _____ the difference between the highest and lowest values.
2. The formula for the variance of an entire population is _____.
3. A _____ is a set of data extracted from the entire population.
4. _____ is defined as the value below which a given percentage falls under.
5. _____ divide the entire set into four equal parts.
6. _____ is defined as half of the distance between the third and the first quartile.
7. The interquartile range (IQR) is the difference between the upper and lower quartile of a given data set and is also called a _____.
8. The box plot is said to be _____ if the median is equidistant from the maximum and minimum values.
9. A _____ is a means to represent data in a graphical format.
10. _____ is a measure of the relationship between two random variables and to what extent, they change together.

**ANSWERS**

1. Range
2. $s^2 = \frac{\sum (X_i - \mu)^2}{n}$
3. Sample
4. Percentile
5. Quartiles
6. Quartile deviation
7. Midspread
8. Symmetric
9. Scatter plot
10. Covariance

Rahul Publications

UNIT IV

**PREDICTIVE ANALYTICS :**

Trend Analysis, Regression Analysis- Least Square Method, Assessing the Fit of Simple Linear Regression, Coefficient of Determination, Introduction to Data Mining- Definition, Methods of Data Mining, Applications of Data Mining.

## 4.1 PREDICTIVE ANALYTICS

**Q1. What is predictive analysis ?**

*Ans :*

Predictive Analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. It does not tell you what will happen in the future.

Predictive analytics is the branch of the Advanced Analytics which is used to make predictions about unknown future events. It uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future.

Predictive Analytics uses historical data to predict future events. Typically, historical data is used to build a mathematical model that captures important trends. That predictive model is then used on current data to predict what will happen next, or to suggest actions to take for optimal outcomes.

**Q2. How predictive analysis works ?**

*Ans :*

Predictive Analytics is the process of using data analytics to make predictions based on data. This process uses data along with analysis, statistics and machine learning techniques to create a predictive model for forecasting future events.

The term "predictive analytics" describes the application of a statistical or machine learning technique to create a quantitative prediction about the future. Frequently, supervised machine learning techniques are used to predict a future value (How long can this machine run before requiring maintenance?) or to estimate a probability (How likely is this customer to default on a loan?).

Predictive Analytics starts with a business goal to use data to reduce waste, save time or cut costs. The process harnesses heterogeneous, often massive, data sets into models that can generate clear, actionable outcomes to support achieving that goal, such as less material waste, less stocked inventory, and manufactured product that meets specifications.

**Q3. Explain the Applications of Predictive analysis.**

*Ans :* **(Imp.)**

Predictive analytics is widely applicable in different areas of an organization. In order to have a clear image of application of predictive analytics, industry wise role of predictive analytics is discussed below,

**1. Airline industry**

In the airline industry, proper planning can be done by predicting the future demand. By applying predictive analytics, airline operators can predict future demand and design customer centric policies and rewards for loyal customers.

**2. Insurance Industry**

In the insurance sector, predictive analytics is applied to detect any kind of insurance fraud. It facilitates the insurance companies to decode the hidden patterns depending on proper analysis of past data to identify the customer who default the premium payments.

Rahul Publications

**3. Banking Sector**

In banking sector, predictive analytics is used to identify frauds relating to credit card and to predict the future preference of the customers so that specific financial product and services can be designed to fulfill the needs of the bank customers.

**4. Manufacturing Sector**

In this competitive environment, manufacturing organizations are required to provide uninterrupted flow of the products because any delay in production cycle will effect the future prospects of the organization. Delay also results in loss of time and resources which are not affordable. By applying predictive analytics, future breakdown of machines can be predicted so that organization can take corrective actions to prevent uninterrupted manufacturing of products.

In conclusion, we can state that predictive analysis models are playing a very important role in the over all business world

### 4.1.1 Trend Analysis

**Q4. Explain briefly about Trend Analysis.**

*Ans :*

A trend line is a straight line connecting a number of points on a graph. It is used to analyze the specific direction of a group of values set in a presentation. There are two kinds of trend lines, an uptrend with values going higher, and a downtrend where the direction of the line gradually drops to the lower values.

**(i) Predicting the Future**

Trend lines allow businesses to see the difference in various points over a period of time. This helps foretell the possible path the values will take in the future. This can help reveal perform Predictive Trend Line. The Predictive Trend Line add-on displays linear, exponential, and logarithmic trend lines in your dashboard charts. The trend line allows an end user to predict future values/metrics based on historical data. See the referenced wiki page on the regression analysis for more information business departments, such as sales.

By knowing how to add a trend line to your presentation, you can create a graphical representation of the values you have computed. This will enable the user to easily comprehend and analyze the message you are trying to imply.

**(ii) Predictive Trend Line**

The Predictive Trend Line add-on displays linear, exponential, and logarithmic trend lines in your dashboard charts. The trend line allows an end user to predict future values/ metrics based on historical data. See the referenced wiki page on the regression analysis for more information.

**Q5. Describe the components of Trend analysis.**

*Ans :*

**1. Trend Movements**

The trend movements are also known as long-term movements. They specify the direction of a time series graph over a certain period of time.

The trend movement can be represented with the help of a trend curve or a trend line. The trend curve can be determined using the two methods, the weighted moving average method and least squares method.

**2. Cyclic Movements**

These movements are also known as cyclic variations. They typically refer to the long-term periodic (or) non-periodic oscillations of a trend line on curve.

In other words, it refers to the cycles that may (or) may not follow the same sequence after equal time intervals.

**3. Seasonal Movements**

The seasonal movements are also known as seasonal variations. They refer to the events that occur every year.

For example, increase in the sales of crackers before Deepawali festival.

Rahul Publications

From the above example, we can identify the seasonal movements (increase in sales of crackers) that occur during a certain time period (month of Deepawali).

**4. Random Movements**

These movements can also be known as irregular movements. They refer to the changes that occur randomly because of some unplanned events.

**Example:** damages caused by natural disasters, calamities etc.

## 4.2 Regression Analysis

**Q6. What is Regression Analysis?**

*Ans :*

**Regression**

Regression refers to the average relationship between two or more variables. One of these variables is called the dependent or the explained variable and the other variable is the independent or the explaining variable.

Regression is one of the statistical method that is used to estimate the unknown value of one variable from the known value of the related variable.

**Regression Analysis**

1. Regression studies 'nature' of relationship between the variables.
2. Regression means stepping back or returning to average value and is a mathematical measure expressing the average relationship between the variables.
3. Regression clearly indicates the cause and effect relationship between the variables.
4. In regression, the variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
5. Regression coefficients are absolute measures indicating the change in the value of one variable for a unit change in the value of the other variable.
6. Regression analysis is very helpful in predicting and estimating value of one variable given the value of another variable. Regression coefficients are asymmetric i.e., $b_{xy} \neq b_{yx}$.
7. The range of $b_{xy}$ and $b_{yx}$ is not restricted. Regression coefficients cannot be directly compared from correlation coefficient. There is no such thing in regression.

**Q7. Explain different types and utility of regression analysis.**

*Ans :* **(Imp.)**

**Types of Regression**

**a) Simple Regression**

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

**b) Multiple Regression**

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

**c) Linear Regression**

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

**d) Nonlinear Regression**

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

**Applications / Utility of Regression Test**

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

**Example**

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

Rahul Publications

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

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

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

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

**Q8. Explain the limitations of regression analysis.**

*Ans :*

**Limitations**

Some of the limitations of regression analysis are as follows :

1. Regression analysis assumes that linear relationship exists among the related variables. But in the 'area of social sciences, linear relationship may not exist among the related variables.

2. When regression analysis is used to evaluate the value of dependent variable based on independent variable, it is assumed that the static conditions of relationship exist between 'them. These statistic conditions do not exist in social sciences, so, this assumption "minimizes the use' of regression analysis in social science.

3. The value of dependent variable can be evaluated based on independent variable by using regression analysis but only upto some limits. If the circumstances go beyond the limits, then resists would be inaccurate.

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

**Q9. Explain about regression equation.**

*Ans :*

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

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

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

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

**Formation of Regression Equations**

There are two ways of forming regression equations as follows,

a) Normal equation and

b) Regression coefficient.

**Formation of Regression Equation through Normal Equation**

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

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

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

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

Rahul Publications

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

**Regression Equation of Y on X**

$Y = a + bX$

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

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

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

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

**Regression Equation of X on Y**

$Yc = a + by$

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

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

$\Sigma xy = a\Sigma y + b\Sigma y2$ ... (2)

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

b) Regression Coefficients

To estimate values of population parameter b0 and b1, under certain assumptions, the fitted or estimated regression equation representing the straight line regression model is written as :

$= a + bx$

where

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

a or $b_0$ = y–intercept that represents average value of

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

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

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

The regression coefficient 'b' is also denoted as :

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

### 4.2.1 Regression by Method of Least Square

**Q10. Discuss briefly about least square regression.**

*Ans :* **(Imp.)**

Least-squares regression is defined as the platform for best fitting the regression line and the observed values. In regression analysis, an assumption is made that in the sample data every value of dependent variable ' F 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 'F and the number of miles travelled is represented by independent variable X. The travel time 'F 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 Y is given by,

Rahul Publications

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

In the above equation, $\beta_0$ and $\beta_1$ refers to population parameters representing the intercept and slope, respectively. If X= 0, the intercept '$\beta_0$' refers to expected value of Y and the slope '$\beta_1$' 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,

$$Y = \beta_0 + \beta_1 X + \in$$

The expected value of 'Y' for a given value of $\beta_0$ and $\beta_1$ are can 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 $\beta_0$ and $\beta_1$, respectively.

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

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

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

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,

$$e_k = y_k - \hat{y}_k \qquad ...(3)$$

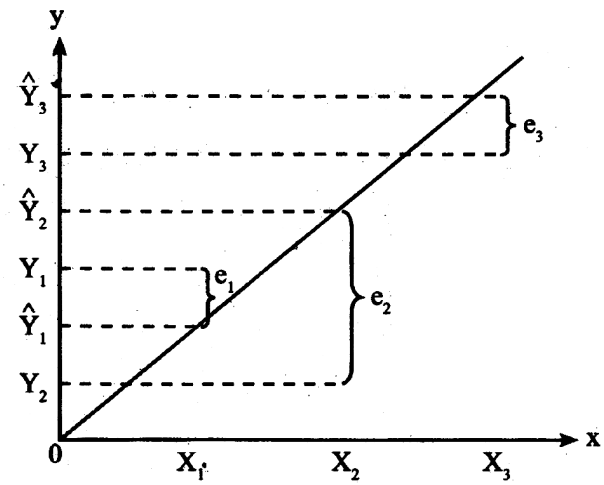



**Fig: Errors of Observation Values**

Rahul Publications

By adding the squares of the errors, we get,

$$\sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n} (Y_k - \hat{Y}_K)^2$$

$$= \sum_{k=1}^{n} (Y_k - [b_0 + b_1 X_k])^2 \qquad ....(4)$$

The best fitting regression line is obtained by finding the best values of the slope and intercept that minimizes the sum of squares of the errors.

In the above equation, $X_K$ and $Y_K$ refers to values of sample data and $b_0$ + $b_1$ are unknowns.

Differential calculations is applied on equation (4) to show that the values of $b_0$ and $b_1$ that minimizes the sum of squares of errors is,

$$\boxed{b_1 = \frac{\sum_{k=1}^{n} X_K Y_k - n\overline{XY}}{\sum_{k=1}^{n} X_K^2 - n\overline{X}^2}}$$

$$\boxed{b_0 = \overline{Y} - b_1 \overline{X}}$$

In excel, least-squares coefficients $b_0$ + $b_1$ are determined using the functions INTERCEPT (y, x) and SLOPE (y, x).

## PROBLEMS

**1. Find the Least Square method for following data. Find the production in 2005?**

| Year | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|
| Production (` 000) | 10 | 19 | 14 | 17 | 15 |

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

The straight line trend is $y_e = a + bc$

By solving the normal equations, we get a and b

Normal equations are,

$\Sigma y = Na + b\Sigma x$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

| Year (x) | Production (Y) | Deviation from mid | $x^2$ | xy |
|---|---|---|---|---|
| 1997 | 10 | 15 + 0.8 × (–2) = 13.4 | 4 | –20 |
| 1998 | 19 | 15 + 0.8 ×(–1) = 14.2 | 1 | –19 |
| 1999 | 14 | 15 + 0.8 × (0) = 15 | 0 | 0 |
| 2000 | 17 | 15 + 0.8 × (1) = 15.8 | 1 | 17 |
| 2001 | 15 | 15 + 0.8 × (0) = 16.6 | 4 | 30 |
| N = 5 | $\Sigma y$ = 75 | $\Sigma x$ = 0 | $\Sigma x^2$ = 10 | $\Sigma xy$ = 8 |

Since, $\Sigma x = 0$,

$$a = \frac{\Sigma y}{N} = \frac{75}{5} = 15$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{8}{10} = 0.8$$

$\therefore$ The straight line trend $y_e = a + bx$ is $y_e = 15 + 0.8\,x$

Estimation of production for the year 2005

For year 2005, deviation $x = 6$

$\therefore y_e = a + bx$

$= 15 + 0.8\,(6)$

$= 15 + 4.8$

$= 19.8$

For the year 2005, the production for the year is estimated to be ` 19.8

### 4.2.2 Assessing the Fit of Simple Linear Regression

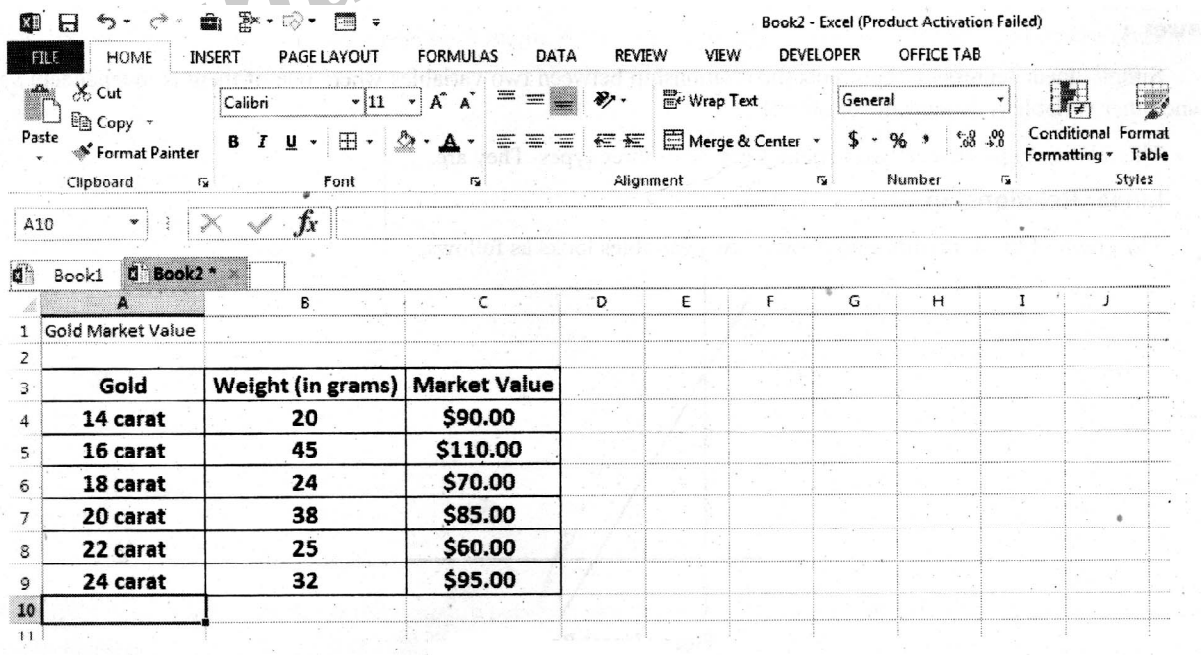
**Q11. Explain the concept of simple linear regression by using MS Excel.**

*Ans :*

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

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

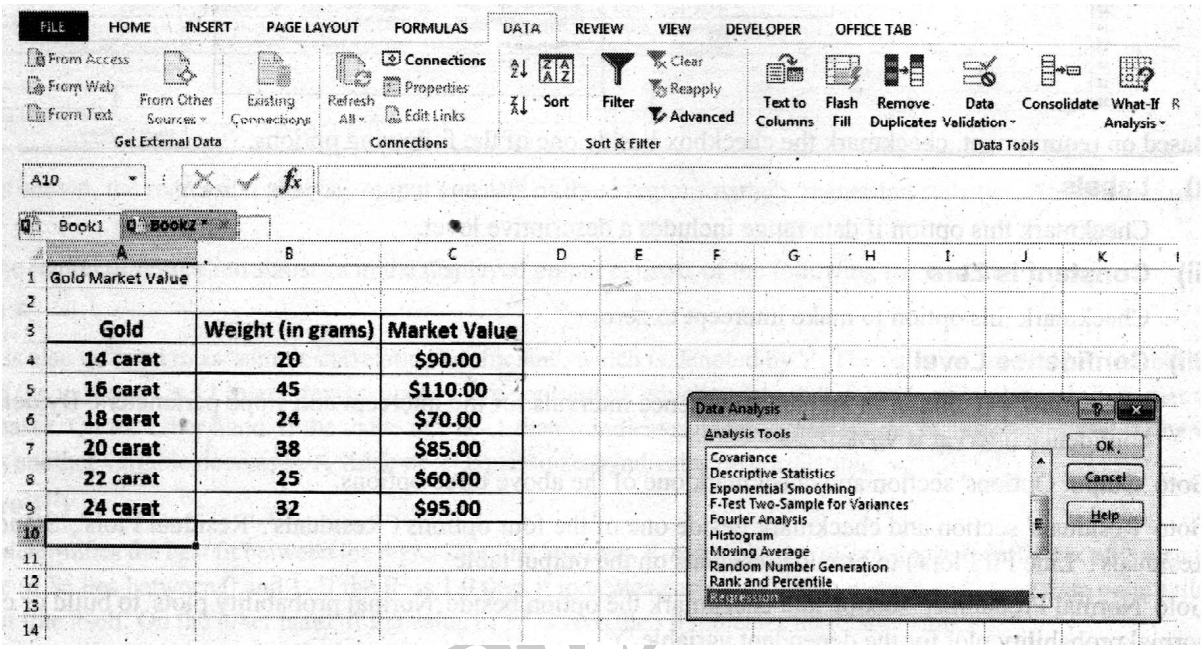
1. Select the data wherein user want to apply regression.



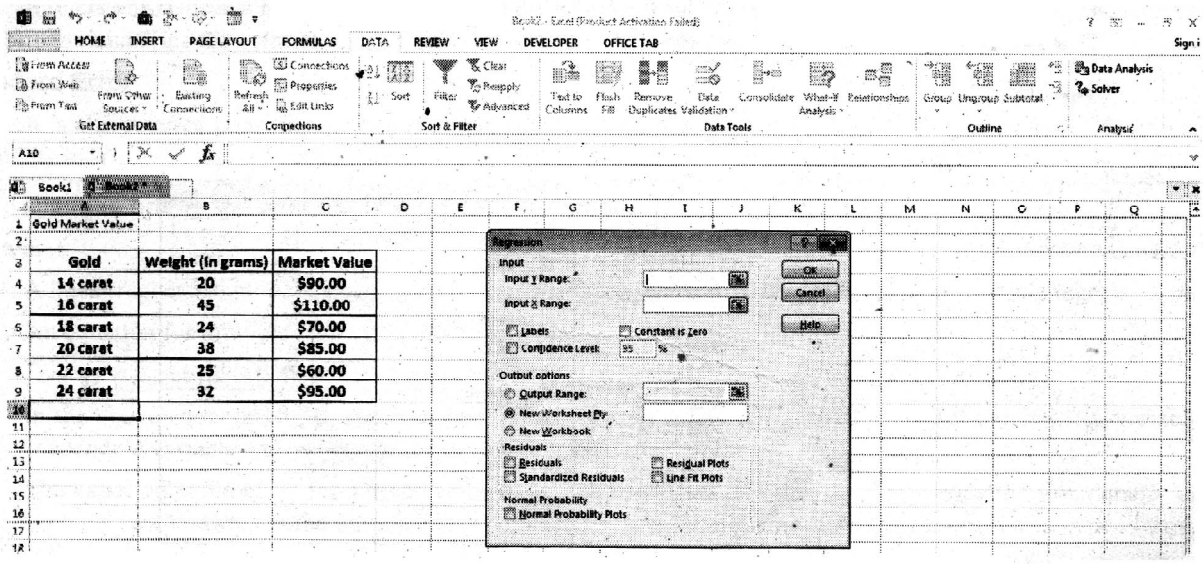

hul Publications

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

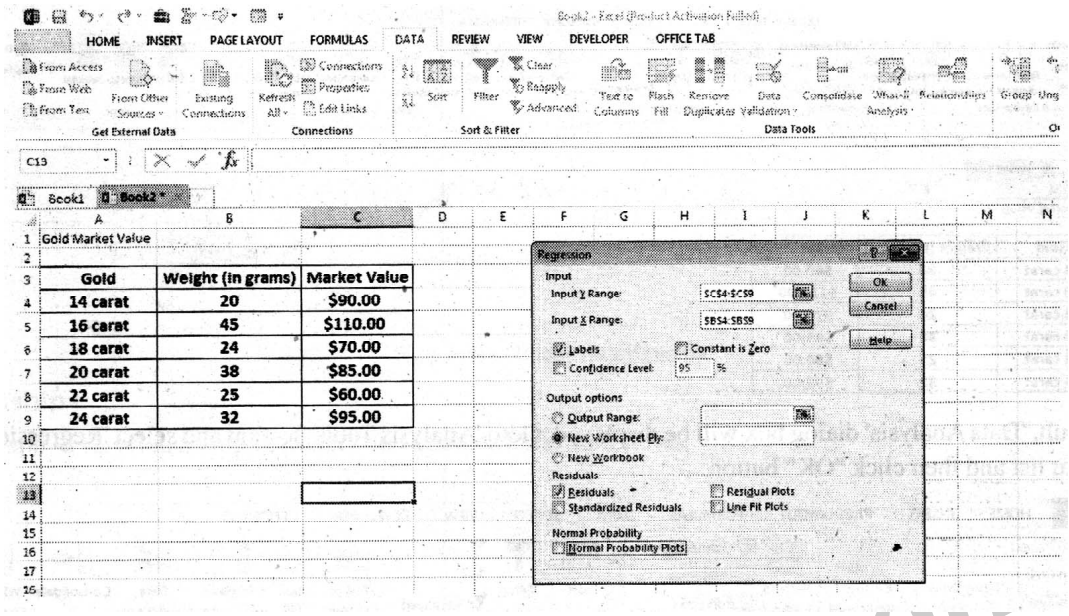
3. As a result, 'Data Analysis' dialog box will be displayed. Goto 'Analysis Tools' section and select 'Regression' option from the menu list and then click "OK" button.




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




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




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

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

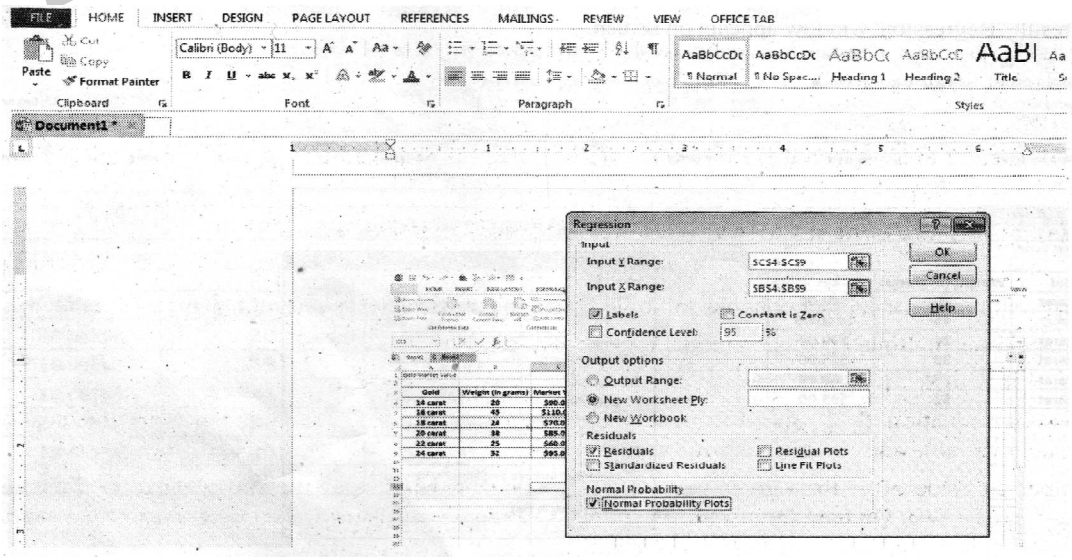
**ii) Constant is Zero :** Checkmark this option to make intercept to zero.

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

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

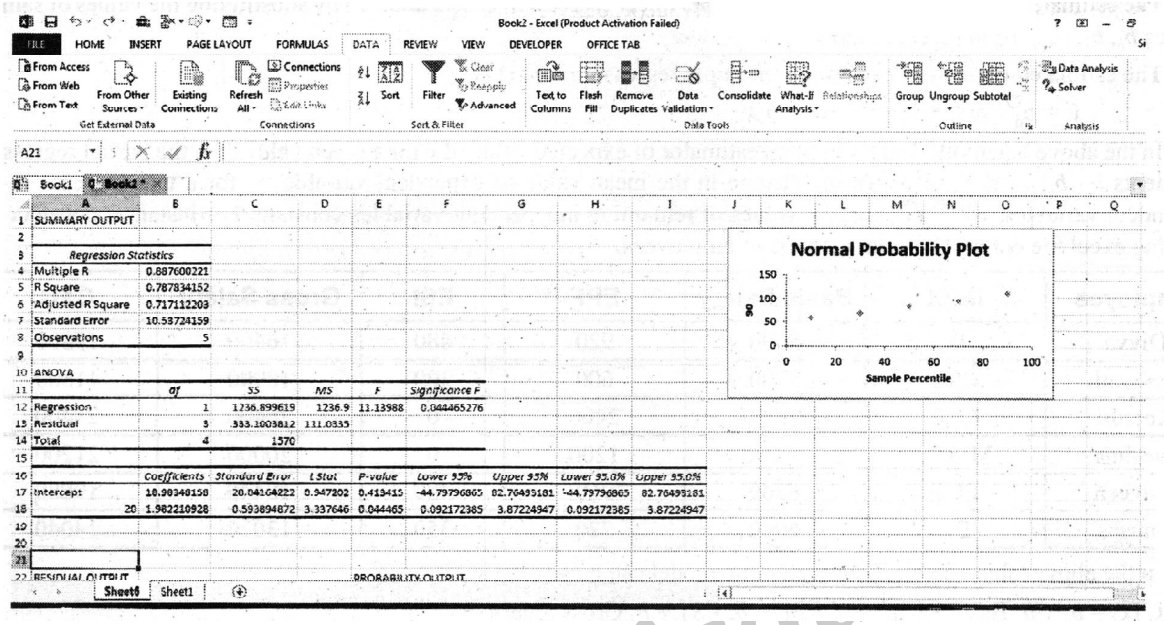
8. Goto 'Residuals' section and checkmark beside one of the four options ('Residuals', 'Residual Plots', 'Standardized Residuals', 'Line Fit Plots') to provide residuals on the output table.

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




Rahul Publications

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




| SUMMARY OUTPUT | | | | | | |
|---|---|---|---|---|---|---|
| Regression Statistics | | | | | | |
| Multiple R | 0.887600221 | | | | | |
| R Square | 0.787834152 | | | | | |
| Adjusted R Square | 0.717112203 | | | | | |
| Standard Error | 10.58724159 | | | | | |
| Observations | 5 | | | | | |
| ANOVA | | | | | | |
| | df | SS | MS | F | Significance F | |
| Regression | 1 | 1236.899619 | 1236.9 | 11.13988 | 0.044465276 | |
| Residual | 3 | 333.1003812 | 111.0335 | | | |
| Total | 4 | 1570 | | | | |
| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
| Intercept | 10.90348158 | 20.04164222 | 0.547202 | 0.413415 | -44.79796865 | 82.70495181 | -44.79796865 | 82.70495181 |
| 20 | 1.982210928 | 0.593894872 | 3.337646 | 0.044465 | 0.092172385 | 3.87224947 | 0.092172385 | 3.87224947 |

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

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

### i) Multiple R

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

### ii) R-Square(R2)

It determines the best fit between the regression line and data. R-sqaure is also referred to as 'coefficient of determination'. The value of R2 lies between 0 and 1. If the R2 is 1.0 then it indicates perfect fit where in each and every data point falls on the regression line itself. On the other hand, if the value of R2 is zero then it indicates no relationship.

### iii) Adjusted R Square

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

### iv) Standard Error

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

**Q12. Discuss briefly the concept of multiple regression using excel.**

*Ans :* **(Imp.)**

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,

In the above equation, $\beta_0$, $\beta_1$ specifies population parameters, $X_1$, $X_2$, .... Xp specifies independent-variables, Y defines dependent variable and '' 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 $\beta_0, \beta_1$ ,.... $\beta_q$ 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 $\beta_0, \beta_1, .....\beta_p$ .

The estimated regression equation in multiple regression model is,

$$= b_0 + b_1 x_1 + b_2 x_2 + .... + b_p x_p$$

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,

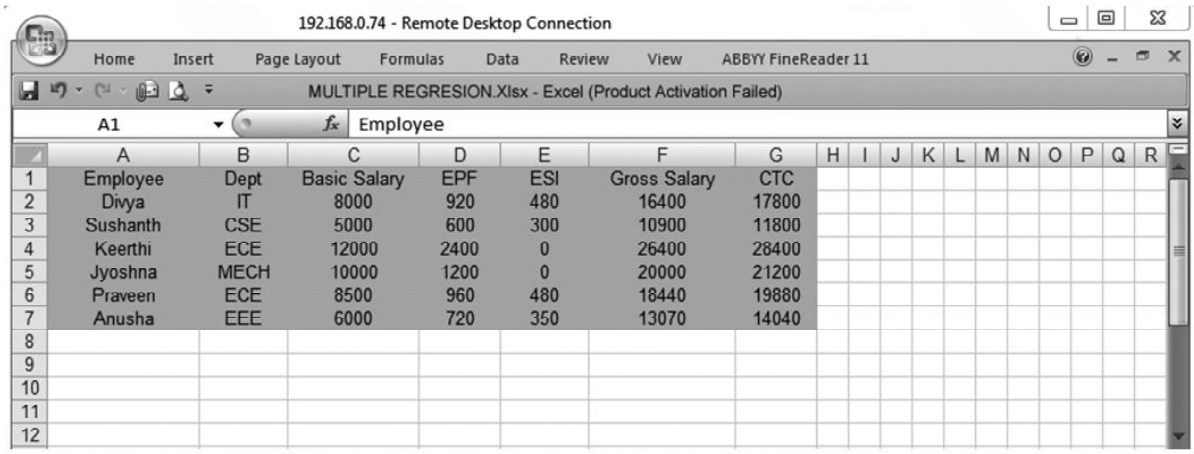
$$CTC = b_0 + b_x \text{ Basic Salary} + b_2 \text{ EPF} + b_3 \text{ ESI} + b_4 \text{ Gross Salary}$$

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,

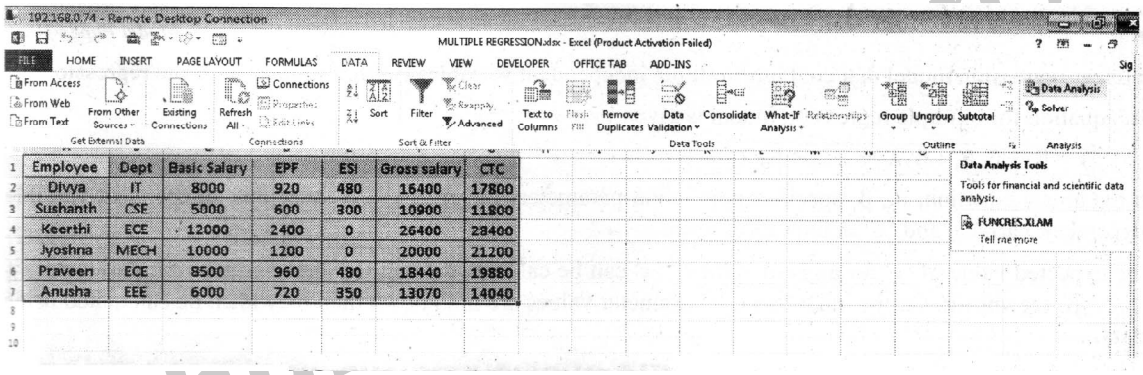
1. Select the data wherein user want to apply regression.

Rahul Publications



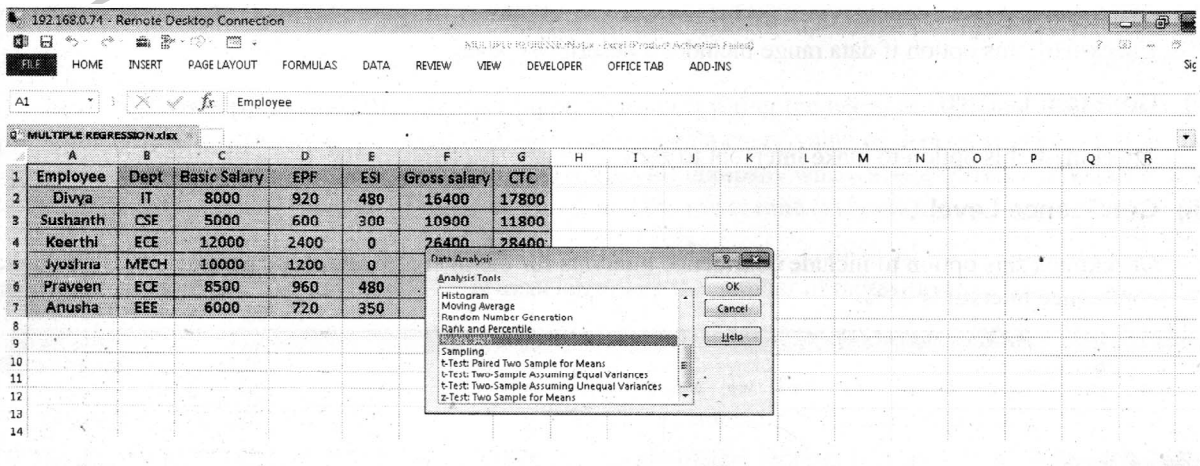

| Employee | Dept | Basic Salary | EPF | ESI | Gross Salary | CTC |
|---|---|---|---|---|---|---|
| Divya | IT | 8000 | 920 | 480 | 16400 | 17800 |
| Sushanth | CSE | 5000 | 600 | 300 | 10900 | 11800 |
| Keerthi | ECE | 12000 | 2400 | 0 | 26400 | 28400 |
| Jyoshna | MECH | 10000 | 1200 | 0 | 20000 |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

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



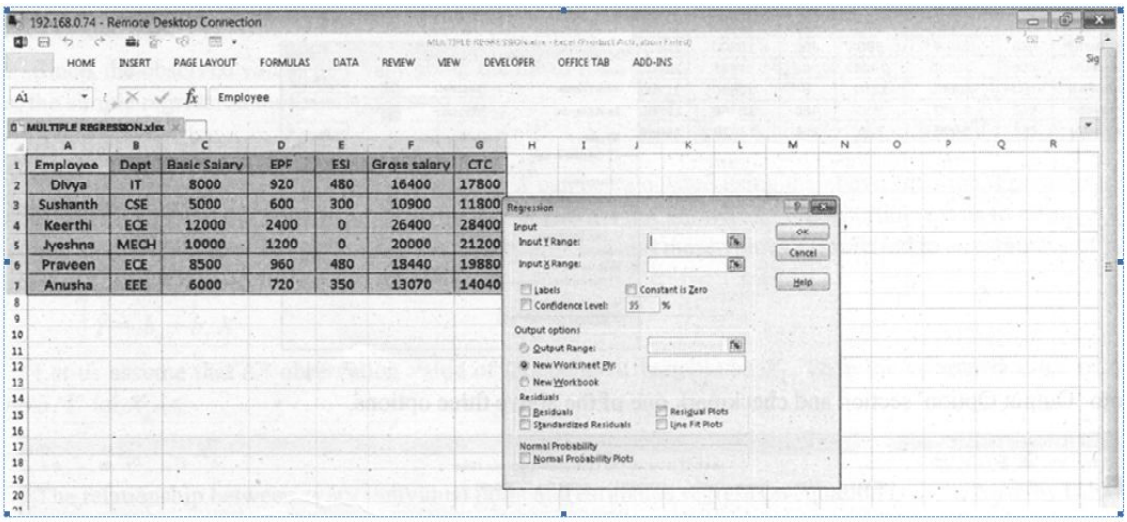

| Employee | Dept | Basic Salary | EPF | ESI | Gross salary | CTC |
|---|---|---|---|---|---|---|
| Divya | IT | 8000 | 920 | 480 | 16400 | 17800 |
| Sushanth | CSE | 5000 | 600 | 300 | 10900 | 11800 |
| Keerthi | ECE | 12000 | 2400 | 0 | 26400 | 28400 |
| Jyoshna | MECH | 10000 | 1200 | 0 | 20000 |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

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



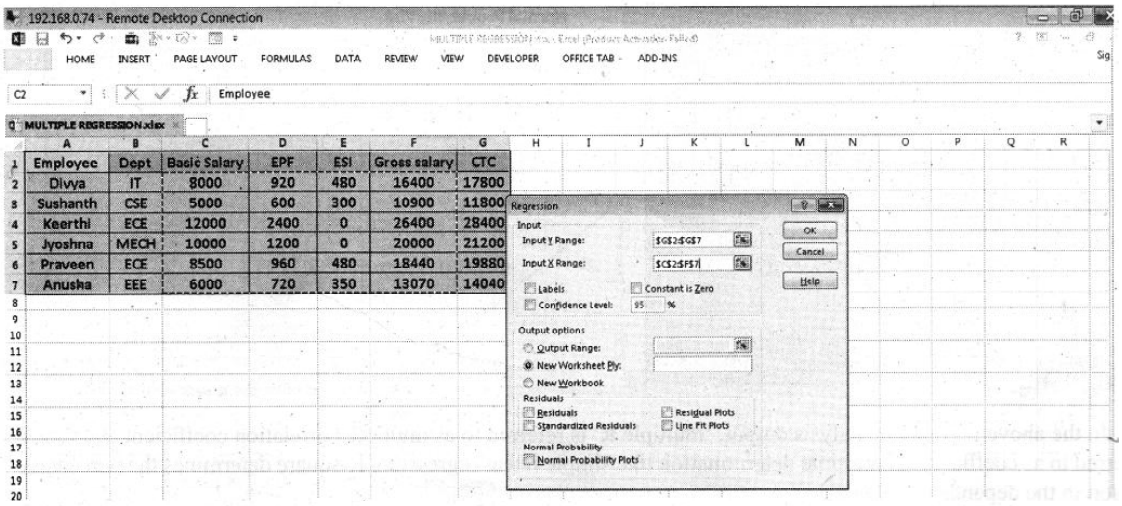

| Employee | Dept | Basic Salary | EPF | ESI | Gross salary | CTC |
|---|---|---|---|---|---|---|
| Divya | IT | 8000 | 920 | 480 | 16400 | 17800 |
| Sushanth | CSE | 5000 | 600 | 300 | 10900 | 11800 |
| Keerthi | ECE | 12000 | 2400 | 0 | 26400 | 28400 |
| Jyoshna | MECH | 10000 | 1200 | 0 | | |
| Praveen | 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 IJC.




| 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**

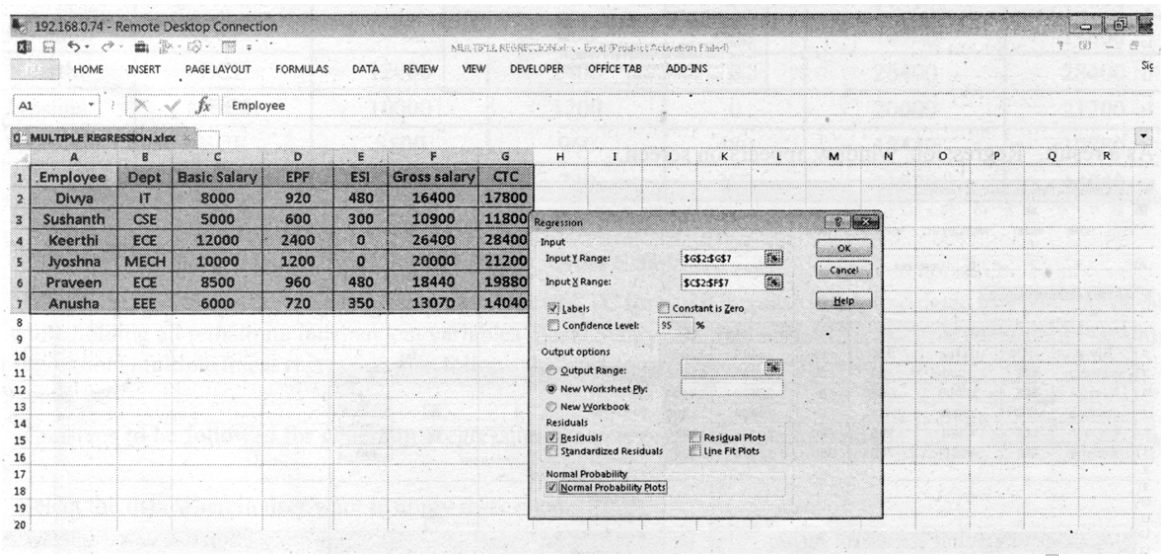
Checkmark this option if data range includes a descriptive level.

**(ii) Constant is Zero**

Checkmark this option to make intercept to zero.

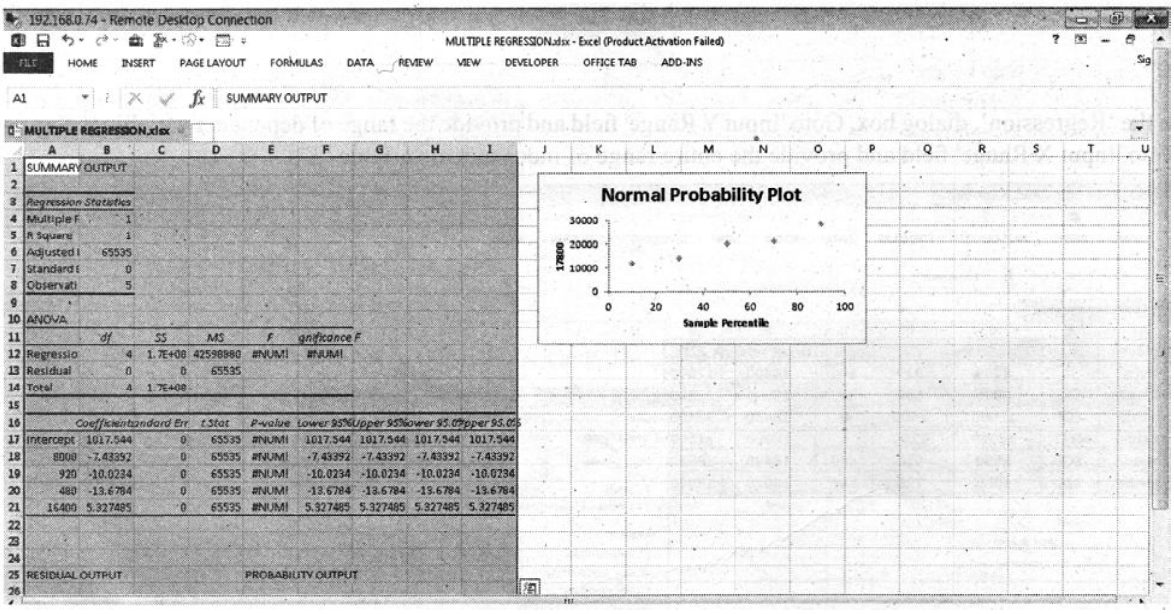
**(iii) Confidence Level**

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 R | 1 | | | | | | | |
| R Square | 1 | | | | | | | |
| Adjusted R | 65535 | | | | | | | |
| Standard E | 0 | | | | | | | |
| Observati | 5 | | | | | | | |
| ANOVA | | | | | | | | |
| | df | SS | MS | F | gnificance F | | | |
| Regressio | 4 | 1.7E+08 | 42598880 | #NUM! | #NUM! | | | |
| Residual | 0 | 0 | 65535 | | | | | |
| Total | 4 | 1.7E+08 | | | | | | |
| | Coefficients | tandard 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.

### 4.2.3 Coefficient of Determination

**Q13. Explain briefly about Coefficient of Determination.**

*Ans :*

The coefficient of determination method is used to predict and explain the future outcomes of a model. This method is also known as R squared. This method also acts like a guideline which helps in measuring the model's accuracy

The coefficient of determination or R squared method is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set.

(i) The coefficient of determination is the square of the correlation (r), thus it ranges from 0 to 1.

(ii) With linear regression, the coefficient of determination is equal to the square of the correlation between the x and y variables.

(iii) If R2 is equal to 0, then the dependent variable cannot be predicted from the independent variable.

(iv) If R2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error.

(v) If R2 is between 0 and 1, then it indicates the extent that the dependent variable can be predictable. If R2 of 0.10 means, it is 10 percent of the variance in the y variable is predicted from the x variable. If 0.20 means, 20 percent of the variance in the y variable is predicted from the x variable, and so on.

**Properties**

- It helps to get the ratio of how a variable which can be predicted from the other one, varies.
- If we want to check how clear it is to make predictions from the data given, we can determine the same by this measurement.
- It helps to find Explained variation / Total Variation
- It also lets us know the strength of the association (linear) between the variables.
- If the value of r2 gets close to 1, The values of y become close to the regression line and similarly if it goes close to 0, the values get away from the regression line.
- It helps in determining the strength of association between different variables.

Rahul Publications

**PROBLEMS**

**2. Find the Regression Analysis of following data.**

| X | 2 | 6 | 4 | 8 | 10 |
|---|---|---|---|---|---|
| Y | 4 | 10 | 6 | 15 | 5 |

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

**Finding the Regression Lines**

| X | Y | $x - \bar{x}$ (x) $\bar{x} = 6$ | $y - \bar{y}$(y) $\bar{y} = 8$ | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|---|---|
| 2 | 4 | –4 | –4 | 16 | 16 | 16 |
| 6 | 10 | 0 | 2 | 0 | 4 | 0 |
| 4 | 6 | –2 | –2 | 4 | 4 | 4 |
| 8 | 15 | 2 | 7 | 4 | 49 | 14 |
| 10 | 5 | 4 | –3 | 16 | 9 | 12 |
| $\Sigma x = 30$ | $\Sigma y = 40$ | $\Sigma x = 0$ | $\Sigma y = 0$ | $\Sigma x^2 = 40$ | $\Sigma y^2 = 82$ | $\Sigma xy = 46$ |

$$\bar{x} = \frac{\Sigma x}{N} = \frac{30}{5} = 6$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{40}{5} = 8$$

**(i) Regression Line x on y**

$$x - \bar{x} = \frac{\Sigma xy}{\Sigma y^2}(y - \bar{y})$$

$$x - 6 = \frac{46}{82}(y - 8)$$

$x - 6 = 0.560\ (y - 8)$

$x - 6 = 0.560y - 4.4$

$x = 0.560y - 4.4$

$x = 0.560y - 4.4 + 6$

$$\boxed{x = 0560y + 1.6}$$

**(ii) Regression Line y on x**

$$y - \bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x - \bar{x})$$

$$y - 8 = \frac{46}{40}(x - 6)$$

$y - 8 = 1.15\ (x - 6)$

$y - 8 = 1.15x - 6.9$

$y = 1.15x - 6.9$

$y = 1.15x - 6.9 + 8$

$$\boxed{y = 1.15x + 1.1}$$

**3. The following table gives the aptitude test scores and productivity induces of 10 workers selected at random**

| Aptitude Scores(x) | 60 | 62 | 65 | 70 | 72 | 48 | 53 | 73 | 65 | 82 |
|---|---|---|---|---|---|---|---|---|---|---|
| Productivity Index (y) | 68 | 60 | 62 | 80 | 85 | 40 | 52 | 62 | 60 | 81 |

**(i) Obtain regression equation y on x.**

**(ii) Estimate the Productivity index of a worker whose test scores is 92.**

Rahul Publications

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

**Finding the Regression Lines**

| X | x – 65 x | $x^2$ | y | y – 65 y | $y^2$ | xy |
|---|---|---|---|---|---|---|
| 60 | – 5 | 25 | 68 | 3 | 9 | –15 |
| 62 | – 3 | 9 | 60 | –5 | 25 | 15 |
| 65 | 0 | 0 | 62 | –3 | 9 | 0 |
| 70 | 5 | 25 | 80 | 15 | 225 | 75 |
| 72 | 7 | 49 | 85 | 20 | 400 | 140 |
| 48 | – 17 | 289 | 40 | – 25 | 625 | 425 |
| 53 | – 12 | 144 | 52 | – 13 | 169 | 156 |
| 73 | 8 | 64 | 62 | –3 | 9 | –24 |
| 65 | 0 | 0 | 60 | –5 | 25 | 0 |
| 82 | 17 | 289 | 81 | 16 | 256 | 272 |
| $\Sigma X = 650$ | $\Sigma x = 0$ | $\Sigma x^2 = 894$ | $\Sigma y = 650$ | $\Sigma Y = 0$ | $\Sigma y^2 = 1752$ | $\Sigma XY = 1044$ |

$$\bar{X} = \frac{\Sigma X}{N} = \frac{650}{10} = 65, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{650}{10} = 65$$

**(i) Regression Equation of Y on X**

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$b_{yx} = \frac{\gamma\sigma_x}{\sigma_y} = \frac{\Sigma XY}{\Sigma x^2} = \frac{1044}{894}$$

$= 1.168$

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$Y - 65 = 1.168 (X - 65)$

$Y - 65 = 1.168 x - 75.92$

$Y = 1.168x - 75.92 + 65$

$Y = 1.168x - 10.92$

**(ii) Estimate the Productivity Index of a Worker Whose Test Score is 92**

$Y = 1.168 x - 10.92$

$Y = 1.168 (92) - 10.92$

$Y = 107.456 - 10.92$

$Y = 96.536.$

Rahul Publications

## 4.3 Introduction to Data Mining

### 4.3.1 Definition

**Q14. Define Data Mining. Explain the scope of Data Mining.**

*Ans :* **(Imp.)**

**Meaning**

Data mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Thus, data mining should have been more appropriately named as knowledge mining which emphasis on mining from large amounts of data. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The key properties of data mining are Automatic discovery of patterns Prediction of likely outcomes Creation of actionable information Focus on large datasets and databases

**Scope**

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

1. **Automated prediction of trends and behaviors**

   Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

2. **Automated discovery of previously unknown patterns**

   Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

**Q15. Explain different tasks of data mining.**

*Ans :*

Data mining involves six common classes of tasks:

1. **Anomaly detection (Outlier/change/ deviation detection)**

   The identification of unusual data records, that might be interesting or data errors that require further investigation.

2. **Association rule learning (Dependency modelling)**

   Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

3. **Clustering**

   Is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Rahul Publications

**4. Classification**

Is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

**5. Regression**

Attempts to find a function which models the data with the least error.

**6. Summarization**

Providing a more compact representation of the data set, including visualization and report generation.

**Q16. Explain the process of data mining.**

*Ans :* **(Imp.)**

Data Mining is a process of discovering various models, summaries, and derived values from a given collection of data. The general experimental procedure adapted to data-mining problems involves the following steps:

**1. State the problem and formulate the hypothesis**

Most data-based modeling studies are performed in a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement. Unfortunately, many application studies tend to focus on the data-mining technique at the expense of a clear problem statement. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be several hypotheses formulated for a single problem at this stage. The first step requires the combined expertise of an application domain and a data-mining model. In practice, it usually means a close interaction between the data-mining expert and the application expert. In successful data-mining applications, this cooperation does not stop in the initial phase; it continues during the entire data-mining process.

**2. Collect the data**

This step is concerned with how the data are generated and collected. In general, there are two distinct possibilities. The first is when the data-generation process is under the control of an expert (modeler): this approach is known as a designed experiment. The second possibility is when the expert cannot influence the data- generation process: this is known as the observational approach. An observational setting, namely, random data generation, is assumed in most data-mining applications. Typically, the sampling distribution is completely unknown after data are collected, or it is partially and implicitly given in the data-collection procedure. It is very important, however, to understand how data collection affects its theoretical distribution, since such a priori knowledge can be very useful for modeling and, later, for the final interpretation of results. Also, it is important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. If this is not the case, the estimated model cannot be successfully used in a final application of the results.

**3. Preprocessing the data**

In the observational setting, data are usually "collected" from the existing databases, data warehouses, and data marts. Data preprocessing usually includes at least two common tasks:

(i) Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such nonrepresentative samples can seriously affect the model produced later. There are two strategies for dealing with outliers:

(ii) Detect and eventually remove outliers as a part of the preprocessing phase, or b. Develop robust modeling methods that are insensitive to outliers.

(iii) Scaling, encoding, and selecting features

Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the

Rahul Publications

range [0, 1] and the other with the range [“100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently. Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application-specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling. These two classes of preprocessing tasks are only illustrative examples of a large spectrum of preprocessing activities in a data-mining process. Data-preprocessing steps should not be considered completely independent from other data-mining phases. In every iteration of the data-mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, a good preprocessing method provides an optimal representation for a data-mining technique by incorporating a priori knowledge in the form of application-specific scaling and encoding.

**4. Estimate the model**

The selection and implementation of the appropriate data-mining technique is the main task in this phase. This process is not straightforward; usually, in practice, the implementation is based on several models, and selecting the best one is an additional task. The basic principles of learning and discovery from data are given in Chapter 4 of this book. Later, Chapter 5 through 13 explain and analyze specific techniques that are applied to perform a successful learning process from data and to develop an appropriate model.

**5. Interpret the model and draw conclusions**

In most cases, data-mining models should help in decision making. Hence, such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex “black-box” models. Note that the goals of accuracy of the model and accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high dimensional models. The problem of interpreting these models, also very important, is considered a separate task, with specific techniques to validate the results. A user does not want hundreds of pages of numeric results. He does not understand them; he cannot summarize, interpret, and use them for successful decision making.

Rahul Publications

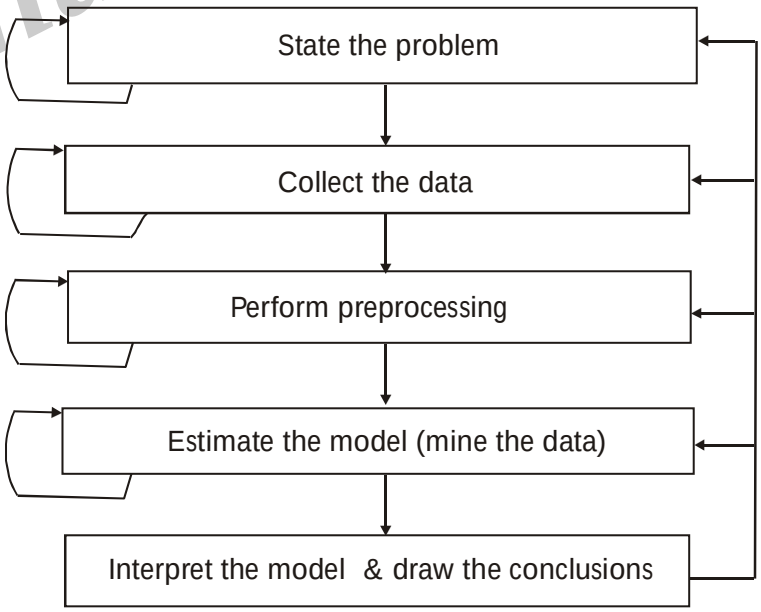



**Fig. : Process of Data Mining**

**Q17. What are the functionalities of date mining.**

*Ans :* **(Imp.)**

Data mining functionalities are used to represent the type of patterns that have to be discovered in data mining tasks. In general, data mining tasks can be classified into two types including descriptive and predictive. Descriptive mining tasks define the common features of the data in the database and the predictive mining tasks act inference on the current information to develop predictions.

There are various data mining functionalities which are as follows -

- **Data characterization**

  It is a summarization of the general characteristics of an object class of data. The data corresponding to the user-specified class is generally collected by a database query. The output of data characterization can be presented in multiple forms.

- **Data discrimination**

  It is a comparison of the general characteristics of target class data objects with the general characteristics of objects from one or a set of contrasting classes. The target and contrasting classes can be represented by the user, and the equivalent data objects fetched through database queries.

- **Association Analysis**

  It analyses the set of items that generally occur together in a transactional dataset. There are two parameters that are used for determining the association rules .

  - It provides which identifies the common item set in the database.
  - Confidence is the conditional probability that an item occurs in a transaction when another item occurs.

- **Classification**

  Classification is the procedure of discovering a model that represents and distinguishes data classes or concepts, for the objective of being able to use the model to predict the class of objects whose class label is anonymous. The derived model is established on the analysis of a set of training data (i.e., data objects whose class label is common).

- **Prediction**

  It defines predict some unavailable data values or pending trends. An object can be anticipated based on the attribute values of the object and attribute values of the classes. It can be a prediction of missing numerical values or increase/decrease trends in time-related information.

- **Clustering**

  It is similar to classification but the classes are not predefined. The classes are represented by data attributes. It is unsupervised learning. The objects are clustered or grouped, depends on the principle of maximizing the intraclass similarity and minimizing the intraclass similarity.

- **Outlier analysis**

  Outliers are data elements that cannot be grouped in a given class or cluster. These are the data objects which have multiple behaviour from the general behaviour of other data objects. The analysis of this type of data can be essential to mine the knowledge.

- **Evolution analysis**

  It defines the trends for objects whose behaviour changes over some time.

### 4.3.2 Methods of Data Mining

**Q18. Explain various Methods of Data Mining.**

**(OR)**

**Explain the various approaches of Date mining**

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

There are many methods used for Data Mining, but the crucial step is to select the appropriate form from them according to the business or the problem statement. These methods help in predicting the future and then making decisions accordingly.

Rahul Publications

**1. Association**

It is used to find a correlation between two or more items by identifying the hidden pattern in the data set and hence also called relation analysis. This method is used in market basket analysis to predict the behavior of the customer.

Suppose, the marketing manager of a supermarket wants to determine which products are frequently purchased together.

As an example,

Buys (x,"beer") -> buys(x, "chips") [support = 1%, confidence = 50%]

Here x represents a customer buying beer and chips together.

Confidence shows certainty that if a customer buys a beer, there is a 50% chance that he/she will also accept the chips.

Support means that 1% of all the transactions under analysis showed that beer and chips were bought together.

Many similar examples like bread and butter or computer and software can be considered.

There are two types of Association Rules:

Single dimensional association rule: These rules contain a single attribute that is repeated.

Multidimensional association rule: These rules contain multiple attributes that are repeated.

**2. Classification**

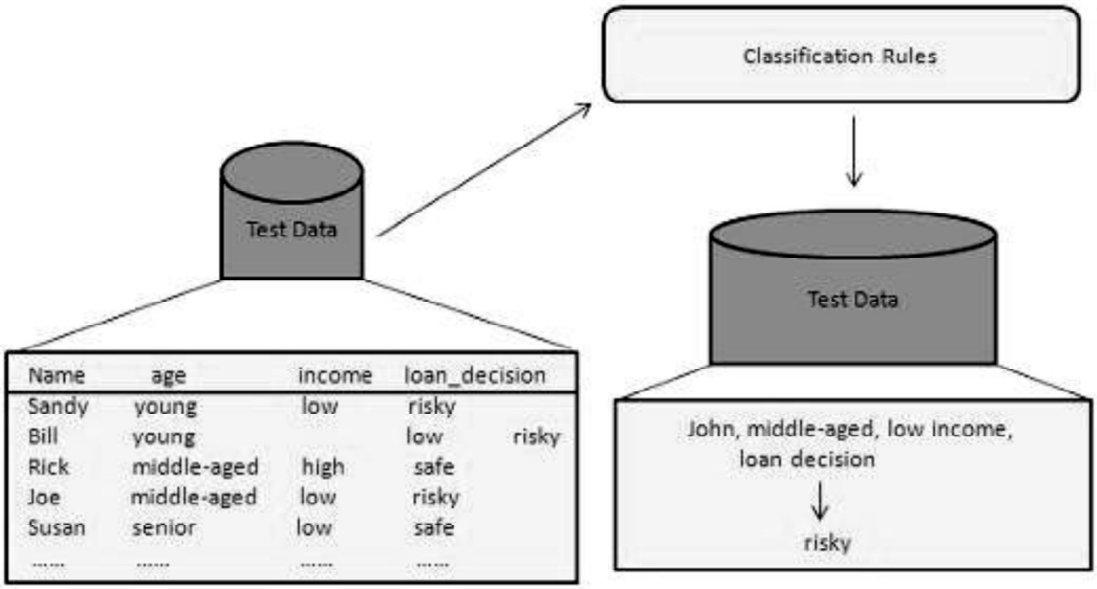
This data mining method is used to distinguish the items in the data sets into classes or groups. It helps to predict the behaviour of entities within the group accurately. It is a two-step process:

Learning step (training phase): In this, a classification algorithm builds the classifier by analyzing a training set.

Classification step: Test data are used to estimate the accuracy or precision of the classification rules.

For example, a banking company uses to identify loan applicants at low, medium or high credit risks. Similarly, a medical researcher analyzes cancer data to predict which medicine to prescribe to the patient.

Rahul Publications



### 3. Clustering Analysis

Clustering is almost similar to classification, but in this cluster are made depending on the similarities of data items. Different groups have dissimilar or unrelated objects. It is also called data segmentation as it partitions huge data sets into groups according to the similarities.

Various clustering methods are used:

Hierarchical Agglomerative methods

Grid-Based Methods

Partitioning Methods

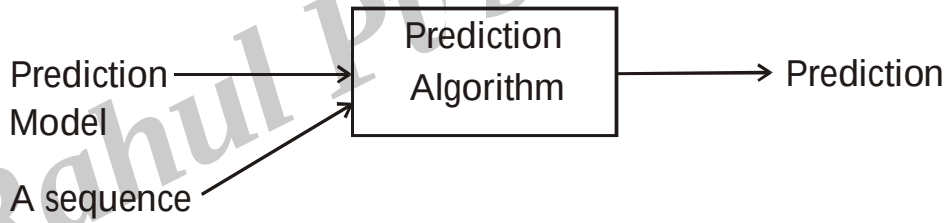
Model-Based Methods

Density-Based Methods

### 4. Prediction

This method is used to predict the future based on the past and present trends or data set. Prediction is mostly used to combine other mining methods such as classification, pattern matching, trend analysis, and relation.

For example, if the sales manager would like to predict the amount of revenue that each item would generate based on past sales data. It models a continuous-valued function that indicates missing numeric data values.




Rahul Publications

### 5. Sequential patterns or Pattern tracking

This method is used to identify patterns that frequently occur over a certain period of time.

For example, a clothing company's sales manager sees that sales of jackets seem to increase just before the winter season, or sales in bakery increase during Christmas or New Year's eve.

### 6. Decision Trees

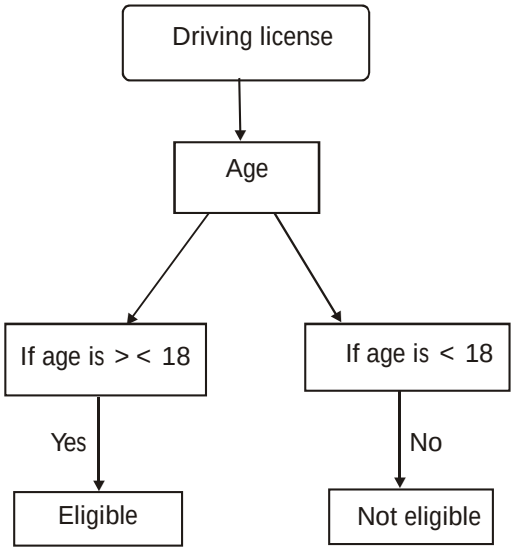
A decision tree is a tree structure (as its name suggests), where

Each internal node represents a test on the attribute.

Branch denotes the result of the test.

Terminal nodes hold the class label.

The topmost node is the root node which has a simple question that has two or more answers. Accordingly, the tree grows, and a flow chart like structure is generated.




In this decision, tree government classifies citizens below age 18 or above age 18. This would help them to decide whether a license must be issued to a particular city or not.

**Q19. What is data exploration ? Explain the Steps of Data Exploration and Preparation**

*Ans :*

Data exploration is an informative search used by data consumers to form true analysis from the information gathered. Often, data is gathered in a non-rigid or controlled manner in large bulks. For true analysis, this unorganized bulk of data needs to be narrowed down. This is where data exploration is used to analyze the data and information from the data to form further analysis.

Data often converges in a central warehouse called a data warehouse. This data can come from various sources using various formats. Relevant data is needed for tasks such as statistical reporting, trend spotting and pattern spotting. Data exploration is the process of gathering such relevant data.

**Steps of Data Exploration and Preparation**

Remember the quality of your inputs decide the quality of your output. So, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here. With my personal estimate, data exploration, cleaning and preparation can take up to 70% of your total project time.

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

Rahul Publications

Finally, we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

### 1. Variable Identification

First, identify Predictor (Input) and Target (output) variables. Next, identify the data type and category of the variables.

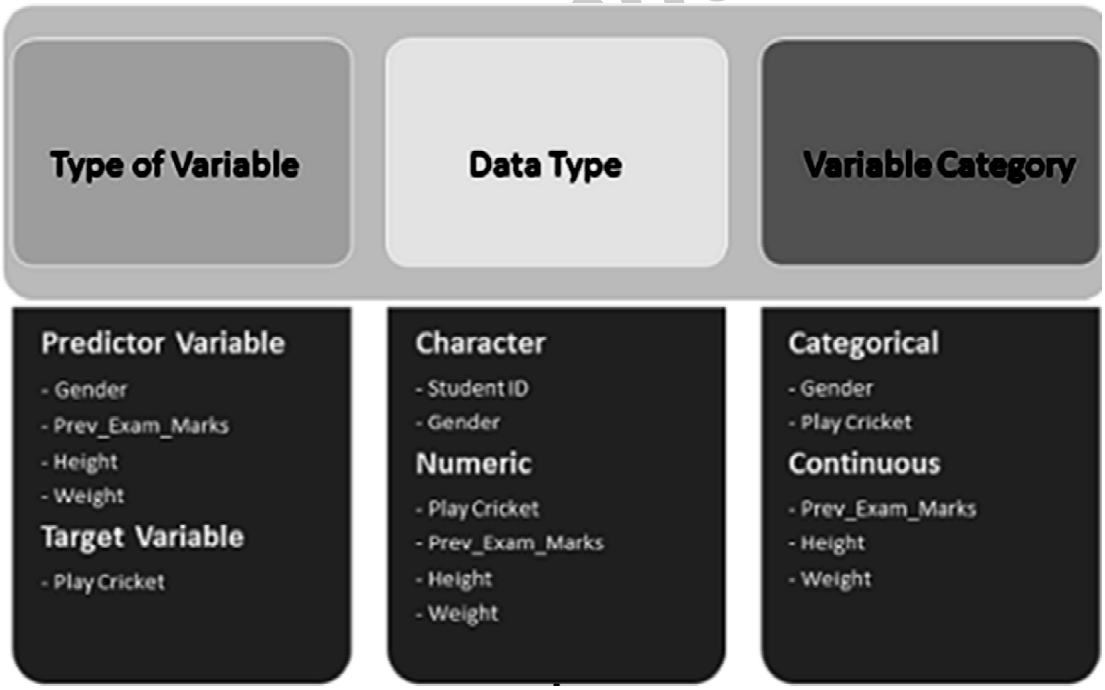
Let's understand this step more clearly by taking an example.

**Example**

Suppose, we want to predict, whether the students will play cricket or not (refer below data set). Here you need to identify predictor variables, target variable, data type of variables and category of variables. Below, the variables have been defined in different category:

| Student ID | Gender | Prev Exam Marks | Height (cm) | Weight Caregory (kgs) | Play Cricket |
|---|---|---|---|---|---|
| S001 | M | 65 | 178 | 61 | 1 |
| S002 | F | 75 | 174 | 56 | 0 |
| S003 | M | 45 | 163 | 62 | 1 |
| S004 | M | 57 | 175 | 70 | 0 |
| S005 | F | 59 | 162 | 67 | 0 |

lications

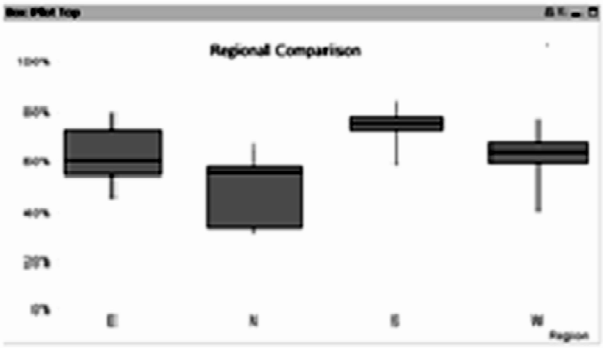


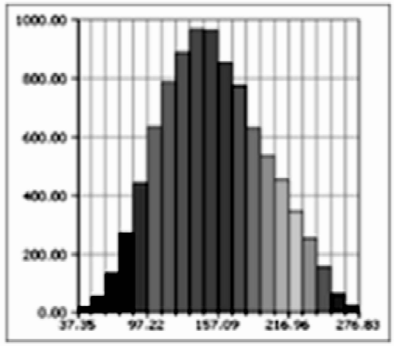

### 2. Univariate Analysis

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. Let's look at these methods and statistical measures for categorical and continuous variables individually:

**(i) Continuous Variables:** In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods as shown below:

| Central Tendency | Measure of Dispersion | Visualization Methods |
|---|---|---|
| Mean | Range | Histogram |
| Median | Quartile | Box Plot |
| Mode | IQR | |
| Min | Variance | |
| Max | Standard Deviation | |
| | Skewness and Kurtosis | |




**Note:** Univariate analysis is also used to highlight missing and outlier values. In the upcoming part of this series, we will look at methods to handle missing and outlier values. To know more about these methods, you can refer course descriptive statistics from Udacity.

**(ii) Categorical Variables:** For categorical variables, we'll use frequency table to understand distribution of each category. We can also read as percentage of values under each category. It can be measured using two metrics, Count and Count% against each category. Bar chart can be used as visualization.

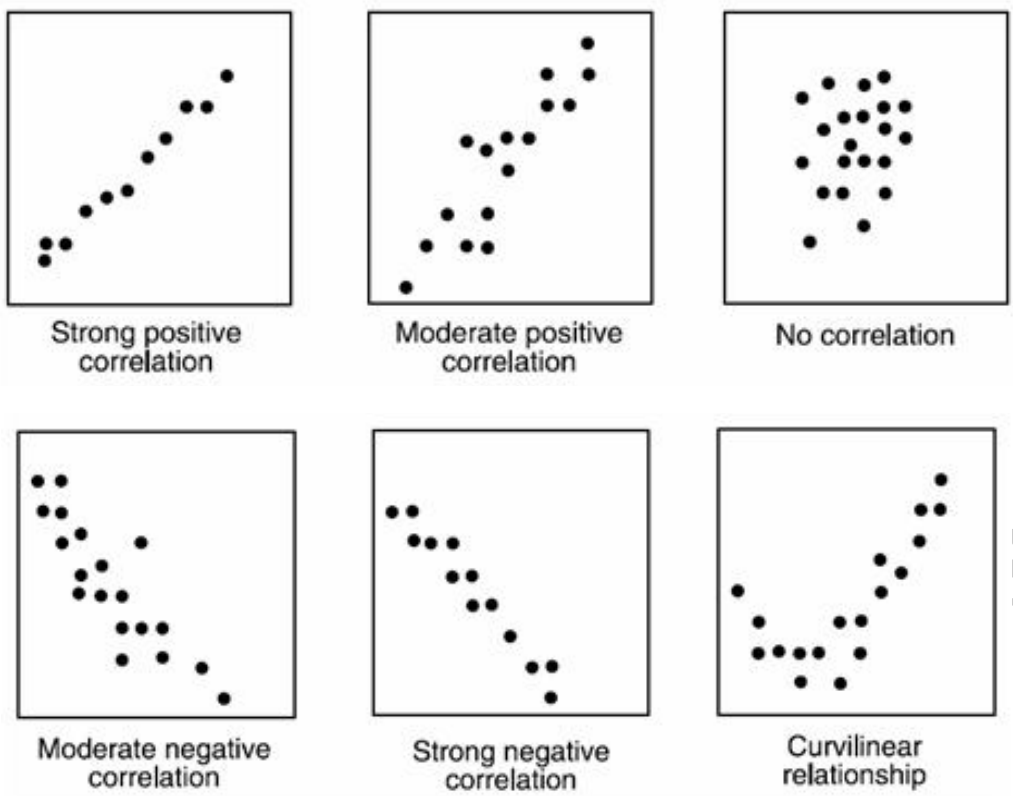
### Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Let's understand the possible combinations in detail:

### Continuous and Continuous

While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.




Scatter plot shows the relationship between two variable but does not indicates the strength of relationship amongst them. To find the strength of the relationship, we use Correlation. Correlation varies between -1 and +1.

- –1: perfect negative linear correlation
- +1:perfect positive linear correlation and
- 0: No correlation

Correlation can be derived using following formula:

**Correlation = Covariance(X,Y) / SQRT( Var(X)* Var(Y))**

Various tools have function or functionality to identify correlation between variables. In Excel, function CORREL() is used to return the correlation between two variables and SAS uses procedure PROC CORR to identify the correlation. These function returns Pearson Correlation value to identify the relationship between two variables:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **X** | 65 | 72 | 78 | 65 | 72 | 70 | 65 | 68 |
| **Y** | 72 | 69 | 79 | 69 | 84 | 75 | 60 | 73 |

| Metrics | Formula | Value |
|---|---|---|
| Co-Variance (X,Y) | = COVAR(E6:L6,E7:L7) | 18.77 |
| Variance (X) | = VAR.P(E6:L6) | 18.48 |
| Variance (Y) | = VAR.P(E7:L7) | 45.23 |
| Correlation | = G10/SQRT(G11*G12) | 0.65 |

In above example, we have good positive relationship(0.65) between two variables X and Y.

### Categorical and Categorical

To find the relationship between two categorical variables, we can use following methods:

- **Two-way table**

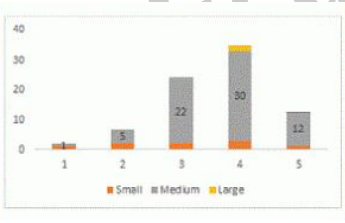
  We can start analyzing the relationship by creating a two-way table of count and count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.

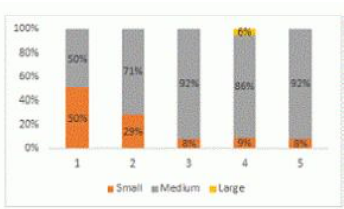
- **Stacked Column Chart:** This method is more of a visual form of Two-way table.

Product Category

| Frequency<br>Row Pct | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| Small | 1<br>11.11 | 2<br>22.22 | 2<br>22.22 | 3<br>33.33 | 1<br>11.11 | 9 |
| Medium | 1<br>1.43 | 5<br>7.14 | 22<br>31.43 | 30<br>42.86 | 12<br>17.14 | 70 |
| Large | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 2<br>100.00 | 0<br>0.00 | 2 |
| Total | 2 | 7 | 24 | 35 | 13 | 81 |

Frequency Missing = 77




### Chi-Square Test

This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

Probability of 0: It indicates that both categorical variable are dependent

Probability of 1: It shows that both variables are independent.

Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence. The chi-square test statistic for a test of independence of two categorical variables is found by:

$X^2 = \Sigma(O - E)^2 / E$ where O represents the observed frequency. E is the expected frequency under the null hypothesis and computed by:

$$E = \frac{\text{row total} \times \text{column total}}{\text{Sample size}}$$

From previous two-way table, the expected count for product category 1 to be of small size is 0.22. It is derived by taking the row total for Size (9) times the column total for Product category (2) then dividing by the sample size (81). This is procedure is conducted for each cell. Statistical Measures used to analyze the power of relationship are:

- Cramer's V for Nominal Categorical Variable
- Mantel-Haenszed Chi-Square for ordinal categorical variable.

Different data science language and tools have specific methods to perform chi-square test. In SAS, we can use Chisq as an option with Procfreq to perform this test.

**Categorical and Continuous**

While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance we can perform Z-test, T-test or ANOVA.

**Z-Test/ T-Test**

Either test assess whether mean of two groups are statistically different from each other or not. If the probability of Z is small then the difference of two averages is more significant. The T-test is very similar to Z-test but it is used when number of observation for both categories is less than 30.

$$Z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$S^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_3^2}{N_1 + N_2 - 2}$$

Where

- $\bar{X}_1, \bar{X}_{2:\text{ Averages}}$
- $S_1^2, S_{2:\text{Variances}}^2$
- $N_1, N_{2:\text{ Counts}}$
- t : has t distribution with $N_1 + N_2 - 2$ degree of freedom

**ANOVA**

It assesses whether the average of more than two groups is statistically different.

**Example**

Suppose, we want to test the effect of five different exercises. For this, we recruit 20 men and assign one type of exercise to 4 men (5 groups). Their weights are recorded after a few weeks. We need to find out whether the effect of these exercises on them is significantly different or not. This can be done by comparing the weights of the 5 groups of 4 men each.

Till here, we have understood the first three stages of Data Exploration, Variable Identification, Uni-Variate and Bi-Variate analysis. We also looked at various statistical and visual methods to identify the relationship between variables.

Now, we will look at the methods of Missing values Treatment. More importantly, we will also look at why missing values occur in our data and why treating them is necessary.

**Q20. Explain the Data Reduction in Data mining.**

*Ans :*

Data reduction is the process of minimizing the amount of data that needs to be stored in a data storage environment. Data reduction can increase storage efficiency and reduce costs.

(a) Data reduction can be achieved using several different types of technologies. The best-known data reduction technique is data deduplication, which eliminates redundant data on storage systems.

Rahul Publications

(b) The duplication process typically occurs at the storage block level. The system analyzes the storage to see if duplicate blocks exist, and gets rid of any redundant blocks. The remaining block is shared by any file that requires a copy of the block. If an application attempts to modify this block, the block is copied prior to modification so that other files that depend on the block can continue to use the unmodified version, thereby avoiding file corruption. Some storage arrays track which blocks are the most heavily shared.

(c) Those blocks that are shared by the largest number of files may be moved to a memory- or flash storage-based cache so they can be read as efficiently as possible.

(d) Data reduction strategies applied on huge data set. Complex data and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.

(e) Data reduction techniques can be applied to obtain a reduces data should be more efficient yet produce the same analytical results.

**Q21. Explain the Strategies for data reduction ?**

*Ans :*

The following are the Strategies for reduction

1. **Data cube aggregation,** where aggregation operations are applied to the data in the construction of a data cube.

2. **Attribute subset selections**, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed,

3. **Dimensionality reduction**, where encoding mechanism are used to reduce the data set size.

4. **Numerosity reductions**, where the data are replaced or estimated by alternative, smaller data representations such as parametric models or non parametric method such as clustering, sampling, and the use of histograms.

5. **Discretization and concept hierarchy** generation, where raw data values for attributes are replaced by range or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction.

**Q22. Explain the techniques used in Data Reduction.**

*Ans :*

**i) Dimensionality Reduction**

Dimensionality Reduction ensures the reduction of the number of attributes or random variables in the data set. Dimensionality reduction is a series of techniques in machine learning and statistics to reduce the number of random variables to consider. It involves feature selection and feature extraction. Dimensionality reduction makes analyzing data much easier and faster for machine learning algorithms without extraneous variables to process, making machine learning algorithms faster and simpler in turn.

**ii) Sample Numerosity Reduction**

Replaces the original data by an alternative smaller data representation This is a technique of choosing smaller forms or data representation to reduce the volume of data.

**These techniques may be parametric or nonparametric**.

**a) Parametric**

For parametric methods, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. (Outliers may also be stored.)

Example: Log-linear models, which estimate discrete multidimensional probability distributions.

Rahul Publications

**b) Nonparametric**

Nonparametric methods are used for storing reduced representations of the data include histograms, clustering, and sampling.

Regression and Log-Linear Models

- Regression and log-linear models can be used to approximate the given data.
- In (simple) linear regression, the data are modeled to fit a straight line.
- Multiple linear regression is an extension of (simple) linear regression, which allows a response variable y to be modeled as a linear function of two or more predictor variables.
- Log-linear models approximate discrete multidimensional probability distributions.
- Log-linear models can be used to estimate the probability of each point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations.
- This allows a higher-dimensional data space to be constructed from lower dimensional spaces.
- Log-linear models are therefore also useful for dimensionality reduction and data smoothing
- Regression and log-linear models can both be used on sparse data, although their application may be limited.
- While both methods can handle skewed data, regression does exceptionally well. Regression can be computationally intensive when applied to high dimensional data, whereas log-linear models show good scalability for up to 10 or so dimensions.

**iii) Cardinality Reduction**

Transformations applied to obtain a reduced representation of the original data.

The term cardinality refers to the uniqueness of data values contained in a particular column (attribute) of a database table. The lower the cardinality, the more duplicated elements in a column. Thus, a column with the lowest possible cardinality would have the same value for every row. SQL databases use cardinality to help determine the optimal query plan for a given query.

**Q23. Explain the Classification of Data Mining.**

*Ans :*

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

- Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups.
- Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups.
- For example, we can apply classification in the application that given all records of employees who left the company, predict who will probably leave the company in a future period. In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.
- In technical term, classification in data mining defines as assigning an object to a certain class based on its similarity to previous examples of other objects.
- The classification process comes under the predictive method. With classification, new samples of data are classified into known classes.

Rahul Publications

- The classification is the initial process of data mining and use algorithms like decision trees, Bayesian classifiers. For classification the data required must be already labeled one.

**Examples of classification are:**

1. A marketing manager of a company needs to analyze the customer with available profile that who will buy a new computer.
2. A bank officer wants to predict that which loan applicants are risky or which are safe.
   - A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time.
   - In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on.
   - Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.
   - The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values:
   - Example, high credit rating or low credit rating. Multi-class targets have more than two values: for example, low, medium, high, or unknown credit rating
   - In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target.
   - Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.
   - Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model.

**Q24. Explain various issues relating to data classification.**

*Ans :*

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities:

- **Data Cleaning**

  Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.

- **Relevance Analysis**

  Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.

- **Data Transformation and Reduction**

  The data can be transformed by any of the following methods:

  **(i) Normalization**

  The data is transformed using normalization. It involves scaling all values for given attribute in order to make them fall within a small specified range. It is used when in the learning step, the neural networks or the methods involving measurements are used.

Rahul Publications

**(ii) Generalization**

The data can also be transformed by generalizing it to the higher concept. For this purpose, we can use the concept hierarchies.

**Q25. Explain the Association in Data Mining ?**

*Ans :*

- It is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rules are often used to analyze sales transactions.
- Association analysis is about discovering relationship among huge data sets. Just like the famous market basket analysis which gives a relationship between {Diapers -> beer}. It says that whenever a person buys diapers he/she also buys beer.
- Besides market basket, association rules can be applied to Bioinformatics , web mining and medical analysis.

**There are two key issues that need to be addressed while applying this ;**

- First detecting the pattern
- Some of the detected patterns can be spurious and may be happening only by chance.
- The strength of a association rule can be measured in terms of support and confidence.
- Support determines how often a rule is applicable to the data set while confidence determines how frequently items in Y appear in transactions that contain X.
- Use packages like a rules, a rules CBA , a rules Sequences in R

Ex :

- library ("a rules")
- data ("Adult")

- rules <- apriori(Adult, parameter = list(supp = 0.5, conf = 0.9, target = "rules"))

**Q26. Explain in detail about cause and effect Modelling.**

*Ans :*

The managers of a company are preferably interested in results like profit, customer satisfaction, retention, production yield etc. The lagging measures describe about the already happened things and external business results Including customer satisfaction, market share and profit. Whereas, leading measures describe about the things to happen in future such as about productivity, turnover, internal metrics, employee satisfaction etc. The lagging measure of customer satisfaction for instance might be with regard to service or sales transactions. And the leading measure would be the behavior of sales representative, accuracy of billing etc.

In case if an employee is not satisfied, then their behavior towards the customers might not be good leading to customer dissatisfaction. Therefore, this must be explained to the managers through business analytics so that they can workout to improvise the employee satisfaction. Thus better means should be provided to know the controllable factors that influence the performance measures of business which cannot be directly controlled by the managers. Such influences can be identified by the correlation analysis which inturn leads to cause and effect development models. With this, managers can make better decisions which influence the future results.

A measure of linear relationship among two variables is called correlation. A strong relationship is determined by high values of correlation coefficient.

Rahul Publications

### 4.4.3 Applications of Data Mining

**Q27. Explain the various applications of Data Mining.**

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

**1. Financial Analysis**

The banking and finance industry relies on high-quality, reliable data. In loan markets, financial and user data can be used for a variety of purposes, like predicting loan payments and determining credit ratings. And data mining methods make such tasks more manageable.

Classification techniques facilitate the separation of crucial factors that influence customers' banking decisions from the irrelevant ones. Further, multidimensional clustering techniques allow the identification of customers with similar loan payment behaviours. Data analysis and mining can also help detect money laundering and other financial crimes.

**2. Telecommunication Industry**

Expanding and growing at a fast pace, especially with the advent of the internet. Data mining can enable key industry players to improve their service quality to stay ahead in the game.

Pattern analysis of spatiotemporal databases can play a huge role in mobile telecommunication, mobile computing, and also web and information services. And techniques like outlier analysis can detect fraudulent users. Also, OLAP and visualization tools can help compare information, such as user group behaviour, profit, data traffic, system overloads, etc.

**3. Intrusion Detection**

Global connectivity in today's technology-driven economy has presented security challenges for network administration. Network resources can face threats and actions that intrude on their confidentiality or integrity. Therefore, detection of intrusion has emerged as a crucial data mining practice. It encompasses association and correlation analysis, aggregation techniques, visualization, and query tools, which can effectively detect any anomalies or deviations from normal behaviour.

**4. Retail Industry**

The organized retail sector holds sizable quantities of data points covering sales, purchasing history, delivery of goods, consumption, and customer service. The databases have become even larger with the arrival of e-commerce marketplaces.

In modern-day retail, data warehouses are being designed and constructed to get the full benefits of data mining. Multidimensional data analysis helps deal with data related to different types of customers, products, regions, and time zones. Online retailers can also recommend products to drive more sales revenue and analyze the effectiveness of their promotional campaigns. So, from noticing buying patterns to improving customer service and satisfaction, data mining opens many doors in this sector.

**5. Energy Industry**

Big Data is available even in the energy sector nowadays, which points to the need for appropriate data mining techniques. Decision tree models and support vector machine learning are among the most popular approaches in the industry, providing feasible solutions for decision-making and management. Additionally, data mining can also achieve productive gains by predicting power outputs and the clearing price of electricity.

**6. Spatial Data Mining**

Geographic Information Systems (GIS) and several other navigation applications make use of data mining to secure vital information and understand its implications. This new trend includes extraction of geographical, environment, and astronomical data,

Rahul Publications

including images from outer space. Typically, spatial data mining can reveal aspects like topology and distance.

**7. Biological Data Analysis**

Biological data mining practices are common in genomics, proteomics, and biomedical research. From characterizing patients' behaviour and predicting office visits to identifying medical therapies for their illnesses, data science techniques provide multiple advantages.

Some of the data mining applications in the Bioinformatics field are:

Semantic integration of heterogeneous and distributed databases

- Association and path analysis
- Use of visualization tools
- Structural pattern discovery
- Analysis of genetic networks and protein pathways

**8. Other Scientific Applications**

Fast numerical simulations in scientific fields like chemical engineering, fluid dynamics, climate, and ecosystem modeling generate vast datasets. Data mining brings capabilities like data warehouses, data preprocessing, visualization, graph-based mining, etc.

**9. Manufacturing Engineering**

System-level designing makes use of data mining to extract relationships between portfolios and product architectures. Moreover, the methods also come in handy for predicting product costs and span time for development.

**10. Criminal Investigation**

Data mining activities are also used in Criminology, which is a study of crime characteristics. First, text-based crime reports need to be converted into word processing files. Then, the identification and crime-machining process would take place by discovering patterns in massive stores of data.

**11. Counter-Terrorism**

Sophisticated mathematical algorithms can indicate which intelligence unit should play the headliner in counter-terrorism activities. Data mining can even help with police administration tasks, like determining where to deploy the workforce and denoting the searches at border crossings.

Rahul Publications

# Short Question and Answers

**1. What is predictive analysis ?**

*Ans :*

Predictive Analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. It does not tell you what will happen in the future.

Predictive analytics is the branch of the Advanced Analytics which is used to make predictions about unknown future events. It uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future.

Predictive Analytics uses historical data to predict future events. Typically, historical data is used to build a mathematical model that captures important trends. That predictive model is then used on current data to predict what will happen next, or to suggest actions to take for optimal outcomes.

**2. Trend Analysis.**

*Ans :*

A trend line is a straight line connecting a number of points on a graph. It is used to analyze the specific direction of a group of values set in a presentation. There are two kinds of trend lines, an uptrend with values going higher, and a downtrend where the direction of the line gradually drops to the lower values.

**(i) Predicting the Future**

Trend lines allow businesses to see the difference in various points over a period of time. This helps foretell the possible path the values will take in the future. This can help reveal perform Predictive Trend Line. The Predictive Trend Line add-on displays linear, exponential, and logarithmic trend lines in your dashboard charts. The trend line allows an end user to predict future values/metrics based on historical data. See the referenced wiki page on the regression analysis for more information business departments, such as sales.

By knowing how to add a trend line to your presentation, you can create a graphical representation of the values you have computed. This will enable the user to easily comprehend and analyze the message you are trying to imply.

**(ii) Predictive Trend Line**

The Predictive Trend Line add-on displays linear, exponential, and logarithmic trend lines in your dashboard charts. The trend line allows an end user to predict future values/ metrics based on historical data. See the referenced wiki page on the regression analysis for more information.

**3. Regression Analysis**

*Ans :*

1. Regression studies 'nature' of relationship between the variables.
2. Regression means stepping back or returning to average value and is a mathematical measure expressing the average relationship between the variables.
3. Regression clearly indicates the cause and effect relationship between the variables.
4. In regression, the variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
5. Regression coefficients are absolute measures indicating the change in the value of one variable for a unit change in the value of the other variable.

**4. Limitations of regression analysis.**

*Ans :*

1. Regression analysis assumes that linear relationship exists among the related variables. But in the 'area of social sciences,

Rahul Publications

linear relationship may not exist among the related variables.

2. When regression analysis is used to evaluate the value of dependent variable based on independent variable, it is assumed that the static conditions of relationship exist between 'them. These statistic conditions do not exist in social sciences, so, this assumption "minimizes the use' of regression analysis in social science.

3. The value of dependent variable can be evaluated based on independent variable by using regression analysis but only upto some limits. If the circumstances go beyond the limits, then resists would be inaccurate.

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

**5. Regression equation.**

*Ans :*

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

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

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

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

**6. Least square regression.**

*Ans :*

Least-squares regression is defined as the platform for best fitting the regression line and the observed values. In regression analysis, an assumption is made that in the sample data every value of dependent variable ' F 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 'F and the number of miles travelled is represented by independent variable X. The travel time 'F 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 Y is given by,

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

In the above equation, $\beta_0$ and $\beta_1$ refers to population parameters representing the intercept and slope, respectively. If X = 0, the intercept '$\beta_0$' refers to expected value of Y and the slope '$\beta_1$' 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,

$$Y = \beta_0 + \beta_1 X + \in$$

**7. Coefficient of Determination.**

*Ans :*

The coefficient of determination method is used to predict and explain the future outcomes of a model. This method is also known as R squared. This method also acts like a guideline which helps in measuring the model's accuracy

Rahul Publications

The coefficient of determination or R squared method is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set.

(i) The coefficient of determination is the square of the correlation (r), thus it ranges from 0 to 1.

(ii) With linear regression, the coefficient of determination is equal to the square of the correlation between the x and y variables.

(iii) If R2 is equal to 0, then the dependent variable cannot be predicted from the independent variable.

(iv) If R2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error.

(v) If R2 is between 0 and 1, then it indicates the extent that the dependent variable can be predictable. If R2 of 0.10 means, it is 10 percent of the variance in the y variable is predicted from the x variable. If 0.20 means, 20 percent of the variance in the y variable is predicted from the x variable, and so on.

**8. Define Data Mining.**

*Ans :*

**Meaning**

Data mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Thus, data mining should have been more appropriately named as knowledge mining which emphasis on mining from large amounts of data. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The key properties of data mining are Automatic discovery of patterns Prediction of likely outcomes Creation of actionable information Focus on large datasets and databases

**9. Applications of Data Mining.**

*Ans :*

**1. Financial Analysis**

The banking and finance industry relies on high-quality, reliable data. In loan markets, financial and user data can be used for a variety of purposes, like predicting loan payments and determining credit ratings. And data mining methods make such tasks more manageable.

Classification techniques facilitate the separation of crucial factors that influence customers' banking decisions from the irrelevant ones. Further, multidimensional clustering techniques allow the identification of customers with similar loan payment behaviours. Data analysis and mining can also help detect money laundering and other financial crimes.

**2. Telecommunication Industry**

Expanding and growing at a fast pace, especially with the advent of the internet. Data mining can enable key industry players to improve their service quality to stay ahead in the game.

Pattern analysis of spatiotemporal databases can play a huge role in mobile tele-communication, mobile computing, and also web and information services. And techniques like outlier analysis can detect fraudulent users. Also, OLAP and visualization tools can help compare information, such as user group behaviour, profit, data traffic, system overloads, etc.

**3. Intrusion Detection**

Global connectivity in today's technology-driven economy has presented security challenges for network administration. Network resources can face threats and actions that intrude on their confidentiality or integrity. Therefore, detection of intrusion has emerged as a crucial data mining practice.

It encompasses association and correlation analysis, aggregation techniques, visualization,

Rahul Publications

and query tools, which can effectively detect any anomalies or deviations from normal behaviour.

**4. Retail Industry**

The organized retail sector holds sizable quantities of data points covering sales, purchasing history, delivery of goods, consumption, and customer service. The databases have become even larger with the arrival of e-commerce marketplaces.

In modern-day retail, data warehouses are being designed and constructed to get the full benefits of data mining. Multidimensional data analysis helps deal with data related to different types of customers, products, regions, and time zones. Online retailers can also recommend products to drive more sales revenue and analyze the effectiveness of their promotional campaigns. So, from noticing buying patterns to improving customer service and satisfaction, data mining opens many doors in this sector.

**5. Energy Industry**

Big Data is available even in the energy sector nowadays, which points to the need for appropriate data mining techniques. Decision tree models and support vector machine learning are among the most popular approaches in the industry, providing feasible solutions for decision-making and management. Additionally, data mining can also achieve productive gains by predicting power outputs and the clearing price of electricity.

**10. Functionalities of Data Mining.**

*Ans :*

There are various data mining functionalities which are as follows

- **Data characterization**

It is a summarization of the general characteristics of an object class of data. The data corresponding to the user-specified class is generally collected by a database query. The output of data characterization can be presented in multiple forms.

- **Data discrimination**

It is a comparison of the general characteristics of target class data objects with the general characteristics of objects from one or a set of contrasting classes. The target and contrasting classes can be represented by the user, and the equivalent data objects fetched through database queries.

- **Association Analysis**

It analyses the set of items that generally occur together in a transactional dataset. There are two parameters that are used for determining the association rules "

- It provides which identifies the common item set in the database.
- Confidence is the conditional probability that an item occurs in a transaction when another item occurs.

- **Classification**

Classification is the procedure of discovering a model that represents and distinguishes data classes or concepts, for the objective of being able to use the model to predict the class of objects whose class label is anonymous. The derived model is established on the analysis of a set of training data (i.e., data objects whose class label is common).

Rahul Publications

# Choose the Correct Answers

1. _____ is an essential process where intelligent methods are applied to extract data patterns. [ b ]
   (a) Data warehousing (b) Data mining
   (c) Text mining (d) Data selection
2. Which of the following is not a data mining functionality? [ c ]
   (a) Characterization and Discrimination (b) Classification and regression
   (c) Selection and interpretation (d) Clustering and Analysis
3. _____ is the process of finding a model that describes and distinguishes data classes or concepts. [ b ]
   (a) Data Characterization (b) Data Classification
   (c) Data discrimination (d) Data selection
4. Strategic value of data mining is _____. [ c ]
   (a) Cost-sensitive (b) Work-sensitive
   (c) Time-sensitive (e) Technical-sensitive
5. _____ is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. [ a ]
   (a) Data Characterization (b) Data Classification
   (c) Data discrimination (d) Data selection
6. If the two series move in reverse directions and the variations in their values are always proportionate, it is said to be: [ c ]
   (a) Negative correlation (b) Positive correlation
   (c) Perfect negative correlation (d) Perfect positive correlation
7. If one item is fixed and unchangeable and the other item varies, the correlation coefficient will be: [ c ]
   (a) Positive (b) Negative
   (c) Zero (d) Undecided
8. A process by which we estimate the value of dependent variable on the basis of one or more independent variables is called: [ b ]
   (a) Correlation (b) Regression
   (c) Residual (d) Slope
9. The slope of the regression line of Y on X is also called the: [ d ]
   (a) Correlation coefficient of X on Y (b) Correlation coefficient of Y on X
   (c) Regression coefficient of X on Y (d) Regression coefficient of Y on X
10. In simple linear regression, the numbers of unknown constants are: [ b ]
   (a) One (b) Two
   (c) Three (d) Four

Rahul Publications

# Fill in the blanks

1. _____ Analytics is the process of using data analytics to make predictions based on data.
2. The regression analysis confined to the study of only two variables at a time is termed as _____ regression.
3. The regression analysis for studying more than two variables at a time is termed as _____ regression.
4. _____ average methods take the average of past actuals and project it forward.
5. Exponential smoothing is a more advanced form of _____ forecasting.
6. _____ is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
7. A data mining project starts with the understanding of the _____ problem.
8. _____ experts build the data model for the modeling process.
9. _____ is an informative search used by data consumers to form true analysis from the information gathered.
10. _____ Analytics is the practice of extracting information from existing data sets in order to determine patterns.

**ANSWERS**

1. Predictive
2. Simple
3. Multiple
4. Moving
5. Time series
6. Data mining
7. Business
8. Domain
9. Data exploration
10. Predictive

Rahul Publications

# FACULTY OF MANAGEMENT

## BBA III Year VI - Semester(CBCS) Examination

## MODEL PAPER - I

## BASIC BUSINESS ANALYTICS

**Time : 3 Hours]** **[Max. Marks : 60**

### PART - A (5 × 3 = 15)

**Note:** Answer any **FIVE** questions.

**ANSWERS**

1. Define business analytics. **(Unit - I, SQA-1)**
2. State the advantages of Big Data. **(Unit - I, SQA-7)**
3. What is population variance? **(Unit - II, SQA-3)**
4. Population. **(Unit - II, SQA-1)**
5. What is Range? **(Unit - III, SQA-2)**
6. Coefficient of Variation. **(Unit - III, SQA-5)**
7. Define Data Mining. **(Unit - IV, SQA-8)**
8. Regression equation. **(Unit - IV, SQA-5)**

### PART - B (3 × 15 = 45)

**Note:** Answer all the questions.

9. (a) Explain different types of Business Analytical Methods. **(Unit - I, Q.No.2)**

OR

(b) Describe the various stages of Big Data Life Cycle. **(Unit - I, Q.No.12)**

10. (a) (i) Explain about population and sample data. **(Unit - II, Q.No.2)**

(ii) What is Time Series Data and Cross Sectional Data? Explain with an examples. **(Unit - II, Q.No.10)**

OR

(b) (i) Explain how to calculate population variance in Excel. **(Unit - III, Q.No.6)**

(ii) Define Quartile? How to find the quartile deviation in statistics? Explain with an examples. **(Unit - III, Q.No.12)**

11. (a) (i) Find the Regression Analysis of following data.

| | | | | | |
|---|---|---|---|---|---|
| X | 2 | 6 | 4 | 8 | 10 |
| Y | 4 | 10 | 6 | 15 | 5 |

**(Unit - IV, Prob.2)**

(ii) Explain the various applications of Data Mining. **(Unit - IV, Q.No.27)**

OR

(b) (i) Explain different types and utility of regression analysis. **(Unit - IV, Q.No.7)**

(ii) Explain various issues relating to data classification. **(Unit - IV, Q.No.24)**

# FACULTY OF MANAGEMENT

## BBA III Year VI - Semester(CBCS) Examination

## MODEL PAPER - II

## BASIC BUSINESS ANALYTICS

**Time : 3 Hours]** **[Max. Marks : 60**

### PART - A (5 × 3 = 15)

**Note:** Answer any **FIVE** questions.

**ANSWERS**

1. What is Big Data? **(Unit - I, SQA-6)**
2. Descriptive Analytics. **(Unit - I, SQA-2)**
3. Sample. **(Unit - II, SQA-2)**
4. Categorical Data **(Unit - II, SQA-5)**
5. What is standard deviation? **(Unit - III, SQA-4)**
6. Percentiles **(Unit - III, SQA-6)**
7. What is predictive analysis? **(Unit - IV, SQA-1)**
8. Functionalities of Data Mining. **(Unit - IV, SQA-10)**

### PART - B (3 × 15 = 45)

**Note:** Answer all the questions.

9. (a) Describe briefly about the role of business analytics in current business environment. **(Unit - I, Q.No.6)**

OR

(b) What are the major types of Big data applications? **(Unit - I, Q.No.16)**

10. (a) (i) Write about various sources of data and how to collect them. **(Unit - II, Q.No.11)**

(ii) How to calculate variance in Excel ? Explain. **(Unit - II, Q.No.5)**

OR

(b) (i) What is standard deviation? Explain with an examples how is Standard Deviation calculated. **(Unit - III, Q.No.7)**

(ii) What is the empirical rule? How and where the empirical rule is used ? Explain with an example. **(Unit - III, Q.No.13)**

11. (a) (i) The following table gives the aptitude test scores and productivity induces of 10 workers selected at random

| Aptitude Scores(x) | 60 | 62 | 65 | 70 | 72 | 48 | 53 | 73 | 65 | 82 |
|---|---|---|---|---|---|---|---|---|---|---|
| Productivity Index (y) | 68 | 60 | 62 | 80 | 85 | 40 | 52 | 62 | 60 | 81 |

(i) Obtain regression equation y on x.

(ii) Estimate the Productivity index of a worker whose test scores is 92.

**(Unit - IV, Prob.3)**

(ii) What is Regression Analysis? **(Unit - IV, Q.No.6)**

OR

(b) (i) Explain the various approaches of Date mining. **(Unit - IV, Q.No.18)**

(ii) Explain the Classification of Data Mining. **(Unit - IV, Q.No.23)**

# FACULTY OF MANAGEMENT

## BBA III Year VI - Semester(CBCS) Examination

## MODEL PAPER - III

## BASIC BUSINESS ANALYTICS

**Time : 3 Hours]** **[Max. Marks : 60**

### PART - A (5 × 3 = 15)

**Note:** Answer any **FIVE** questions.

**ANSWERS**

1. Diagnostic Analytics **(Unit - I, SQA-3)**
2. What are the differences between descriptive analytics? **(Unit - I, SQA-9)**
3. Types of Data **(Unit - II, SQA-10)**
4. Time Series Data. **(Unit - II, SQA-6)**
5. Quartiles. **(Unit - III, SQA-7)**
6. What is correlation coefficient? **(Unit - III, SQA-11)**
7. Applications of Data Mining. **(Unit - IV, SQA-9)**
8. Least square regression. **(Unit - IV, SQA-6)**

### PART - B (3 × 15 = 45)

**Note:** Answer all the questions.

9. (a) Explain the different models in business analytics? **(Unit - I, Q.No.5)**

OR

(b) Explain the role of big data in competing food apps Swiggy and Zomato. **(Unit - I, Q.No.17)**

10. (a) (i) Explain the utility of measures of central tendency in understanding performance of business organization. **(Unit - II, Q.No.12)**

(ii) Explain the Relationship between Mean, Median and Mode with examples. **(Unit - II, Q.No.13)**

OR

(b) (i) What is variance? Explain how to find variance in various modes with an examples. **(Unit - III, Q.No.4)**

(ii) What is a Scatter Plot? Explain how to draw scatter plot for the given example. **(Unit - III, Q.No.18)**

11. (a) (i) Find the Least Square method for following data. Find the production in 2005?

| Year | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|
| Production (` 000) | 10 | 19 | 14 | 17 | 15 |

**(Unit - IV, Prob.1)**

(ii) Explain the Applications of Predictive analysis. **(Unit - IV, Q.No.3)**

OR

(b) (i) Explain the concept of simple linear regression by using MS Excel. **(Unit - IV, Q.No.11)**

(ii) Explain briefly about Coefficient of Determination. **(Unit - IV, Q.No.13)**