மனோன்மணியம் சுந்தரனார் பல்கலைக்கழகம்

Manonmaniam Sundaranar University

Reaccredited with 'A' Grade (CGPA 3.13 Out of 4.0) by NAAC (3rd Cycle)

Tirunelveli - 627 012, Tamilnadu, India.

DIRECTORATE OF DISTANCE

&

CONTINUING EDUCATION

BUSINESS ANALYTICS

EDITED BY

Dr.M.NITHYA M.COM.,M.Phil.,M.B.A.,TN-SET.,Ph.D

ASSISTANT PROFESSOR (T)

DEPARTMENT OF COMMERCE

MANONMANIAM SUNDARANAR UNIVERISTY

TIRUNELVELI

M. COM (II SEMESTER) UNDER CBCS

ELECTIVE 1 (b)

22 / TG –Colleges / M.Com / Semester II / Ppr.no.11 / Elective –1(b)

BUSINESSANALYTICS

L	T	P	C
4	0	0	3

Objectives

1. To enable students to learn the basics of business data analytics platforms
2. To teach quantitative analysis including sampling etc
3. To learn advanced statistical techniques such as multivariate analysis etc
4. To gain an understanding of the nuances of data mining
5. To teach the techniques of regression analysis

UNIT I Introduction to Data Analytics Platform - Visualizing Data - Describing and Summarizing Da - Challenges of Conventional Systems - Intelligent Data Analysis - Analytic Methodologies or Techniq es Used in Logical Analysis

UNIT II Quantitative Analysis - Sampling Methods and Estimation – Probability Distribution - Descriptive Statistics - Inferential Statistics - Hypothesis Testing, Explanatory and Predictive Models, nd Fact-Based Management to Drive Decisions and Actions - Tools - Analysis vs Reporting.

UNIT III One-Sample Tests - Two Independent Samples Tests - K Related Samples Tests - Measure of Correlation and Association - Multivariate Nonparametric Test for Interdependence - Probability nd Decision Making Under Uncertainty - Normal, Binomial, Poisson, and Exponential Distributions

UNIT IV Data Mining - Importing Data into Excel - SQL - Analysis of Variance and Experimental De gn - Statistical Process Control - Statistical Reporting - Foundations, Methods, Interpretations in Excel – – STATA – PSPP – EVIEWS – Machine Learning.

UNIT V Regression Analysis - Estimating Relationships - Linear versus Nonlinear Relationshi - Statistical Inference - Time Series Forecasting - Introduction to Optimization and Simulation Modeli – Optimization and Simulation Model - Decision Support System

Learning Outcome :

After the completion of the course, the students must be able to:

1. Gain an understanding of the basics of business data analytics platforms
2. Gainknowledge of quantitative analysis including sampling etc
3. Learnadvanced statistical techniques such as multivariate analysis etc
4. Describe the nuance of data mining
5. Gain knowledge of techniques of regression analysis

References :

1. Bowerman, B. (2016). Business Statistics in Practice: Using Data, Modeling, and Anal tics. McGraw-Hill Higher Education
2. Christian Albright, Wayne L. Winston (2015). Business Analytics : Data Analysis and De sion Making 5th Edition, CENGAGE
3. Cliff, T. (2014). Exploratory Data Analysis in Business and Economics: An Introduction sing SPSS. Stata, and Excel: Springer, New York, New York, 215
4. Gert H. N. Laursen, JesperThorlund (2018). Business Analytics for Managers, 2ed: Taking Bu ness Intelligence Beyond Reporting, Wiley

UNIT – I - INTRODUCTION TO DATA ANALYTICS PLATFORM

Introduction

A data analyst will extract raw data, organize it, and then analyze it, transforming it from incomprehensible numbers into coherent, intelligible information. Having interpreted the data, the data analyst will then pass on their findings in the form of suggestions or recommendations about what the company's next steps should be. Most companies are collecting loads of data all the time—but, in its raw form, this data doesn't really mean anything. This is where data analytics comes in. Data analytics is the process of analyzing raw data in order to draw out meaningful, actionable insights, which are then used to inform and drive smart business decisions.

Data analytics helps you to make sense of the past and to predict future trends and behaviors; rather than basing your decisions and strategies on guesswork, you're making informed choices based on what the data is telling you. Armed with the insights drawn from the data, businesses and organizations are able to develop a much deeper understanding of their audience, their industry, and their company as a whole—and, as a result, are much better equipped to make decisions and plan ahead.

As data is becoming more prominent by the minute, organizations are becoming data-driven, which means adopting methods to collect more data. This data is then sorted, stored, and then analyzed to derive logical and valuable information. Data analytics makes the process possible.

Data analytics is the process that refers to deriving valuable insights and information from data using quantitative and qualitative methods. It helps businesses and even in science - researchers use it to verify their theories, for example.

What kind of data can a company collect? There are three primary kinds.

- **First-party data:** The data a company collects about its customers.
- **Second-party data:** The data a company gets from a known organization that collected it originally.
- **Aggregated data:** The data a company buys from a marketplace.

So, now that you know what Data Analytics is, let's cover a brief evolution of it.

Evolution of Data Analytics

Data analytics has become the next big thing in both large companies and small startups. The process of data analytics has evolved. Let's take a journey through the evolution of data analytics.

Data Analytics and Statistics: Statistics has a pretty long history. Like, for example, taxation, governments carried out planning activities for the creation of censuses. It was possible with the use of statistics. Data analytics stemmed from statistics, which analyzed the obtained data.

Data Analysis and Computing: Technology advancements were game-changers to how businesses adopted data analytics. In 1890, Herman Hollerith invented the "Tabulating Machine" to reduce the time taken to create the Census. This machine was highly useful in finishing the 1890 US Census in only 18 months.

Data mining: Data mining got introduced in the 1990s, which is a process that discovers patterns in large data files. When data analytics moved from traditional methods to more modern means, you could obtain more positive results.

Google Web Search: When the Google search engine came into the picture, big data could be analyzed and processed quickly. It played an essential part in the evolution of Data Analytics because the search engine was more automated, scalable, and high-performing.

Data Processing: Today, Python & R are the leading technologies in data analytics. They are open-sourced and are capable of integrating with big data platforms and visualization tools. Businesses prefer R when the primary goal is exploratory analysis or modeling. At the same time, enterprises prefer Python to develop applications that have an embedded analytics engine.

Predictive Modeling: Some advanced data analytics techniques that the data scientists and organizations are using are: Random Forests, Matrix Factorization, TensorFlow, Simulated Annealing, etc.

Visualization: Many organizations are adopting more open-source technologies for their business. Few examples are D3 and Angular. This decision relies on several factors like cost, customization options, visual appeal, and interactivity.

Types of Data Analytics

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

1. Descriptive Analytics

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

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

2. Diagnostic Analytics

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

Identifying any unexpected changes in the data

Data related to the changes is collected.

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

3. Predictive Analytics

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

4. Prescriptive Analytics

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

Role of Data Analytics

Data analytics combines statistics, IT, and business. The main objective of a data analyst is to discover patterns in data. By doing this, the efficiency and performance of an organization improve.

We can explain the role of data analytics with the following points:

- It can retrieve masked information from data. This information can then be evaluated according to business needs.
- It helps in generating reports from the collected data and then given to the team involved.
- With data analytics, you can perform market analysis. It helps in understanding the strengths and weaknesses of the competitors.
- It improves customer experience by understanding their requirements.

Difference Between Data Analysis and Data Analytics

What makes data analytics different from data analysis? Let's find out.

Here are the six key differences between them both:

Structure: Data analysis includes defining data, investigation, cleaning, transforming the data to give meaningful results. Data analytics generally has a collection of data and investigation.

Key idea: While data analysis is a specialized form of data analytics, on the other hand, data analytics is a 'general' form of analytics.

Tools: The available data analysis tools are Google Fusion Tables, NodeXL, Tableau Public, etc. The available tools for data analytics are Python, SAS, Apache Spark, Excel, etc.

Purpose: Data analysis can be used for descriptive analysis, exploratory analysis, predictive analysis, etc. On the other hand, data analytics is used to find customer patterns, market trends, hidden patterns, etc.

How does Data Analytics Work?

The data analytics process takes place in the following steps:

Step 1: Analytics process

The first step is data collection. Data scientists identify the required information first and then work with the other data engineers and IT specialists to assemble and categorize it. Data integration routines combine the data from different sources and convert them into a standard format. It is then loaded into an analytics system.

Step 2: Data profiling and Data cleansing

In this next step, data quality issues are fixed. It includes running data profiling and data cleansing tasks. After, data governance policies are applied to make sure the data follows corporate standards.

Step 3: Predictive modeling tools

In this step, the data scientist builds an analytics model using predictive modeling tools. The model is "trained," which means its accuracy is tested, revised, and tested again. Finally, you can run the model in production mode.

Step 4: Communication of results

The data analytics process's last step includes communicating the results obtained from the analytical model to end-users. Charts and infographics are used to make this step easier.

Big Data Analytics

In simple words, big data analytics evaluate large data sets that contain different types of data. Hence, the name - Big Data. It helps identify hidden patterns in the data, market trends, customer preferences and demands, and other useful information. Big data analytics allows businesses to do well-informed companies so they can reap profits for their organization.

Now that we have understood the critical ideas of data analytics let's move on to the field's career prospects and future.

Top Applications of Data Analytics

Data analytics is applied in the following top areas:

- Security
- Transportation
- Fraud and risk detection
- Delivery logistics
- Customer interactions

- Real-estate and city planning
- Healthcare

Data Analytics in Future

It will be no surprise that data analytics tools will evolve rapidly in the coming years. According to recent studies, the field will have one of the fastest-growing rates. Here's what we can expect how the future of data analytics can look like:

More companies will adopt data analytics. What makes this field extremely useful is that businesses can retrieve data and create helpful reports without understanding the underlying algorithms. Data analytics increases the efficiency of the organization, and more companies will continue to adopt data analytics.

It will be a challenge to hire more data specialists. Even today, there is a shortage of skilled data analytics and scientists in the industry. Experts in the fields need to plan correctly to address that challenge by creating funding for training or creating programs to help the candidates.

There will be growth in machine learning. Machine learning and artificial intelligence (AI) strategies and tools will become more advanced. Businesses can adopt these strategies to create new products and services with new increased value.

As data keeps growing, it will be challenging to manage it. Finding a way to manage large sets of company data continually will become a challenge. New threats of security issues, privacy, time, and resources problems will arise, and solutions need to be found to address these challenges.

Today, data analytics is driving some of the significant companies to success. In a competitive world, gaining insights from extensive data contributes to ultimate growth. So, in summary, companies taking advantage of the benefits of data analytics and its advancements certainly stand out from the crowd without a doubt.

In this article, we have covered all the major concepts under data analytics. We hope you have understood its importance by now and how it helps businesses grow to become successful.

Five main steps that a data analyst will follow when tackling a new project:

Step 1: Define the question(s) you want to answer

The first step is to identify why you are conducting analysis and what question or challenge you hope to solve. At this stage, you’ll take a clearly defined problem and come up with a relevant question or hypothesis you can test. You’ll then need to identify what kinds of data you’ll need and where it will come from.

For example: A potential business problem might be that customers aren't subscribing to a paid membership after their free trial ends. Your research question could then be "What strategies can we use to boost customer retention?"

Step 2: Collect the data

With a clear question in mind, you're ready to start collecting your data. Data analysts will usually gather structured data from primary or internal sources, such as CRM software or email marketing tools. They may also turn to secondary or external sources, such as open data sources. These include government portals, tools like Google Trends, and data published by major organizations such as UNICEF and the World Health Organization.

Step 3: Clean the data

Once you've collected your data, you need to get it ready for analysis—and this means thoroughly cleaning your dataset. Your original dataset may contain duplicates, anomalies, or missing data which could distort how the data is interpreted, so these all need to be removed. Data cleaning can be a time-consuming task, but it's crucial for obtaining accurate results.

Step 4: Analyze the data

Now for the actual analysis! How you analyze the data will depend on the question you're asking and the kind of data you're working with, but some common techniques include regression analysis, cluster analysis, and time-series analysis (to name just a few). We'll go over some of these techniques in the next section. This step in the process also ties in with the four different types of analysis we looked at in section three (descriptive, diagnostic, predictive, and prescriptive).

Step 5: Visualize and share your findings

This final step in the process is where data is transformed into valuable business insights. Depending on the type of analysis conducted, you'll present your findings in a way that others can understand—in the form of a chart or graph, for example.

At this stage, you'll demonstrate what the data analysis tells you in regards to your initial question or business challenge, and collaborate with key stakeholders on how to move forwards. This is also a good time to highlight any limitations to your data analysis and to consider what further analysis might be conducted.

Data analytics techniques

Before we introduce some key data analytics techniques, let’s quickly distinguish between the two different types of data you might work with: quantitative and qualitative. Quantitative data is essentially anything measurable—for example, the number of people who answered “yes” to a particular question on a survey, or the number of sales made in a given year. Qualitative data, on the other hand, cannot be measured, and comprises things like what people say in an interview or the text written as part of an email.

Data analysts will usually work with quantitative data; however, there are some roles out there that will also require you to collect and analyze qualitative data, so it’s good to have an understanding of both. With that in mind, here are some of the most common data analytics techniques:

- Regression analysis: This method is used to estimate or “model” the relationship between a set of variables. You might use this to see if certain variables (a movie star’s number of Instagram followers and how much her last five films grossed on average) can be used to accurately predict another variable (whether or not her next film will be a big hit). Regression analysis is mainly used to make predictions. Note, however, that on their own, regressions can only be used to determine whether or not there is a relationship between a set of variables—they can’t tell you anything about cause and effect.
- Factor analysis: Sometimes known as dimension reduction, this technique helps data analysts to uncover the underlying variables that drive people’s behavior and the choices they make. Ultimately, it condenses the data in many variables into a few “super-variables”, making the data easier to work with. For example: If you have three different variables which represent customer satisfaction, you might use factor analysis to condense these variables into just one all-encompassing customer satisfaction score.
- Cohort analysis: A cohort is a group of users who have a certain characteristic in common within a specified time period—for example, all customers who purchased using a mobile device in March may be considered as one distinct cohort. In cohort analysis, customer data is broken up into smaller groups or cohorts; so, instead of treating all customer data the same, companies can see trends and patterns over time that relate to particular cohorts. In recognizing these patterns, companies are then able to offer a more targeted service.
- Cluster analysis: This technique is all about identifying structures within a dataset. Cluster analysis essentially segments the data into groups that are internally

homogenous and externally heterogeneous—in other words, the objects in one cluster must be more similar to each other than they are to the objects in other clusters. Cluster analysis enables you to see how data is distributed across a dataset where there are no existing predefined classes or groupings. In marketing, for example, cluster analysis may be used to identify distinct target groups within a larger customer base.

- Time-series analysis: In simple terms, time-series data is a sequence of data points which measure the same variable at different points in time. Time-series analysis, then, is the collection of data at specific intervals over a period of time in order to identify trends and cycles, enabling data analysts to make accurate forecasts for the future. If you wanted to predict the future demand for a particular product, you might use time-series analysis to see how the demand for this product typically looks at certain points in time.

These are just a few of the many techniques that data analysts will use, and we've only scratched the surface in terms of what each technique involves and how it's used. Some other common techniques include Monte Carlo simulations, dispersion analysis, discriminant analysis, and text or content analysis (the latter being a technique for analyzing qualitative data). We've covered seven of the most useful data analysis techniques in this guide.

Data analytics tools

Now let's take a look at some of the tools that a data analyst might work with. If you're looking to become a data analyst, you'll need to be proficient in at least some of the tools listed below—but, if you've never even heard of them, don't let that deter you! Like most things, getting to grips with the tools of the trade is all part of the learning curve.

- Microsoft Excel is a software program that enables you to organize, format, and calculate data using formulas within a spreadsheet system. Microsoft Excel may be used by data analysts to run basic queries and to create pivot tables, graphs, and charts. Excel also features a macro programming language called Visual Basic for Applications (VBA).
- Tableau is a popular business intelligence and data analytics software which is primarily used as a tool for data visualization. Data analysts use Tableau to simplify raw data into visual dashboards, worksheets, maps, and charts. This helps to make the data accessible and easy to understand, allowing data analysts to effectively share their insights and recommendations.

- SAS is a command-driven software package used for carrying out advanced statistical analysis and data visualization. Offering a wide variety of statistical methods and algorithms, customizable options for analysis and output, and publication-quality graphics, SAS is one of the most widely used software packages in the industry.
- RapidMiner is a software package used for data mining (uncovering patterns), text mining, predictive analytics, and machine learning. Used by both data analysts and data scientists alike, RapidMiner comes with a wide range of features—including data modeling, validation, and automation.
- Power BI is a business analytics solution that lets you visualize your data and share insights across your organization. Similar to Tableau, Power BI is primarily used for data visualization. While Tableau is built for data analysts, Power BI is a more general business intelligence tool.

What are the Types of Data Analysis?

The Data Analysis technique you decide to go with depends on the kind of information you have. Accordingly, there are two main Data Analysis techniques, namely Qualitative and Quantitative. Let's see what each means and how they can be of use to your business.

1. Quantitative Data Analysis

This type of Data Analysis leans more toward the statistical nature of your data. It generally tells you what is happening and whether the trends are showing a rise or fall. Below are the types of Quantitative Data Analysis:

- Descriptive Analysis: This type of Data Analysis lets you see the patterns and trends in a particular set of data. It includes processes such as calculating frequencies, percentages, and measures of central tendency, including mean, mode, and median.
- Inferential Analysis: It is used when the differences and correlations between particular data sets need to be examined. The processes involved include ANOVA, t-Tests, and Chi-Square.

2. Qualitative Data Analysis

While Quantitative Analysis focuses on numeric data, Qualitative is the complete opposite, dealing with non-numeric data such as audio, video recordings, images, texts, and transcripts. In addition, Qualitative Data generally tells you how your data is changing.

What are the Components of Data Analytics?

Data Analytics components refer to the different techniques you can use for processing any set of data. They include:

- Text Analytics: This is the technique used in autocorrect in phones and software such as Microsoft Word. It involves analyzing large amounts of text to come up with Algorithms. Applications include Linguistic Analysis and Pattern Recognition.
- Data Mining: One of the most critical Data Mining applications is determining behavioral patterns in inpatient data during clinical trials. As the name suggests, Data Mining breaks large chunks of data into smaller pieces that fit a specific purpose.
- Business Intelligence: This is one of the essential processes for any successful business. It involves transforming data into actionable strategies for a particular commercial entity. For example, this is the process behind product placement and pricing in most companies.

Data Visualization

While Data Analytics is more involved in bringing some form of structure into unorganized data, Data Visualization deals with picturing the information to develop trends and conclusions. In Data Visualization, information is organized into charts, graphs, and other forms of visual representations. This simplifies otherwise complicated information and makes it accessible to all the involved stakeholders to make critical business decisions.

Types of Data Visualization Techniques

Like Data Analytics, the type of Data Visualization technique you choose will largely depend on the type of data to be modeled and the purpose. It is worth noting that some Visualizations are manually created while others are automated. Below are some of the popular Visualization techniques:

- Histograms: This is a Graphical Visualization Tool that organizes a set of data into a range of frequencies. It bears key similarities to a Bar Graph and organizes information in a way that makes it easy to interpret.
- Graphs: These are excellent tools for analyzing the time series relationship in a particular set of data. For instance, a company’s annual profits could be analysed based on each month using a graph.
- Fever Charts: A Fever Chart is an indispensable tool for any business since it shows how data changes over time. For instance, a particular product’s performance could be analyzed based on its yearly profits.

- Heat map Visualization: This tool is based on the psychological fact that the human brain interprets colors much faster than numbers. It is a graph that uses numerical data points highlighted in light or warm colors to represent high or low-value points.
- Infographics: Infographics are effective when analyzing complex datasets. They take large amounts of data and organize it into an easy to interpret format.

These were some of the most popular Visualization you can leverage to level up your Data Analytics and Visualization workflows.

Advantages of Data Analytics and Visualization

Data Analytics and Visualization are a crucial elements of the business decision-making process. It helps the stakeholders to recognize patterns in the data and devise profitable business strategies. Below are some of the benefits of Data Analytics and Visualization:

- Better Decision Making: By using skilled Data Analysts and the right software, companies can identify market trends and make better business decisions to Boost Sales and Profits.
- Better Insights: Companies can get better insights into their Customer Base- using Data Analytics and Visualization, companies can break large customer data down into smaller sets that can be used to understand the Client Base better.
- Improving Productivity and Revenue Growth: By looking at the results from Data Analytics and Visualization, companies get to know which areas they need to invest in and what processes need to be automated for better efficiency.
- Noting Changes in Market Behaviour: With a real-time Data Analytics and Visualization Dashboard, company stakeholders can quickly identify changes in market behavior and make appropriate business decisions.
- Analysing Different Markets: Using Data Analytics and Visualization techniques, companies can analyse different markets and decide which ones to place attention on and which ones to avoid.
- Business Trends: This is one of the most valuable applications of Data Analytics and Visualization. It allows businesses to examine the present and past trends to make predictions that determine the way forward for the business.
- Data Relationships: This is one of the most obvious benefits of Data Analytics and Visualization. It helps companies note the relationships between independent data sets and make business decisions based on these findings.

Challenges for Conventional System

Business analytics can be possible only on large volume of data. It is sometime difficult obtain large volume of data and not question its integrity.

1. Executive Ownership – Business Analytics requires buy-in from senior leadership and a clear corporate strategy for integrating predictive models.
2. IT Involvement – Technology infrastructure and tools must be able to handle the data and Business Analytics processes.
3. Available Production Data vs. Cleansed Modeling Data – Watch for technology infrastructure that restricts available data for historical modeling, and know the difference between historical data for model development and real-time data in production Business analytics can be possible only on large volume of data. It is sometime difficult obtain large volume of data and not question its integrity.

Challenges of Big Data

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

1. Lack of proper understanding of Big Data

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

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

2. Data growth issues

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

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

3. Confusion while Big Data tool selection

Companies often get confused while selecting the best tool for Big Data analysis and storage. Is HBase or Cassandra the best technology for data storage? Is Hadoop Map Reduce good enough or will Spark be a better option for data analytics and storage?

These questions bother companies and sometimes they are unable to find the answers. They end up making poor decisions and selecting inappropriate technology. As a result, money, time, efforts and work hours are wasted.

4. Lack of data professionals

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

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

5. Securing data

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

6. Integrating data from a variety of sources

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

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

Intelligent Data Analysis

Intelligent Data Analysis (IDA) is one of the major issues in artificial intelligence and information. Intelligent data analysis discloses hidden facts that are not known previously and provides potentially important information or facts from large quantities of data (White, 2008). It also helps in making a decision. Based on machine learning, artificial intelligence, recognition of pattern, and records and visualization technology mainly, IDA helps to obtain useful information, necessary data and interesting models from a lot of data available online in order to make the right choices.

Intelligent data analysis helps to solve a problem that is already solved as a matter of routine. If the data is collected for the past cases together with the result that was finally achieved, such data can be used to revise and optimize the presently used strategy to arrive at a conclusion.

In certain cases, if some questions arise for the first time, and have only a little knowledge about it, data from the related situations helps us to solve the new problem or any unknown relationships can be discovered from the data to gain knowledge in an unfamiliar area.

Steps Involved In IDA:

IDA, in general, includes three stages: (1) Preparation of data; (2) data mining; (3) data validation and explanation (Keim & Ward, 2007). The preparation of data involves opting for the required data from the related data source and incorporating it into a data set that can be used for data mining.

The main goal of intelligent data analysis is to obtain knowledge. Data analysis is the process of a combination of extracting data from data set, analyzing, classification of data, organizing, reasoning, and so on. It is challenging to choose suitable methods to resolve the complexity of the process.

Regarding the term visualization, we have moved away from visualization to use the term charting. The term analysis is used for the method of incorporating, influencing, filtering and scrubbing the data, which certainly contains, but is not limited to interrelating with their data through charts.

The Goal of Data Analysis:

Data analysis need not essentially involve arithmetic or statistics. While it is true that analysis often involves one or both, and that many analytical pursuits cannot be handled without them, much of the data analysis that people perform in the course of their work involves at most mathematics no more complicated than the calculation of the mean of a set of values. The essential activity of analysis is a comparison (of values, patterns, etc.), which can often be done by simply using our eyes.

The aim of the analysis is not to find out appealing information in the data. Rather, this is only a vital part of the process (Berthold & Hand, 2003). The aim is to make sense of data (i.e., to understand what it means) and then to make decisions based on the understanding that is achieved. Information in and of itself is not useful. Even understanding information in and of it is not useful. The aim of data analysis is to make better decisions.

The process of data analysis starts with the collection of data that can add to the solution of any given problem, and with the organization of that data in some regular form. It involves identifying and applying a statistical or deterministic schema or model of the data that can be manipulated for explanatory or predictive purposes. It then involves an interactive or automated solution that explores the structured data in order to extract information – a solution to the business problem – from the data.

Describing Data

Using descriptive statistics to describe our data, we can explain and provide a quick summary of our data. We can do this in several ways. First, assuming multiple trials of a test were completed, we can produce a quantitative summary of the the individual subject's performance by reporting the average or highest score achieved. Consider an example where you are doing some sprint testing and every athlete completes 2 trials. Should you report the average of the two trials or just the peak value? The answer likely depends on the specific situation.

We can also describe the characteristics of the performance of every subject in our sample by reporting the score that all the other scores are centered around and how much the scores vary around this number. For example, we might be tracking steps as a measure of physical activity and the average number of steps taken in our sample was 980. That would be the center of our data. If the standard deviation is 39 steps, then we know that our data do not vary too much. If our standard deviation was 856, then we would know that our data varies quite a bit more since the standard deviation is nearly the same size as the average. We will dive deeper into this concept later in this chapter.

Finally, we can describe how the data are distributed. If all the distribution is plotted, like the example 40-yard dash sprint data shown in the figure above, we can describe its shape. While it isn't perfect, this is mostly symmetrical and not skewed to one side or the other. Notice the black line in the middle of the plot depicting the average of the data and the two dotted lines on either side depicting one standard deviation above and below the average. Most of the data fall within this range.

UNIT – II – QUANTITATIVE ANALYSIS

Sampling Methods

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.

It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

For example, if a drug manufacturer would like to research the adverse side effects of a drug on the country’s population, it is almost impossible to conduct a research study that involves everyone. In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug’s behavior.

Types of sampling: sampling methods

Sampling in market action research is of two types – probability sampling and non-probability sampling. Let’s take a closer look at these two methods of sampling.

1. **Probability sampling:** Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.
2. **Non-probability sampling:** In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

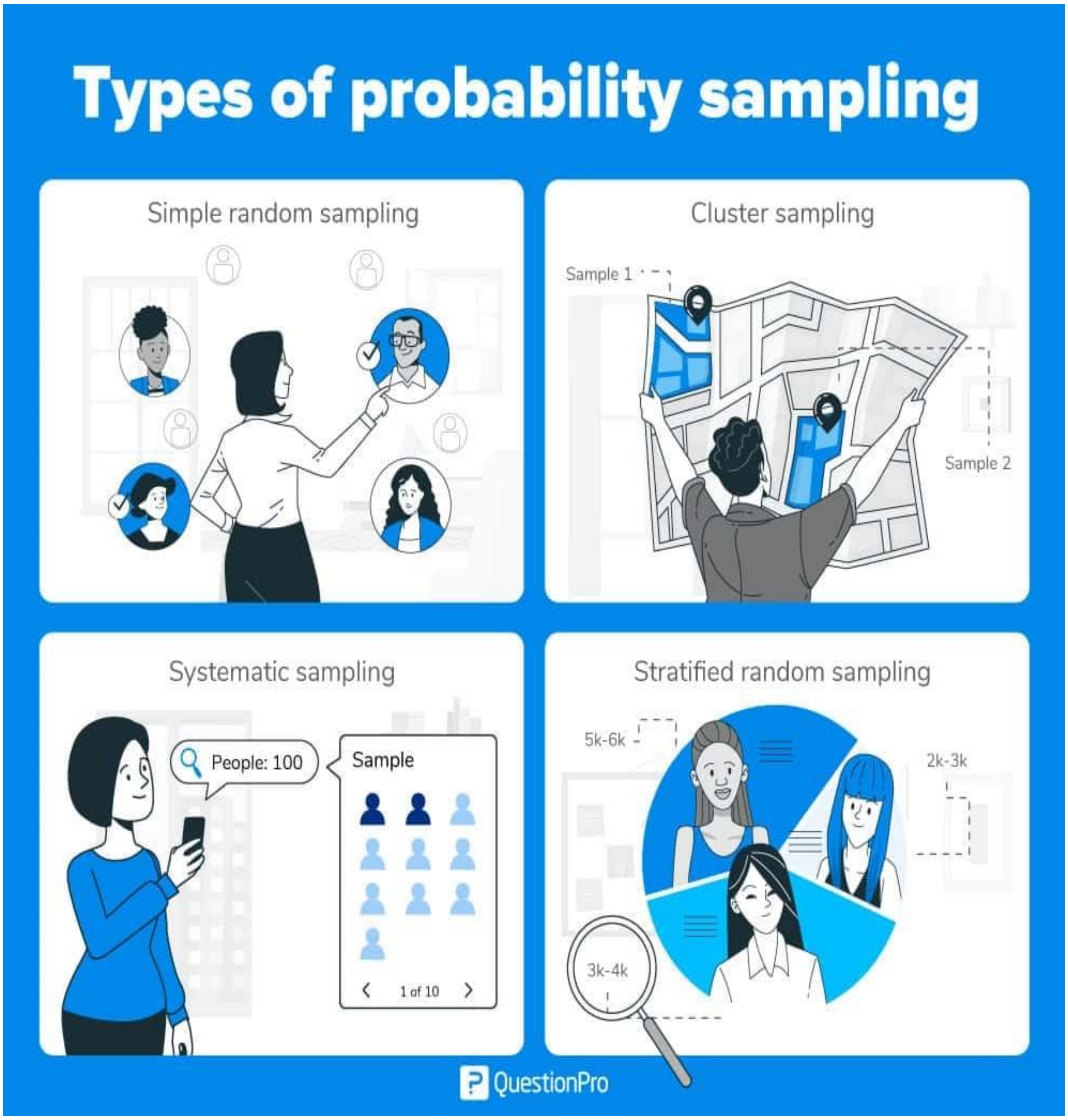
In this blog, we discuss the various probability and non-probability sampling methods that you can implement in any market research study.

Types of probability sampling with examples:

Probability sampling is a sampling technique in which researchers choose samples from a larger population using a method based on the theory of probability. This sampling method considers every member of the population and forms samples based on a fixed process.

For example, in a population of 1000 members, every member will have a 1/1000 chance of being selected to be a part of a sample. Probability sampling eliminates sampling bias in the population and gives all members a fair chance to be included in the sample.

There are four types of probability sampling techniques:



- **Simple random sampling:** One of the best probability sampling techniques that helps in saving time and resources, is the Simple Random Sampling method. It is a reliable

method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.
For example, in an organization of 500 employees, if the HR team decides on conducting team building activities, it is highly likely that they would prefer picking chits out of a bowl. In this case, each of the 500 employees has an equal opportunity of being selected.

- **Cluster sampling:** Cluster sampling is a method where the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inference from the feedback.
For example, if the United States government wishes to evaluate the number of immigrants living in the Mainland US, they can divide it into clusters based on states such as California, Texas, Florida, Massachusetts, Colorado, Hawaii, etc. This way of conducting a survey will be more effective as the results will be organized into states and provide insightful immigration data.
- **Systematic sampling:** Researchers use the systematic sampling method to choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.
For example, a researcher intends to collect a systematic sample of 500 people in a population of 5000. He/she numbers each element of the population from 1-5000 and will choose every 10th individual to be a part of the sample (Total population/ Sample Size = 5000/500 = 10).
- **Stratified random sampling:** Stratified random sampling is a method in which the researcher divides the population into smaller groups that don’t overlap but represent the entire population. While sampling, these groups can be organized and then draw a sample from each group separately.
For example, a researcher looking to analyze the characteristics of people belonging to different annual income divisions will create strata (groups) according to the annual family income. Eg – less than $20,000, $21,000 – $30,000, $31,000 to $40,000, $41,000 to $50,000, etc. By doing this, the researcher concludes the characteristics of people belonging to different income groups. Marketers can analyze which income groups to target and which ones to eliminate to create a roadmap that would bear fruitful results.

Uses of probability sampling

There are multiple uses of probability sampling:

- **Reduce Sample Bias:** Using the probability sampling method, the bias in the sample derived from a population is negligible to non-existent. The selection of the sample mainly depicts the understanding and the inference of the researcher. Probability sampling leads to higher quality data collection as the sample appropriately represents the population.
- **Diverse Population:** When the population is vast and diverse, it is essential to have adequate representation so that the data is not skewed towards one demographic. For example, if Square would like to understand the people that could make their point-of-sale devices, a survey conducted from a sample of people across the US from different industries and socio-economic backgrounds helps.
- **Create an Accurate Sample:** Probability sampling helps the researchers plan and create an accurate sample. This helps to obtain well-defined data.

Types of non-probability sampling with examples

The non-probability method is a sampling method that involves a collection of feedback based on a researcher or statistician's sample selection capabilities and not on a fixed selection process. In most situations, the output of a survey conducted with a non-probable sample leads to skewed results, which may not represent the desired target population. But, there are situations such as the preliminary stages of research or cost constraints for conducting research, where non-probability sampling will be much more useful than the other type.

Four types of non-probability sampling explain the purpose of this sampling method in a better manner:

- **Convenience sampling:** This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as convenience sampling, because of the researcher's ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it's purely done based on proximity and not representativeness. This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.

For example, startups and NGOs usually conduct convenience sampling at a mall to distribute leaflets of upcoming events or promotion of a cause – they do that by standing at the mall entrance and giving out pamphlets randomly.

- **Judgmental or purposive sampling:** Judgemental or purposive samples are formed by the discretion of the researcher. Researchers purely consider the purpose of the study, along with the understanding of the target audience. For instance, when researchers want to understand the thought process of people interested in studying for their master's degree. The selection criteria will be: "Are you interested in doing your masters in …?" and those who respond with a "No" are excluded from the sample.
- **Snowball sampling:** Snowball sampling is a sampling method that researchers apply when the subjects are difficult to trace. For example, it will be extremely challenging to survey shelterless people or illegal immigrants. In such cases, using the snowball theory, researchers can track a few categories to interview and derive results. Researchers also implement this sampling method in situations where the topic is highly sensitive and not openly discussed—for example, surveys to gather information about HIV Aids. Not many victims will readily respond to the questions. Still, researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.
- **Quota sampling:** In Quota sampling, the selection of members in this sampling technique happens based on a pre-set standard. In this case, as a sample is formed based on specific attributes, the created sample will have the same qualities found in the total population. It is a rapid method of collecting samples.

Uses of non-probability sampling

Non-probability sampling is used for the following:

- **Create a hypothesis:** Researchers use the non-probability sampling method to create an assumption when limited to no prior information is available. This method helps with the immediate return of data and builds a base for further research.
- **Exploratory research:** Researchers use this sampling technique widely when conducting qualitative research, pilot studies, or exploratory research.
- **Budget and time constraints:** The non-probability method when there are budget and time constraints, and some preliminary data must be collected. Since the survey design is not rigid, it is easier to pick respondents at random and have them take the survey or questionnaire.

For any research, it is essential to choose a sampling method accurately to meet the goals of your study. The effectiveness of your sampling relies on various factors. Here are some steps expert researchers follow to decide the best sampling method.

- Jot down the research goals. Generally, it must be a combination of cost, precision, or accuracy.
- Identify the effective sampling techniques that might potentially achieve the research goals.
- Test each of these methods and examine whether they help in achieving your goal.
- Select the method that works best for the research.

ESTIMATION

In most statistical research studies, population parameters are usually unknown and have to be estimated from a sample. As such the methods for estimating the population parameters assume animportant role in statistical anlysis.

The random variables (such as $\bar{X}$ and $\square_s^2$) used to estimate population parameters, such as

□ and $\square_p^2$ are conventionally called as '*estimators*', while specific values of these (such as $\bar{X} \square 105$

or $\square_s^2 \square 21.44$) are referred to as '*estimates*' of the population parameters. The estimate of a population parameter may be one single value or it could be a range of values. In the former case itis referred as *point estimate*, whereas in the latter case it is termed as *interval estimate*. The

researcher usually makes these two types of estimates through sampling analysis. While making estimates of population parameters, the researcher can give only the best point estimate or else he shall have to speak in terms of intervals and probabilities for he can never estimate with certainty the exact values of population parameters. Accordingly he must know the various properties of a good estimator so that he can select appropriate estimators for his study. He must know that a good estimator possesses the following properties:

(i) An estimator should on the average be equal to the value of the parameter being estimated.This is popularly known as the *property* of *unbiasedness*. An estimator is said to be unbiased if the expected value of the estimator is equal to the parameter being estimated.

The sample mean $(\bar{X})$ is he most widely used estimator because of the fact that it provides an unbiased estimate of the population mean (□).

(ii) An estimator should have a relatively small variance. This means that the most efficient estimator, among a group of unbiased estimators, is one which has the smallest variance. This property is technically described as the *property of efficiency*.

(iii) An estimator should use as much as possible the information available from the sample. This property is known as the *property of sufficiency*.

(iv) An estimator should approach the value of population parameter as the sample size becomes larger and larger. This property is referred to as the *property of consistency*.

Keeping in view the above stated properties, the researcher must select appropriate estimator(s) for his study. We may now explain the methods which will enable us to estimate with reasonable accuracy the population mean and the population proportion, the two widely used concepts.

ESTIMATING THE POPULATION MEAN (□)

So far as the point estimate is concerned, the sample mean $\bar{X}$ is the best estimator of the population mean, □ , and its sampling distribution, so long as the sample is sufficiently large, approximates the normal distribution. If we know the sampling distribution of X , we can make statements about any estimate that we may make from the sampling information. Assume that we take a sample of 36

students and find that the sample yields an arithmetic mean of 6.2 i.e., $\bar{X}$ □ 6.2 . Replace these student names on the population list and draw another sample of 36 randomly and let us assume that we get a mean of 7.5 this time. Similarly a third sample may yield a mean of 6.9; fourth a mean of 6.7, and so on. We go on drawing such samples till we accumulate a large number of means of samples of 36. Each such sample mean is a separate point estimate of the population mean. When such means are presented in the form of a distribution, the distribution happens to be quite close to normal. This is a characteristic of a distribution of sample means (and also of other sample statistics). Even if the population is not normal, the sample means drawn from that population are dispersed around the parameter in a distribution that is generally close to normal; the mean of the distribution of sample means is equal to the population mean.[5] This is true in case of large samples as per the dictates of the central limit theorem. This relationship between a population distribution and a distribution of sample

mean is critical for drawing inferences about parameters. The relationship between the dispersion ofa population distribution and that of the sample mean can be stated as under:

$$\sigma_{X} = \frac{\square\, p}{\sqrt{n}}$$

where $\square_{X}$ = standard error of mean of a given sample size

$\square\, p$ = standard deviation of the population

n = size of the sample.

How to find $\square\, p$ when we have the sample data only for our analysis? The answer is that we mustuse some best estimate of $\square\, p$ and the best estimate can be the standard deviation of the sample,

$\square_{s}$. Thus, the standard error of mean can be worked out as under:[6]

$$\sigma_{\bar{X}} = \frac{\square\, s}{\sqrt{n}}$$

where $\square_{s} \square \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}}$

With the help of this, one may give interval estimates about the parameter in probabilistic terms(utilising the fundamental characteristics of the normal distribution). Suppose we take one sample of 36 items and work out its mean $(\bar{X})$ to be equal to 6.20 and its standard deviation $(\square_{s})$ to be equal to 3.8, Then the best point estimate of population mean $(\square)$ is 6.20. The standard error of mean $(\square_{\bar{X}})$ would be $3.8/\sqrt{36} \square 3.8/6 \square 0.663$. If we take the interval estimate of $\square$ to be $\bar{X} \square 1.96\,(\square_{\bar{X}})$ or $6.20 \square 1.24$ or from 4.96 to 7.44, it means that there is a 95 per cent chance that the population mean is within 4.96 to 7.44 interval. In other words, this means that if we were to take a complete census of all items in the population, the chances are 95 to 5 that we would find the population mean lies between 4.96 to 7.44*. In case we desire to have an estimate that will hold for a much smaller range, then we must either accept a smaller degree of confidence in the results or take a sample large enough to provide this smaller interval with adequate confidence levels. Usually we think of increasing the sample size till we can secure the desired interval estimate and the degree of confidence.

Illustration 1

From a random sample of 36 New Delhi civil service personnel, the mean age and the sample standard deviation were found to be 40 years and 4.5 years respectively. Construct a 95 per cent confidence interval for the mean age of civil servants in New Delhi.

Solution: The given information can be written as under:

$$n = 36$$

$$\bar{X} \ □ \ 40 \text{ years}$$

$$□_s \ □ \ 4.5 \text{ years}$$

and the standard variate, z, for 95 per cent confidence is 1.96 (as per the normal curve area table).

Thus, 95 per cent confidence inteval for the mean age of population is:

$$X \ □ \ z \frac{□ \ s}{\sqrt{n}}$$

or $$40 \ □ \ 1.96 \frac{4.5}{\sqrt{36}}$$

or $$40 \ □ \ (1.96)(0.75)$$

or $$40 \ □$$

1.47 Years

Illustration 2

In a random selection of 64 of the 2400 intersections in a small city, the mean number of

[6] To make the sample standard deviation an unbiased estimate of the population, it is necessary to divide □ $(X_i$ □ $\bar{X})^2$ by (n – 1) and not by simply (n).

[*] In case we want to change the degree of confidence in the interval estimate, the same can be done using the table of areas under the normal curve.

scooteraccidents per year was 3.2 and the sample standard deviation was 0.8.

(1) Make an estimate of the standard deviation of the population from the sample standard deviation.

(2) Work out the standard error of mean for this finite population.

(3) If the desired confidence level is .90, what will be the upper and lower limits of the confidence interval for the mean number of accidents per intersection per year?

Solution: The given information can be written as under: *N* = 2400 (This means that population is finite)*n* = 64

$\bar{X}$ □ 3.2

□ $_s$ □ 0.8

and the standard variate (*z*) for 90 per cent confidence is 1.645 (as per the normal curve area table).

Now we can answer the given questions thus:

(1) The best point estimate of the standard deviation of the population is the standard deviationof the sample itself.

Hence,

□ $\hat{}_p$ □ □ $_s$ □ 0.8

(2) Standard error of mean for the given finite population is as follows:

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Thus, 90 per cent confidence interval for population mean is

$$\bar{X} \pm t\frac{s}{\sqrt{n}}$$

$$= 36.8 \pm 2.353\frac{2.8}{\sqrt{4}} = 36.8 \pm (2.353)(1.4)$$

$$= 36.8 \pm 3.294 \text{ tons per shift.}$$

ESTIMATING POPULATION PROPORTION

So far as the point estimate is concerned, the sample proportion (*p*) of units that have a particular characteristic is the best estimator of the population proportion $(\hat{p})$ and its sampling distribution, so long as the sample is sufficiently large, approximates the normal distribution. Thus, if we take a random sample of 50 items and find that 10 per cent of these are defective i.e., *p* = .10, we can use this sample proportion (*p* = .10) as best estimator of the population proportion $(\hat{p} = p = .10)$. In case we want to construct confidence interval to estimate a population proportion, we should use the binomial distribution with the mean of population $(\mu) = n \cdot p$, where *n* = number of trials, *p* = probability of a success in any of the trials and population standard deviation $= \sqrt{npq}$. As the sample size increases, the binomial distribution approaches normal distribution which we can use for our purpose of estimating a population proportion. The mean of the sampling distribution of the proportion of successes $(\square\, p)$ is taken as equal to p and the standard deviation for the proportion of

successes, also known as the standard error of proportion, is taken as equal $\sqrt{pq/n}$. But when population proportion is unknown, then we can estimate the population parameters by substituting the corresponding sample statistics p and q in the formula for the standard error of proportion to obtain the estimated standard error of the proportion as shown below:

$$\square_{p} \square \sqrt{\frac{pq}{}}$$

Using the above estimated standard error of proportion, we can work out the confidence interval for population proportion thus:

$$p \square z \square \sqrt{\frac{pq}{}}$$

where

p = sample proportion of successes;

$q = 1 - p$;

n = number of trials (size of the sample);

z = standard variate for given confidence level (as per normal curve area table).

Illustration 4

A market research survey in which 64 consumers were contacted states that 64 per cent of all consumers of a certain product were motivated by the product's advertising. Find the confidence limits for the proportion of consumers motivated by advertising in the population, given a confidencelevel equal to 0.95.

Solution: The given information can be written as under:

$n = 64$

p = 64% or .64

$q = 1 - p = 1 - .64 = .36$

and the standard variate (z) for 95 per cent confidence is 1.96 (as per the normal curve area table).

Thus, 95 per cent confidence interval for the proportion of consumers motivated by advertising in the population is:

$$p \;\square\; z \;\square \sqrt{pq}$$

$$\square\ .64 \;\square\; 1.96\sqrt{(0.64)(0.36)}$$

$$\square\ .64 \;\square\; (1.96)(.06)$$

$$\square\ .64 \;\square\; .1176$$

Thus, lower confidence limit is

52.24%upper confidence limit

is 75.76%

For the sake of convenience, we can summarise the formulae which give confidence intevalswhile estimating population mean (□)and the population proportion ($\hat{p}$)as shown in the following

table.

Summarising Important Formulae Concerning Estimation

	In case of infinite population	*In case of finite population*[*]
Estimating population mean (□) when we know □ p	X □ □ p z □ n	X □ □ □ N □ z □ p nN □ n 1
Estimating population mean (□) when we do not know □ p	X □ z □ □ s n	X □ z □ □ N □ n s □ N □ 1 n

	In case of infinite population	*In case of finite population*[*]
and use σ_s as the best estimateof σ_p and sample is large (i.e.,		
Estimating population mean (μ) when we do not know σ_p and use σ_s as the best	$\bar{X} \pm t \cdot \frac{\sigma_s}{\sqrt{n}}$	$\bar{X} \pm t \cdot \frac{\sigma_s}{\sqrt{}} \times \sqrt{N-n}$
Estimating the population proportion $(\hat{p})$ when *p* is not	$p \pm z \cdot \sqrt{\frac{pq}{n}}$	$p \pm z \cdot \sqrt{\frac{pq}{n}} \times \sqrt{\frac{N-n}{N-1}}$

Hypothesis is usually considered as the principal instrument in research. Its main function is to suggest new experiments and observations. In fact, many experiments are carried out with the deliberate object of testing hypotheses. Decision-makers often face situations wherein they are interested in testing hypotheses on the basis of available information and then take decisions on thebasis of such testing. In social science, where direct knowledge of population parameter(s) is rare, hypothesis testing is the often used strategy for deciding whether a sample data offer such support for a hypothesis that generalisation can be made. Thus hypothesis testing enables us to make probability statements about population parameter(s). The hypothesis may not be proved absolutely, but in practiceit is accepted if it has withstood a critical testing. Before we explain how hypotheses are tested through different tests meant for the purpose, it will be appropriate to explain clearly the meaning of a hypothesis and the related concepts for better understanding of the hypothesis testing techniques.

Hypothesis Testing

WHAT IS A HYPOTHESIS?

Ordinarily, when one talks about hypothesis, one simply means a mere assumption or some suppositionto be proved or disproved. But for a researcher hypothesis is a formal question that he intends to resolve. Thus a hypothesis may be defined as a proposition or a set of proposition set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable. For example, consider statements like the following ones:

"Students who receive counselling will show a greater increase in creativity than students not receiving counselling" Or

"the automobile *A* is performing as well as automobile *B*."

These are hypotheses capable of being objectively verified and tested. Thus, we may conclude thata hypothesis states what we are looking for and it is a proposition which can be put to a test to determine its validity.

Characteristics of hypothesis: Hypothesis must possess the following characteristics:

(i) Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.

(ii) Hypothesis should be capable of being tested. In a swamp of untestable hypotheses, many a time the research programmes have bogged down. Some prior study may be done by researcher in order to make hypothesis a testable one. A hypothesis "is testable if other deductions can be made from it which, in turn, can be confirmed or disproved by observation."[1]

(iii) Hypothesis should state relationship between variables, if it happens to be a relational hypothesis.

(iv) Hypothesis should be limited in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and he should develop such hypotheses.

(v) Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesishas nothing to do with its significance.

(vi) Hypothesis should be consistent with most known facts i.e., it must be consistent with a substantial body of established facts. In other words, it should be one which judges acceptas being the most likely.

(vii) Hypothesis should be amenable to testing within a reasonable time. One should not use even an excellent hypothesis, if the same cannot be tested in reasonable time for one cannot spend a life-time collecting data to test it.

(viii) Hypothesis must explain the facts that gave rise to the need for explanation. This means that by using the hypothesis plus other known and accepted generalizations, one should beable to deduce the original problem condition. Thus hypothesis must actually explain what it claims to explain; it should have empirical reference.

BASIC CONCEPTS CONCERNING TESTING OF HYPOTHESES

Basic concepts in the context of testing of hypotheses need to be explained.

(a) *Null hypothesis and alternative hypothesis:* In the context of statistical analysis, we often talk about null hypothesis and alternative hypothesis. If we are to compare method *A* with method *B* about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the null hypothesis. As against this, we may think that the method *A* issuperior or the method *B* is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H_0 and the alternative hypothesis as H_a. Suppose we want

to test the hypothesis that the population mean (μ) is equal to the hypothesised mean $(\mu_{H_0}) = 100$.

If our sample results do not support this null hypothesis, we should conclude that something else is true. What we conclude rejecting the null hypothesis is known as alternative hypothesis. In other words, the set of alternatives to the null hypothesis is referred to as the alternative hypothesis. If weaccept H_0, then we are rejecting H_a and if we reject H_0, then we are accepting H_a. For

$H_0 : \mu = \mu_{H_0} = 100$, we may consider three possible alternative hypotheses as follows*:

Alternative hypothesis	*To be read as follows*
H_a : □ □ □ H 0	(The alternative hypothesis is that the population mean is not equal to 100 i.e., it may be more or less than 100)
H_a : □ □ □ H 0	(The alternative hypothesis is that the population mean is greater than 100)
H_a : □ □ □ H 0	(The alternative hypothesis is that the population mean is less than 100)

The null hypothesis and the alternative hypothesis are chosen before the sample is drawn (the researcher must avoid the error of deriving hypotheses from the data that he collects and then testing the hypotheses from the same data). In the choice of null hypothesis, the following considerations are usually kept in view:

(a) Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis is the one which one wishes to disprove. Thus, a null hypothesis represents the hypothesis we are trying to reject, and alternative hypothesis represents all other possibilities.

(b) If the rejection of a certain hypothesis when it is actually true involves great risk, it is taken as null hypothesis because then the probability of rejecting it when it is true is □ (the level of significance) which is chosen very small.

(c) Null hypothesis should always be specific hypothesis i.e., it should not state about or approximately a certain value.

Generally, in hypothesis testing we proceed on the basis of null hypothesis, keeping the alternative hypothesis in view. Why so? The answer is that on the assumption that null hypothesis is true, one can assign the probabilities to different possible sample results, but this cannot be done if we proceed with the alternative hypothesis. Hence the use of null hypothesis (at times also known as statistical hypothesis) is quite frequent.

(b) *The level of significance:* This is a very important concept in the context of hypothesis testing.It is always some percentage (usually 5%) which should be chosen wit great care, thought and reason. In case we take the significance level at 5 per cent, then this implies that H_0 will be rejected

when the sampling result (i.e., observed evidence) has a less than 0.05 probability of occurring if H_0 is true. In other words, the 5 per cent level of significance means that researcher is willing to take as much as a 5 per cent risk of rejecting the null hypothesis when it (H_0) happens to be true. Thus the significance level is the maximum value of the probability of rejecting H_0 when it is true and is usuallydetermined in advance before testing the hypothesis.

(c) *Decision rule or test of hypothesis:* Given a hypothesis H_0 and an alternative hypothesis H_a, we make a rule which is known as decision rule according to which we accept H_0 (i.e., reject H_a) orreject H_0 (i.e., accept H_a). For instance, if (H_0 is that a certain lot is good (there are very few defective items in it) against H_a) that the lot is not good (there are too many defective items in it), then we must decide the number of items to be tested and the criterion for accepting or rejecting thehypothesis. We might test 10 items in the lot and plan our decision saying that if there are none or only 1 defective item among the 10, we will accept H_0 otherwise we will reject H_0 (or accept H_a). This sort of basis is known as decision rule.

(d) *Type I and Type II errors:* In the context of testing of hypotheses, there are basically two typesof errors we can make. We may reject H_0 when H_0 is true and we may accept H_0 when in fact H_0 is not true. The former is known as Type I error and the latter as Type II error. In other words, Type Ierror means rejection of hypothesis which should have been accepted and Type II error means accepting the hypothesis which should have been rejected. Type I error is denoted by □ (alpha) known as □ error, also called the level of significance of test; and Type II error is denoted by □ (beta) known as □ error. In a tabular form the said two errors can be presented as follows:

	Decision	
	Accept H_0	Reject H_0
H_0 (true)	Correct decision	Type I error(□ error)
H_0 (false)	Type II error(□ error)	Correct decision

The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject H_0 when H_0 is true. We can control Type I error just by fixing it at a lower level. For instance, if we fix it at 1 per cent, we will say that the maximum probability of committing Type I error would only be 0.01.

But with a fixed sample size, *n*, when we try to reduce Type I error, the probability of committingType II error increases. Both types of errors cannot be reduced simultaneously. There is a trade-off between two types of errors which means that the probability of making one type of error can onlybe reduced if we are willing to increase the probability of making the other type of error. To deal with this trade-off in business situations, decision-makers decide the appropriate level of Type I error by examining the costs or penalties attached to both types of errors. If Type I error involves the time and trouble of reworking a batch of chemicals that should have been accepted, whereas Type II error means taking a chance that an entire group of users of this chemical compound will be poisoned, then

in such a situation one should prefer a Type I error to a Type II error. As a result one must set very high level for Type I error in one's testing technique of a given hypothesis.[2] Hence, in the testing ofhypothesis, one must make all possible effort to strike an adequate balance between Type I and TypeII errors.

(e) *Two-tailed and One-tailed tests:* In the context of hypothesis testing, these two terms are quite important and must be clearly understood. A two-tailed test rejects the null hypothesis if, say, the sample mean is significantly higher or lower than the hypothesized value of the mean of the population. Such a test is appropriate when the null hypothesis is some specified value and the alternative hypothesis is a value ^{0} not equal to the specified ^{0} value of the null hypothesis. Symbolically, the two-

tailed test is appropriate when we have $_0$ and

H_0 : □ □ □$_H$ H_a : □ □ □$_H$ which may mean □ □ □ H

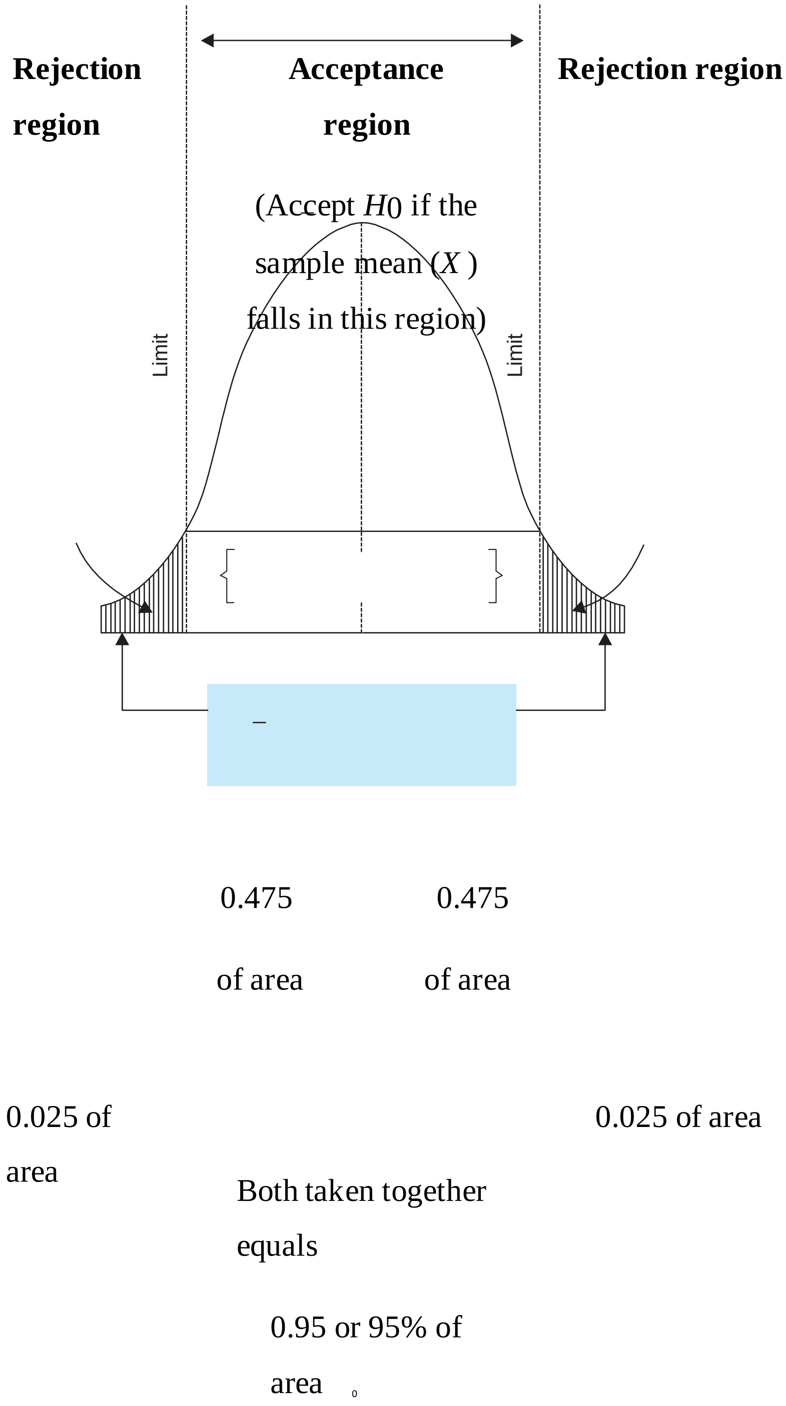
or □ □ $_0$.

□ H

Thus, in a two-tailed test, there are two rejection regions*, one on each tail of the curve

which can be illustrated as under:

Acceptance and rejection regions in case of a two-tailed test (with 5% significance level)


Rejection region
Acceptance region
Rejection region
(Accept H_0 if the sample mean ($\bar{X}$) falls in this region)
Limit
Limit
0.475 of area
0.475 of area
0.025 of area
0.025 of area
Both taken together equals 0.95 or 95% of area

$Z = –1.96$ Mathematically we can state:

$$\textit{Acceptance Region } A : |Z| \;\square\; 1.96$$

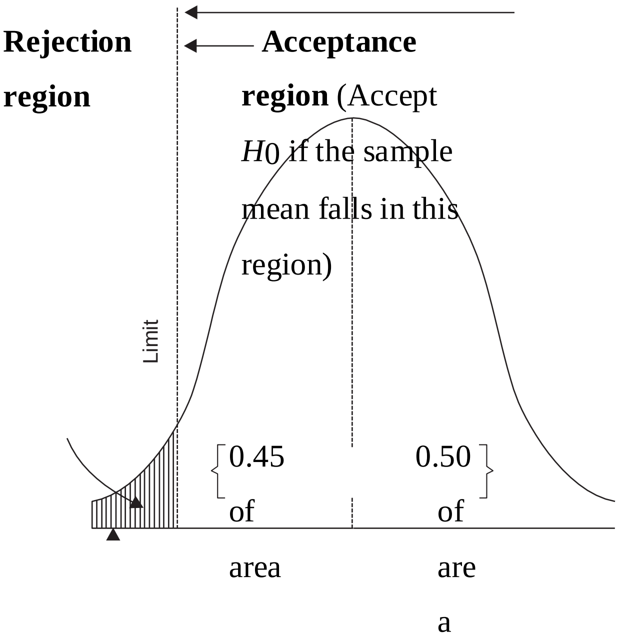
$$\textit{Rejection Region } R : |Z| > 1.96$$

If the significance level is 5 per cent and the two-tailed test is to be applied, the probability of the rejection area will be 0.05 (equally splitted on both tails of the curve as 0.025) and that of the acceptance region will be 0.95 as shown in the above curve. If we take □ □ 100 and if our sample mean deviates significantly from 100 in either direction, then we shall reject the null hypothesis; but if the sample mean does not deviate significantly from □ , in that case we shall accept the null

hypothesis.

But there are situations when only one-tailed test is considered appropriate. A *one-tailed test* would be used when we are to test, say, whether the population mean is either lower than or higher than some hypothesized value. For instance, if our H_0 : □ □ □$_H$ and H_a : □ □ □$_H$, then we are interested in what is known as left-tailed test (wherein there is one rejection region only on the lefttail) which can be illustrated as below:

Acceptance and rejection regions in case of one tailed test (left-tail) with 5% significance


Rejection region
Acceptance region (Accept *H*0 if the sample mean falls in this region)
Limit
0.45 of area
0.50 of area
0.05 of area

0.95 or 95% of area

Both taken together
equals

Z □= −1.645 ₀

Reject H_0 if the sample mean ($\bar{X}$) falls in this region

Mathematically we can
state:

$$\textit{Acceptance Region} \;\; A : Z \; □ \; □ \; 1.645$$

$$\textit{Rejection Region} \;\; R : Z \; □\_□ 1.645$$

Z = 1.96

If our □ □ 100 and if our sample mean deviates significantly from100 in the lower direction, we shall reject H_0, otherwise we shall accept H_0 at a certain level of significance. If the significance level in the given case is kept at 5%, then the rejection region will be equal to 0.05 of area in the left tail as has been shown in the above curve.

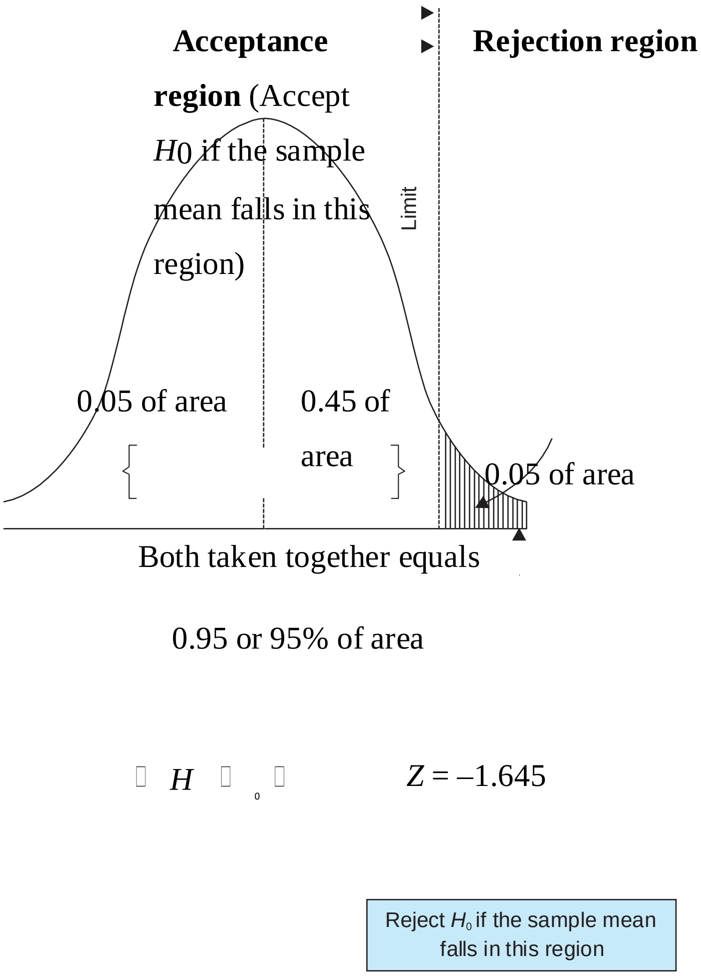
In case our H_0 : □ □ $_0$ and H_a : □ □ , we are then interested in what is known as one-

□ H □ H

0

tailed test (right tail) and the rejection region will be on the right tail of the curve as shown below:

Acceptance and rejection
regions in case of one-
tailed test (right tail) with
5% significance level


Acceptance region (Accept H_0 if the sample mean falls in this region)
Rejection region
Limit
0.05 of area
0.45 of area
0.05 of area
Both taken together equals 0.95 or 95% of area
H_0
$Z = -1.645$
Reject H_0 if the sample mean falls in this region

PROCEDURE FOR HYPOTHESIS TESTING

To test a hypothesis means to tell (on the basis of the data the researcher has collected) whether ornot the hypothesis seems to be valid. In hypothesis testing the main question is: whether to accept thenull hypothesis or not to accept the null hypothesis? Procedure for hypothesis testing refers to all those steps that we undertake for making a choice between the two actions i.e., rejection and acceptance of a null hypothesis. The various steps involved in hypothesis testing are stated below:

(i) *Making a formal statement:* The step consists in making a formal statement of the null hypothesis (H_0) and also of the alternative hypothesis (H_a). This means that hypotheses should be clearly stated, considering the nature of the research problem. For instance, Mr. Mohan of the Civil Engineering Department wants to test the load bearing capacity of an old bridge which must be more than 10 tons, in that case he can state his hypotheses as under:

Null hypothesis H_0 : □ □ 10 tons

Alternative Hypothesis H_a : □ □ 10 tons

Take another example. The average score in an aptitude test administered at the national level is 80.To evaluate a state's education system, the average score of 100 of the state's students selected onrandom basis was 75. The state wants to know if there is a significant difference between the local scores and the national scores. In such a situation the hypotheses may be stated as under:

Null hypothesis H_0 : □ □ 80

Alternative Hypothesis H_a : □ □ 80

The formulation of hypotheses is an important step which must be accomplished with due care inaccordance with the object and nature of the problem under consideration. It also indicates whether we should use a one-tailed test or a two-tailed test. If H_a is of the type greater than (or of the type lesser than), we use a one-tailed test, but when H_a is of the type "whether greater or smaller" then we use a two-tailed test.

(ii) *Selecting a significance level:* The hypotheses are tested on a pre-determined level of significance and as such the same should be specified. Generally, in practice, either 5% level or 1% level is adopted for the purpose. The factors that affect the level of significance are: (a) the magnitude of thedifference between sample means; (b) the size of the samples; (c) the variability of measurements within samples; and (d) whether the hypothesis is directional or non-directional (A directional hypothesis is one which predicts the direction of the difference between, say, means). In brief, the level of significance must be adequate in the context of the purpose and nature of enquiry.

(iii) *Deciding the distribution to use:* After deciding the level of significance, the next step in hypothesis testing is to determine the appropriate sampling distribution. The choice generally remains between normal distribution and the *t*-distribution. The rules for selecting the correct distribution are similar to those which we have stated earlier in the context of estimation.

(iv) *Selecting a random sample and computing an appropriate value:* Another step is to select a random sample(s) and compute an appropriate value from the sample data concerning the test statistic utilizing the relevant distribution. In other words, draw a sample to furnish empirical data.

(v) *Calculation of the probability:* One has then to calculate the probability that the sample result would diverge as widely as it has from expectations, if the null hypothesis were in fact true.

(vi) *Comparing the probability:* Yet another step consists in comparing the probability thus calculated with the specified value for □ , the significance level. If the calculated probability is equal to or smaller than the □ value in case of one-tailed test (and □ /2 in case of two-tailed test), then reject the null hypothesis (i.e., accept the alternative hypothesis), but if the calculated probability is greater, then accept the null hypothesis. In case we reject $H_{0,}$ we run a risk of (at most the level of significance) committing an error of Type I, but if we accept H_0, then we run some risk (the size of which cannot be specified as long as the H_0 happens to be vague rather than specific) of committing an error of Type II.

FLOW DIAGRAM FOR HYPOTHESIS TESTING

The above stated general procedure for hypothesis testing can also be depicted in the from of a flow-chart for better understanding as shown in below

FLOW DIAGRAM FOR HYPOTHESIS TESTING

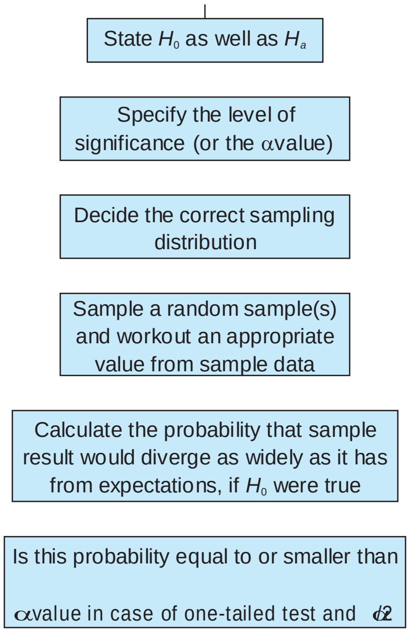

State H_0 as well as H_a
Specify the level of significance (or the α value)
Decide the correct sampling distribution
Sample a random sample(s) and workout an appropriate value from sample data
Calculate the probability that sample result would diverge as widely as it has from expectations, if H_0 were true
Is this probability equal to or smaller than α value in case of one-tailed test and α2

Illustration 1

A certain chemical process is said to have produced 15 or less pounds of waste material for every 60 lbs. batch with a corresponding standard deviation of 5 lbs. A random sample of 100 batches gives an average of 16 lbs. of waste per batch. Test at 10 per cent level whether the average quantity of waste per batch has increased. Compute the power of the test for □ = 16 lbs. If we raise the level of significance to 20 per cent, then how the power of the test for □ = 16 lbs. would be affected?

Solution: As we want to test the hypothesis that the average quantity of waste per batch of 60 lbs. is 15 or less pounds against the hypothesis that the waste quantity is more than 15 lbs., we can writeas under:

$$H_0 : □ \; \bar{□} \; 15 \text{ lbs.}$$

$$H_a : □ \; □ \; 15 \text{ lbs.}$$

As H_a is one-sided, we shall use the one-tailed test (in the right tail because H_a is of more than type) at 10% level for finding the value of standard deviate (z), corresponding to .4000 area of normal curve which comes to 1.28 as per normal curve area table.* From this we can find the limit of □ for accepting H_0 as under:

Accept H_0 if $\bar{X} \square 15 \square 1.28 (\square_p / \sqrt{n})$

or $\bar{X} \square 15 \square 1.28 \ (5/\sqrt{100})$

or $\bar{X} \square 15.64$

at 10% level of significance otherwise accept H_a.

But the sample average is 16 lbs. which does not come in the acceptance region as above. We, therefore, reject H_0 and conclude that average quantity of waste per batch has increased. For finding the power of the test, we first calculate □ and then subtract it from one. Since □ is a conditional probability which depends on the value of $\bar{\square}$, we take it as 16 as given in the question. We can now

write □ = p (Accept H_0 : □ □ 15 □ □ 16) . Since we have already worked out that H_0 is accepted if $\bar{X} \square 15.64$

(at 10% level of significance), therefore □ □ p ($\bar{X}$

HYPOTHESIS TESTING OF MEANS

Mean of the population can be tested presuming different situations such as the population may be normal or other than normal, it may be finite or infinite, sample size may be large or small, variance of the population may be known or unknown and the alternative hypothesis may be two-sided or one- sided. Our testing technique will differ in different situations. We may consider some of the important situations.

1. *Population normal, population infinite, sample size may be large or small but variance of the population is known, H_a may be one-sided or two-sided*:

   In such a situation *z*-test is used for testing hypothesis of mean and the test statistic *z* isworked our as under:

$$z \square \frac{\bar{X} \square \square_{H_0}}{\sigma_p \,\big|\, \sqrt{n}}$$

2. *Population normal, population finite, sample size may be large or small but varianceof the population is known, H_a may be one-sided or two-sided*:

   In such a situation *z*-test is used and the test statistic *z* is worked out as under (usingfinite population multiplier):

$$z = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_p / \sqrt{n}\right) \times \left[\sqrt{\left(N-n\right) / \left(N-1\right)}\right]}$$

3. *Population normal, population infinite, sample size small and variance of thepopulation unknown, H_a may be one-sided or two-sided:*

   In such a situation *t*-test is used and the test statistic *t* is worked out as under:

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s / \sqrt{n}} \quad \text{with d.f.} = (n - 1)$$

and $$\sigma_s = \sqrt{\frac{\sum \left(X_i - \bar{X}\right)^2}{\left(n-1\right)}}$$

4. *Population normal, population finite, sample size small and variance of the populationunknown, and H_a may be one-sided or two-sided:*

   In such a situation *t*-test is used and the test statistic '*t*' is worked out as under (usingfinite population multiplier):

$$t = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_s / \sqrt{n}\right) \times \sqrt{\left(N-n\right) / \left(N-1\right)}} \quad \text{with d.f.} = (n - 1)$$

15.64 16)

TESTING OF HYPOTHESIS

Test of Significance for Large Samples

The test of significance for the large samples can be explained by the following assumptions:

1. The random sampling distribution of statistics is approximately normal.

2. Sampling values are sufficiently close to the population value and can be used for the calculation of standard error of estimate.

The Standard Error Of Mean.

In the case of large samples, when we are testing the significance of statistic, the concept of standard error is used. It measures only sampling errors. Sampling errors are involved in estimating a population parameter from a sample, instead of including all the essential information in the population.

when standard deviation of the population is known, the formula is

$$S.E.\ \overline{X} = \frac{\sigma^{P}}{\sqrt{n}}$$

Where,

$$S.E.\ \overline{X} =$$

The standard error of the mean, σ_p = Standard deviation of the population, and n = Number of observations in the sample.

(ii) when standard deviation of population is not known, we have to use the standard deviation of the sample in calculating standard error of mean.

The formula is

$$S.E.\ \overline{X} = \frac{\sigma\ (\text{sample})}{\sqrt{n}}$$

Where,

σ = Standard deviation of the sample, and n = Sample size

Illustration:1

Let us take the hypothesis that there is no significant difference between the sample mean and the hypothetical population mean.

$$\text{S.E. } \bar{X} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = \frac{20}{10} = 2$$

$$\frac{\text{Difference}}{\text{S.E.}\bar{X}} = \frac{125 - 116}{2} = \frac{9}{2} = 4.5$$

Since, the difference is more than 2.58 S.E.(1% level) it could not have arisen due to fluctuations of sampling. Hence the mean weight of students in the population could not be 125 lbs.

Test Of Significance For Small Samples

If the sample size is less than 30, then those samples may be regarded as small samples. As a rule, the methods and the theory of large samples are not applicable to the small samples. The small samples are used in testing a given hypothesis, to find out the observed values, which could have arisen by sampling fluctuations from some values given in advance. In a small sample, the investigator's estimate will vary widely from sample to sample. An inference drawn from a smaller sample result is less precise than the inference drawn from a large sample result.

t-distribution will be employed, when the sample size is 30 or less and the population standard deviation is unknown.

The formula is

$$t = \frac{(\bar{X} - \mu)}{\sigma} \times \sqrt{n}$$

where,

$$\sigma = \sqrt{\Sigma(X - X)^2/n - 1}$$

Illustration:2

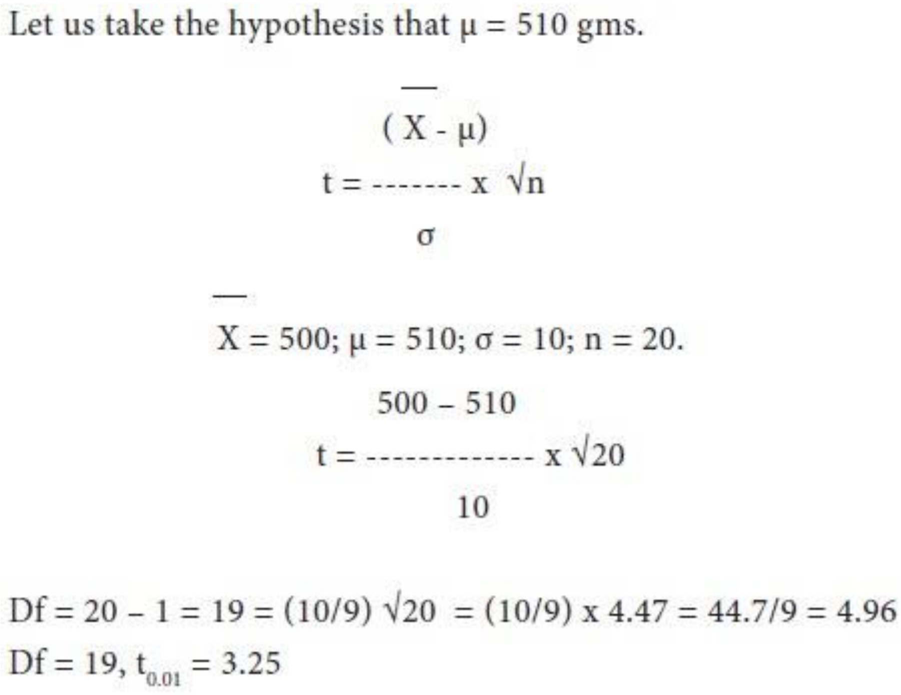
The following results are obtained from a sample of 20 boxes of mangoes:

Mean weight of contents = 490gms,

Standard deviation of the weight = 9 gms.

Could the sample come from a population having a mean of 500 gms?

Solution:



Let us take the hypothesis that μ = 510 gms.

$$t = \frac{(\bar{X} - \mu)}{\sigma} \times \sqrt{n}$$

$\bar{X} = 500$; $\mu = 510$; $\sigma = 10$; $n = 20$.

$$t = \frac{500 - 510}{10} \times \sqrt{20}$$

Df = 20 – 1 = 19 = (10/9) √20 = (10/9) x 4.47 = 44.7/9 = 4.96

Df = 19, $t_{0.01}$ = 3.25

What is predictive modeling?

Predictive modeling is a mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data. It is a crucial component of predictive analytics, a type of data analytics which uses current and historical data to forecast activity, behavior and trends.

Examples of predictive modeling include estimating the quality of a sales lead, the likelihood of spam or the probability someone will click a link or buy a product. These capabilities are often baked into various business applications, so it is worth understanding the mechanics of predictive modeling to troubleshoot and improve performance.

Although predictive modeling implies a focus on forecasting the future, it can also predict outcomes (e.g., the probability a transaction is fraudulent). In this case, the event has already happened (fraud committed). The goal here is to predict whether future analysis will find the transaction *is* fraudulent. Predictive modeling can also forecast future requirements or facilitate what-if analysis.

"Predictive modeling is a form of data mining that analyzes historical data with the goal of identifying trends or patterns and then using those insights to predict future outcomes," explained Donncha Carroll a partner in the revenue growth practice of Axiom Consulting Partners. "Essentially, it asks the question, 'have I seen this before' followed by, 'what typically comes after this pattern.'"

Fact-Based Management

Fact-Based Management or Evidence-Based Management gained popularity during the 90's and 00's, but seems to have been put aside for more humanities-inspired leadership theories, which certainly have merit, if they are build on top of a good foundation of execution.

Fact-Based Management within ordinary business operation demands a little pragmatic adjustment, but it is a necessary and sufficient prerequisite for understanding and acting as management in a company.

In recent years there has been a strange focus on Management vs. Leadership, which in it's outset is a fine idea, however also one that is based on a flawed foundation.

Management is the task of managing the business within the boundaries set for the operation, whereas Leadership entails developing the business through moving the boundaries.

So the difference between Management and Leadership has nothing to do with the humane, inspiring and empathetic Leader vs. the mechanistic, tough and brutal Manager, which is the way most feel-good self-proclaimed management gurus present it.

There is nothing that stands in the way of a humane and empathetic approach to being a manager - in fact that is most often the most efficient way of managing people.

Leadership and Management are thus 2 equally important aspects of managing a company - without a proper executed operation, everything falls apart, including the Leadership!

Fact-Based Management is utilizing all available information and data for decision-making in the company - whether it be in operations, in business development or devising a new strategy.

These facts can be more or less accurate, but should always be logically grounded in evidence, one way or another, and should be sorted in a hierarchy, so that the most valid facts are valued most.

From time to time there will be no real facts to base a decision on, and here decisions must be based on gut-feeling, sporadic indications and logic in combination with common sense. It is not bad decision-making to base decisions on an incomplete basis, and sometimes it can also take too long to gather and analyse all available data. However in most circumstances, in small and medium sized businesses, there will be a lot of facts to sweep and analyze, and most decisions are not so urgent that it is not worthwhile to do the math...

Fact-Based Management is a practice which requires the embrace of all managers in the organization to get the full benefit released. That does not imply that all managers can be equally proficient in practising Fact-Based Management, however the aspiration should be the same for all levels of management.

It can be a hard process to change a culture to accept Fact-Based Management, as it requires hard work and focus to find and apply the relevant facts to decision-making. Regardless of the right well implemented business intelligence and other fact-gathering and -organizing tools, it requires an energy consuming mental process to gather, sort and analyse available data, which we all as human beings try to avoid, if we can. This makes it paramount to keep the reins tight and secure follow-through in actions - otherwise the organization takes the easy way out, until they discover the benefits.

When a management only uses values, attitude, gut-feeling and non-verbalized experience, the company can do well if the company is healthy and the business model is sound. So why bother with the high-maintenance Fact-Based Management, demanding hard work and focus?

The essence is that, beyond the responsibility for maximizing value for the shareholders, the management has a responsibility for the company in it self - and what is a company first and foremost; a group of people working together. People.

When crisis sets in, then it matters how efficient and well managed the company is. At that moment it pays off to have a well oiled and efficient motor moving the operation of the business, which will perform better or even optimal. In this situation Leadership is crucial and without the stable and sound basis of the Fact-Based Management and an organization embracing it, the Leadership is doomed. It is in the time of crisis that Leadership shall prove its worth and bring the company back on track - and without Fact-Based Management, success will be a mere fluke and not the result of capable management - or leadership for that matter.

Types of predictive models

There are many ways of classifying predictive models and in practice multiple types of models may be combined for best results. The most salient distinction is between unsupervised versus supervised models.

- Unsupervised models use traditional statistics to classify the data directly, using techniques like logistic regression, time series analysis and decision trees.
- Supervised models use newer machine learning techniques such as neural networks to identify patterns buried in data that has already been labeled.

The biggest difference between these approaches is that with supervised models more care must be taken to properly label data sets upfront.

"The application of different types of models tends to be more domain-specific than industry-specific," said Scott Buchholz, government and public services CTO and emerging technology research director at Deloitte Consulting.

In certain cases, for example, standard statistical regression analysis may provide the best predictive power. In other cases, more sophisticated models are the right approach. For example, in a hospital, classic statistical techniques may be enough to identify key constraints for scheduling, but neural networks, a type of deep learning, may be required to optimize patient assignment to doctors.

Once data scientists gather this sample data, they must select the right model. Linear regressions are among the simplest types of predictive models. Linear models take two variables that are correlated -- one independent and the other dependent -- and plot one on the x-axis and one on the y-axis. The model applies a best fit line to the resulting data points. Data scientists can use this to predict future occurrences of the dependent variable.

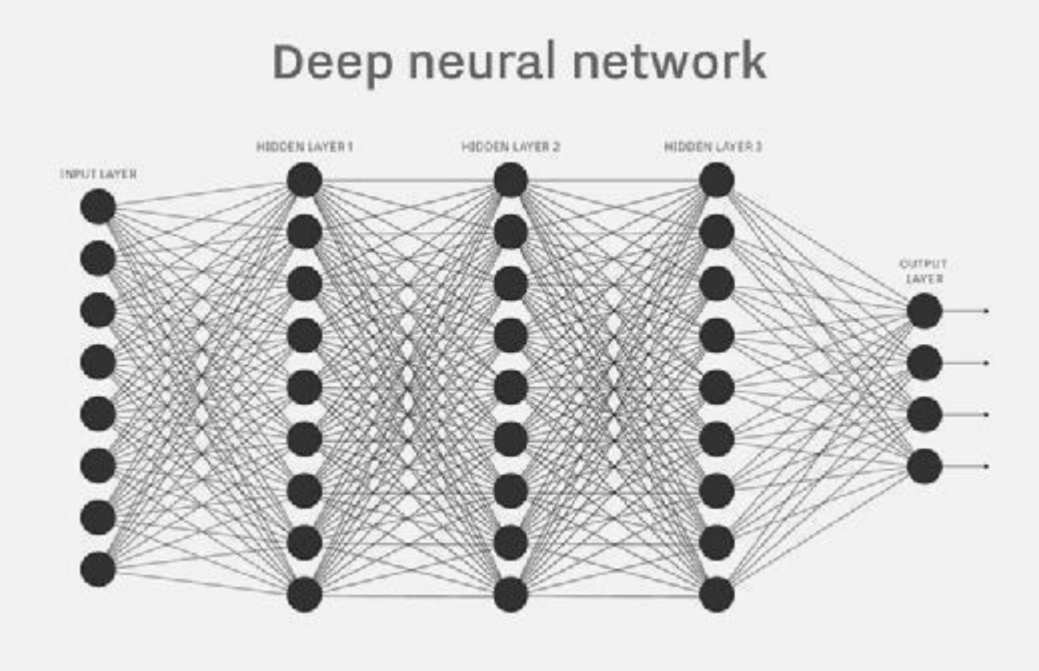
Some of the most popular methods include the following:

- **Decision trees.** Decision tree algorithms take data (mined, open source, internal) and graph it out in branches to display the possible outcomes of various decisions. Decision trees classify response variables and predict response variables based on past decisions, can be used with incomplete data sets and are easily explainable and accessible for novice data scientists.
- **Time series analysis.** This is a technique for the prediction of events through a sequence of time. You can predict future events by analyzing past trends and extrapolating from there.
- **Logistic regression.** This method is a statistical analysis method that aids in data preparation. As more data is brought in, the algorithm's ability to sort and classify it improves and therefore predictions can be made.
- **Neural networks.** This technique reviews large volumes of labeled data in search of correlations between variables in the data. Neural networks form the basis of many of today's examples of artificial intelligence (AI), including image recognition, smart assistants and natural language generation.

The most complex area of predictive modeling is the neural network. This type of machine learning model independently reviews large volumes of labeled data in search of correlations between variables in the data. It can detect even subtle correlations that only emerge after reviewing millions of data points. The algorithm can then make inferences about unlabeled data files that are similar in type to the data set it trained on.

Common algorithms for predictive modeling

- **Random Forest.** This algorithm combines unrelated decision trees and uses classification and regression to organize and label vast amounts of data.
- **Gradient boosted model.** Similar to Random Forest, this algorithm uses several decision trees, but in this method, each tree corrects the flaws of the previous one and builds a more accurate picture.
- **K-Means.** This algorithm groups data points in a similar fashion as clustering models and is popular in devising personalized retail offers. It create personalized offers by seeking out similarities among large groups of customers.
- **Prophet.** A forecasting procedure, this algorithm is especially effective when dealing with capacity planning. This algorithm deals with time series data and is relatively flexible.



A neural network is a type of predictive model that independently reviews large volumes of labeled data in search of correlations between variables in the data.

What are the uses of predictive modeling?

Predictive modeling is often associated with meteorology and weather forecasting, but predictive models have many applications in business. Today's predictive analytics techniques can discover patterns in the data to identify upcoming risks and opportunities for an organization.

"Almost anywhere a smart human is regularly making a prediction in a historically data rich environment is a good use case for predicative analytics," Buchholz said. "After all, the model has no ego and won't get bored."

One of the most common uses of predictive modeling is in online advertising and marketing. Modelers use web surfers' historical data, to determine what kinds of products users might be interested in and what they are likely to click on.

Bayesian spam filters use predictive modeling to identify the probability that a given message is spam.

In fraud detection, predictive modeling is used to identify outliers in a data set that point toward fraudulent activity. In customer relationship management, predictive modeling is used to target messaging to customers who are most likely to make a purchase.

Carroll said that predictive modeling is widely used in predictive maintenance, which has become a huge industry generating billions of dollars in revenue. One of the more notable examples can be found in the airline industry where engineers use IoT devices to remotely monitor performance of aircraft components like fuel pumps or jet engines.

These tools enable preemptive deployment of maintenance resources to increase equipment utilization and limit unexpected downtime. "These actions can meaningfully improve operational efficiency in a world that runs just in time where surprises can be very expensive," Caroll said.

Other areas where predictive models are used include the following:

- capacity planning
- change management
- disaster recovery
- engineering
- physical and digital security management
- city planning

How to build a predictive model

Building a predictive model starts with identifying historical data that's representative of the outcome you are trying to predict.

"The model can infer outcomes from historical data but cannot predict what it has never seen before," Carroll said. Therefore, the volume and breadth of information used to train the model is critical to securing an accurate prediction for the future.

The next step is to identify ways to clean, transform and combine the raw data that leads to better predictions.

Skill is required in not only finding the appropriate set of raw data but also transforming it into data features that are most appropriate for a given model. For example, calculations of time-boxed weekly averages may be more useful and lead to better algorithms than real-time levels.

It is also important to weed out data that is coincidental or not relevant to a model. At best, the additional data will slow the model down, and at worst, it will lead to less accurate models.

This is both an art and a science. The art lies in cultivating a gut feeling for the meaning of things and intuiting the underlying causes. The science lies in methodically applying algorithms to consistently achieve reliable results, and then evaluating these algorithms over time. Just because a spam filter works on day one does not mean marketers will not tune their messages, making the filter less effective.

Analyzing representative portions of the available information -- sampling -- can help speed development time on models and enable them to be deployed more quickly.

Benefits of predictive modeling

Phil Cooper, group VP of products at Clari, a RevOps software startup, said some of the top benefits of predictive modeling in business include the following:

- **Prioritizing resources.** Predictive modeling is used to identify sales lead conversion and send the best leads to inside sales teams; predict whether a customer service case will be escalated and triage and route it appropriately; and predict whether a customer will pay their invoice on time and optimize accounts receivable workflows.
- **Improving profit margins.** Predictive modeling is used to forecast inventory, create pricing strategies, predict the number of customers and configure store layouts to maximize sales.
- **Optimizing marketing campaigns.** Predictive modeling is used to unearth new customer insights and predict behaviors based on inputs, allowing organizations to tailor marketing strategies, retain valuable customers and take advantage of cross-sell opportunities.
- **Reducing risk.** Predictive analytics can detect activities that are out of the ordinary such as fraudulent transactions, corporate spying or cyber attacks to reduce reaction time and negative consequences.

The techniques used in predictive modeling are probabilistic as opposed to deterministic. This means models generate probabilities of an outcome and include some uncertainty.

"This is a fundamental and inherent difference between data modeling of historical facts versus predicting future events [based on historical data] and has implications for how this information is communicated to users," Cooper said. Understanding this difference is a critical necessity for transparency and explainability in how a prediction or recommendation was generated.

Challenges of predictive modeling

Here are some of the challenges related to predictive modeling.

Data preparation. One of the most frequently overlooked challenges of predictive modeling is acquiring the correct amount of data and sorting out the right data to use when developing algorithms. By some estimates, data scientists spend about 80% of their time on this step. Data collection is important but limited in usefulness if this data is not properly managed and cleaned.

Once the data has been sorted, organizations must be careful to avoid overfitting. Over-testing on training data can result in a model that appears very accurate but has memorized the key points in the data set rather than learned how to generalize.

Technical and cultural barriers. While predictive modeling is often considered to be primarily a mathematical problem, users must plan for the technical and organizational barriers that might prevent them from getting the data they need. Often, systems that store useful data are not connected directly to centralized data warehouses. Also, some lines of business may feel that the data they manage is their asset, and they may not share it freely with data science teams.

Choosing the right business case. Another potential obstacle for predictive modeling initiatives is making sure projects address significant business challenges. Sometimes, data scientists discover correlations that seem interesting at the time and build algorithms to investigate the correlation further. However, just because they find something that is statistically significant does not mean it presents an insight the business can use. Predictive modeling initiatives need to have a solid foundation of business relevance.

Bias. "One of the more pressing problems everyone is talking about, but few have addressed effectively, is the challenge of bias," Carroll said. Bias is naturally introduced into the system through historical data since past outcomes reflect existing bias.

Nate Nichols, distinguished principal at Narrative Science, a natural language generation tools provider, is excited about the role that new explainable machine learning methods such as LIME or SHAP could play in addressing concerns about bias and promoting trust.

"People trust models more when they have some understanding of what the models are doing, and trust is paramount for predictive analytic capabilities," Nichols said. Being able to provide explanations for the predictions, he said, is a huge positive differentiator in the increasingly crowded field of predictive analytic products.

Predictive modeling versus predictive analytics

Predictive modeling is but one aspect in the larger predictive analytics process cycle. This includes collecting, transforming, cleaning and modeling data using independent variables, and then reiterating if the model does not quite fit the problem to be addressed.

"Once data has been gathered, transformed and cleansed, then predictive modeling is performed on the data," said Terri Sage, chief technology officer at 1010data, an analytics consultancy.

Collecting data, transforming and cleaning are processes used for other types of analytic development.

"The difference with predictive analytics is the inclusion and discarding of variables during the iterative modeling process," Sage explained.

This will differ across various industries and use cases, as there will be diverse data used and different variables discovered during the modeling iterations.

For example, in healthcare, predictive models may ingest a tremendous amount of data pertaining to a patient and forecast a patient's response to certain treatments and prognosis. Data may include the patient's specific medical history, environment, social risk factors, genetics -- all which vary from person to person. The use of predictive modeling in healthcare marks a shift from treating patients based on averages to treating patients as individuals.

Similarly, with marketing analytics, predictive models might use data sets based on a consumer's salary, spending habits and demographics. Different data and modeling will be used for banking and insurance to help determine credit ratings and identify fraudulent activities.

Predictive modeling tools

Before deploying a predictive model tool, it is crucial for your organization to ask questions and sort out the following: Clarify who will be running the software, what the use case will be for these tools, what other tools will your predictive analytics be interacting with, as well as the budget.

Different tools have different data literacy requirements, are effective in different use cases, are best used with similar software and can be expensive. Once your organization has clarity on these issues, comparing tools becomes easier.

- **Sisense.** A business intelligence software aimed at a variety of companies that offers a range of business analytics features. This requires minimal IT background.
- **Oracle Crystal Ball.** A spread sheet-based application focused on engineers, strategic planners and scientists across industries that can be used for predictive modeling, forecasting as well as simulation and optimization.
- **IBM SPSS Predictive Analytics Enterprise.** A business intelligence platform that supports open source integration and features descriptive and predictive analysis as well as data preparation.
- **SAS Advanced Analytics.** A program that offers algorithms that identify the likelihood of future outcomes and can be used for data mining, forecasting and econometrics.

The future of predictive modeling

There are three key trends that will drive the future of data modeling.

1. First, data modeling capabilities are being baked into more business applications and citizen data science tools. These capabilities can provide the appropriate guardrails and templates for business users to work with predictive modeling.
2. Second, the tools and frameworks for low-code predictive modeling are making it easier for data science experts to quickly cleanse data, create models and vet the results.
3. Third, better tools are coming to automate many of the data engineering tasks required to push predictive models into production. Carroll predicts this will allow more organizations to shift from simply building models to deploying them in ways that deliver on their potential value.

Analytics vs reporting

Analytics is the technique of examining data and reports to obtain actionable insights that can be used to comprehend and improve business performance. Business users may gain insights from data, recognize trends, and make better decisions with analytics.

On the one hand, analytics is about finding value or making new data to help you decide. This can be performed either manually or mechanically. Next-generation analytics uses new technologies like AI or machine learning to make predictions about the future based on past and present data.

Analytics vs reporting: Key Differences & Importance

Analytics and reporting can help a business improve operational efficiency and production in several ways. Analytics is the process of making decisions based on the data presented, while reporting is used to make complicated information easier to understand.

Analytics and reporting are often referred to as the same. Although both take in data as input and present it in charts, graphs, or dashboards, they have several key differences. This post will cover analytics and reporting, key differences, and its importance in business.

The steps involved in data analytics are as follows:

- Developing a data hypothesis
- Data collection and transformation
- Creating analytical models to analyze and provide insights
- Utilization of data visualization, trend analysis, deep dives, and other tools.
- Making decisions based on data and insights

On the other hand, reporting is the process of presenting data from numerous sources clearly and simply. The procedure is always carefully set out to report correct data and avoid misunderstandings.

Today's reporting applications offer cutting-edge dashboards with advanced data visualization features. Companies produce a variety of reports, such as financial reports, accounting reports, operational reports, market studies, and more. This makes it easier to see how each function is operating quickly.

In general, the procedures needed to create a report are as follows:

- Determining the business requirement
- Obtaining and compiling essential data
- Technical data translation
- Recognizing the data context
- Building dashboards for reporting
- Providing real-time reporting
- Allowing users to dive down into reports

Key differences between analytics vs reporting

Differences between analytics and reporting can significantly benefit your business. If you want to use both to their full potential and not miss out on essential parts of either one knowing the difference between the two is important. Some key differences are:

Analytics	**Reporting**
Analytics is the method of examining and analysing summarized data to make business decisions.	Reporting is an action that includes all the needed information and data and is put together in an organized way.
Questioning the data, understanding it, investigating it, and presenting it to the end users are all part of analytics.	Identifying business events, gathering the required information, organizing, summarizing, and presenting existing data are all part of reporting.
The purpose of analytics is to draw conclusions based on data.	The purpose of reporting is to organize the data into meaningful information.
Analytics is used by data analysts, scientists, and business people to make effective decisions.	Reporting is provided to the appropriate business leaders to perform effectively and efficiently within a firm.

Analytics and reporting can be used to reach a number of different goals. Both of these can be very helpful to a business if they are used correctly.

Importance of analytics vs reporting

A business needs to understand the differences between analytics and reporting. Better data knowledge through analytics and reporting helps businesses in decision-making and action inside the organization. It results in higher value and performance. Analytics is not really possible without reporting, but analytics is more than just reporting. Both tools are made for sharing important information that will help business people make better decisions

Transforming data into insights

Analytics assists businesses in converting information into insights, whereas reporting transforms data into information. Analytics aims to take the data and figure out what it means.

Analytics examines report data to determine why and how to fix organizational problems. Analysts begin by asking questions that may arise as they examine how the data in the reports has been structured. A qualified analyst can make recommendations to improve business performance once the data analysis is complete.

Analytics and reporting go hand in hand, and you can’t have one without the other. The raw data are the first step in the whole process. The data then needs to be put together to make it look like accurate information. Reports can be comprehensive and employ a range of technologies. Still, their main objective is always to make it simpler for analysts to understand what is actually happening within the organization.

Reporting and analytics have distinct differences. Reporting focuses on arranging and presenting facts, while analytics provides actionable insights. However, both are important and connected. Your implementation plans will stay on track if everyone on your team agrees on what they mean when they talk about analytics or reporting.

Organizations all around the world are utilizing knowledge management systems and solutions such as Insights Hub to manage data better, reduce the time it takes to obtain insights, and increase the utilization of historical data while cutting costs and increasing ROI.

Analysis vs reporting

Reporting:

- Once data is collected, it will be organized using tools such as graphs and tables.
- The process of organizing this data is called reporting.
- Reporting translates raw data into information.
- Reporting helps companies to monitor their online business and be alerted when data falls outside of expected ranges.
- Good reporting should raise questions about the business from its end users.

Analysis:

- Analytics is the process of taking the organized data and analyzing it.
- This helps users to gain valuable insights on how businesses can improve their performance.

- Analysis transforms data and information into insights.
- The goal of the analysis is to answer questions by interpreting the data at a deeper level and providing actionable recommendations.

UNIT – III – ONE - SAMPLE TESTS

One-Sample t-test

Test for population mean

Hypothesis test

Formula: $$t = \frac{\bar{x} - \Delta}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is the sample mean, Δ is a specified value to be tested, *s* is the sample standard deviation, and *n* is the size of the sample. Look up the significance level of the *z*-value in the standard normal table (Table 2 in "Statistics Tables").

When the standard deviation of the sample is substituted for the standard deviation of the population, the statistic does not have a normal distribution; it has what is called the *t*-distribution(see Table 3 in "Statistics Tables"). Because there is a different *t*-distribution for each sample size, it is not practical to list a separate area-of -the-curve table for each one. Instead, critical *t*-values for common alpha levels (0.10, 0.05, 0.01, and so forth) are usually given in a single table for a range of sample sizes. For very large samples, the *t*-distribution approximates the standard normal (*z*) distribution. In practice, it is best to use *t*-distributions any time the population standard deviation is not known.

Values in the *t*-table are not actually listed by sample size but by degrees of freedom *(df)*. The number of degrees of freedom for a problem involving the *t*-distribution for sample size *n* is simply *n* – 1 for a one-sample mean problem.

A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor have 90 percent confidence that the mean score for the class on the test would be above 70?

null hypothesis: H_0: μ = 70

alternative hypothesis: H_a : μ > 70

First, compute the sample mean and standard deviation:

$$\begin{array}{r} 62 \\ 92 \\ 75 \\ 68 \\ 83 \\ +\ 95 \\ \hline 475 \end{array} \qquad \begin{array}{l} \bar{x} = \frac{475}{6} = 79.17 \\ s = 13.17 \end{array}$$

Next, compute the *t*-value:

$$t = \frac{79.17 - 70}{\frac{13.17}{\sqrt{6}}} = \frac{9.17}{5.38} = 1.71$$

To test the hypothesis, the computed *t*-value of 1.71 will be compared to the critical value in the *t*-table. But which do you expect to be larger and which do you expect to be smaller? One way to reason about this is to look at the formula and see what effect different means would have on the computation. If the sample mean had been 85 instead of 79.17, the resulting *t*-value would have been larger. Because the sample mean is in the numerator, the larger it is, the larger the resulting figure will be. At the same time, you know that a higher sample mean will make it more likely that the professor will conclude that the math proficiency of the class is satisfactory and that the null hypothesis of less-than-satisfactory class math knowledge can be rejected. Therefore, it must be true that the larger the computed *t*-value, the greater the chance that the null hypothesis can be rejected. It follows, then, that if the computed *t*-value is larger than the critical *t*-value from the table, the null hypothesis can be rejected.

A 90 percent confidence level is equivalent to an alpha level of 0.10. Because extreme values in one rather than two directions will lead to rejection of the null hypothesis, this is a one-tailed test, and you do not divide the alpha level by 2. The number of degrees of freedom for the problem is 6 – 1 = 5. The value in the *t*-table for $t_{.10,5}$ is 1.476. Because the computed *t*-value of 1.71 is larger than the critical value in the table, the null hypothesis can be rejected, and the professor has evidence that the class mean on the math test would be at least 70.

Note that the formula for the one-sample *t*-test for a population mean is the same as the *z*-test, except that the *t*-test substitutes the sample standard deviation *s* for the population standard deviation σ and takes critical values from the *t*-distribution instead of the *z*-distribution. The *t*-distribution is particularly useful for tests with small samples ($n < 30$).

A Little League baseball coach wants to know if his team is representative of other teams in scoring runs. Nationally, the average number of runs scored by a Little League team in a game is 5.7. He chooses five games at random in which his team scored 5 , 9, 4, 11, and 8 runs. Is it likely that his team's scores could have come from the national distribution? Assume an alpha level of 0.05.

Because the team's scoring rate could be either higher than or lower than the national average, the problem calls for a two-tailed test. First, state the null and alternative hypotheses:

null hypothesis: H_0: $\mu = 5.7$

alternative hypothesis: H_a : $\mu \neq 5.7$

Next compute the sample mean and standard deviation:

$$\begin{array}{r} 5 \\ 9 \\ 4 \\ 11 \\ \underline{+\ 8} \\ 37 \end{array} \qquad \begin{array}{l} \bar{x}=\frac{37}{5}=7.4 \\ s=2.88 \end{array}$$

Next, the *t*-value:

$$t=\frac{7.4-5.7}{\frac{2.88}{\sqrt{5}}}=\frac{1.7}{1.29}=1.32$$

Now, look up the critical value from the *t*-table(Table 3 in "Statistics Tables"). You need to know two things in order to do this: the degrees of freedom and the desired alpha level. The degrees of freedom is 5 – 1 = 4. The overall alpha level is 0.05, but because this is a two-tailed test, the alpha level must be divided by two, which yields 0.025. The tabled value for $t_{.025,4}$is 2.776. The computed *t* of 1.32 is smaller, so you cannot reject the null hypothesis that the mean of this team is equal to the population mean. The coach cannot conclude that his team is different from the national distribution on runs scored.

Formula: $(a,b)=\bar{x}\pm t_{\alpha/2,df}\cdot\frac{s}{\sqrt{n}}$

where *a* and *b* are the limits of the confidence interval, $\bar{x}$ is the sample mean, $t_{\alpha/2,df}$ is the value from the *t*-table corresponding to half of the desired alpha level at *n* – 1 degrees of freedom, *s* is the sample standard deviation, and *n* is the size of the sample.

Using the previous example, what is a 95 percent confidence interval for runs scored per team per game?

First, determine the *t*-value. A 95 percent confidence level is equivalent to an alpha level of 0.05. Half of 0.05 is 0.025. The *t*-value corresponding to an area of 0.025 at either end of the *t*-distribution for 4 degrees of freedom ($t_{.025,4}$) is 2.776. The interval may now be calculated:

$$\begin{aligned}(a,b)&=7.4\pm2.78\frac{2.88}{\sqrt{5}}\\&=7.4\pm3.58\\&=(3.82,10.98)\end{aligned}$$

The interval is fairly wide, mostly because *n* is small.

One-sample *t*-test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

What if my data isn't nearly normally distributed?

If your sample sizes are very small, you might not be able to test for normality. You might need to rely on your understanding of the data. When you cannot safely assume normality, you can perform a *nonparametric* test that doesn't assume normality.

Using the one-sample *t*-test

The sections below discuss what we need for the test, checking our data, performing the test, understanding test results and statistical details.

What do we need?

For the one-sample *t*-test, we need one variable.

We also have an idea, or hypothesis, that the mean of the population has some value. Here are two examples:

- A hospital has a random sample of cholesterol measurements for men. These patients were seen for issues other than cholesterol. They were not taking any medications for high cholesterol. The hospital wants to know if the unknown mean cholesterol for patients is different from a goal level of 200 mg.
- We measure the grams of protein for a sample of energy bars. The label claims that the bars have 20 grams of protein. We want to know if the labels are correct or not.

One-sample t*-test assumptions*

For a valid test, we need data values that are:

- Independent (values are not related to one another).
- Continuous.
- Obtained via a simple random sample from the population.

Also, the population is assumed to be normally distributed.

One-sample *t*-test example

Imagine we have collected a random sample of 31 energy bars from a number of different stores to represent the population of energy bars available to the general consumer. The labels on the bars claim that each bar contains 20 grams of protein.

Table 1: Grams of protein in random sample of energy bars

Energy Bar - Grams of Protein						
20.70	27.46	22.15	19.85	21.29	24.75	
20.75	22.91	25.34	20.33	21.54	21.08	
22.14	19.56	21.10	18.04	24.12	19.95	
19.72	18.28	16.26	17.46	20.53	22.12	
25.06	22.44	19.08	19.88	21.39	22.33	25.79

If you look at the table above, you see that some bars have less than 20 grams of protein. Other bars have more. You might think that the data support the idea that the labels are correct. Others might disagree. The statistical test provides a sound method to make a decision, so that everyone makes the same decision on the same set of data values.

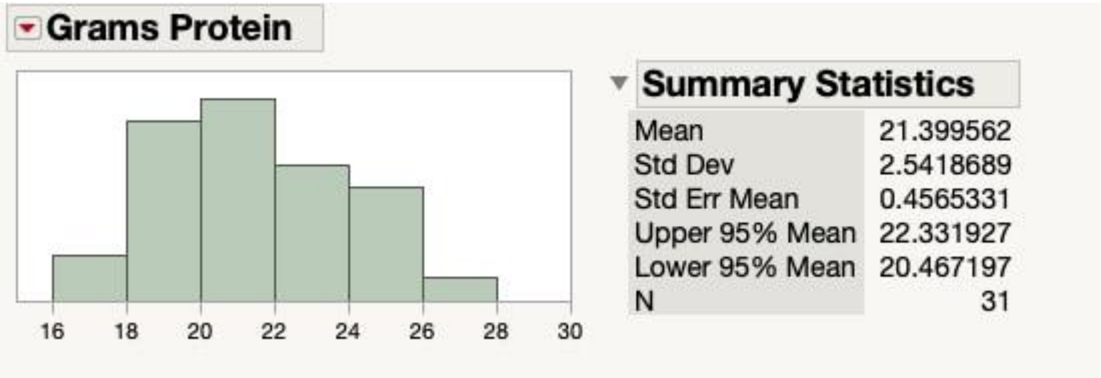
Checking the data

Let’s start by answering: Is the *t*-test an appropriate method to test that the energy bars have 20 grams of protein ? The list below checks the requirements for the test.

- The data values are independent. The grams of protein in one energy bar do not depend on the grams in any other energy bar. An example of dependent values would be if you collected energy bars from a single production lot. A sample from a single lot is representative of that lot, not energy bars in general.
- The data values are grams of protein. The measurements are continuous.
- We assume the energy bars are a simple random sample from the population of energy bars available to the general consumer (i.e., a mix of lots of bars).
- We assume the population from which we are collecting our sample is normally distributed, and for large samples, we can check this assumption.

We decide that the *t*-test is an appropriate method.

Before jumping into analysis, we should take a quick look at the data. The figure below shows a histogram and summary statistics for the energy bars.



Summary Statistics

Mean	21.399562
Std Dev	2.5418689
Std Err Mean	0.4565331
Upper 95% Mean	22.331927
Lower 95% Mean	20.467197
N	31

Figure 1: Histogram and summary statistics for the grams of protein in energy bars

From a quick look at the histogram, we see that there are no unusual points, or *outliers*. The data look roughly bell-shaped, so our assumption of a normal distribution seems reasonable.

From a quick look at the statistics, we see that the average is 21.40, above 20. Does this average from our sample of 31 bars invalidate the label's claim of 20 grams of protein for the unknown entire population mean? Or not?

How to perform the one-sample *t*-test

For the *t*-test calculations we need the average, standard deviation and sample size. These are shown in the summary statistics section of Figure 1 above.

We round the statistics to two decimal places. Software will show more decimal places, and use them in calculations. (Note that Table 1 shows only two decimal places; the actual data used to calculate the summary statistics has more.)

We start by finding the difference between the sample average and 20:

$21.40-20 = 1.4021.40-20 = 1.40$

Next, we calculate the standard error for the mean. The calculation is:

Standard Error for the mean = $s\sqrt{\boxed{n}}=2.54\sqrt{\boxed{31}}=0.456sn=2.5431=0.456$

This matches the value in Figure 1 above.

We now have the pieces for our test statistic. We calculate our test statistic as:

t=Difference Standard Error=1.400.456=3.07t=Difference Standard Error=1.400.456=3.07

To make our decision, we compare the test statistic to a value from the *t*-distribution.

Statistical details

Let's look at the energy bar data and the 1-sample *t*-test using statistical terms.

Our null hypothesis is that the underlying population mean is equal to 20. The null hypothesis is written as:

$Ho:\mu=20Ho:\mu=20$

The alternative hypothesis is that the underlying population mean is not equal to 20. The labels claiming 20 grams of protein would be incorrect. This is written as:

$Ha:\mu\neq20Ha:\mu\neq20$

This is a two-sided test. We are testing if the population mean is different from 20 grams in either direction. If we can reject the null hypothesis that the mean is equal to 20 grams, then we make a practical conclusion that the labels for the bars are incorrect. If we cannot reject the null hypothesis, then we make a practical conclusion that the labels for the bars may be correct.

We calculate the average for the sample and then calculate the difference with the population mean, mu:

$\overline{x}-\mu\bar{x}-\mu$

We calculate the standard error as:

$s\sqrt{\boxed{n}}sn$

The formula shows the sample standard deviation as *s* and the sample size as *n*.

The test statistic uses the formula shown below:

$\overline{x}-\mu s/\sqrt{\boxed{n}}\bar{x}-\mu s/n$

We compare the test statistic to a *t* value with our chosen alpha value and the degrees of freedom for our data. Using the energy bar data as an example, we set α = 0.05. The degrees of freedom (*df*) are based on the sample size and are calculated as:

$df=n-1=31-1=30df=n-1=31-1=30$

Statisticians write the *t* value with α = 0.05 and 30 degrees of freedom as:

t0.05,30t0.05,30

The *t* value for a two-sided test with α = 0.05 and 30 degrees of freedom is +/- 2.042. There are two possible results from our comparison:

- The test statistic is less extreme than the critical *t* values; in other words, the test statistic is not less than -2.042, or is not greater than +2.042. You fail to reject the null hypothesis that the mean is equal to the specified value. In our example, you would be unable to conclude that the label for the protein bars should be changed.
- The test statistic is more extreme than the critical *t* values; in other words, the test statistic is less than -2.042, or is greater than +2.042. You reject the null hypothesis that the mean is equal to the specified value. In our example, you conclude that either the label should be updated or the production process should be improved to produce, on average, bars with 20 grams of protein.

Testing for normality

The normality assumption is more important for small sample sizes than for larger sample sizes.

Normal distributions are symmetric, which means they are "even" on both sides of the center. Normal distributions do not have extreme values, or outliers. You can check these two features of a normal distribution with graphs. Earlier, we decided that the energy bar data was "close enough" to normal to go ahead with the assumption of normality. The figure below shows a normal quantile plot for the data, and supports our decision.

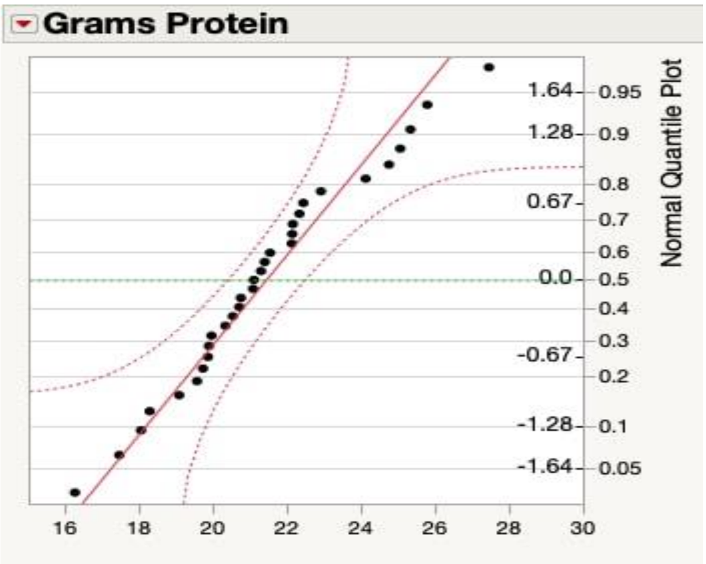
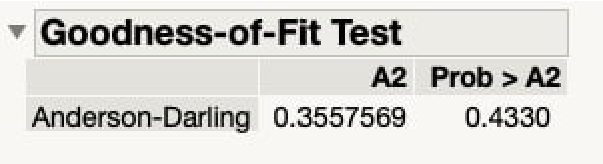


Figure 4: Normal quantile plot for energy bar data

You can also perform a formal test for normality using software. The figure below shows results of testing for normality with JMP software. We cannot reject the hypothesis of a normal distribution.



Goodness-of-Fit Test

	A2	Prob > A2
Anderson-Darling	0.3557569	0.4330

Figure 5: Testing for normality using JMP software

We can go ahead with the assumption that the energy bar data is normally distributed.

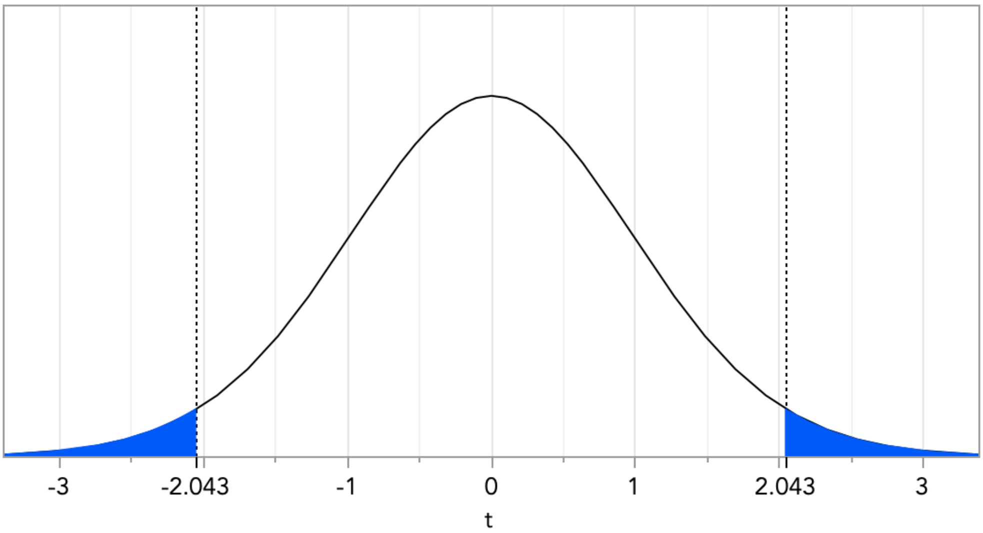
What if my data are not from a Normal distribution?

If your sample size is very small, it is hard to test for normality. In this situation, you might need to use your understanding of the measurements. For example, for the energy bar data, the company knows that the underlying distribution of grams of protein is normally distributed. Even for a very small sample, the company would likely go ahead with the *t*-test and assume normality.

What if you know the underlying measurements are not normally distributed? Or what if your sample size is large and the test for normality is rejected? In this situation, you can use a nonparametric test. Nonparametric analyses do not depend on an assumption that the data values are from a specific distribution. For the one-sample *t*-test, the one possible nonparametric test is the Wilcoxon Signed Rank test.

Understanding p-values

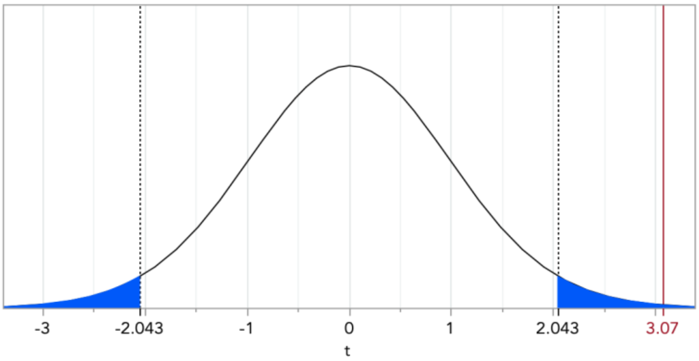
Using a visual, you can check to see if your test statistic is more extreme than a specified value in the distribution. The figure below shows a *t*-distribution with 30 degrees of freedom.



t-distribution with 30 degrees of freedom and α = 0.05

Since our test is two-sided and we set α = 0.05, the figure shows that the value of 2.042 “cuts off” 5% of the data in the tails combined.

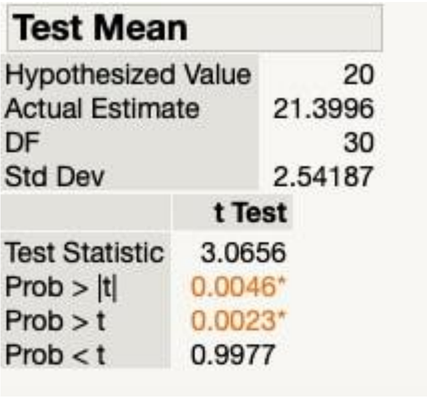
The next figure shows our results. You can see the test statistic falls above the specified critical value. It is far enough “out in the tail” to reject the hypothesis that the mean is equal to 20.



Our results displayed in a t-distribution with 30 degrees of freedom

Putting it all together with Software

You are likely to use software to perform a *t*-test. The figure below shows results for the 1-sample *t*-test for the energy bar data from JMP software.



```
Test Mean
Hypothesized Value        20
Actual Estimate      21.3996
DF                        30
Std Dev              2.54187
                  t Test
Test Statistic    3.0656
Prob > |t|        0.0046*
Prob > t          0.0023*
Prob < t          0.9977
```

One-sample t-test results for energy bar data using JMP software

The software shows the null hypothesis value of 20 and the average and standard deviation from the data. The test statistic is 3.07. This matches the calculations above.

The software shows results for a two-sided test and for one-sided tests. We want the two-sided test. Our null hypothesis is that the mean grams of protein is equal to 20. Our alternative hypothesis is that the mean grams of protein is not equal to 20. The software shows a *p*-value of 0.0046 for the two-sided test. This *p*-value describes the likelihood of seeing a sample average as extreme as 21.4, or more extreme, when the underlying population mean is actually 20; in other words, the probability of observing a sample mean as different, or even more different from 20, than the mean we observed in our sample. A *p*-value of 0.0046 means there is about 46 chances out of 10,000. We feel confident in rejecting the null hypothesis that the population mean is equal to 20.

Two independent samples t-tests:

In statistics, the two sample t-test for independent samples is a type of hypothesis test that can be used to determine whether the means of two populations are statistically different given the two samples are independent and have normal distributions. As data scientists, it is important to understand how to use the two sample t-test for independent samples so that you can correctly analyze your data. In this blog post, we will discuss the **two sample t-test** for **independent samples** in detail, including the formula and examples.

Two-Sample Test

A two-sample T-test is defined as statistical hypothesis testing technique in which **two independent sample**s are compared to determine if the means of two populations are statistically different. The two-sample T-test is used when the standard deviations of the populations to be compared are unknown and the sample size is small. The size of sample 30 or less is considered as small sample. That said, the size of the sample is not a strict condition for using T-test. The two-sample T-test is used when the **two samples are independent** and have **normal distributions**. In order to use a two-sample T-test as described in this blog, you need to have two **independent samples**. The independent samples mean that the *two samples cannot be from the same group of people and they cannot be related in any way*. However, two-sample T-test can also be used for pairwise comparisons when the "two" samples represent the same items tested in different scenarios. The pairwise t-test will be dealt with in different blog.

Let's say you want to know if two different brands of batteries have the same average life. You could take a battery from each brand, use them until they die, and record the results. This would be an extremely time-consuming process, and it's not very likely that you'd get a large enough sample size to draw any conclusions. Another option is to use a two-sample T-test. This test allows you to compare the averages of two groups without having to measure the batteries' life spans yourself.

The following are a few real-life examples where two-sample T-test for independent samples can be used:

- Comparing the average test scores of two classes from two different schools
- Comparing the average weights of two different or independent groups of people
- Determining whether the medication have the same efficacy on two different or independent groups of people
- Compare whether the effect of vaccination on two different groups

T-statistics when population variances or standard deviations are unequal

The formula for T-statistics is different based on whether the populations' standard deviations are same / equal or different. When the standard deviations of populations are not equal, the following formula is used to calculate the T-statistics and degrees of freedom.

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where **X̄1** is mean of first sample, **X̄2** is mean of second sample, **µ1** is the mean of first population, **µ2** is the mean of second population, **s1** is the standard deviation of first sample, **s2** is the standard deviation of second sample, **n1** is the size of the first sample, **n2** is the size of the second sample.

The degrees of freedom can be calculated as the sum of two sample sizes minus two.

Degrees of freedom, **df** = **n1 + n2 – 2**

A confidence interval for the difference between two means specifies a range of values within which the difference between the means of the two populations may lie. The difference between the means of two populations can be estimated based on the following formula:

Difference in population means = Difference in sample means +/- T*standard error

$$(\overline{x}_1 - \overline{x}_2) \pm t^*\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

In above formula, the standard error is the square root term.

T-statistics when population variances or standard deviations are equal

In case, the two populations' standard deviations are equal, the formula termed as **pooled t-statistics** is used based on the usage of **pooled standard deviations** of the two samples. The following is the formula for the **pooled t-statistics:**

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In the above formula, **Sp** is termed as pooled standard deviation. The formula for pooled variance can be calculated based on the following:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

The value for the degree of freedom can be calculated as the sum of two sample sizes minus two.

Degrees of freedom, **df** = **n1 + n2 – 2**

When two-sample T-test instead of two-sample Z-test?

When the population standard deviations are known and the sample size is large, we go for two-sample Z-test for comparing the two different populations. The sample size greater than 30 is considered to be large sample size. Otherwise, a two-sample T-test is known with T-distribution and a given degrees of freedom.

Two-sample T-test: Examples

Lets say we need to compare the performance of two call centers in terms of average call lengths and find out if the difference is statistically significant or the difference is a chance occurrence. To start with, we will need to formulate the null and alternate hypothesis.

Null hypothesis, H0: There is no difference between the average call length between two call centers.

Alternate hypothesis, Ha: There is a difference between the average call length and hence the performance.

We randomly select 20 calls from each call center and calculate the average call lengths. The two call centers seem to have different average call lengths. Is this difference statistically significant?

First, we need to calculate the two sample means and standard deviations:

Call Center A: Sample mean, **X̄1** = 122 seconds, SD, **S1** = 15 seconds, **n1** = 20

Call Center B: Sample mean, **X̄2** = 135 seconds, SD, **S2** = 20 seconds, **n2** = 20

Next, we use a two-sample t-test to determine if the difference between two sample means is statistically significant. We will use a 95% confidence level and $\alpha = 0.05$.

The two-sample t-statistic is calculated as the following assuming that the standard deviations of the population is not same and the population mean is same.

t = ((135 – 122) – 0)/SQRT((20*20/20) + ((15*15)/20))

t = 13/SQRT(20 + 11.25)

t = 13/SQRT(31.25)

t = 2.3256

The value of degrees of freedom can be calculated as the following:

Degree of freedom, df = n1 + n2 -2 = 20 + 20 – 2 = 38

The critical value of a two-tailed T-test with degrees of freedom as 38 and level of significance as 0.05 comes out to be **2.0244**. Since the current t-value of 2.3256 is greater than the critical value of 2.0244, one can reject the null hypothesis that there is no difference between the performances in terms of the call length time. Thus, based on the given evidence, the alternate hypothesis stands as true.

PROBABILITY

If an experiment is repeated under essentially homogeneous and similar conditions, two possible conclusions can be arrived. They are: the results are unique and the outcome can be predictable and result is not unique but may be one of the several possible outcomes. In this context, it is better to understand various terms pertaining to probability before examining the probability theory. The main terms are explained as follows:

Random Experiment

An experiment which can be repeated under the same conditions and the outcome cannot be predicted under any circumstances is known as random experiment. For example: An unbiased coin is tossed. Here we are not in a position to predict whether head or tail is going to occur. Hence, this type of experiment is known as random experiment.

Sample Space

A set of possible outcomes of a random experiment is known as sample space. For example, in the case of tossing of an unbiased coin twice, the possible outcomes are HH, HT, TH and TT. This can be represented in a sample space as S= (HH, HT, TH, TT).

An Event

Any possible outcomes of an experiment are known as an event. In the case of tossing of an unbiased coin twice, HH is an event. An event can be classified into two. They are: (a) Simple events, and (ii) Compound events. Simple event is an event which has only one sample point in the sample space. Compound event is an event which has more than one sample point in the sample space. In the case of tossing of an unbiased coin twice HH is a simple event and TH and TT are the compound events.

Complementary Event

A and A' are the complementary event if A' consists of all those sample point which is not included in A. For instance, an unbiased dice is thrown once. The probability of an odd number turns up are complementary to an even number turns up. Here, it is worth mentioning that the probability of sample space is always is equal to one. Hence, the P (A') = 1 - P (A).

Mutually Exclusive Events

A and B are the two mutually exclusive events if the occurrence of A precludes the occurrence of B. For example, in the case of tossing of an unbiased coin once, the occurrence of head precludes the occurrence of tail. Hence, head and tail are the mutually exclusive event in the case of tossing of an unbiased coin once. If A and B are mutually exclusive events, then the probability of occurrence of A or B is equal to sum of their individual probabilities. Symbolically, it can be presented as:

P (A U B) = P (A) + P (B)

If A and B is joint sets, then the addition theorem of probability can be stated as:

P (A U B) = P(A) + P(B) - P(AB)

Independent Event

A and B are the two independent event if the occurrence of A does not influence the occurrence of B. In the case of tossing of an unbiased coin twice, the occurrence of head in the first toss does not influence the occurrence of head or tail in the toss. Hence, these two events are called independent events. In the case of independent event, the multiplication theorem can be stated as the probability of A and B is the product of their individual probabilities. Symbolically, it can be presented as:-

P (A B) = P (A) * P (B)

Addition Theorem of Probability

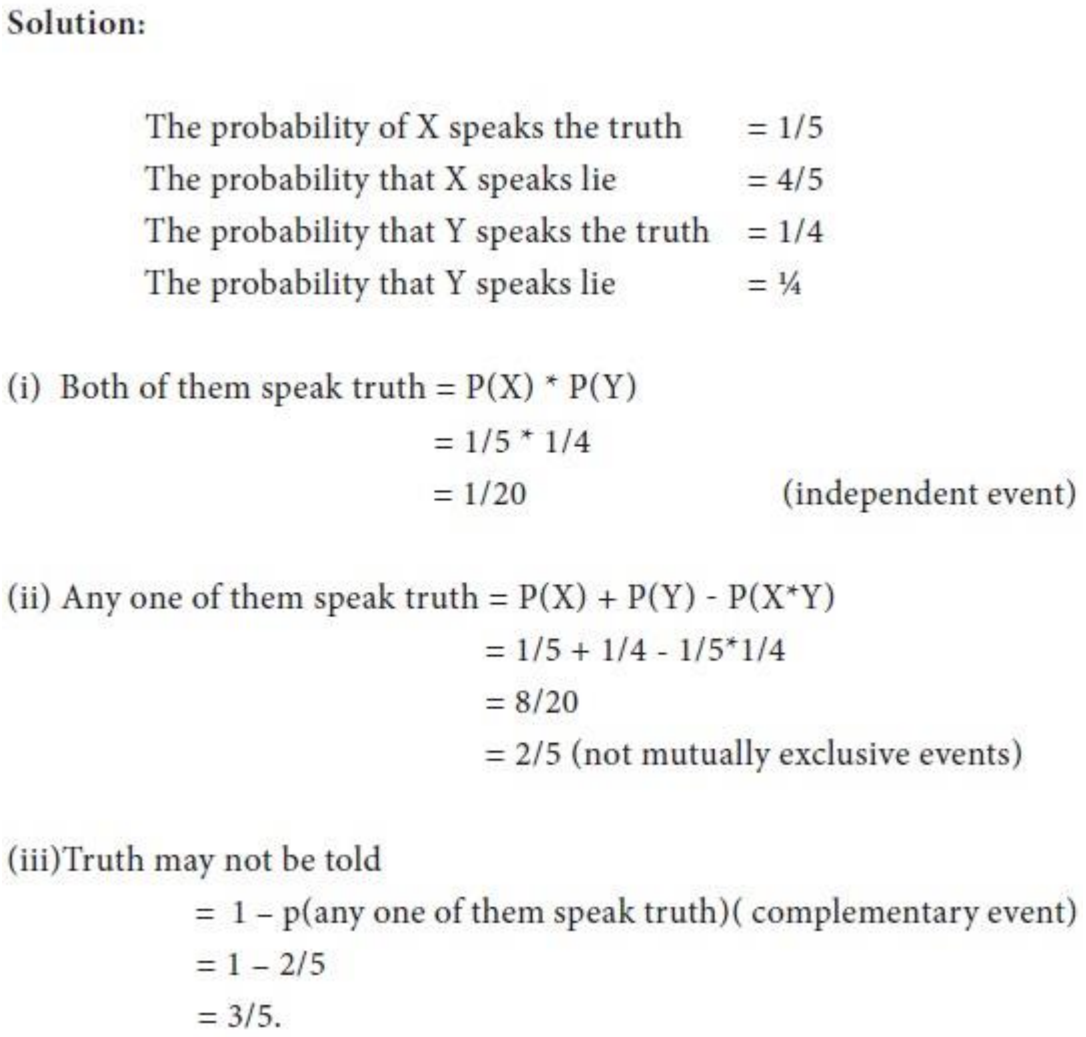
Let A and B be the two mutually exclusive events, then the probability of A or B is equal to the sum of their individual probabilities. (for detail refer mutually exclusive events)

Multiplication Theorem of Probability

Let A and B be the two independent events, then the probability of A and B is equal to the product of their individual probabilities. (for details refer independent events)

Example: The odds that person X speaks the truth are 4:1 and the odds that Y speaks the truth are. Find the probability that:-

1. Both of them speak the truth,
2. Any one of them speak the truth and
3. Truth may not be told.



Solution:

The probability of X speaks the truth = 1/5
The probability that X speaks lie = 4/5
The probability that Y speaks the truth = 1/4
The probability that Y speaks lie = ¼

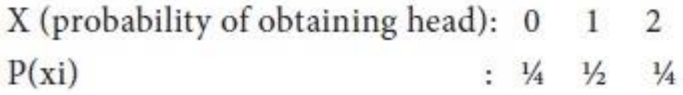
(i) Both of them speak truth = P(X) * P(Y)
= 1/5 * 1/4
= 1/20 (independent event)

(ii) Any one of them speak truth = P(X) + P(Y) - P(X*Y)
= 1/5 + 1/4 - 1/5*1/4
= 8/20
= 2/5 (not mutually exclusive events)

(iii)Truth may not be told
= 1 – p(any one of them speak truth)(complementary event)
= 1 – 2/5
= 3/5.

Probability Distribution

If X is discrete random variable which takes the values of x_1, x_2,x_3….. X_n and the corresponding probabilities are p_1, p_2, ………p_n, then, X follows the probability distribution. The two main properties of probability distribution are: (i) P(Xi) is always greater than or equal to zero and less than or equal to one, and (ii) the summation of probability distribution is always equal to one. For example, tossing of an unbiased coin twice. Then the probability distribution is:



X (probability of obtaining head):	0	1	2
P(xi) :	¼	½	¼

Expectation of probability

Let X be the discrete random variable which takes the value of x_1, x_2,…… x_n then the respective probability is p_1, p_2, ……….. p_n, then the expectation of probability distribution is $p_1x_1 + p_2x_2 +$ ………….. $+ p_nx_n$. In the above example, the expectation of probability distribution is (0* ¼ +1*1/2+2*¼) =1.

Tags : Research Methodology - Statistical Analysis

BINOMIALDISTRIBUTION

The binomial distribution also known as 'Bernoulli Distribution' is associated with the name of a Swiss mathematician, James Bernoulli who is also known as Jacques or Jakon (1654 – 1705). Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives. It can be explained as follows:

1. If an experiment is repeated under the same conditions for a fixed number of trials, say, n.
2. In each trial, there are only two possible outcomes of the experiment. Let us define it as "success" or "failure". Then the sample space of possible outcomes of each experiment is:
3.S = [failure, success]
4. The probability of a success denoted by p remains constant from trial to trial and the probability of a failure denoted by q which is equal to (1 – p).
5. The trials are independent in nature i.e., the outcomes of any trial or sequence of trials do not affect the outcomes of subsequent trials. Hence, the multiplication theorem of probability can be applied for the occurrence of success and failure. Thus, the probability of success or failure is p.q. 6. Let us assume that we conduct an experiment in n times. Out of which x times be the success and failure is (n-x) times. The occurrence of success or failure in successive trials is mutually exclusive events. Hence, we can apply addition theorem of probability.
7. Based on the above two theorems, the probability of success or failure is

$$P(X) = {}^{n}C_{x} p^{x} q^{n-x}$$

$$\frac{n\,!}{x\,!\,(n-x)\,!}\,.\; P^{x} q^{n-x}$$

where p = probability of success in a single trail, q = 1 – p, n = Number of trials and x = no. of successes in n trials.

Thus, for an event A with probability of occurrence p and non-occurrence q, if n trials are made, probability distribution of the number of occurrences of A will be as set. If we want to obtain the probable frequencies of the various outcomes in n sets of N trials, the following expression shall be used: N(p + q)n

$$N(p + q)^n = Np^n + {}^nC_1p^{n-1}q + {}^nC2p^{n-2}q^2 + \ldots\ldots + {}^nC_rp^{n-r}q^r + \ldots\ldots q^n.$$

The frequencies obtained by the above expansion are known as expected or theoretical frequencies. On the other hand, the frequencies actually obtained by making experiments are called actual or observed frequencies. Generally, there is some difference between the observed and expected frequencies but the difference becomes smaller and smaller as N increases.

Obtaining Coefficient Of The Binomial Distribution:

The following rules may be considered for obtaining coefficients from the binomial expansion:

1. The first term is qn.,
2. The second term is nC_1qn-1p,
3. In each succeeding term the power of q is reduced by 1 and the power of p is increased by 1.
4. The coefficient of any term is found by multiplying the coefficient of the preceding term by the power of q in that preceding term, and dividing the products so obtained by one more than the power of p in that proceeding term. Thus, when we expand (q + p)n, we will obtain the following:-

$$(p + q)^n = p^n + {}^nC_1p^{n-1}q + {}^nC_2p^{n-2}q^2 + \ldots\ldots + {}^nC_rp^{n-r}q^r + \ldots\ldots q^n.$$

Where, 1, nC_1, nC_2 ……. are called the binomial coefficient. Thus in the expansion of (p + q)4 we will have (p + q)4 = p4 +4p3q +6p2q2 + 4p1q3 + q4 and the coefficients will be 1, 4, 6, 4,1.From the above binomial expansion, the following general relationships should be noted:

1. The number of terms in a binomial expansion is always n + 1,
2. The exponents of p and q, for any single term, when added together, always sum to n.
3. The exponents of p are n, (n – 1), (n – 2),…….1, 0, respectively and the exponents of q are 0, 1,2,……(n– 1), n, respectively.
4. The coefficients for the n + 1 terms of the distribution are always symmetrical in nature.

Properties Of Binomial Distribution

The main properties of binomial distribution are:-

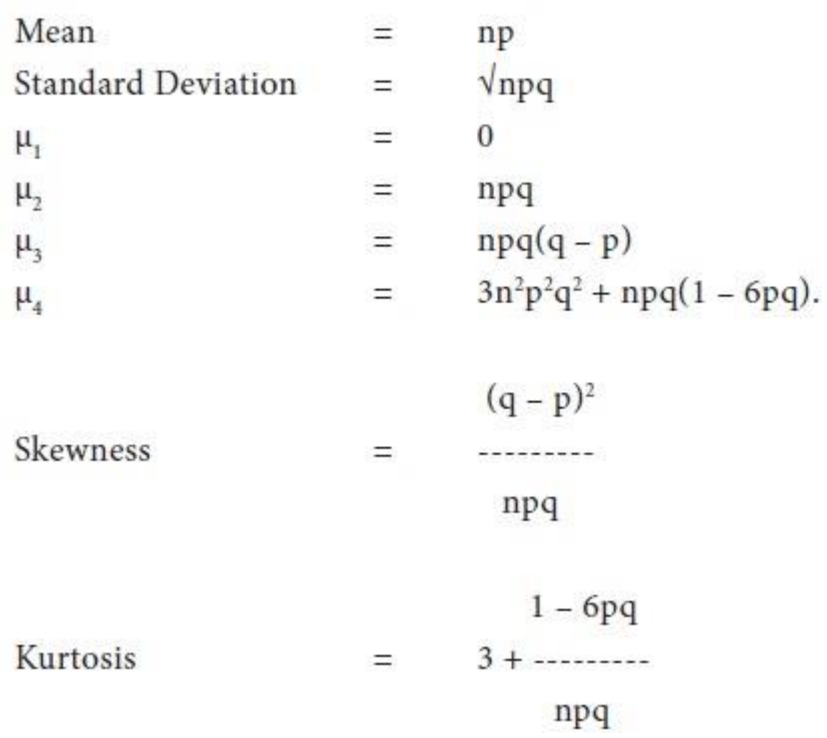
1. The shape and location of binomial distribution changes as p changes for a given n or as n changes for a given p. As p increases for a fixed n, the binomial distribution shifts to the right.
2. The mode of the binomial distribution is equal to the value of x which has the largest probability. The mean and mode are equal if np is an integer.

3. As n increases for a fixed p, the binomial distribution moves to the right, flattens and spreads out.

4. The mean of the binomial distribution is np and it increases as n increases with p held constant. For larger n there are more possible outcomes of a binomial experiment and the probability associated with any particular outcome becomes smaller.

5. If n is larger and if neither p nor q is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized variable given by z = (X – np) / √npq.

6. The various constants of binomial distribution are:



Mean	=	np
Standard Deviation	=	$\sqrt{npq}$
μ_1	=	0
μ_2	=	npq
μ_3	=	$npq(q - p)$
μ_4	=	$3n^2p^2q^2 + npq(1 - 6pq)$.
Skewness	=	$\dfrac{(q - p)^2}{npq}$
Kurtosis	=	$3 + \dfrac{1 - 6pq}{npq}$

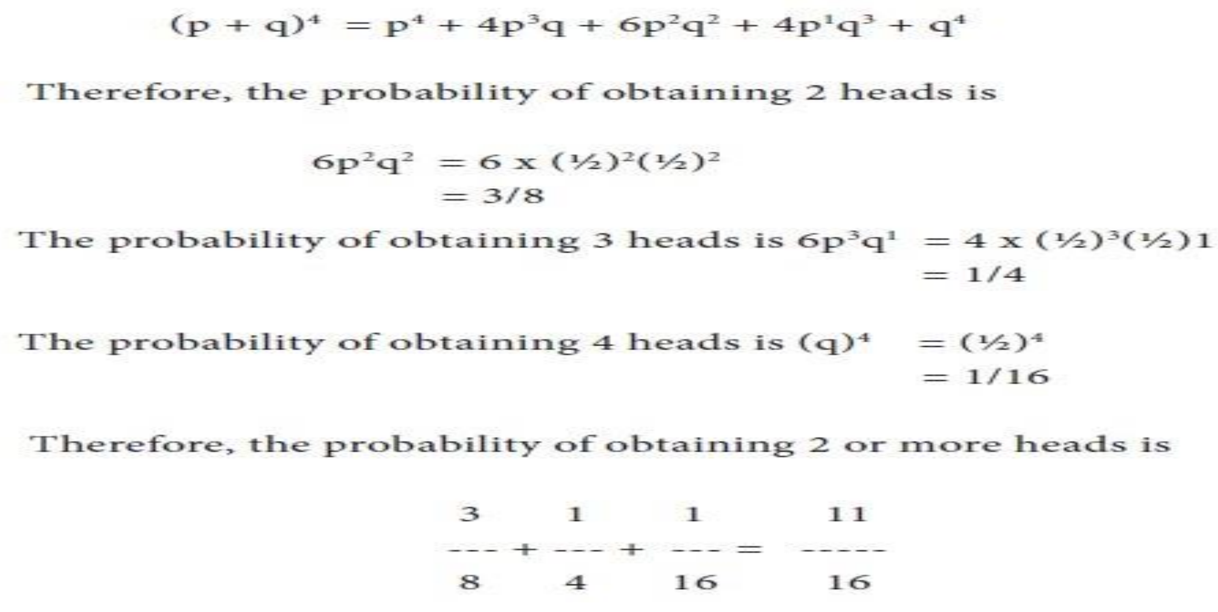
Illustrations:1

A coin is tossed four times. What is the probability of obtaining two or more heads?

Solution:

when a coin is tossed the probabilities of head and tail in case of an unbiased coin are equal, i.e., p = q = ½

The various possibilities for all the events are the terms of the expansion (q+p)4



$$(p + q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4p^1q^3 + q^4$$

Therefore, the probability of obtaining 2 heads is

$$6p^2q^2 = 6 \text{ x } (½)^2(½)^2$$
$$= 3/8$$

The probability of obtaining 3 heads is $6p^3q^1 = 4 \text{ x } (½)^3(½)1$

$$= 1/4$$

The probability of obtaining 4 heads is $(q)^4 = (½)^4$

$$= 1/16$$

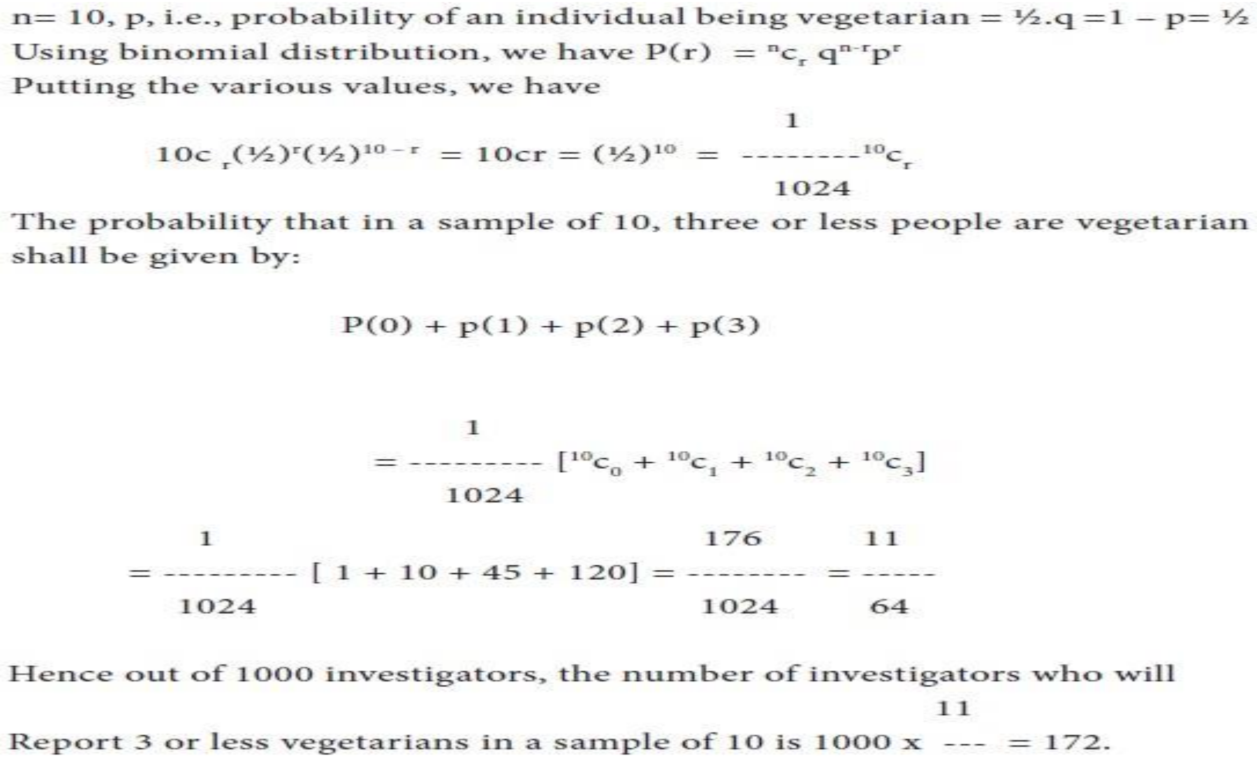
Therefore, the probability of obtaining 2 or more heads is

$$\frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{11}{16}$$

Illustration:2

Assuming that half the population is vegetarian so that the chance of an individual being a vegetarian is ½ and assuming that 100 investigations can take sample of 10 individuals to verify whether they are vegetarians, how many investigation would you expect to report that three people or less were vegetarians?

Solution:



n= 10, p, i.e., probability of an individual being vegetarian = ½.q =1 – p= ½
Using binomial distribution, we have $P(r) = {}^nc_r\, q^{n-r}p^r$
Putting the various values, we have

$$10c_r(½)^r(½)^{10-r} = 10cr = (½)^{10} = \frac{1}{1024}{}^{10}c_r$$

The probability that in a sample of 10, three or less people are vegetarian shall be given by:

$$P(0) + p(1) + p(2) + p(3)$$

$$= \frac{1}{1024}[{}^{10}c_0 + {}^{10}c_1 + {}^{10}c_2 + {}^{10}c_3]$$

$$= \frac{1}{1024}[1 + 10 + 45 + 120] = \frac{176}{1024} = \frac{11}{64}$$

Hence out of 1000 investigators, the number of investigators who will Report 3 or less vegetarians in a sample of 10 is $1000 \text{ x } \frac{11}{---} = 172.$

POISSON DISTRIBUTION

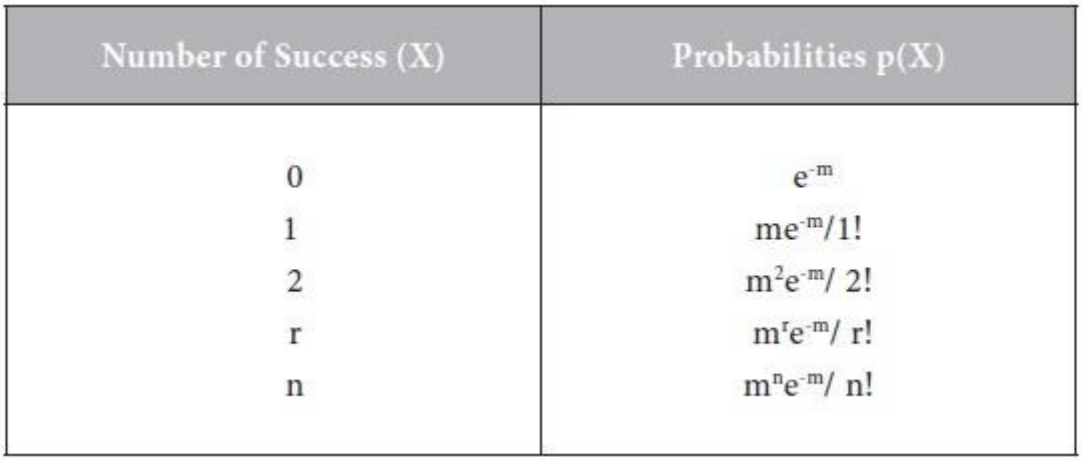
Poisson distribution was derived in 1837 by a French Mathematician Simeon D Poisson (1731 – 1840). In binomial distribution, the values of p and q and n are given. There is a certainty of the total number of events.

But there are cases where p is very small and n is very large and such case is normally related to poisson distribution. For example, persons killed in road accidents, the number of defective articles produced by a quality machine. Poisson distribution may be obtained as a limiting case of binomial probability distribution, under the following condition.

i. P, successes, approach zero (p 0)

ii. np = m is finite.

The poisson distribution of the probabilities of occurrence of various rare events (successes) 0,1,2,…. Are given below:

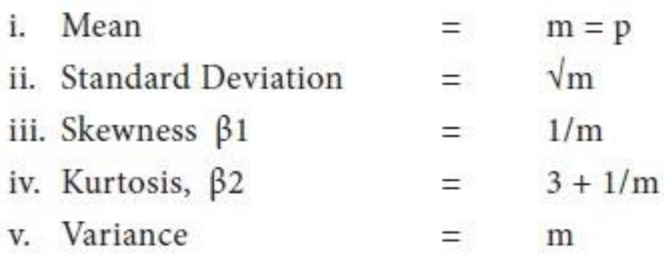


Number of Success (X)	Probabilities p(X)
0	e^{-m}
1	$me^{-m}/1!$
2	$m^2e^{-m}/ 2!$
r	$m^re^{-m}/ r!$
n	$m^ne^{-m}/ n!$

Where, e = 2.718, and m = average number of occurrence of given distribution.

The poisson distribution is a discrete distribution with a parameter m.

The various constants are:



i.	Mean	=	m = p
ii.	Standard Deviation	=	$\sqrt{m}$
iii.	Skewness $\beta 1$	=	1/m
iv.	Kurtosis, $\beta 2$	=	3 + 1/m
v.	Variance	=	m

Illustration:1

A book contains 100 misprints distributed randomly throughout its 100 pages. What is the probability that a page observed at random contains at least two misprints? Assume Poisson Distribution.

Solution:

$$M = \frac{\text{Total number of misprints}}{\text{Total number of pages}} = \frac{100}{100} = 1$$

Probability that a page contains at least two misprints:

$$P(r \geq 2) = 1 - [p(0) + p(1)]$$

$$P(r) = \frac{m^r e^{-m}}{r!}$$

$$P(0) = \frac{1^0 e^{-1}}{0!} = e^{-1} = \frac{1}{e} = \frac{1}{2.7183}$$

$$P(1) = \frac{1^1 e^{-1}}{1!} = e^{-1} = \frac{1}{e} = \frac{1}{2.7183}$$

$$P(0) + p(1) = \frac{1}{2.718} + \frac{1}{2.718} = 0.736$$

$$P(r \geq 2) = 1 - [p(0) + p(1)] = 1 - 0.736 = \mathbf{0.264}$$

Illustration: 2

If the mean of a Poisson distribution is 16, find (1) S.D.(2) B_1 (3) B_2 (4) μ_3 (5) μ_4

Solution:

$m = 16$

1. S.D. $= \sqrt{m} = \sqrt{16} = 4$
2. $\beta_1 = 1/m = 1/16 = 0.625$
3. $\beta_2 = 3 + 1/m = 3 + 0.625 = 3.0625$
4. $\mu_3 = m = 16$
5. $\mu_4 = m + 3m^2 = 16 + 3(16)^2 = 784$

NORMALDISTRIBUTION

The normal distribution was first described by Abraham Demoivre (1667-1754) as the limiting form of binomial model in 1733. Normal distribution was rediscovered by Gauss in 1809 and by Laplace in 1812. Both Gauss and Laplace were led to the distribution by their work on the theory of errors of observations arising in physical measuring processes particularly in astronomy.

The probability function of a Normal Distribution is defined as:

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

Where, X = Values of the continuous random variable, μ = Mean of the normal random variable, e = 2.7183, π =

3.1416

Relation Between Binomial, Poisson And Normal Distributions

Binomial, Poisson and Normal Distribution are closely related to one other. When N is large while the probability P of the occurrence of an event is close to zero so that q = (1-p) the binomial distribution is very closely approximated by the Poisson distribution with m = np.

The Poisson distribution approaches a normal distribution with standardized variable (x – m)/ √m as m increases to infinity.

Normal distribution and its properties

The important properties of the normal distribution are:-

1. The normal curve is "bell shaped" and symmetrical in nature. The distribution of the frequencies on either side of the maximum ordinate of the curve is similar with each other.

2. The maximum ordinate of the normal curve is at x = μ. Hence the mean, median and mode of the normal distribution coincide.

3. It ranges between - ∞ to + ∞

4. The value of the maximum ordinate is 1/ σ√2π.

5. The points where the curve change from convex to concave or vice versa is at X = μ ± σ.

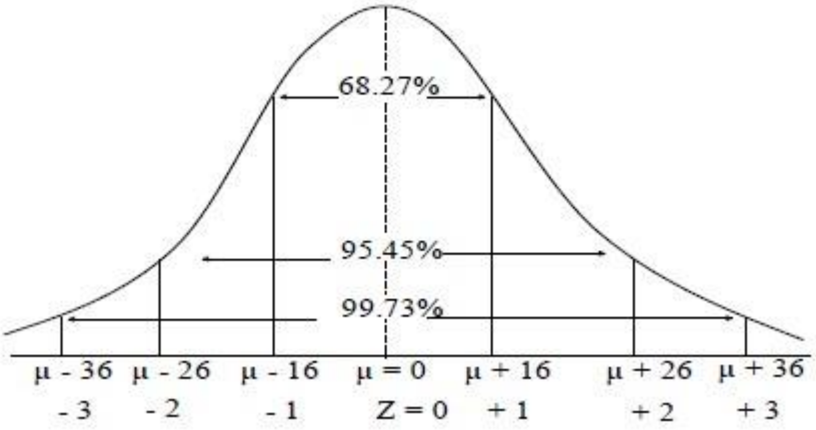
6. The first and third quartiles are equidistant from median.

7. The area under the normal curve distribution are:

a. μ ± 1σ covers 68.27% area;

b. μ ± 2σ covers 95.45% area.

c. μ ± 3σ covers 99.73% area.



8. When μ = 0 and σ = 1, then the normal distribution will be a standard normal curve. The probability function of standard normal curve is

$$P(X) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

The following table gives the area under the normal probability curve for some important value of Z.

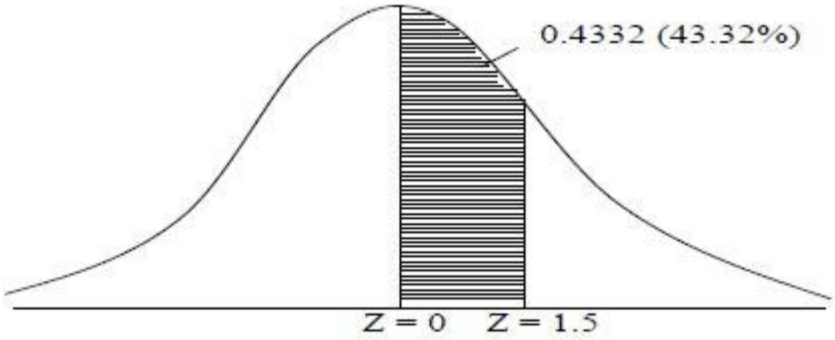
Distance from the mean ordinate in Terms of ± σ	Area under the curve
Z = ± 0.6745	0.50
Z = ± 1.0	0.6826
Z = ± 1.96	0.95
Z = ± 2.00	0.9544
Z = ± 2.58	0.99
Z = ± 3.0	0.9973

9. All odd moments are equal to zero.

10. Skewness = 0 and Kurtosis = 3 in normal distribution.

Illustration:1

Find the probability that the standard normal value lies between 0 and 1.5



As the mean, Z = 0.

To find the area between Z = 0 and Z = 1.5, look the area between 0 and 1.5, from the table. It is 0.4332 (shaded area)

Illustration:2

The results of a particular examination are given below in a summary form:

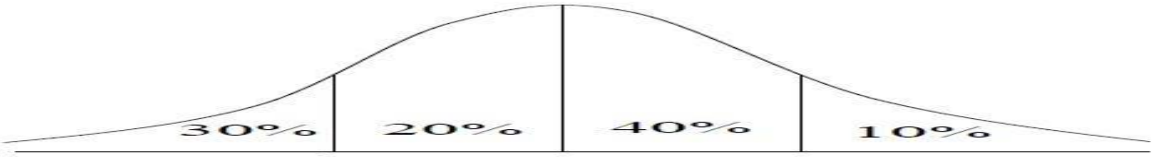
Result	Percentage of candidates
Passed with distinction	10
Passed	60
Failed	30

It is known that a candidate gets plucked if he obtains less than 40 marks, out of 100 while he must obtain at least 75 marks in order to pass with distinction. Determine the mean and standard deviation of the distribution of marks assuming this to be normal.

Solution:

30% students get marks less than 40.

$$Z = \frac{40 - \overline{X}}{\sigma} = -0.52 \text{ (from the table)}$$



$40 - \overline{X} = -0.52\sigma$ ----------- (i)

10% students get more than 75

40% area = $75 - \overline{X} = 1.28$ ------------ (ii)

= $75 - \overline{X} = 1.28\sigma$

Subtract (ii) from (i)

$40 - \overline{X} = -0.52\ \sigma$

$75 - \overline{X} = 1.28\ \sigma$

$-35 = -1.8\ \sigma$

$35 = 1.8\ \sigma$

$1.80\ \sigma = 35$

$$\sigma = \frac{35}{1.80} = 19.4$$

Mean

$40 - \overline{X} = -0.52 \times (19.4)$

$-\overline{X} = -40 - 10.09 = 50.09$

Illustration:3

The scores observed by candidate in a certain test are normally distributed with mean 1000 and standard deviation 200. What per cent of candidates receive scores (i) less than 800, (ii) between 800 and 1200? (the area under the curve between Z = 0 and Z = 1 is 0.34134).

Solution:

$$\overline{X} = 1000; \sigma = 200$$

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

i) For X = 800

$$Z = \frac{800 - 1000}{200} = -1$$

Area between Z = -1 and Z = 0 is 0.34134

Area for Z = -1 = 0.5 – 0.34134 = 0.15866

Therefore, the percentage = 0.15866 x 100 = 15.86%

ii) when, X = 1200,

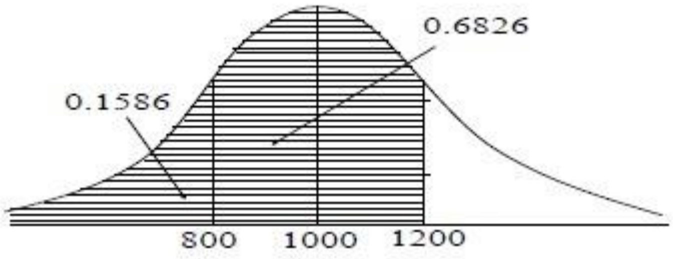
$$Z = \frac{1200 - 1000}{200} = 1$$

Area between Z = 0 and Z = 1 is 0.34134

Area between X = 400 to X = 600

i.e.,

Z = -1 and Z = 1 is 0.34134 + 0.34134 = 0.6826 = 68.26%



SIMPLE CORRELATION - Correlation And Regression Analysis

Correlation

Correlation means the average relationship between two or more variables. When changes in the values of a variable affect the values of another variable, we say that there is a correlation between the two variables. The two variables may move in the same direction or in opposite directions. Simply because of the presence of correlation between two variables, we cannot jump to the conclusion that there is a cause-effect relationship between them. Sometimes, it may be due to chance also.

Simple correlation

We say that the correlation is simple if the comparison involves two variables only.

TYPES OF CORRELATION

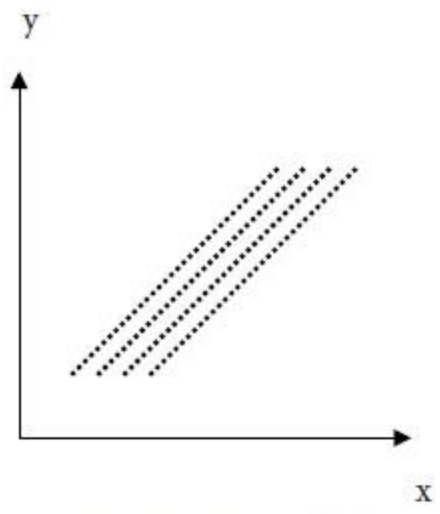
Positive correlation

If two variables x and y move in the same direction, we say that there is a positive correlation between them. In this case, when the value of one variable increases, the value of the other variable also increases and when the value of one variable decreases, the value of the other variable also decreases. Eg. The age and height of a child.

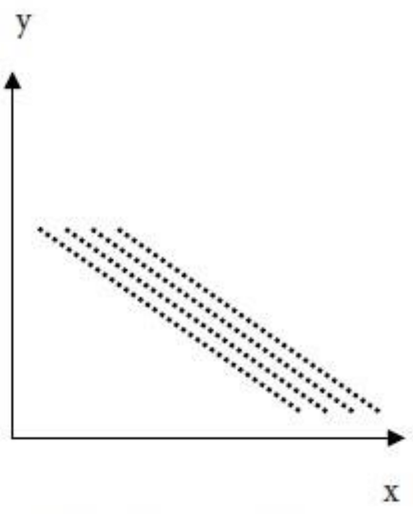
Negative correlation

If two variables x and y move in opposite directions, we say that there is a negative correlation between them. i.e., when the value of one variable increases, the value of the other variable decreases and vice versa. Eg. The price and demand of a normal good.

The following diagrams illustrate positive and negative correlations between x and y.



Positive Correlation



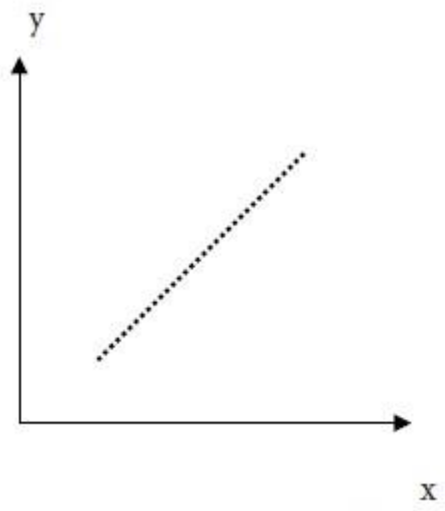
Negative Correlation

Perfect Positive Correlation

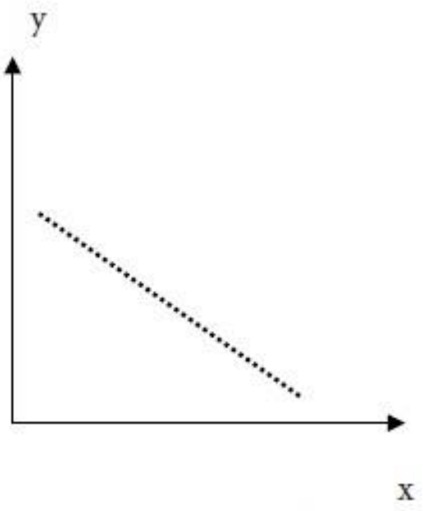
If changes in two variables are in the same direction and the changes are in equal proportion, we say that there is a perfect positive correlation between them.

Perfect Negative Correlation

If changes in two variables are in opposite directions and the absolute values of changes are in equal proportion, we say that there is a perfect negative correlation between them.



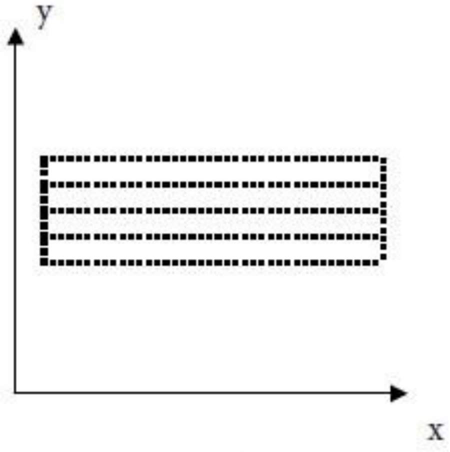
Perfect Positive Correlation



Perfect Negative Correlation

Zero Correlation

If there is no relationship between the two variables, then the variables are said to be independent. In this case the correlation between the two variables is zero.



Zero Correlation

Linear Correlation

If the quantum of change in one variable always bears a constant ratio to the quantum of change in the other variable, we say that the two variables have a linear correlation between them.

Coefficient of Correlation

The coefficient of correlation between two variables X, Y is a measure of the degree of association (i.e., strength of relationship) between them. The coefficient of correlation is usually denoted by ‘r’.

Karl Pearson’s Coefficient of Simple Correlation:

Let N denote the number of pairs of observations of two variables X and Y. The correlation coefficient r between X and Y is defined by

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

This formula is suitable for solving problems with hand calculators. To apply this formula, we have to calculate ∑ X,∑Y, ∑XY, ∑X2, ∑Y2.

PARTIAL CORRELATION

Simple correlation is a measure of the relationship between a dependent variable and another independent variable. For example, if the performance of a sales person depends only on the training that he has received, then the relationship between the training and the sales performance is measured by the simple correlation coefficient r. However, a dependent variable may depend on several variables. For example, the yarn produced in a factory may depend on the efficiency of the machine, the quality of cotton, the efficiency of workers, etc. It becomes necessary to have a measure of relationship in such complex situations. Partial correlation is used for this purpose. The technique of partial correlation proves useful when one has to develop a model with 3 to 5 variables.

Suppose Y is a dependent variable, depending on n other variables X_1, X_2, …, X_n.. Partial correlation is a measure of the relationship between and any one of the variables X_1, X_2,…,X_n, as if the other variables have been eliminated from the situation.

The partial correlation coefficient is defined in terms of simple correlation coefficients as follows: Let $r_{12.\ 3}$ denote the correlation of X_1 and X_2 by eliminating the effect of X_3. Let r_{12} be the simple correlation coefficient between X_1 and X_2. Let r_{13} be the simple correlation coefficient between X_1 and X_3.

Let r_{23} be the simple correlation coefficient between X_2 and X_3.

Then we have

$$r_{12.3} = \frac{r_{12} - r_{13}\ r_{23}}{\sqrt{(1 - r^2_{13})\ (1 - r^2_{23})}}$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}\ r_{32}}{\sqrt{(1 - r^2_{12})\ (1 - r^2_{32})}}$$

and

$$r_{32.1} = \frac{r_{23} - r_{21}\ r_{13}}{\sqrt{(1 - r^2_{21})\ (1 - r^2_{13})}}$$

Problem 1

Given that $r_{12} = 0.6$, $r_{13} = 0.58$, $r_{23} = 0.70$ determine the partial correlation coefficient $r_{12.3}$

Solution:

We have

$$= \frac{0.6 - 0.58 x 0.70}{\sqrt{(1-(0.58)^2)\ (1-(0.70)^2)}}$$

$$= \frac{0.6 - 0.406}{\sqrt{(1-0.3364)\ (1-0.49)}}$$

$$= \frac{0.194}{\sqrt{0.6636 x\ 0.51}}$$

$$= \frac{0.194}{0.8146 x 0.7141}$$

$$= \frac{0.194}{0.5817}$$

$$= 0.3335$$

Problem 2

If $r_{12} = 0.75$, $r_{13} = 0.80$, $r_{23} = 0.70$, find the partial correlation coefficient $r_{13.2}$

Solution:

We have

$$r_{13.2} = \frac{r_{13} - r_{12}\ r_{32}}{\sqrt{(1-r^2{}_{12})\ (1-r^2{}_{32})}}$$

$$= \frac{0.8 - 0.75 X 0.70}{\sqrt{(1-(0.75)^2)\ (1-(0.70)^2)}}$$

$$= \frac{0.8 - 0.525}{\sqrt{(1-0.5625)\ (1-0.49)}}$$

$$= \frac{0.275}{\sqrt{(0.4375)\ (0.51)}}$$

$$= \frac{0.275}{0.6614 X 0.7141}$$

$$= \frac{0.275}{0.4723}$$

$$= 0.5823$$

Properties of Correlation Coefficient

Let r denote the correlation coefficient between two variables. r≥ is interpreted using the following properties:

1. The value of r ranges from – 1.0 to 0.0 or from 0.0 to 1.0

2. A value of r = 1.0 indicates that there exists perfect positive correlation between the two variables.

3. A value of r = - 1.0 indicates that there exists perfect negative correlation between the two variables.

4. A value r = 0.0 indicates zero correlation i.e., it shows that there is no correlation at all between the two variables.

5. A positive value of r shows a positive correlation between the two variables.

6. A negative value of r shows a negative correlation between the two variables.

7. A value of r = 0.9 and above indicates a very high degree of positive correlation between the two variables.

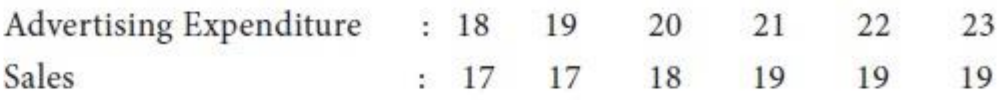
8. A value of - 0.9 ≥ r > - 1.0 shows a very high degree of negative correlation between the two variables.

9. For a reasonably high degree of positive correlation, we require r to be from 0.75 to 1.0.

10. A value of r from 0.6 to 0.75 may be taken as a moderate degree of positive correlation.

Problem 1

The following are data on Advertising Expenditure (in Rupees Thousand) and Sales (Rupees in lakhs) in a company.

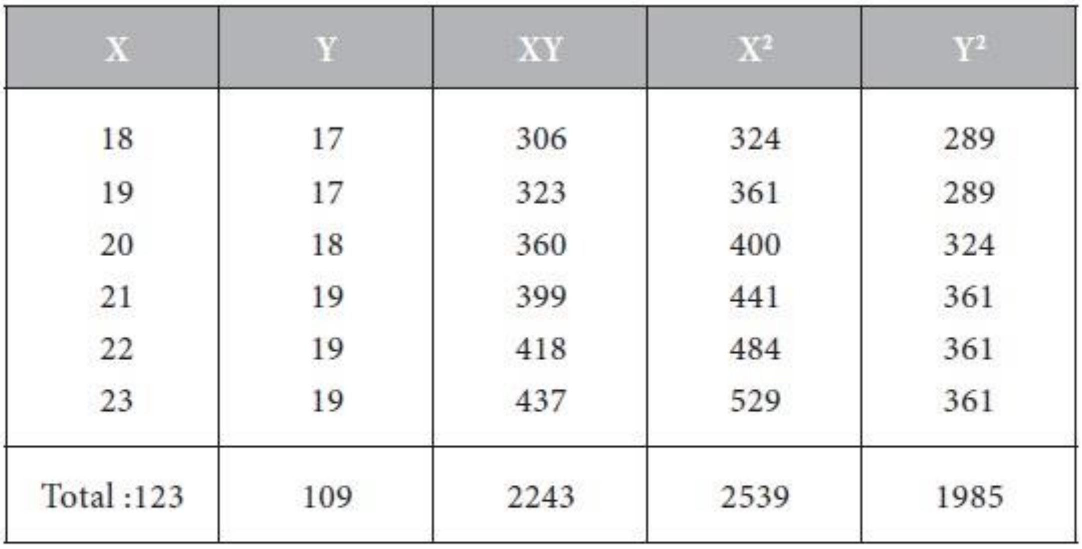


Advertising Expenditure	:	18	19	20	21	22	23
Sales	:	17	17	18	19	19	19

Determine the correlation coefficient between them and interpret the result.

Solution:

We have N = 6. Calculate Σ X, ΣY, ΣXY, ΣY², ΣY² as follows:



X	Y	XY	X^2	Y^2
18	17	306	324	289
19	17	323	361	289
20	18	360	400	324
21	19	399	441	361
22	19	418	484	361
23	19	437	529	361
Total :123	109	2243	2539	1985

The correlation coefficient r between the two variables is calculated as follows:

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r = \frac{6 \times 2243 - 123 \times 109}{\sqrt{6 \times 2539 - (123)^2}\sqrt{6 \times 1985 - (109)^2}}$$

= (13458 – 13407) / {√(15234- 15129) √(11910- 11881)}
=51/{√105 √29} = 51/ (10.247 x 5.365)
= 51/ 54.975
= 0.9277

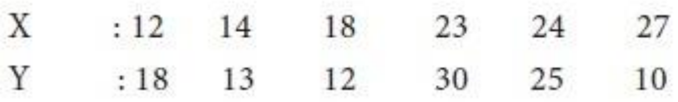
Interpretation

The value of r is 0.92. It shows that there is a high, positive correlation between the two variables ‘Advertising Expenditure’ and ‘Sales’. This provides a basis to consider some functional relationship

between them.

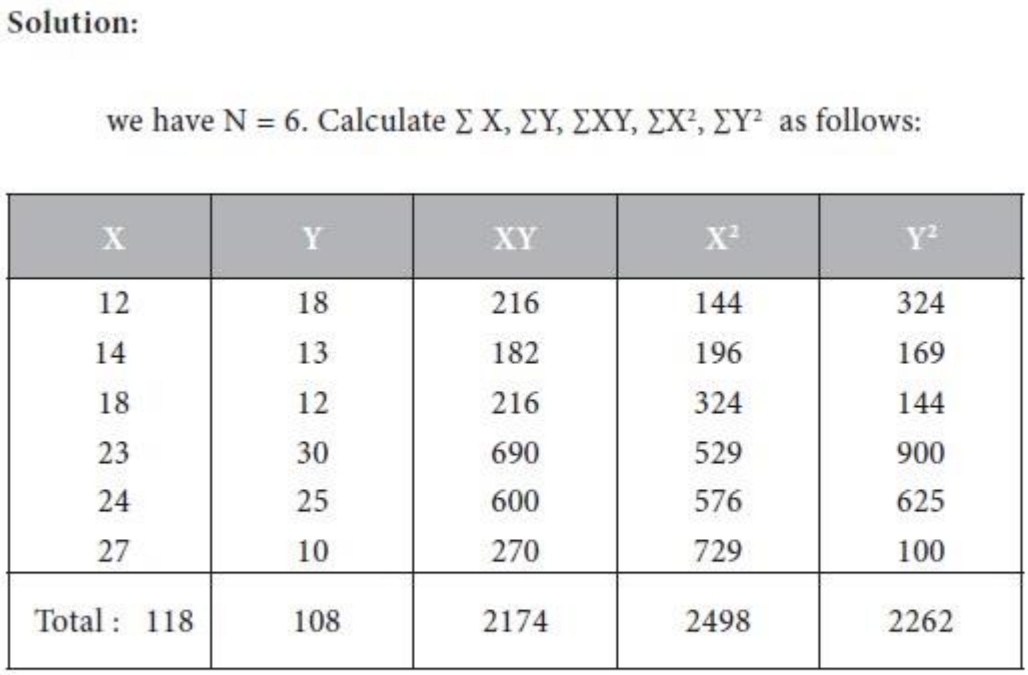
Problem 2

Consider the following data on two variables X and Y.



X	: 12	14	18	23	24	27
Y	: 18	13	12	30	25	10

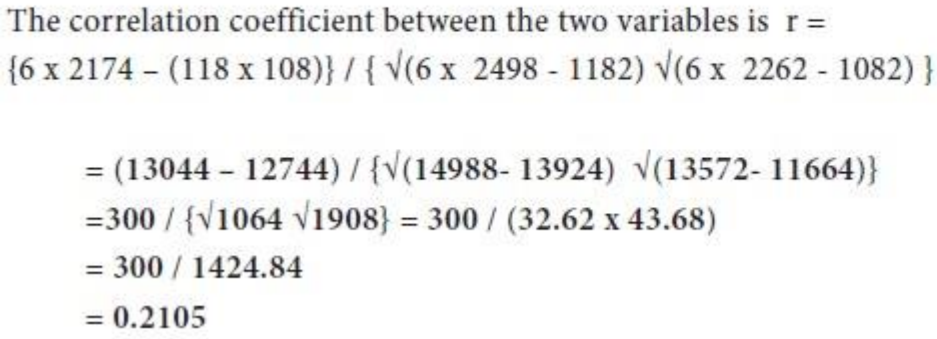
Determine the correlation coefficient between the two variables and interpret the result.



Solution:

we have N = 6. Calculate Σ X, ΣY, ΣXY, ΣX^2, ΣY^2 as follows:

X	Y	XY	X^2	Y^2
12	18	216	144	324
14	13	182	196	169
18	12	216	324	144
23	30	690	529	900
24	25	600	576	625
27	10	270	729	100
Total : 118	108	2174	2498	2262



The correlation coefficient between the two variables is r =
{6 x 2174 – (118 x 108)} / { √(6 x 2498 - 1182) √(6 x 2262 - 1082) }

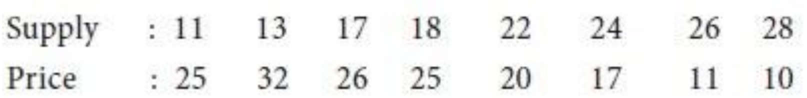
= (13044 – 12744) / {√(14988- 13924) √(13572- 11664)}
=300 / {√1064 √1908} = 300 / (32.62 x 43.68)
= 300 / 1424.84
= 0.2105

Interpretation

The value of r is 0.21. Even though it is positive, the value of r is very less. Hence we conclude that there is no correlation between the two variables X and Y. Consequently we cannot construct any functional relational relationship between them.

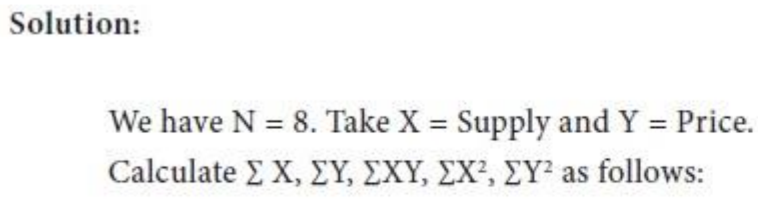
Problem 3

Consider the following data on supply and price. Determine the correlation Coefficient between the two variables and interpret the result.



Supply	: 11	13	17	18	22	24	26	28
Price	: 25	32	26	25	20	17	11	10

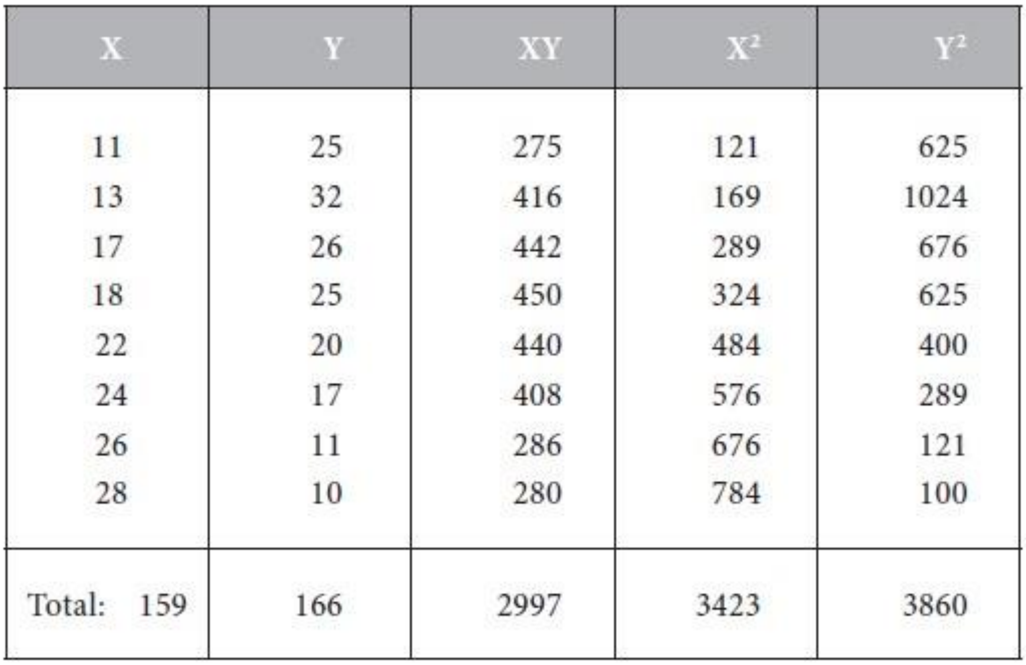
Determine the correlation coefficient between the two variables and interpret the result.



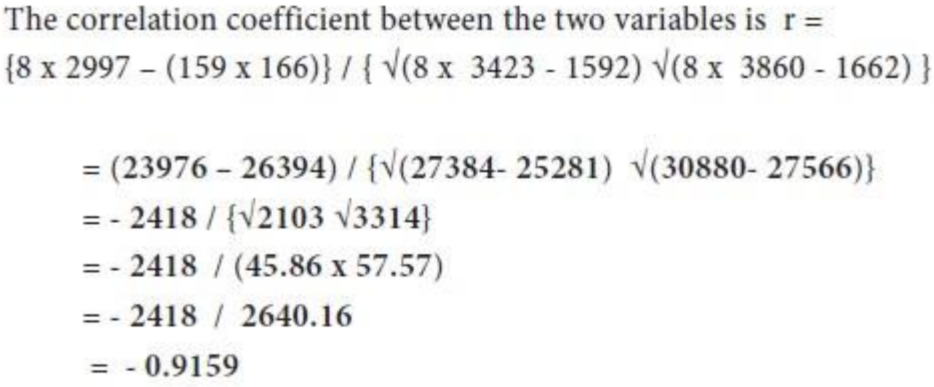
Solution:

We have N = 8. Take X = Supply and Y = Price.

Calculate Σ X, ΣY, ΣXY, ΣX^2, ΣY^2 as follows:



X	Y	XY	X^2	Y^2
11	25	275	121	625
13	32	416	169	1024
17	26	442	289	676
18	25	450	324	625
22	20	440	484	400
24	17	408	576	289
26	11	286	676	121
28	10	280	784	100
Total: 159	166	2997	3423	3860



The correlation coefficient between the two variables is r =
{8 x 2997 – (159 x 166)} / { √(8 x 3423 - 1592) √(8 x 3860 - 1662) }

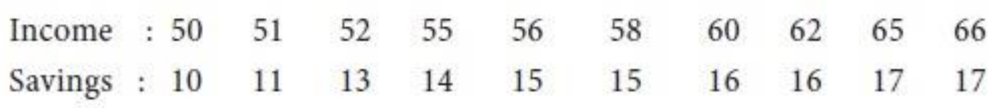
= (23976 – 26394) / {√(27384- 25281) √(30880- 27566)}
= - 2418 / {√2103 √3314}
= - 2418 / (45.86 x 57.57)
= - 2418 / 2640.16
= - 0.9159

Interpretation

The value of r is - 0.92. The negative sign in r shows that the two variables move in opposite directions. The absolute value of r is 0.92 which is very high. Therefore we conclude that there is high negative correlation between the two variables ‘Supply’ and ‘Price’.

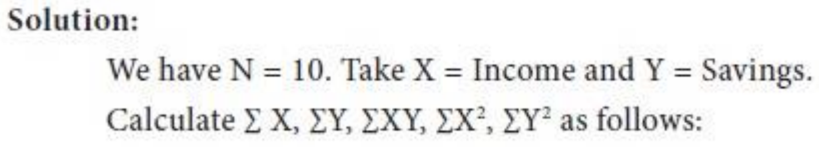
Problem 4

Consider the following data on income and savings in Rs. Thousand.



Income :	50	51	52	55	56	58	60	62	65	66
Savings :	10	11	13	14	15	15	16	16	17	17

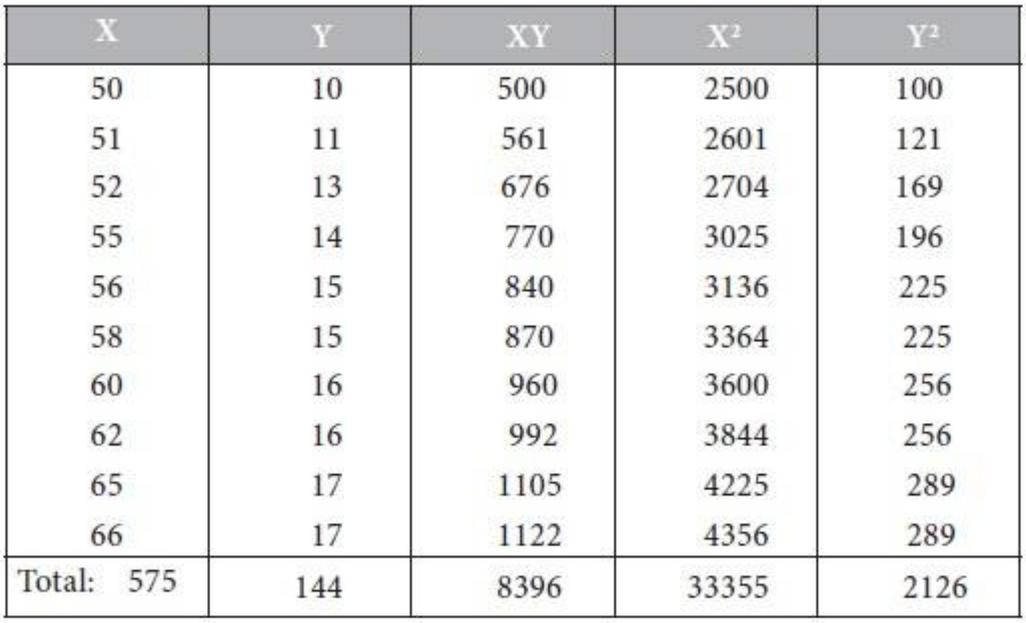
Determine the correlation coefficient between the two variables and interpret the result.



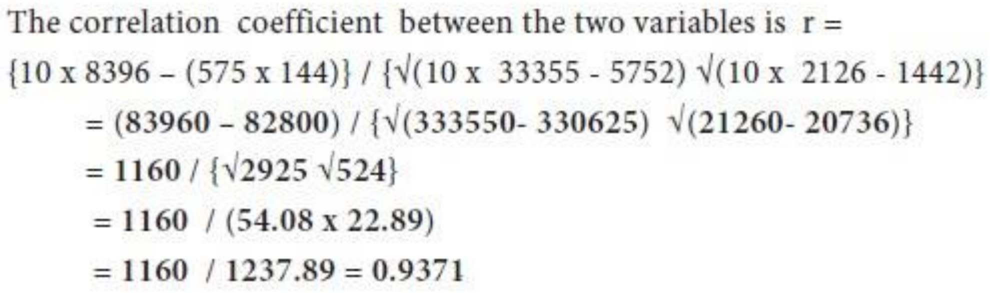
Solution:

We have N = 10. Take X = Income and Y = Savings.

Calculate Σ X, ΣY, ΣXY, ΣX^2, ΣY^2 as follows:



X	Y	XY	X^2	Y^2
50	10	500	2500	100
51	11	561	2601	121
52	13	676	2704	169
55	14	770	3025	196
56	15	840	3136	225
58	15	870	3364	225
60	16	960	3600	256
62	16	992	3844	256
65	17	1105	4225	289
66	17	1122	4356	289
Total: 575	144	8396	33355	2126



The correlation coefficient between the two variables is r =

{10 x 8396 – (575 x 144)} / {√(10 x 33355 - 575²) √(10 x 2126 - 144²)}

= (83960 – 82800) / {√(333550- 330625) √(21260- 20736)}

= 1160 / {√2925 √524}

= 1160 / (54.08 x 22.89)

= 1160 / 1237.89 = 0.9371

Interpretation

The value of r is 0.93. The positive sign in r shows that the two variables move in the same direction. The value of r is very high. Therefore we conclude that there is high positive correlation between the two variables ‘Income’ and ‘Savings’. As a result, we can construct a functional relationship between them.

RANK CORRELATION

Spearman’s rank correlation coefficient

If ranks can be assigned to pairs of observations for two variables X and Y, then the correlation between the ranks is called the **rank correlation coefficient**. It is usually denoted by the **symbol** ρ (rho). It is given by the formula

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N}$$

where

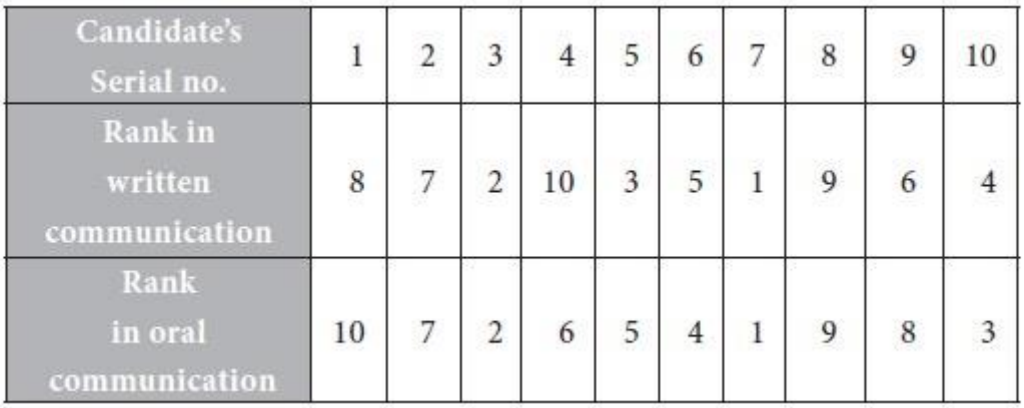
D = difference between the corresponding ranks of X and Y

$= R_X - R_Y$

and N is the total number of pairs of observations of X and Y.

Problem 5

Alpha Recruiting Agency short listed 10 candidates for final selection. They were examined in written and oral communication skills. They were ranked as follows:

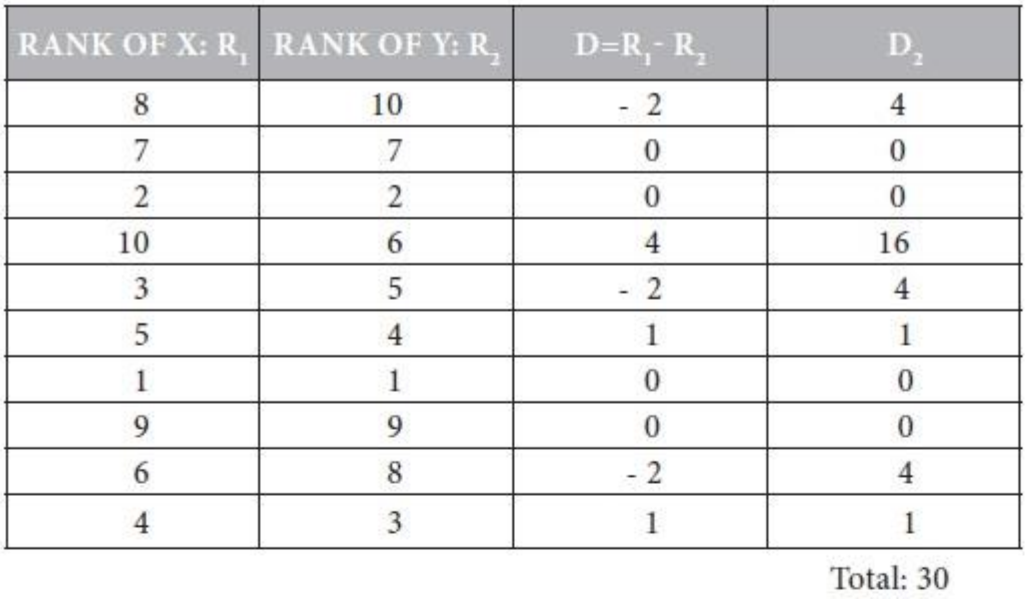


Candidate's Serial no.	1	2	3	4	5	6	7	8	9	10
Rank in written communication	8	7	2	10	3	5	1	9	6	4
Rank in oral communication	10	7	2	6	5	4	1	9	8	3

Find out whether there is any correlation between the written and oral communication skills of the short listed candidates.

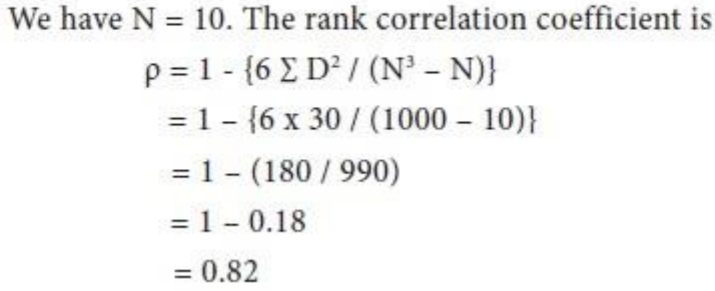
Solution:

Take X = Written Communication Skill and Y = Oral Communication Skill.



RANK OF X: R_1	RANK OF Y: R_2	$D=R_1 - R_2$	D_2
8	10	- 2	4
7	7	0	0
2	2	0	0
10	6	4	16
3	5	- 2	4
5	4	1	1
1	1	0	0
9	9	0	0
6	8	- 2	4
4	3	1	1

Total: 30



We have N = 10. The rank correlation coefficient is

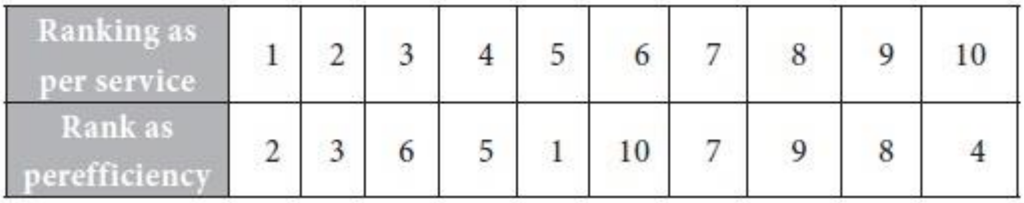
$$\rho = 1 - \{6 \Sigma D^2 / (N^3 - N)\}$$
$$= 1 - \{6 \text{ x } 30 / (1000 - 10)\}$$
$$= 1 - (180 / 990)$$
$$= 1 - 0.18$$
$$= 0.82$$

Inference:

From the value of r, it is inferred that there is a high, positive rank correlation between the written and oral communication skills of the short listed candidates.

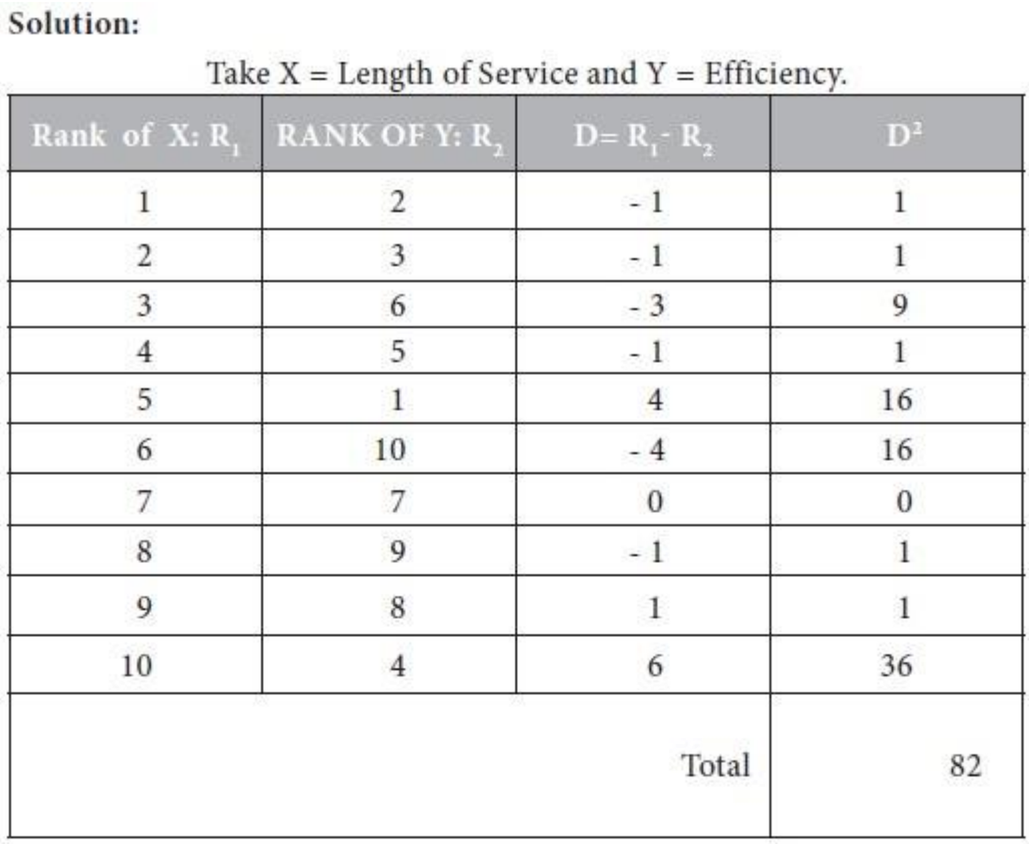
Problem 6

The following are the ranks obtained by 10 workers in abc company on the basis of their length of service and efficiency.



Ranking as per service	1	2	3	4	5	6	7	8	9	10
Rank as perefficiency	2	3	6	5	1	10	7	9	8	4

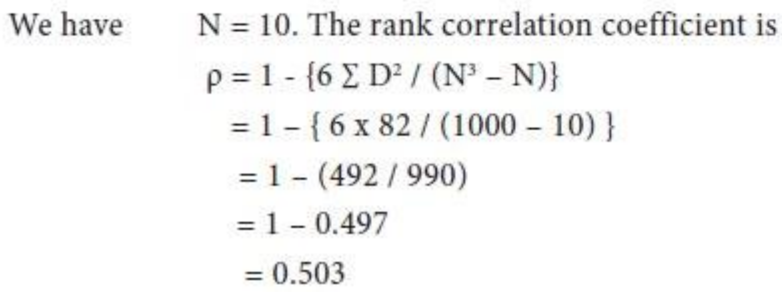
Find out whether there is any correlation between the ranks obtained by the workers as per the two criteria.



Solution:

Take X = Length of Service and Y = Efficiency.

Rank of X: R_1	RANK OF Y: R_2	$D = R_1 - R_2$	D^2
1	2	- 1	1
2	3	- 1	1
3	6	- 3	9
4	5	- 1	1
5	1	4	16
6	10	- 4	16
7	7	0	0
8	9	- 1	1
9	8	1	1
10	4	6	36
		Total	82



We have N = 10. The rank correlation coefficient is

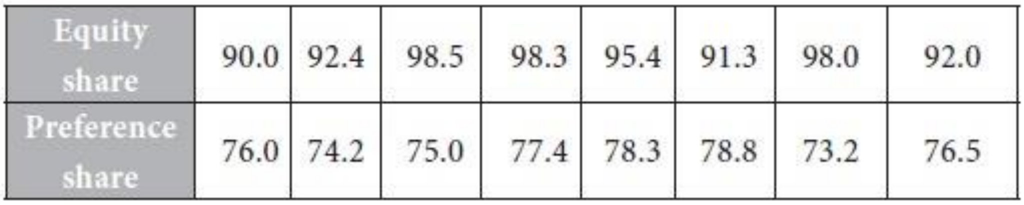
$$\rho = 1 - \{6 \Sigma D^2 / (N^3 - N)\}$$
$$= 1 - \{ 6 \text{ x } 82 / (1000 - 10) \}$$
$$= 1 - (492 / 990)$$
$$= 1 - 0.497$$
$$= 0.503$$

Inference:

The rank correlation coefficient is not high.

Problem 7 (conversion of scores into ranks)

Calculate the rank correlation to determine the relationship between equity shares and preference shares given by the following data on their price.



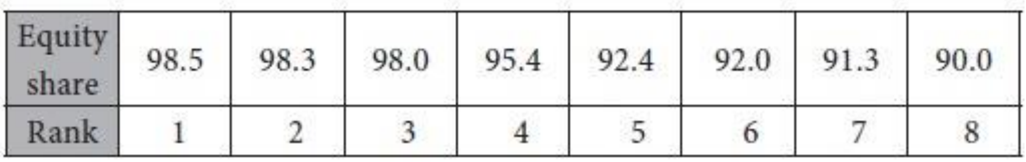
Equity share	90.0	92.4	98.5	98.3	95.4	91.3	98.0	92.0
Preference share	76.0	74.2	75.0	77.4	78.3	78.8	73.2	76.5

Solution:

From the given data on share price, we have to find out the ranks for equity shares and preference shares.

Step 1.

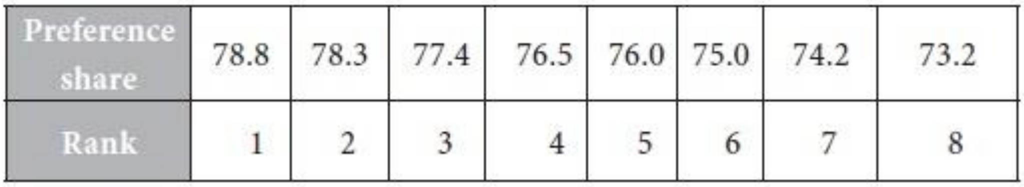
First, consider the equity shares and arrange them in descending order of their price as 1,2,…,8. We have the following ranks:



Equity share	98.5	98.3	98.0	95.4	92.4	92.0	91.3	90.0
Rank	1	2	3	4	5	6	7	8

Step 2.

Next, take the preference shares and arrange them in descending order of their price as 1,2,…,8. We obtain the following ranks:

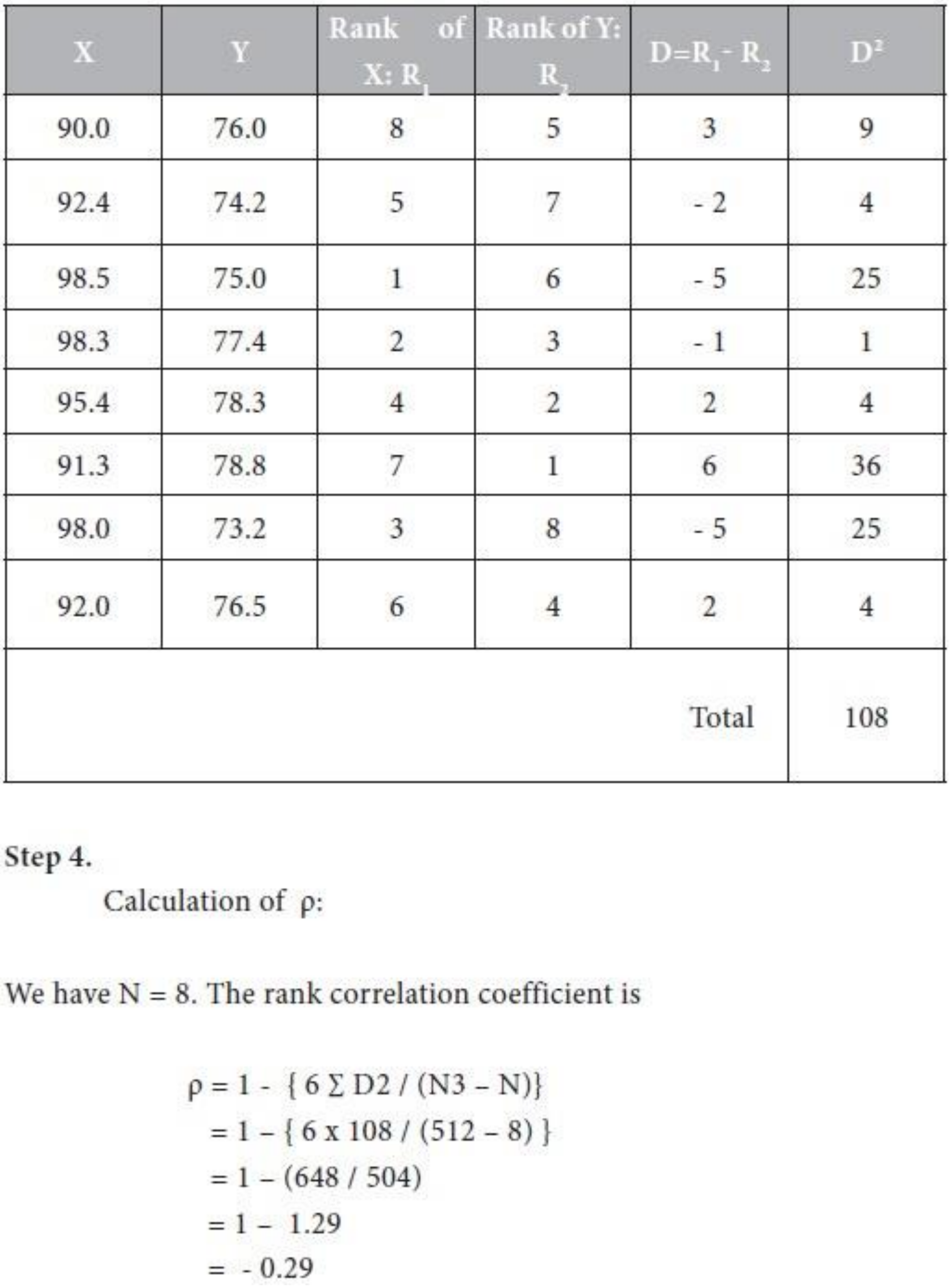


Preference share	78.8	78.3	77.4	76.5	76.0	75.0	74.2	73.2
Rank	1	2	3	4	5	6	7	8

Step 3.

Calculation of D2:

Fit the given data with the correct rank. Take X = Equity share and Y = Preference share. We have the following table:



X	Y	Rank of X: R_1	Rank of Y: R_2	$D=R_1 - R_2$	D^2
90.0	76.0	8	5	3	9
92.4	74.2	5	7	- 2	4
98.5	75.0	1	6	- 5	25
98.3	77.4	2	3	- 1	1
95.4	78.3	4	2	2	4
91.3	78.8	7	1	6	36
98.0	73.2	3	8	- 5	25
92.0	76.5	6	4	2	4
				Total	108

Step 4.

Calculation of ρ:

We have N = 8. The rank correlation coefficient is

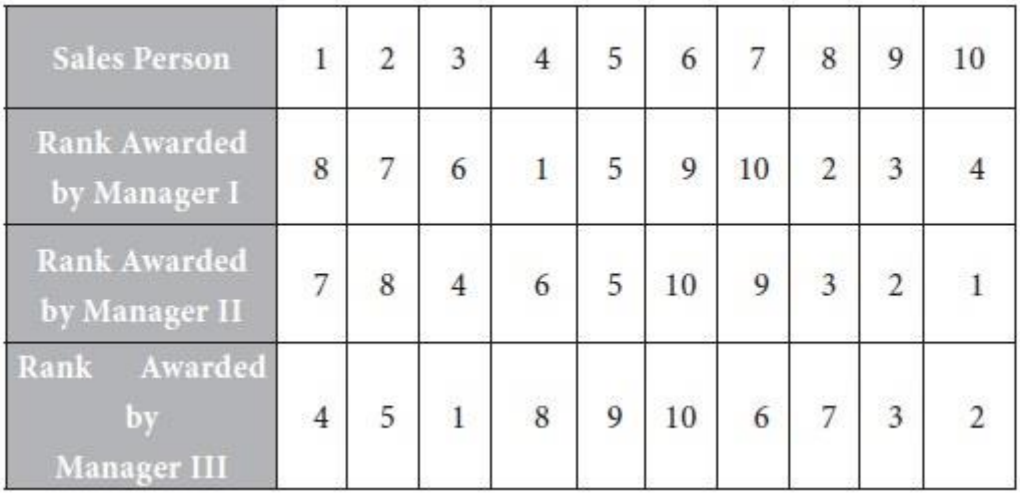
$$\begin{aligned} \rho &= 1 - \{ 6 \Sigma D2 / (N3 - N)\} \\ &= 1 - \{ 6 \text{ x } 108 / (512 - 8) \} \\ &= 1 - (648 / 504) \\ &= 1 - 1.29 \\ &= -0.29 \end{aligned}$$

Inference:

From the value of ρ, it is inferred that the equity shares and preference shares under consideration are negatively correlated. However, the absolute value of ρ is 0.29 which is not even moderate.

Problem 8

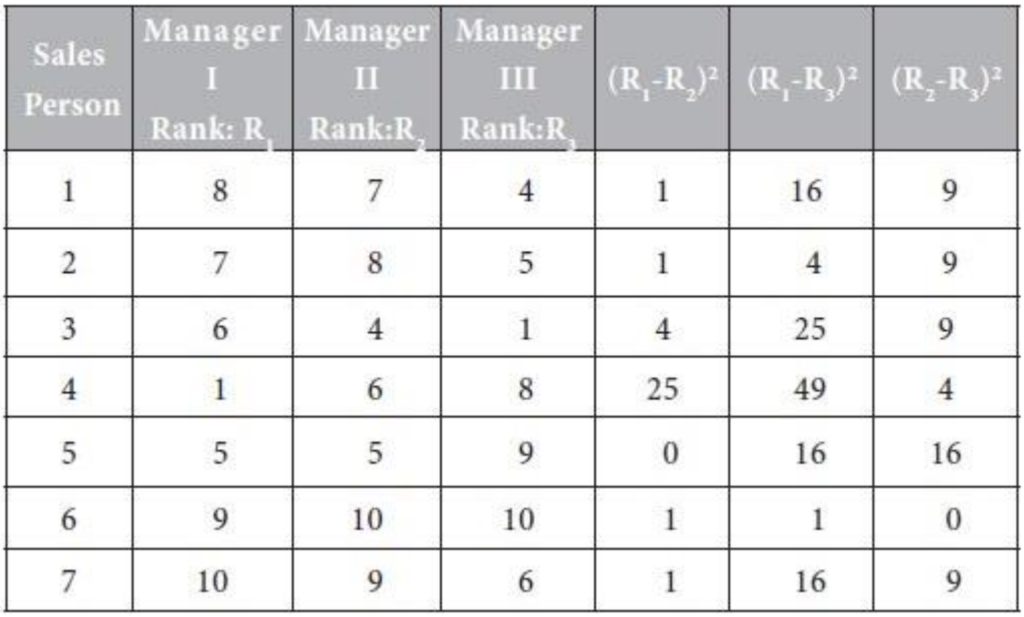
Three managers evaluate the performance of 10 sales persons in an organization and award ranks to them as follows:



Sales Person	1	2	3	4	5	6	7	8	9	10
Rank Awarded by Manager I	8	7	6	1	5	9	10	2	3	4
Rank Awarded by Manager II	7	8	4	6	5	10	9	3	2	1
Rank Awarded by Manager III	4	5	1	8	9	10	6	7	3	2

Determine which two managers have the nearest approach in the evaluation of the performance of the sales persons.

Solution:



Sales Person	Manager I Rank: R_1	Manager II Rank:R_2	Manager III Rank:R_3	$(R_1-R_2)^2$	$(R_1-R_3)^2$	$(R_2-R_3)^2$
1	8	7	4	1	16	9
2	7	8	5	1	4	9
3	6	4	1	4	25	9
4	1	6	8	25	49	4
5	5	5	9	0	16	16
6	9	10	10	1	1	0
7	10	9	6	1	16	9

8	2	3	7	1	25	16
9	3	2	3	1	0	1
10	4	1	2	9	4	1
			Total	44	156	74

We have N = 10. The rank correlation coefficient between mangers I and II is

$$\begin{aligned} \rho &= 1 - \{6 \Sigma D^2 / (N^3 - N)\} \\ &= 1 - \{6 \text{ x } 44 / (1000 - 10)\} \\ &= 1 - (264 / 990) \\ &= 1 - 0.27 \\ &= 0.73 \end{aligned}$$

The rank correlation coefficient between mangers I and III is

$$\begin{aligned} &1 - \{6 \text{ x } 156 / (1000 - 10)\} \\ &= 1 - (936 / 990) \\ &= 1 - 0.95 \\ &= 0.05 \end{aligned}$$

The rank correlation coefficient between mangers II and III is

$$\begin{aligned} &1 - \{6 \text{ x } 74 / (1000 - 10)\} \\ &= 1 - (444 / 990) \\ &= 1 - 0.44 \\ &= 0.56 \end{aligned}$$

Inference:

Comparing the 3 values of ρ, it is inferred that Mangers I and ii have the nearest approach in the evaluation of the performance of the sales persons.

Repeated values: Resolving ties in ranks

When ranks are awarded to candidates, it is possible that certain candidates obtain equal ranks. For example, if two or three, or four candidates secure equal ranks, a procedure that can be followed to resolve the ties is described below.

We follow the **Average Rank Method**. If there are n items, arrange them in ascending order or descending order and give ranks 1, 2, 3, …, n. Then look at those items which have equal values. For such items, take the average ranks.

If there are two items with equal values, their ranks will be two consecutive integers, say s and s + 1. Their average is { s + (s+1)} / 2. Assign this rank to both items. Note that we allow ranks to be fractions also.

If there are three items with equal values, their ranks will be three consecutive integers, say s, s + 1 and s + 2. Their average is { s + (s+1) (s+2) } / 3 = (3s + 3) / 3 = s + 1. Assign this rank to all the three items. A similar procedure is followed if four or more number of items has equal values.

Correction term for ρ when ranks are tied

Consider the formula for rank correlation coefficient. We have

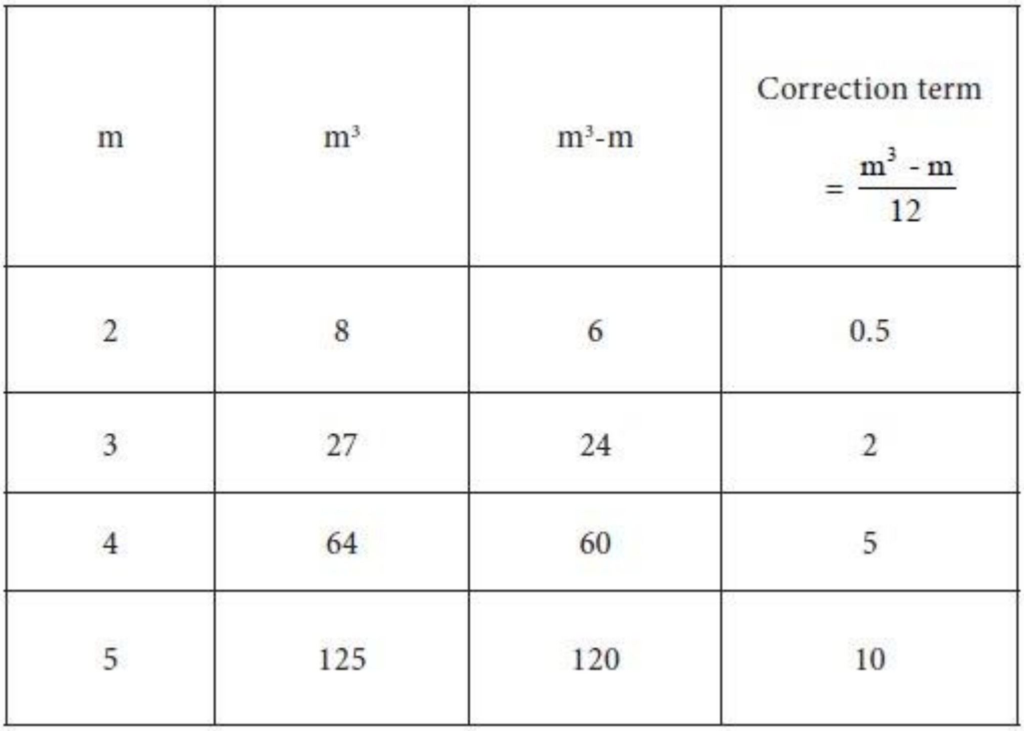
$$\rho = 1 - \frac{6\sum D^2}{N^3 - N}$$

If there is a tie involving m items, we have to add

$$\frac{m^3 - m}{12}$$

to the term D2 in ρ. We have to add as many terms like (m3 – m) / 12 as there are ties.

Let us calculate the correction terms for certain values of m. These are provided in the following table.



m	m^3	m^3-m	Correction term $= \frac{m^3 - m}{12}$
2	8	6	0.5
3	27	24	2
4	64	60	5
5	125	120	10

Illustrative examples:

If there is a tie involving 2 items, then the correction term is 0.5

If there are 2 ties involving 2 items each, then the correction term is 0.5 + 0.5 = 1

If there are 3 ties with 2 items each, then the correction term is 0.5 + 0.5 + 0.5 = 1.5

If there is a tie involving 3 items, then the correction term is 2

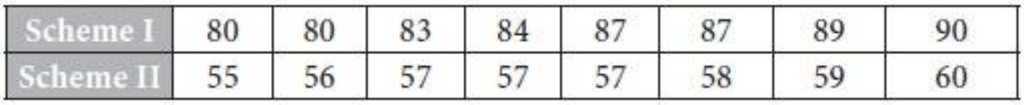
If there are 2 ties involving 3 items each, then the correction term is 2 + 2 = 4

If there is a tie with 2 items and another tie with 3 items, then the correction term is 0.5 + 2 = 2.5

If there are 2 ties with 2 items each and another tie with 3 items, then the correction term is 0.5 + 0.5 + 2 = 3

Problem 9 : ***Resolving ties in ranks***

The following are the details of ratings scored by two popular insurance schemes. Determine the rank correlation coefficient between them.



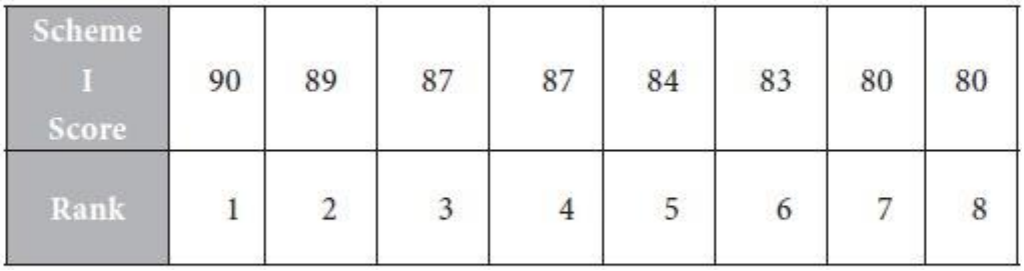
Scheme I	80	80	83	84	87	87	89	90
Scheme II	55	56	57	57	57	58	59	60

Solution:

From the given values, we have to determine the ranks.

Step 1.

Arrange the scores for Insurance Scheme I in descending order and rank them as 1,2,3,…,8.

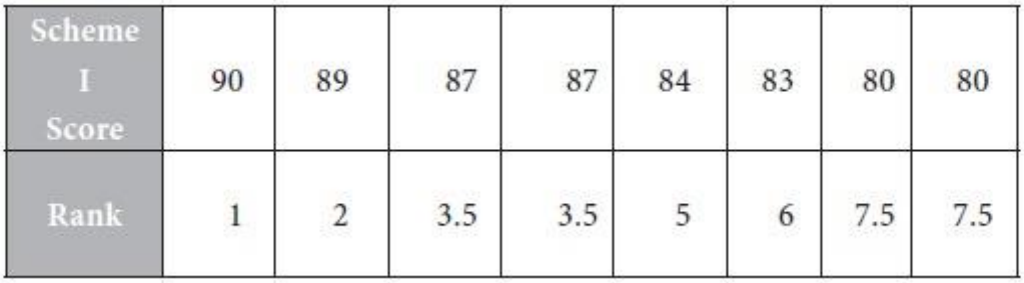


Scheme I Score	90	89	87	87	84	83	80	80
Rank	1	2	3	4	5	6	7	8

The score 87 appears twice. The corresponding ranks are 3, 4. Their average is (3 + 4) / 2 = 3.5. Assign this rank to the two equal scores in Scheme I.

The score 80 appears twice. The corresponding ranks are 7, 8. Their average is (7 + 8) / 2 = 7.5. Assign this rank to the two equal scores in Scheme I.

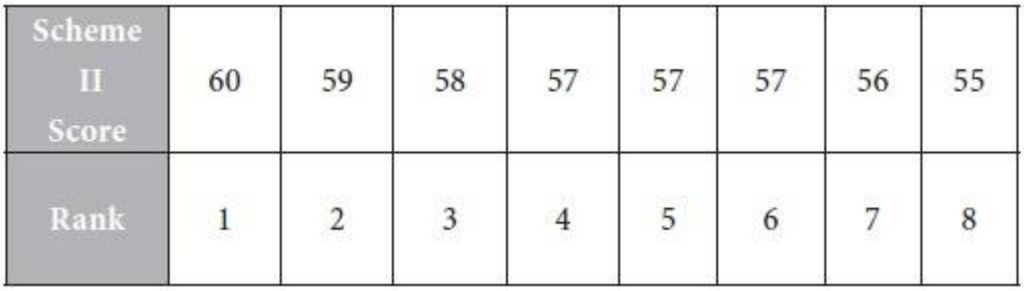
The revised ranks for Insurance Scheme I are as follows:



Scheme I Score	90	89	87	87	84	83	80	80
Rank	1	2	3.5	3.5	5	6	7.5	7.5

Step 2.

Arrange the scores for Insurance Scheme II in descending order and rank them as 1,2,3,…,8.

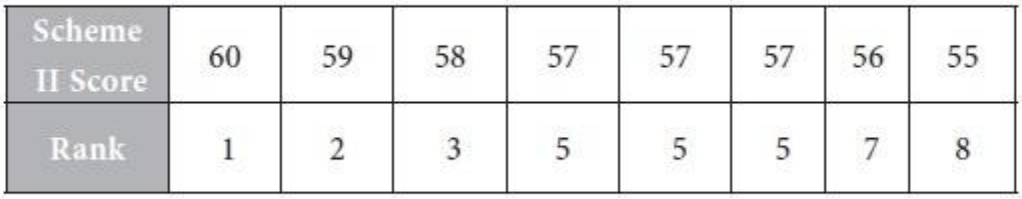


Scheme II Score	60	59	58	57	57	57	56	55
Rank	1	2	3	4	5	6	7	8

The score 57 appears thrice. The corresponding ranks are 4, 5, 6.

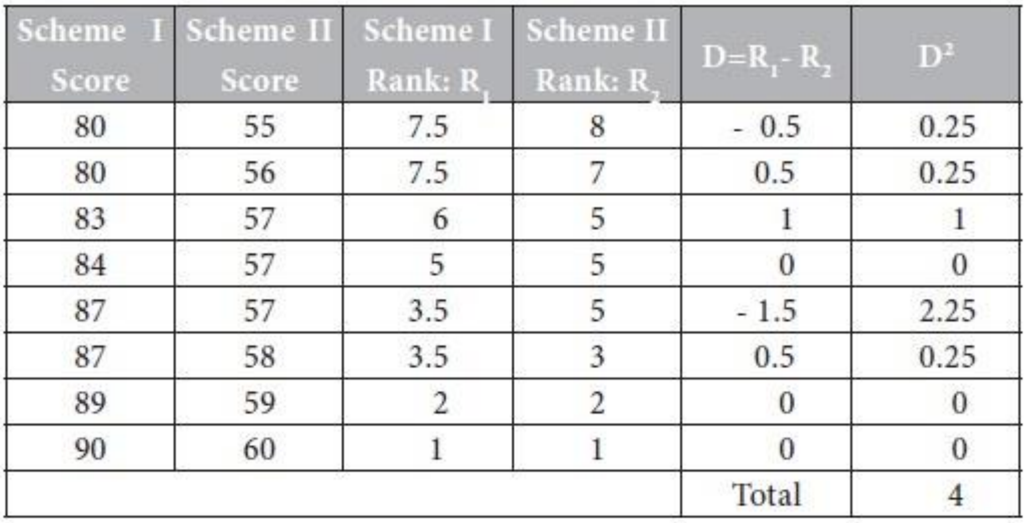
Their average is (4 + 5 + 6) / 3 = 15 / 3 = 5. Assign this rank to the three equal scores in Scheme II.

The revised ranks for Insurance Scheme II are as follows:

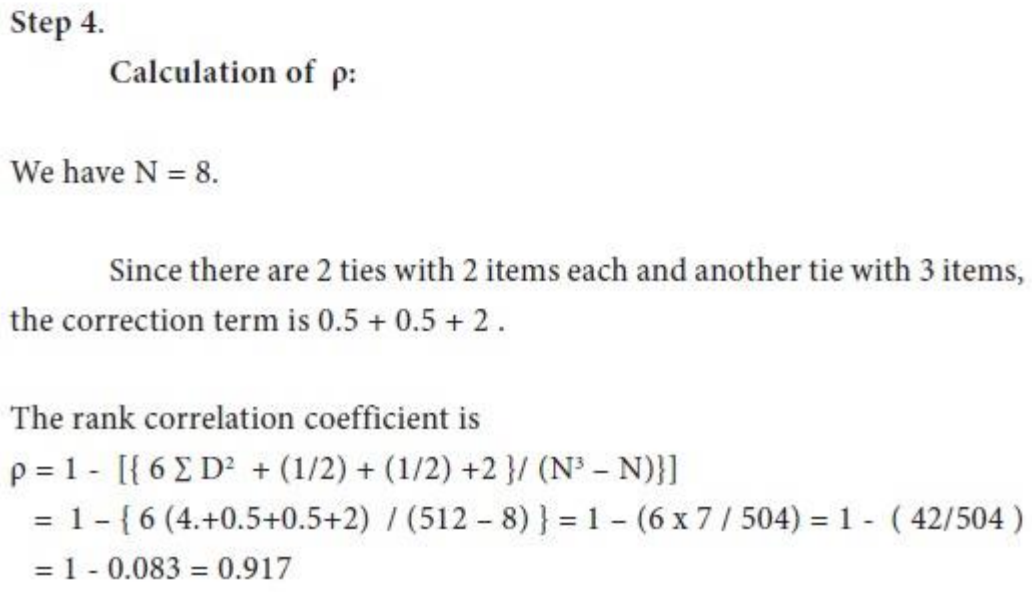


Scheme II Score	60	59	58	57	57	57	56	55
Rank	1	2	3	5	5	5	7	8

Step 3. Calculation of D2: Assign the revised ranks to the given pairs of values and calculate D2 as follows:



Scheme I Score	Scheme II Score	Scheme I Rank: R_1	Scheme II Rank: R_2	$D=R_1- R_2$	D^2
80	55	7.5	8	- 0.5	0.25
80	56	7.5	7	0.5	0.25
83	57	6	5	1	1
84	57	5	5	0	0
87	57	3.5	5	- 1.5	2.25
87	58	3.5	3	0.5	0.25
89	59	2	2	0	0
90	60	1	1	0	0
				Total	4



Step 4.

Calculation of ρ:

We have N = 8.

Since there are 2 ties with 2 items each and another tie with 3 items, the correction term is 0.5 + 0.5 + 2 .

The rank correlation coefficient is

$\rho = 1 - [\{ 6 \Sigma D^2 + (1/2) + (1/2) +2 \}/ (N^3 – N)\}]$

$= 1 – \{ 6 (4.+0.5+0.5+2) / (512 – 8) \} = 1 – (6 \times 7 / 504) = 1 - (42/504)$

$= 1 - 0.083 = 0.917$

Inference:

It is inferred that the two insurance schemes are highly, positively correlated.

REGRESSION ANALYSIS

In the pairs of observations, if there is a cause and effect relationship between the variables X and Y, then the average relationship between these two variables is called regression, which means “stepping back” or “return to the average”. The linear relationship giving the best mean value of a variable corresponding to the other variable is called a **regression line or line of the best fit**. The regression of X on Y is different from the regression of Y on X. Thus, there are two equations of regression and the two regression lines are given as follows:

Regression of Y on X: $Y - \overline{Y} = b_{yx}(X - \overline{X})$

Regression of X on Y: $X - \overline{X} = b_{xy}(Y - \overline{Y})$

Where $\overline{X}$, $\overline{Y}$ are the means of X, Y respectively.

Normal equations

Suppose we have to fit a straight line to the n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n , y_n)$. Suppose the equation of straight line finally comes as

Y = a + b X ....(1)

Where

a, b are constants to be determined. Mathematically speaking, when we require finding the equation of a straight line, two distinct points on the straight line are sufficient. However, a different approach is followed here. We want to include all the observations in our attempt to build a straight line. Then all the n observed points (x, y) are required to satisfy the relation

(1). Consider the summation of all such terms. We get

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

i.e.

$$\Sigma y = an + b (\Sigma x) \qquad (2)$$

To find two quantities a and b, we require two equations. We have obtained one equation i.e., (2). We need one more equation. For this purpose, multiply both sides of (1) by

Result:

Let σ_x, σ_y denote the standard deviations of x, y respectively. We have the following result.

$$b_{yx} = r\frac{\sigma_Y}{\sigma_X} \quad and \quad b_{xy} = r\frac{\sigma_X}{\sigma_Y}$$

$$\therefore r^2 = b_{yx}b_{xy} \quad and\ so \quad r = \sqrt{b_{yx}b_{xy}}$$

Result:

The coefficient of correlation r between X and Y is the square root of the product of the b values in the two regression equations. We can find r by this way also.

Application

The method of regression is very much useful for business forecasting.

x. We obtain

$x\,y = ax + bx^2$.

Consider the summation of all such terms. We get

$\Sigma\, x\, y = \Sigma\,(ax + bx^2\,) = (\Sigma\, a\, x) + (\,\Sigma\, bx^2)$

i.e.,

$$\Sigma\, x\, y = a\,(\Sigma\, x\,) + b\,(\Sigma\, x^2) \;\ldots\ldots\ldots\ldots\ldots\;(3)$$

Equations (2) and (3) are referred to as the normal equations associated with the regression of y on x. Solving these two equations, we obtain

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n \sum X^2 - (\sum X)^2}$$

$$\text{and} \quad b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

Note:

For calculating the coefficient of correlation,

we require $\Sigma X, \Sigma Y, \Sigma\, Xy, \Sigma\, X^2, \Sigma Y^2$.

For calculating the regression of y on x, we require $\Sigma X, \Sigma Y, \Sigma\, Xy, \Sigma\, X^2$: Thus, tabular column is same in both the cases with the difference that ΣY^2. is also required for the coefficient of correlation.

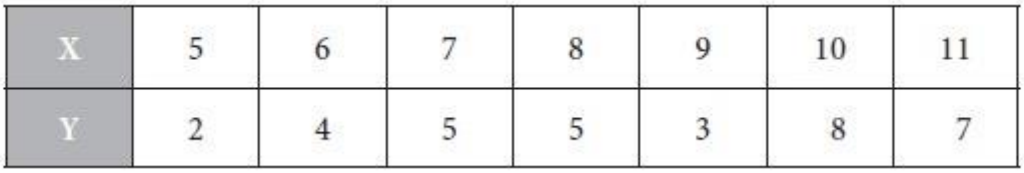
Next, if we consider the regression line of x on y, we get the equation X = a + b y. The expressions for the coefficients can be got by interchanging the roles of X and Y in the previous discussion. Thus, we obtain

$$a = \frac{\sum Y^2 \sum X - \sum Y \sum XY}{n \sum Y^2 - (\sum Y)^2}$$

$$\text{And} \quad b = \frac{n \sum XY - \sum X \sum Y}{n \sum Y^2 - (\sum Y)^2}$$

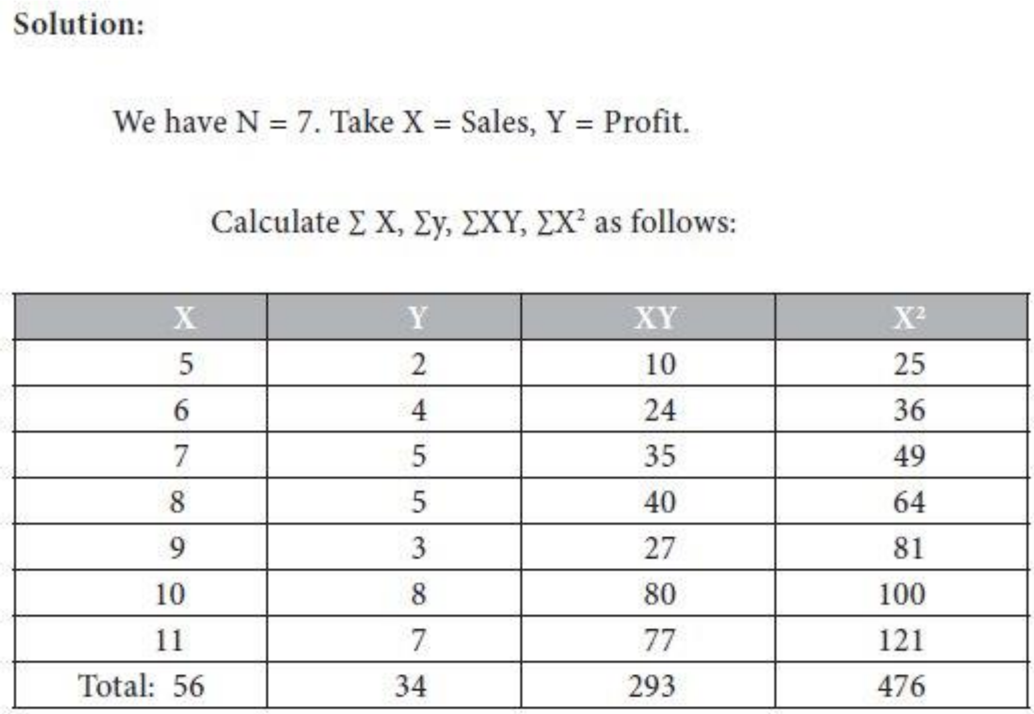
Problem 1

Consider the following data on sales and profit.



X	5	6	7	8	9	10	11
Y	2	4	5	5	3	8	7

Determine the regression of profit on sales.

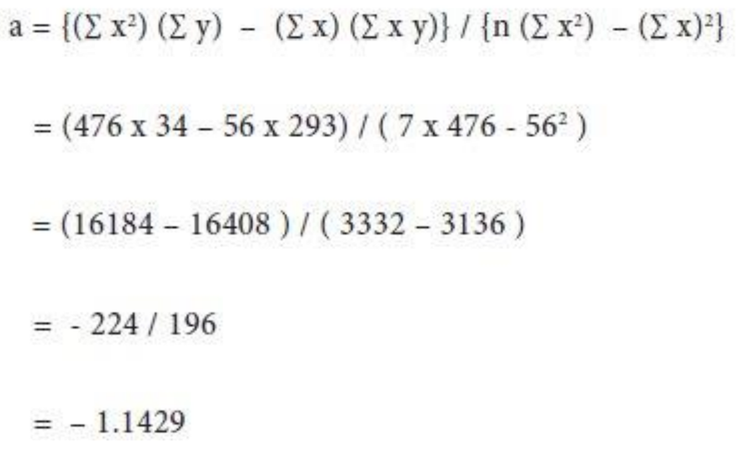


Solution:

We have N = 7. Take X = Sales, Y = Profit.

Calculate Σ X, Σy, ΣXY, ΣX² as follows:

X	Y	XY	X^2
5	2	10	25
6	4	24	36
7	5	35	49
8	5	40	64
9	3	27	81
10	8	80	100
11	7	77	121
Total: 56	34	293	476



$$a = \{(\Sigma x^2)(\Sigma y) - (\Sigma x)(\Sigma x y)\} / \{n(\Sigma x^2) - (\Sigma x)^2\}$$

$$= (476 \times 34 - 56 \times 293) / (7 \times 476 - 56^2)$$

$$= (16184 - 16408) / (3332 - 3136)$$

$$= -224 / 196$$

$$= -1.1429$$

$$b = \{n\ (\Sigma\ x\ y)\ -\ (\Sigma\ x)\ (\Sigma\ y)\}\ /\ \{n\ (\Sigma\ x^2)\ -\ (\Sigma\ x)\ ^2\}$$

$$= (7 \times 293 - 56 \times 34)/\ 196 = (2051 - 1904)/\ 196$$

$$= 147\ /196$$

$$= 0.75$$

The regression of Y on X is given by the equation

$$Y = a + b\ X$$

I.e.,

$$Y = -\ 1.14 + 0.75\ X$$

Problem 2

The following are the details of income and expenditure of 10 households.

Income	40	70	50	60	80	50	90	40	60	60
Expenditure	25	60	45	50	45	20	55	30	35	30

Determine the regression of expenditure on income and estimate the expenditure when the income is 65.

Solution:

We have N = 10. Take X = Income, Y = Expenditure

Calculate Σ X, Σy, ΣXy, ΣX^2 as follows:

X	Y	XY	X^2
40	25	1000	1600
70	60	4200	4900
50	45	2250	2500

60	50	3000	3600
80	45	3600	6400
50	20	1000	2500
90	55	4950	8100
40	30	1200	1600
60	35	2100	3600
60	30	1800	3600
Total: 600	395	25100	38400

$$a = \{(\Sigma x^2)(\Sigma y) - (\Sigma x)(\Sigma xy)\} / \{n(\Sigma x^2) - (\Sigma x)^2\}$$

$$= (38400 \text{ x } 395 - 600 \text{ x } 25100) / (10 \text{ x } 38400 - 600^2)$$

$$= (15168000 - 15060000) / (384000 - 360000)$$

$$= 108000 / 24000$$

$$= 4.5$$

$$b = \{n(\Sigma xy) - (\Sigma x)(\Sigma y)\} / \{n(\Sigma x^2) - (\Sigma x)^2\}$$

$$= (10 \text{ x } 25100 - 600 \text{ x } 395) / 24000$$

$$= (251000 - 237000) / 24000$$

$$= 14000 / 24000$$

$$= 0.58$$

The regression of y on x is given by the equation

$$Y = a + b X$$

i.e.,

$$Y = 4.5 + 0.583 \ X$$

To estimate the expenditure when income is 65:

Take X = 65 in the above equation. Then we get

$$Y = 4.5 + 0.583 \times 65$$

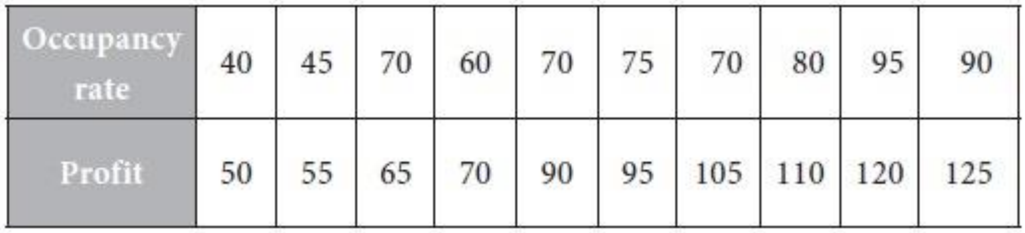
$$= 4.5 + 37.895$$

$$= 42.395$$

$$= 42 \text{ (approximately).}$$

Problem 3

Consider the following data on occupancy rate and profit of a hotel.



Occupancy rate	40	45	70	60	70	75	70	80	95	90
Profit	50	55	65	70	90	95	105	110	120	125

Determine the regressions of

(i) profit on occupancy rate and

(ii) occupancy rate on profit.

Solution:

We have N = 10. Take X = Occupancy Rate, Y = Profit.

Note that in Problems 10 and 11, we wanted only one regression line and so we did not take ∑Y2 . Now we require two regression lines. Therefore, Calculate Σ X, ΣY, ΣXY, ΣX^2, ΣY^2.

X	Y	XY	X2	Y2
40	50	2000	1600	2500
45	55	2475	2025	3025
70	65	4550	4900	4225
60	70	4200	3600	4900
70	90	6300	4900	8100
75	95	7125	5625	9025
70	105	7350	4900	11025
80	110	8800	6400	12100
95	120	11400	9025	14400
90	125	11250	8100	15625
Total: 695	885	65450	51075	84925

The regression line of Y on X:

$$Y = a + b X$$

Where

$$a = \{(\Sigma x^2)(\Sigma y) - (\Sigma x)(\Sigma x y)\} / \{n (\Sigma x^2) - (\Sigma x)^2\}$$

and

$$b = \{n (\Sigma x y) - (\Sigma x)(\Sigma y)\} / \{n (\Sigma x^2) - (\Sigma x)^2\}$$

We obtain

$$a = (51075 \text{ x } 885 - 695 \text{ x } 65450) / (10\text{x}51075 - 695^2)$$

$$= (45201375 - 45487750) / (510750 - 483025)$$

$$= -286375 / 27725$$

$$= -10.329$$

$$b = (10 \text{ x } 65450 - 695 \text{ x } 885) / 27725$$

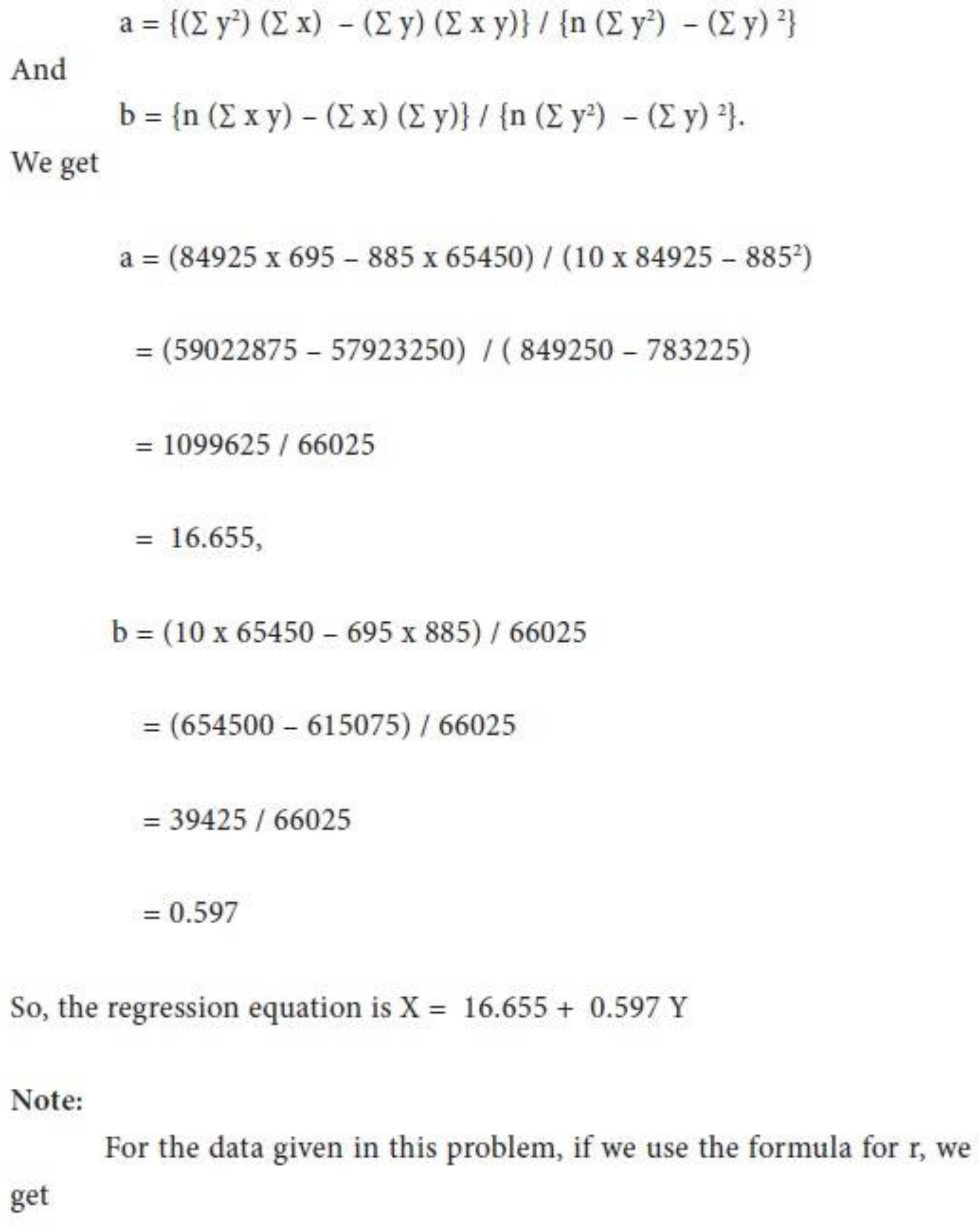
$$= (654500 - 615075) / 27725$$

$$= 39425 / 27725$$

$$= 1.422$$

So, the regression equation is Y = - 10.329 + 1.422 X

Next, if we consider **the regression line of X on Y,** We get the equation X = a + b Y where



a = {(Σ y^2) (Σ x) – (Σ y) (Σ x y)} / {n (Σ y^2) – (Σ y) ^{2}}

And

b = {n (Σ x y) – (Σ x) (Σ y)} / {n (Σ y^2) – (Σ y) ^{2}}.

We get

a = (84925 x 695 – 885 x 65450) / (10 x 84925 – 885^2)

= (59022875 – 57923250) / (849250 – 783225)

= 1099625 / 66025

= 16.655,

b = (10 x 65450 – 695 x 885) / 66025

= (654500 – 615075) / 66025

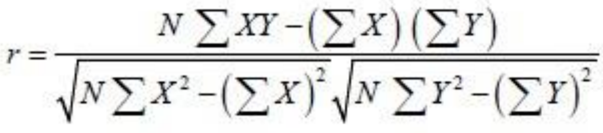
= 39425 / 66025

= 0.597

So, the regression equation is X = 16.655 + 0.597 Y

Note:

For the data given in this problem, if we use the formula for r, we get



$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$= (10 \times 65450 - 695 \times 885) / \{ \sqrt{(10 \times 51075 - 695^2)} \sqrt{(10 \times 84925 - 885^2)} \}$

$= (654500 - 615075) / (\sqrt{27725} \sqrt{66025})$

$= 39425 / 166.508 \times 256.95$

$= 39425 / 42784.23$

$= 0.9214$

However, once we know the two b values, we can find the coefficient of correlation r between X and Y as the square root of the product of the two b values.Thus we obtain

$r = \sqrt{(1.422 \times 0.597)}$
$= \sqrt{0.848934}$
$= 0.9214.$

Note that this agrees with the above value of r.

Decision making under uncertainty

“A decision is the is a conclusion of a process by which one choices between two or more available courses of action for the purpose of attaining a goal”

- A decision an act of choice where in a manager forms a conclusion about what must be done under a given situation. And decision making is a process to arrive at a decision , The process by witch an individual or organization choesse on position or action from many alternatives.
- Three aspects of human behavior are involved in decision making.
- Cognition-Activities of mind associated with knowledge
- Conation- The action of mind implied by such word as willing , desire and aversion.
- Affection- The aspect of mind associated with emotion , feeling , mood and temperament.
- There are various conditions or environment under which the decisions are made .
- Certainty
- Uncertainty
- Risk
- Conflict or competition
- decision making under Uncertainty- When a decision involves condition about which the manager has no information , either about the outcome or the relative chances or any single outcome, he is said to be operating under conditions of uncertainty. Because the manager does not have any information on which he can develop any analysis, the best he can do is to be aware that he has no opportunity of predicting the events. Under these

condition, a number of different criteria have been proposed as possible bases for decision-making.These are as follows.

- Maximizing the maximum possible payoff- the maximum criterion(optimistic).
- Maximizing the minimum possible payoff- the maximum criterion(pessimistic).
- Minimizing the maximum possible regret to the decision maker- The minimax criterion(regret).
- Assuming equally likely probabilities for the occurance of each possible state of nature- The insufficient Criterion(insufficient reasoning).

Decision making under uncertainty

- Maximax criterion – this decision criterion is applied by the most optimist decision maker when he thinks optimistically about the happening of events affecting decision .If this philosophy is followed , the manager will chose that alternative under which it is possible to receive the most favourable pay-off.
- Maximin criterion- this criterion is adopted by the most pessimistic decision maker. The manager believes that worst possible may take place. This pessimism cause the selection of that alternative which maximizes the least favourable payoff.
- Minimax criterion- minimax criterion leads to the minimization of regret. The managerial regret is define as the pay off for each alternative under state of nature of compitative action subtracted from the most favourable payoff which is possible with the happening of the particular event. When manager chooses an alternatives and when a state of nature takes place which does not result in the most favourable payoff, regret takes place.
- Insufficient reason criterion- this proceeding decision criteria assume that without any experience, it is not possible or worthwhile to allocate any probability to the state of situation. In this case also, probability can allocated through there is no criterion for allocating the probability. The situation is refferd to as insufficient reason criterion or la place criterion.

Non-parametric Test

Non-parametric tests are experiments that do not require the underlying population for assumptions. It does not rely on any data referring to any particular parametric group of probability distributions. Non-parametric methods are also called distribution-free tests since they do not have any underlying population. In this article, we will discuss what a non-parametric test is, different methods, merits, demerits and examples of non-parametric testing methods.

What is a Non-parametric Test?

Non-parametric tests are the mathematical methods used in statistical hypothesis testing, which do not make assumptions about the frequency distribution of variables that are to be evaluated. The non-parametric experiment is used when there are skewed data, and it comprises techniques that do not depend on data pertaining to any particular distribution.

The word non-parametric does not mean that these models do not have any parameters. The fact is, the characteristics and number of parameters are pretty flexible and not predefined. Therefore, these models are called distribution-free models.

Non-Parametric T-Test

Whenever a few assumptions in the given population are uncertain, we use non-parametric tests, which are also considered parametric counterparts. When data are not distributed normally or when they are on an ordinal level of measurement, we have to use non-parametric tests for analysis. The basic rule is to use a parametric t-test for normally distributed data and a non-parametric test for skewed data.

Non-Parametric Paired T-Test

The paired sample t-test is used to match two means scores, and these scores come from the same group. Pair samples t-test is used when variables are independent and have two levels, and those levels are repeated measures.

- Difference Between Parametric And Nonparametric
- T-test Formula
- Hypothesis Testing Formula
- Chi-Square Test

Non-parametric Test Methods

The four different techniques of parametric tests, such as Mann Whitney U test, the sign test, the Wilcoxon signed-rank test, and the Kruskal Wallis test are discussed here in detail. We know that the non-parametric tests are completely based on the ranks, which are assigned to the ordered data. The four different types of non-parametric test are summarized below with their uses, null hypothesis, test statistic, and the decision rule.

Kruskal Wallis Test

Kruskal Wallis test is used to compare the continuous outcome in greater than two independent samples.

Null hypothesis, H_0: K Population medians are equal.

Test statistic:

If N is the total sample size, k is the number of comparison groups, Rj is the sum of the ranks in the jth group and nj is the sample size in the jth group, then the test statistic, H is given by:

H=(12N(N+1)∑j=1kRj2nj)−3(N+1)

Decision Rule: Reject the null hypothesis H_0 if H ≥ critical value

Sign Test

The sign test is used to compare the continuous outcome in the paired samples or the two matches samples.

Null hypothesis, H_0: Median difference should be zero

Test statistic: The test statistic of the sign test is the smaller of the number of positive or negative signs.

Decision Rule: Reject the null hypothesis if the smaller of number of the positive or the negative signs are less than or equal to the critical value from the table.

Mann Whitney U Test

Mann Whitney U test is used to compare the continuous outcomes in the two independent samples.

Null hypothesis, H_0: The two populations should be equal.

Test statistic:

If R_1 and R_2 are the sum of the ranks in group 1 and group 2 respectively, then the test statistic "U" is the smaller of:

$U1=n1n2+n1(n1+1)2-R1$

$U2=n1n2+n2(n2+1)2-R2$

Decision Rule: Reject the null hypothesis if the test statistic, U is less than or equal to critical value from the table.

Wilcoxon Signed-Rank Test

Wilcoxon signed-rank test is used to compare the continuous outcome in the two matched samples or the paired samples.

Null hypothesis, H_0: Median difference should be zero.

Test statistic: The test statistic W, is defined as the smaller of W+ or W- .

Where W+ and W- are the sums of the positive and the negative ranks of the different scores.

Decision Rule: Reject the null hypothesis if the test statistic, W is less than or equal to the critical value from the table.

Advantages and Disadvantages of Non-Parametric Test

The advantages of the non-parametric test are:

- Easily understandable
- Short calculations
- Assumption of distribution is not required
- Applicable to all types of data

The disadvantages of the non-parametric test are:

- Less efficient as compared to parametric test
- The results may or may not provide an accurate answer because they are distribution free

Applications of Non-Parametric Test

The conditions when non-parametric tests are used are listed below:

- When parametric tests are not satisfied.
- When testing the hypothesis, it does not have any distribution.
- For quick data analysis.
- When unscaled data is available.

The alternatives to all statistical analyses comparing means are non-parametric analyses. A Parameter is a statistic that describes the population . Non-parametric statistics don't require the population data to be normally distributed.

If the data are not normally distributed, then we can't compare means because there is no center! Non-normal distributions may occur when there are:

- Few people (small N)
- Extreme scores (outliers)
- There's an arbitrary cut-off point on the scale. (Like if a survey asked for ages, but then just said, "17 and below".)

All of the non-parametric statistics for use with
quantitative variables
(means) work with the *ranks* of the variables, rather than the values themselves.

Mann-Whitney U Test

The Mann-Whitney U-test is a non-parametric alternative to an independent samples
tt-test that some people recommend for non-normal data. An independent samples t-test can usually handle if the standard deviations are similar or are not normally distributed, so there's little reason to use the Mann-Whitney U-test unless you have a true ranked variable instead of a
quantitative variable .Despite that fact, this is a behavioral statistics textbook, so we're going to talk about statistical alternative.

You can use the Mann-Whitney when:

- Your data is already in ranks (ordinal), *or*
- When you'd like to use an independent sample t-test, but the data is probably not normally distributed. (When the data is not normally distributed, the mean is sorta… meaningless.)

Formula

The formulas are below, but they are so uncommonly used that they won't be in the Common Formulas page at the back of the textbook. Similarly, there is a critical value table for U-scores, but that will also not be included in the Common Critical Values page.

To calculate this formula, you would need to list out all of the scores in order, and identify which is from which group. Then, you would calculate R1, which is the sum of the *ranks* of all of the scores (not the scores themselves) from the first group.

$U1=(N1*N2)+((N1*(N1+1))2)-R1U1=(N1*N2)+((N1*(N1+1))2)-R1$

$U2=(N1*N2)-U1U2=(N1*N2)-U1$

Mann-Whitney steps:

1. Calculate the *two* formulas,
2. Then compare the *smallest* of the two calculated U values to a critical U from a critical U table.

Interpreting Results

Imagine that I wanted to compare the mean of Exam #1 of two sections of my behavioral statistics classes, one in the morning and one in the evening.

- Research Hypothesis

  : The morning class's average Exam #1 score will be higher than the average Exam #1 score of the evening section.
- Symbols: XM¯>XE¯XM¯>XE¯

UNIT – IV - DATA MINING

Introduction

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

Importing Data into Excel

You might have to use data from various sources for analysis. In Excel, you can import data from different data sources. Some of the data sources are as follows −

- Microsoft Access Database
- Web Page
- Text File
- SQL Server Table
- SQL Server Analysis Cube
- XML File

You can import any number of tables simultaneously from a database.

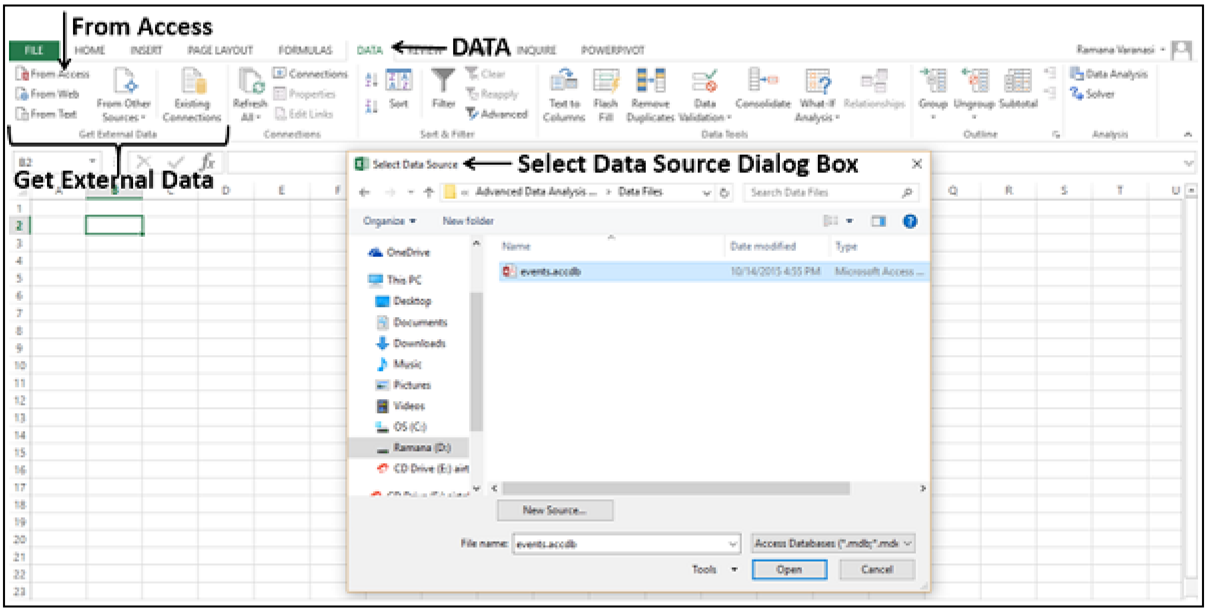
Importing Data from Microsoft Access Database

We will learn how to import data from MS Access database. Follow the steps given below −

Step 1 − Open a new blank workbook in Excel.

Step 2 − Click the DATA tab on the Ribbon.

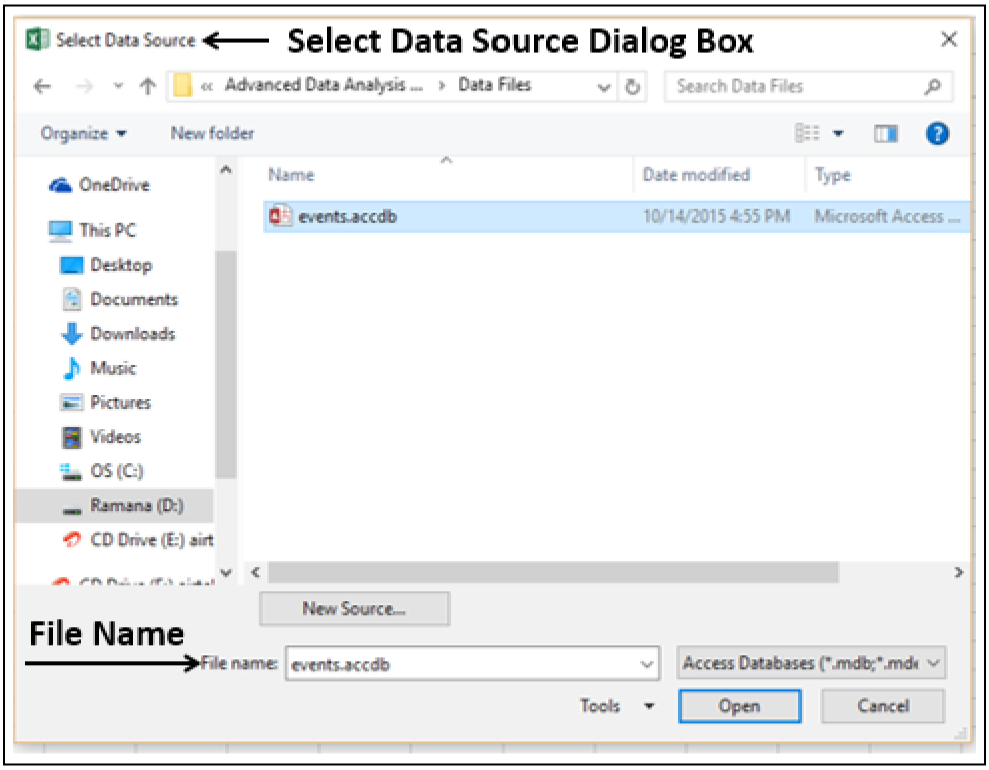
Step 3 − Click From Access in the Get External Data group. The Select Data Source dialog box appears.



File name: events.accdb

Access Databases (*.mdb;*.mde

Step 4 – Select the Access database file that you want to import. Access database files will have the extension .accdb.

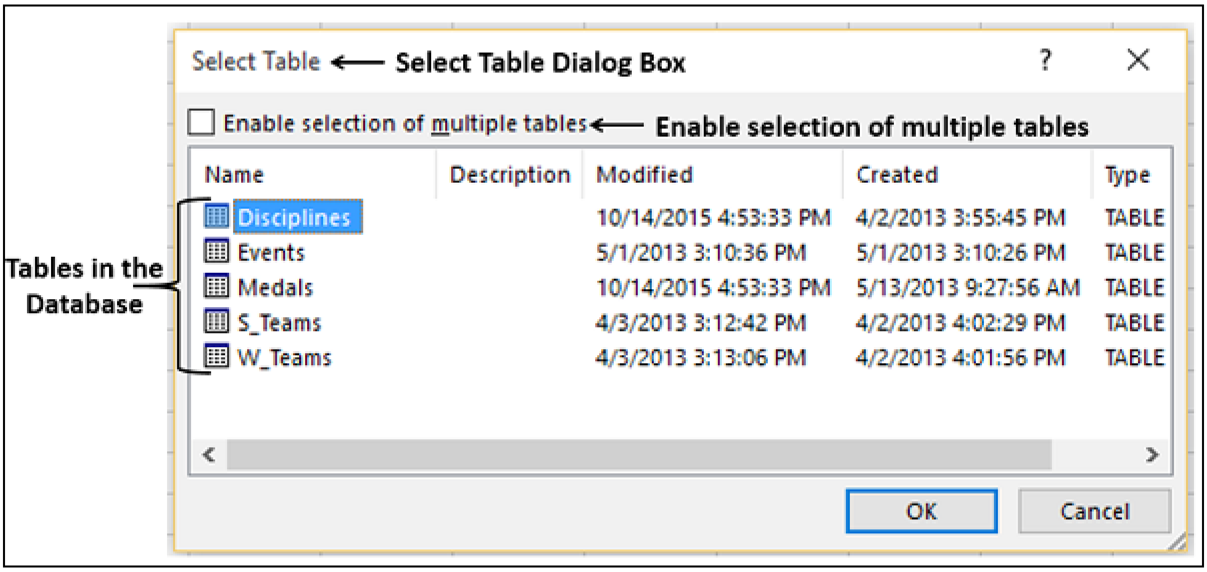


Name	Date modified	Type
events.accdb	10/14/2015 4:55 PM	Microsoft Access ...

File name: events.accdb

Access Databases (*.mdb;*.mde

The Select Table dialog box appears displaying the tables found in the Access database. You can either import all the tables in the database at once or import only the selected tables based on your data analysis needs.

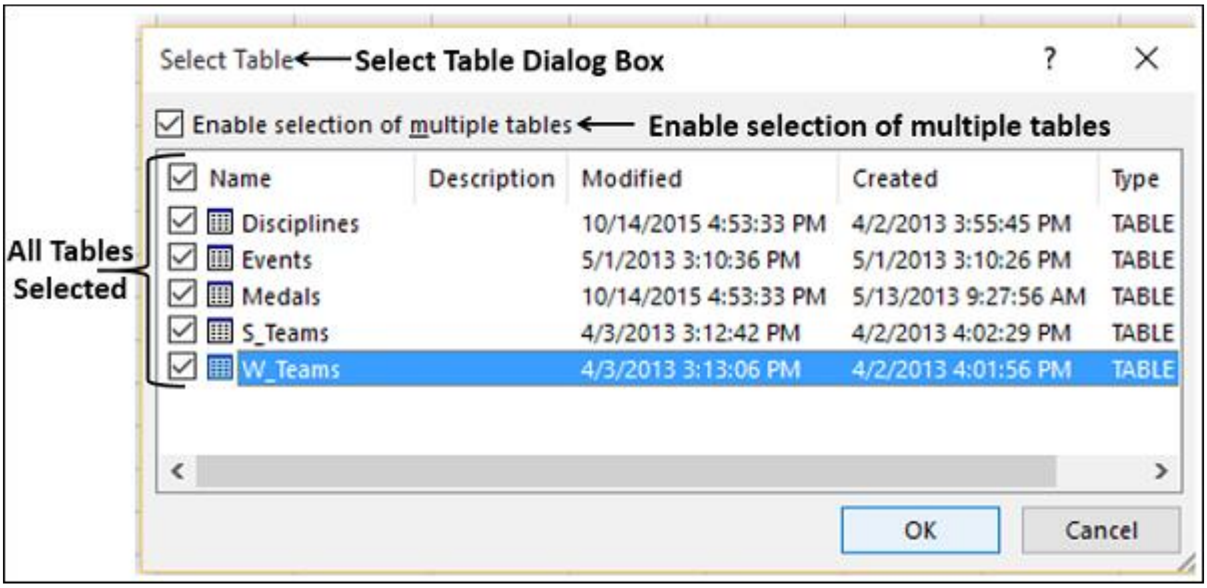


Select Table

□ Enable selection of multiple tables

Name	Description	Modified	Created	Type
Disciplines		10/14/2015 4:53:33 PM	4/2/2013 3:55:45 PM	TABLE
Events		5/1/2013 3:10:36 PM	5/1/2013 3:10:26 PM	TABLE
Medals		10/14/2015 4:53:33 PM	5/13/2013 9:27:56 AM	TABLE
S_Teams		4/3/2013 3:12:42 PM	4/2/2013 4:02:29 PM	TABLE
W_Teams		4/3/2013 3:13:06 PM	4/2/2013 4:01:56 PM	TABLE

Step 5 – Select the Enable selection of multiple tables box and select all the tables.

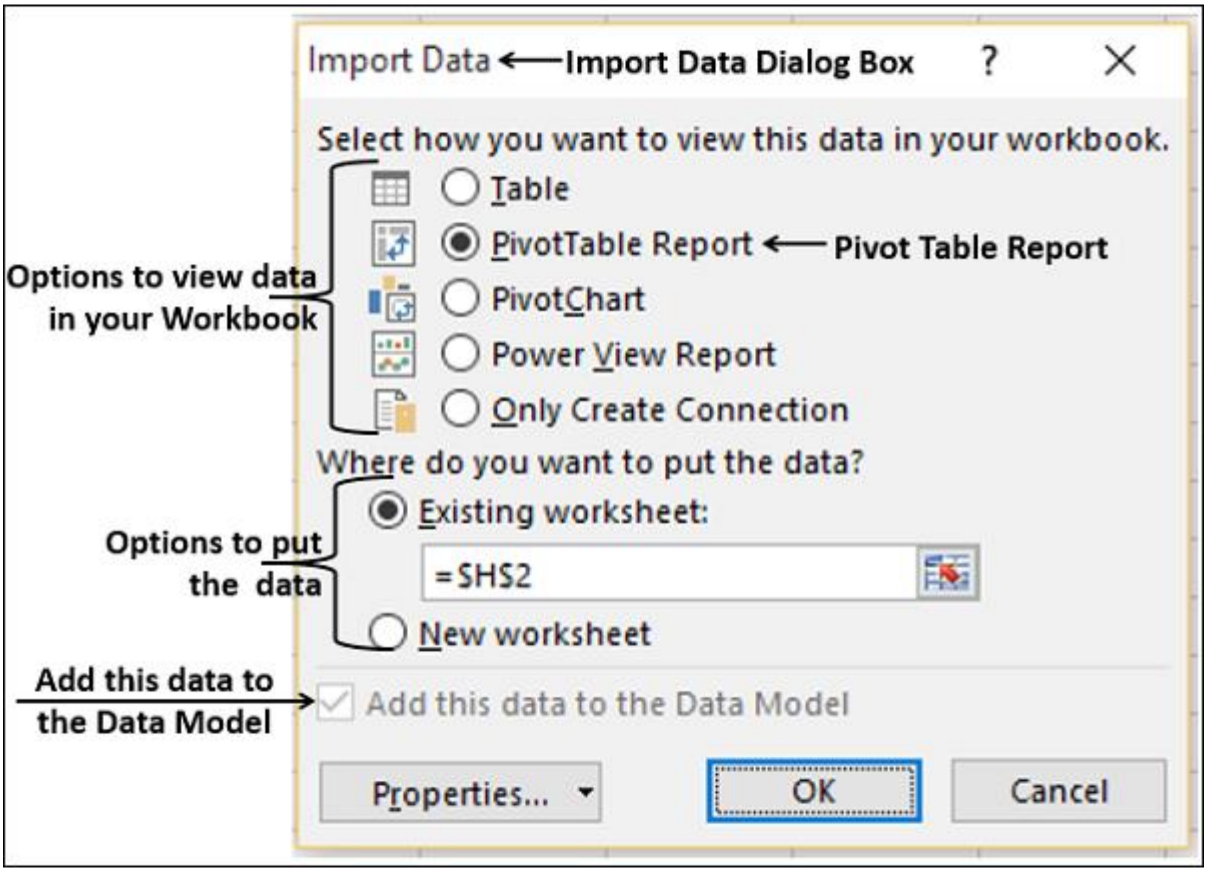


Select Table

☑ Enable selection of multiple tables

☑ Name	Description	Modified	Created	Type
☑ Disciplines		10/14/2015 4:53:33 PM	4/2/2013 3:55:45 PM	TABLE
☑ Events		5/1/2013 3:10:36 PM	5/1/2013 3:10:26 PM	TABLE
☑ Medals		10/14/2015 4:53:33 PM	5/13/2013 9:27:56 AM	TABLE
☑ S_Teams		4/3/2013 3:12:42 PM	4/2/2013 4:02:29 PM	TABLE
☑ W_Teams		4/3/2013 3:13:06 PM	4/2/2013 4:01:56 PM	TABLE

Step 6 – Click OK. The Import Data dialog box appears.



Import Data

Select how you want to view this data in your workbook.

- ○ Table
- ● PivotTable Report
- ○ PivotChart
- ○ Power View Report
- ○ Only Create Connection

Where do you want to put the data?

- ● Existing worksheet:
  =$H$2
- ○ New worksheet

☑ Add this data to the Data Model

As you observe, you have the following options to view the data you are importing in your workbook –

- Table
- PivotTable Report
- PivotChart
- Power View Report

You also have an option - only create connection. Further, PivotTable Report is selected by default.

Excel also gives you the options to put the data in your workbook –

- Existing worksheet
- New worksheet

You will find another check box that is selected and disabled – Add this data to the Data Model. Whenever you import data tables into your workbook, they are automatically added to the Data Model in your workbook. You will learn more about the Data Model in later chapters.

You can try each one of the options to view the data you are importing, and check how the data appears in your workbook –

- If you select Table, Existing worksheet option gets disabled, New worksheet option gets selected and Excel creates as many worksheets as the number of tables you are importing from the database. The Excel tables appear in these worksheets.
- If you select PivotTable Report, Excel imports the tables into the workbook and creates an empty PivotTable for analysing the data in the imported tables. You have an option to create the PivotTable in an existing worksheet or a new worksheet.

  Excel tables for the imported data tables will not appear in the workbook. However, you will find all the data tables in the PivotTable fields list, along with the fields in each table.
- If you select PivotChart, Excel imports the tables into the workbook and creates an empty PivotChart for displaying the data in the imported tables. You have an option to create the PivotChart in an existing worksheet or a new worksheet.

  Excel tables for the imported data tables will not appear in the workbook. However, you will find all the data tables in the PivotChart fields list, along with the fields in each table.
- If you select Power View Report, Excel imports the tables into the workbook and creates a Power View Report in a new worksheet. You will learn how to use Power View Reports for analyzing data in later chapters.

  Excel tables for the imported data tables will not appear in the workbook. However, you will find all the data tables in the Power View Report fields list, along with the fields in each table.
- If you select the option - Only Create Connection, a data connection will be established between the database and your workbook. No tables or reports appear in the workbook. However, the imported tables are added to the Data Model in your workbook by default.

  You need to choose any of these options, based on your intent of importing data for data analysis. As you observed above, irrespective of the option you have chosen, the data is imported and added to the Data Model in your workbook.

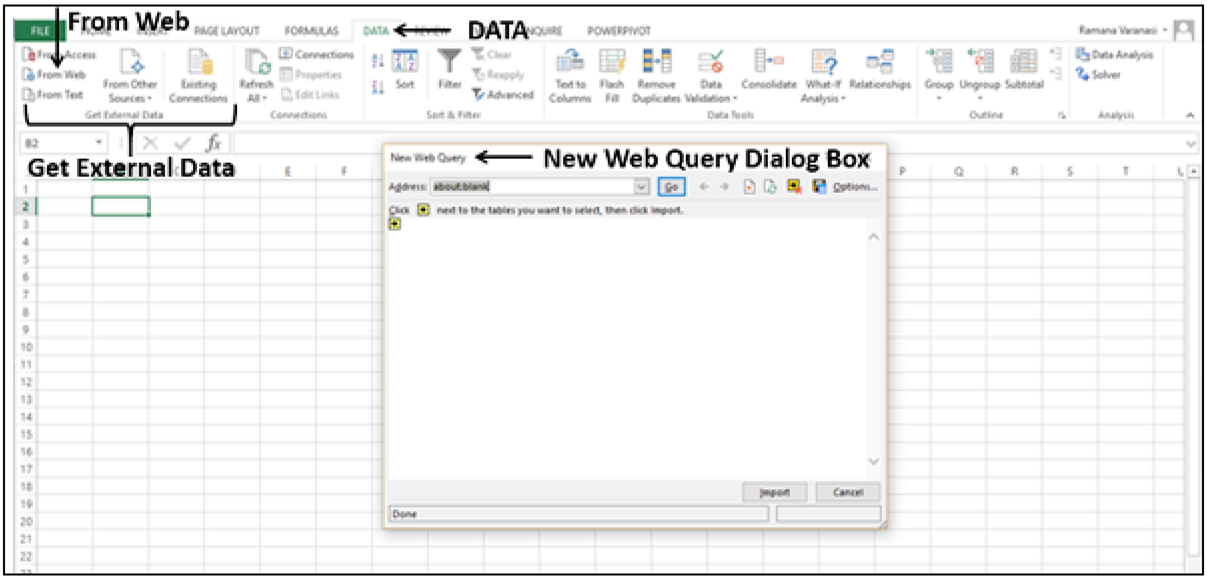
Importing Data from a Web Page

Sometimes, you might have to use the data that is refreshed on a web site. You can import data from a table on a website into Excel.

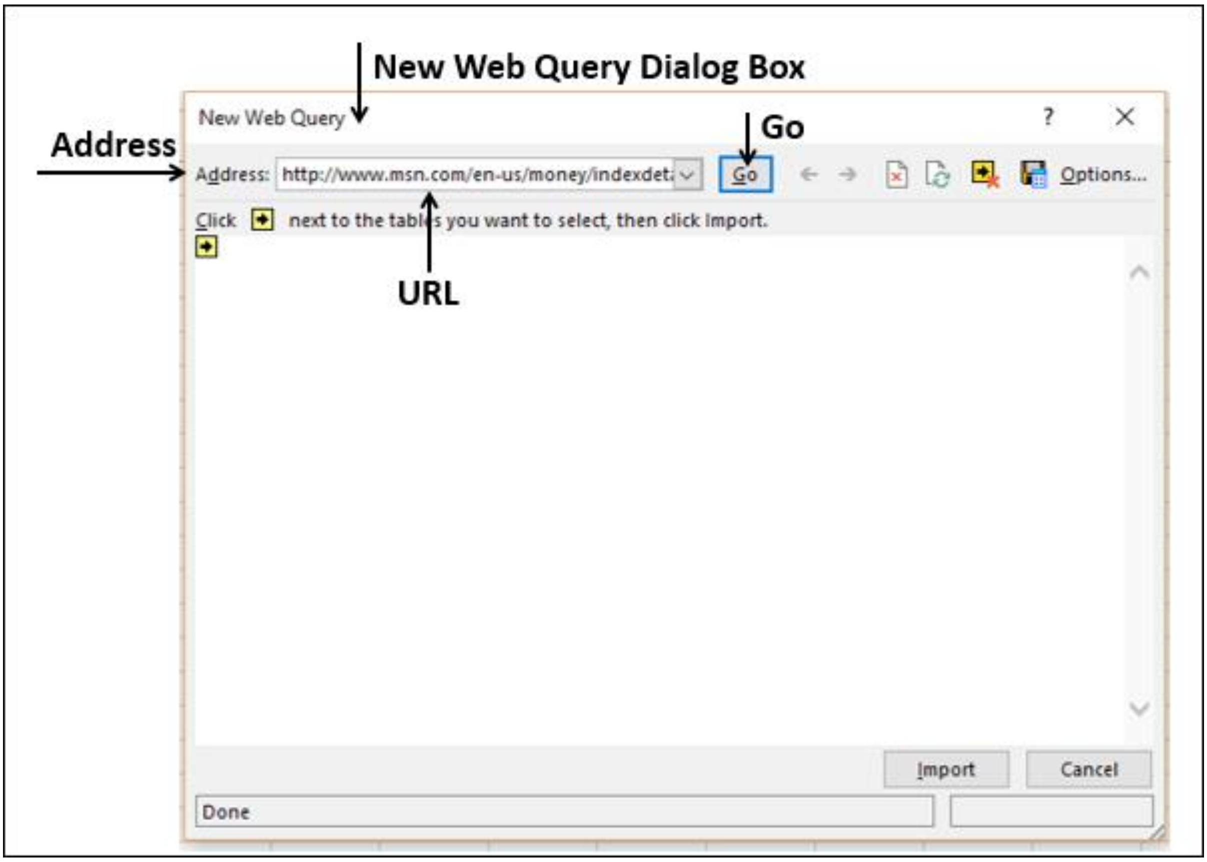
Step 1 − Open a new blank workbook in Excel.

Step 2 − Click the DATA tab on the Ribbon.

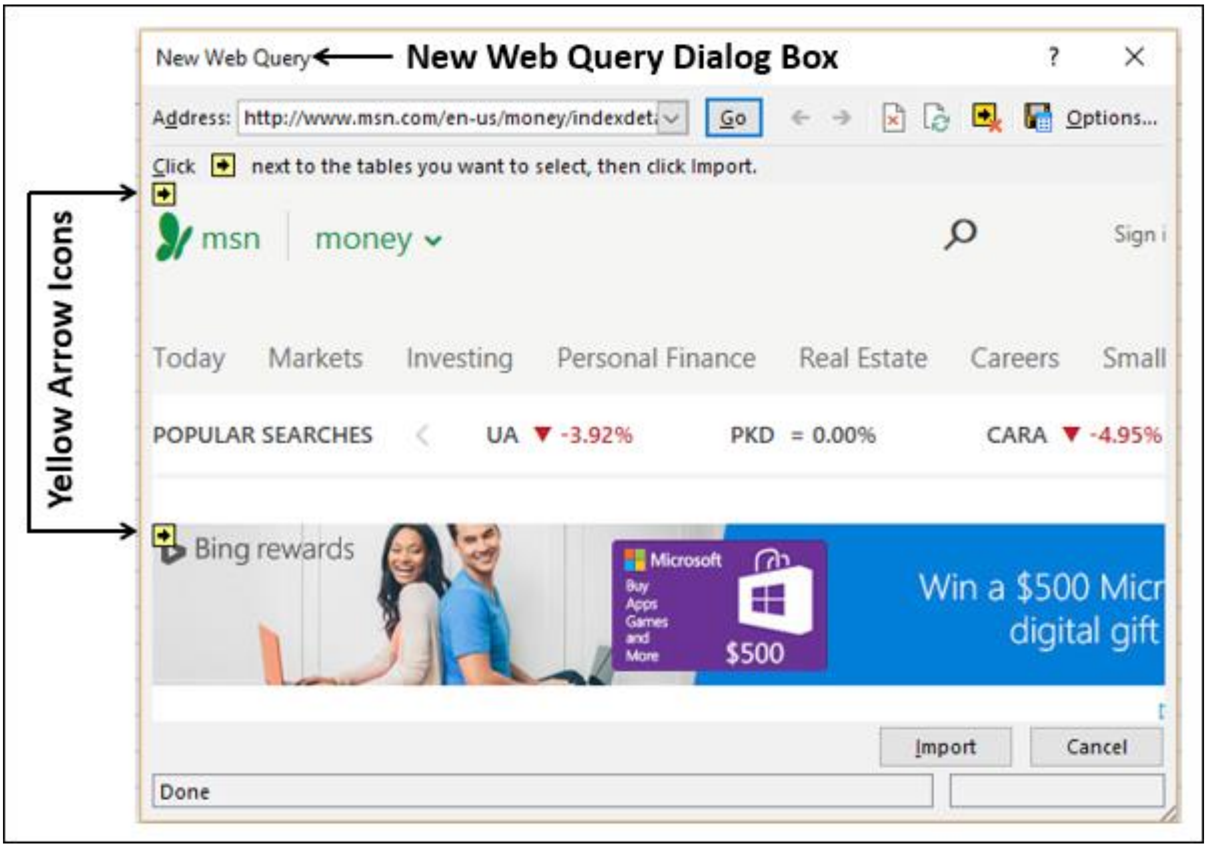
Step 3 − Click From Web in the Get External Data group. The New Web Query dialog box appears.



Step 4 − Enter the URL of the web site from where you want to import data, in the box next to Address and click Go.

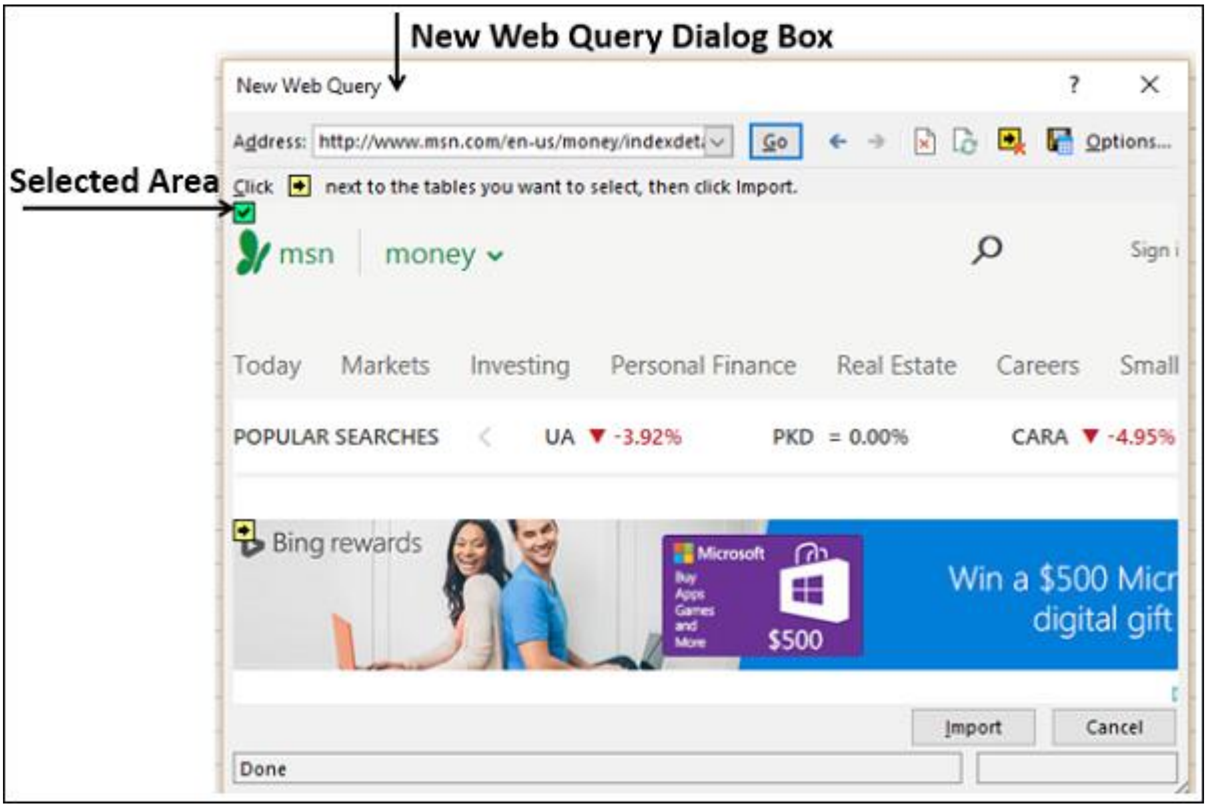

New Web Query Dialog Box
Address
Go
URL
New Web Query
Address: http://www.msn.com/en-us/money/indexdet
Go
Options...
Click next to the tables you want to select, then click Import.
Import
Cancel
Done

Step 5 − The data on the website appears. There will be yellow arrow icons next to the table data that can be imported.



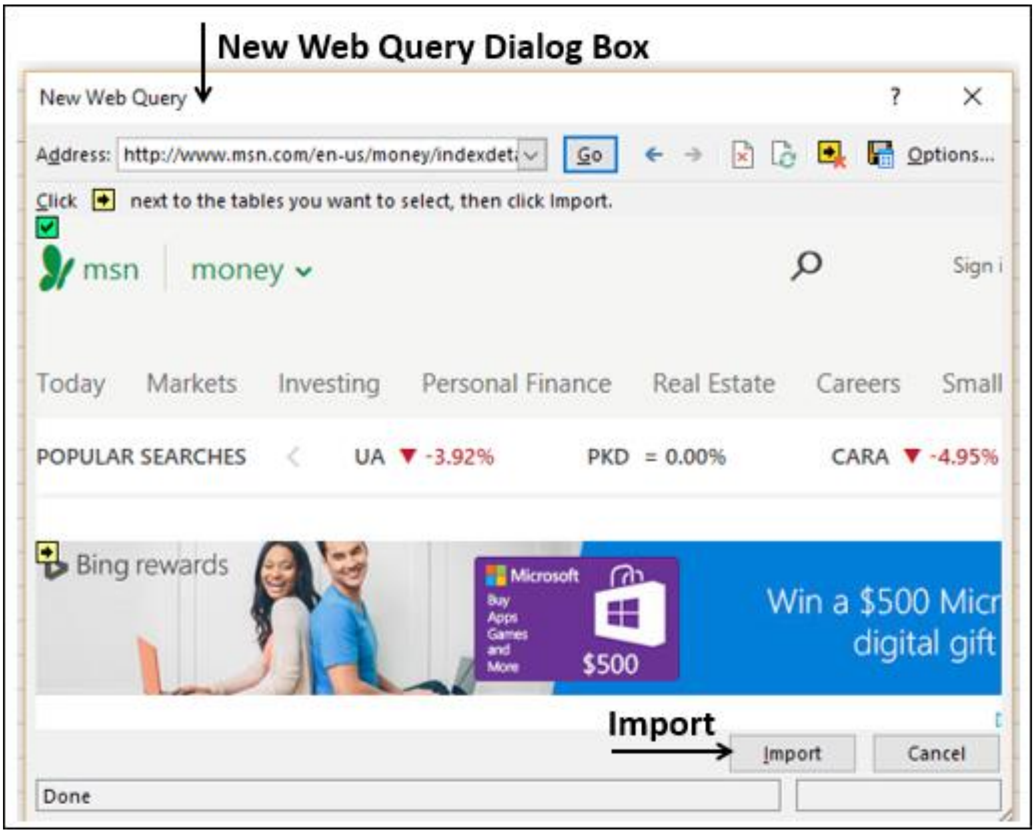
Click ➡ next to the tables you want to select, then click Import.

Step 6 – Click the yellow icons to select the data you want to import. This turns the yellow icons to green boxes with a checkmark as shown in the following screen shot.



Click ➡ next to the tables you want to select, then click Import.

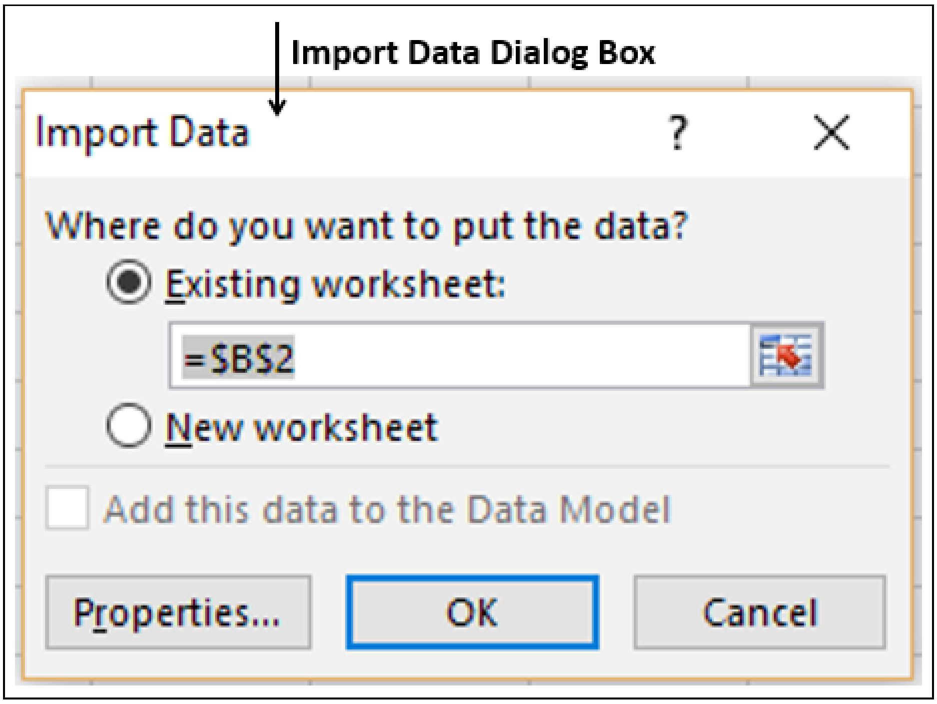
Step 7 – Click the Import button after you have selected what you want.



Address: http://www.msn.com/en-us/money/indexdet

Click next to the tables you want to select, then click Import.

The Import Data dialog box appears.



Where do you want to put the data?

● Existing worksheet:

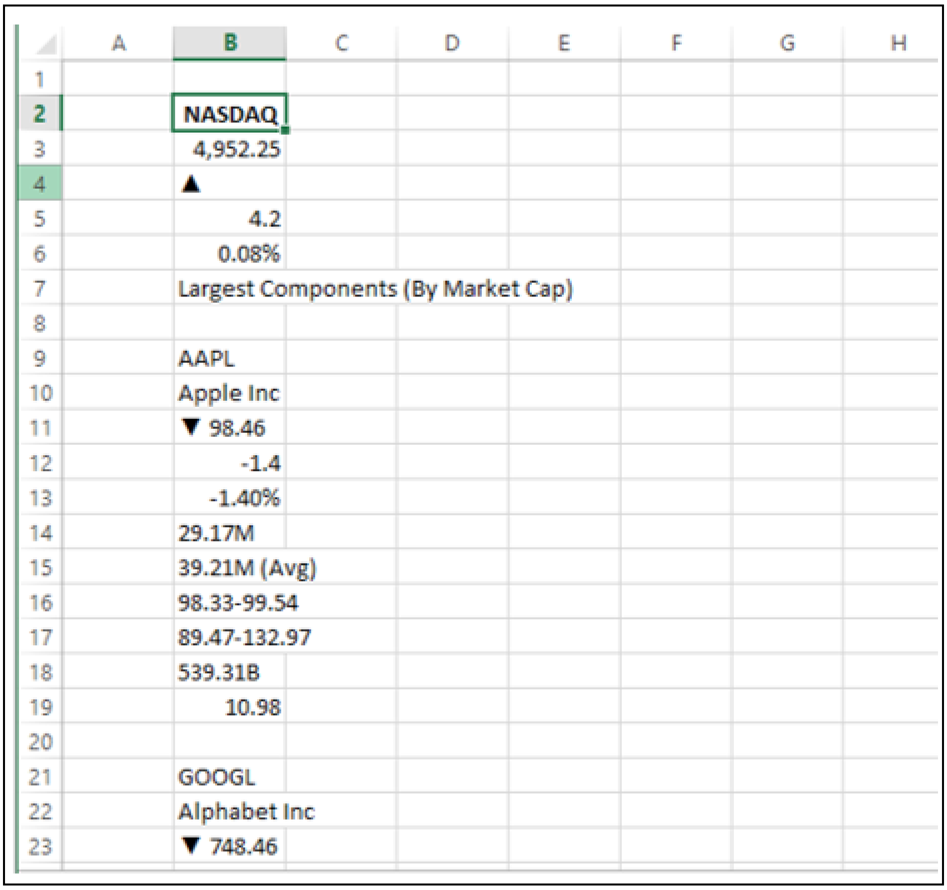
=$B$2

○ New worksheet

□ Add this data to the Data Model

Step 8 – Specify where you want to put the data and click Ok.

Step 9 – Arrange the data for further analysis and/or presentation.



	A	B	C	D	E	F	G	H
1								
2		NASDAQ						
3		4,952.25						
4		▲						
5		4.2						
6		0.08%						
7		Largest Components (By Market Cap)						
8								
9		AAPL						
10		Apple Inc						
11		▼ 98.46						
12		-1.4						
13		-1.40%						
14		29.17M						
15		39.21M (Avg)						
16		98.33-99.54						
17		89.47-132.97						
18		539.31B						
19		10.98						
20								
21		GOOGL						
22		Alphabet Inc						
23		▼ 748.46						

Copy-pasting data from web

Another way of getting data from a web page is by copying and pasting the required data.

Step 1 – Insert a new worksheet.

Step 2 – Copy the data from the web page and paste it on the worksheet.

Step 3 – Create a table with the pasted data.

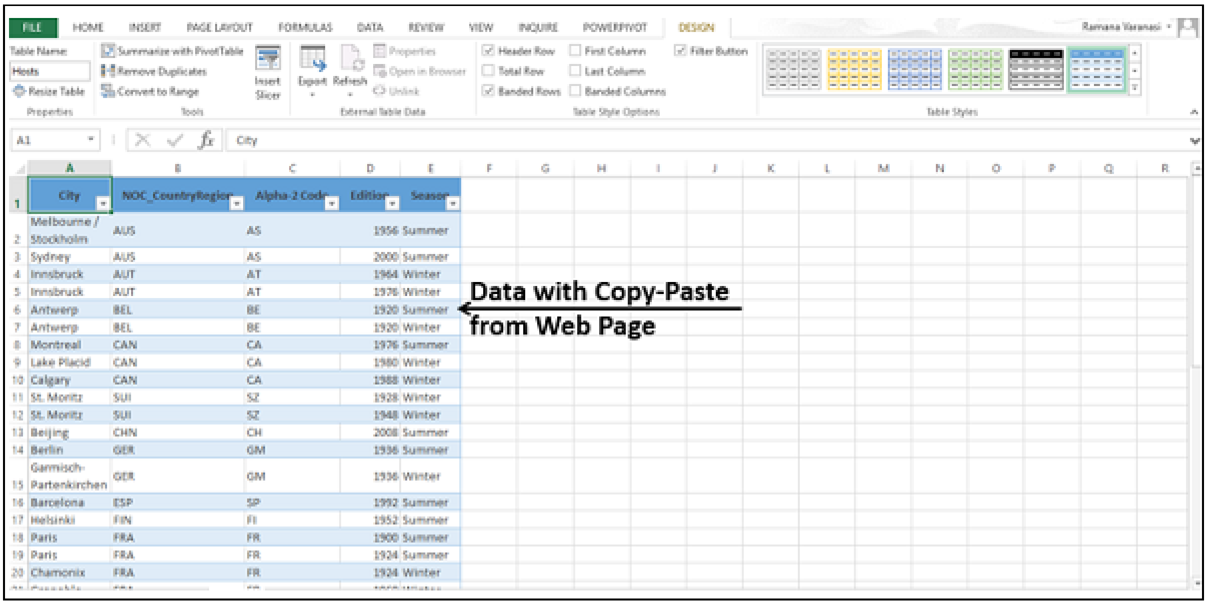


Table Name: Hosts

A1: City

	A	B	C	D	E
1	City	NOC_CountryRegion	Alpha-2 Code	Edition	Season
2	Melbourne / Stockholm	AUS	AS	1956	Summer
3	Sydney	AUS	AS	2000	Summer
4	Innsbruck	AUT	AT	1964	Winter
5	Innsbruck	AUT	AT	1976	Winter
6	Antwerp	BEL	BE	1920	Summer
7	Antwerp	BEL	BE	1920	Winter
8	Montreal	CAN	CA	1976	Summer
9	Lake Placid	CAN	CA	1980	Winter
10	Calgary	CAN	CA	1988	Winter
11	St. Moritz	SUI	SZ	1928	Winter
12	St. Moritz	SUI	SZ	1948	Winter
13	Beijing	CHN	CH	2008	Summer
14	Berlin	GER	GM	1936	Summer
15	Garmisch-Partenkirchen	GER	GM	1936	Winter
16	Barcelona	ESP	SP	1992	Summer
17	Helsinki	FIN	FI	1952	Summer
18	Paris	FRA	FR	1900	Summer
19	Paris	FRA	FR	1924	Summer
20	Chamonix	FRA	FR	1924	Winter

Data with Copy-Paste from Web Page

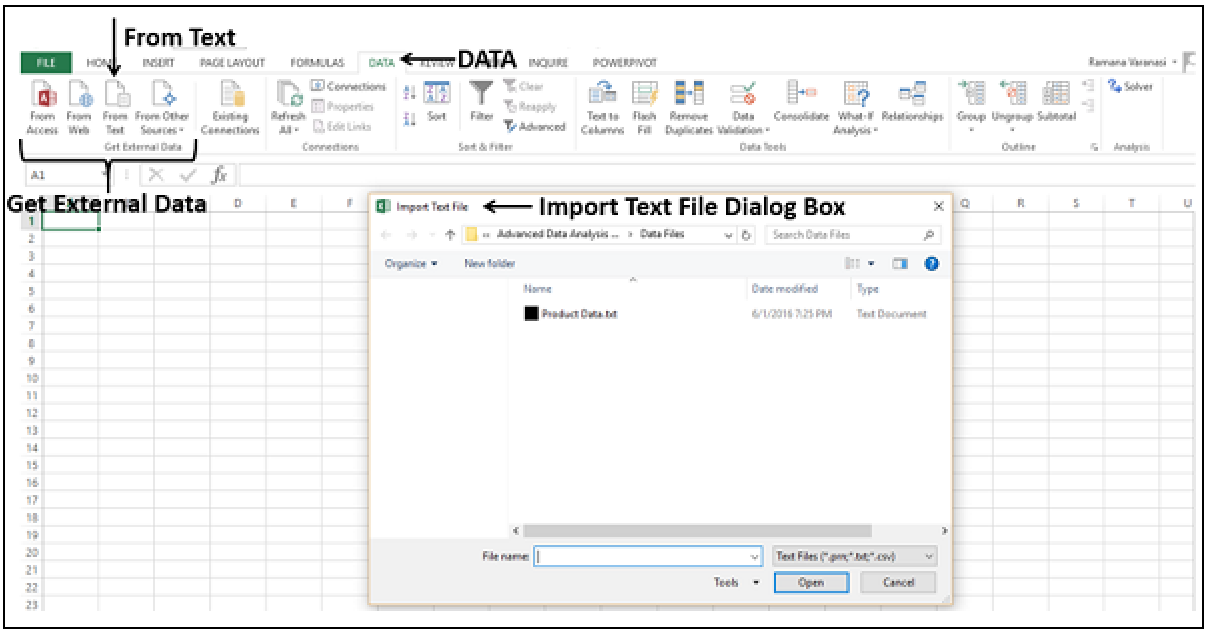
Importing Data from a Text File

If you have data in .txt or .csv or .prn files, you can import data from those files treating them as text files. Follow the steps given below −

Step 1 − Open a new worksheet in Excel.

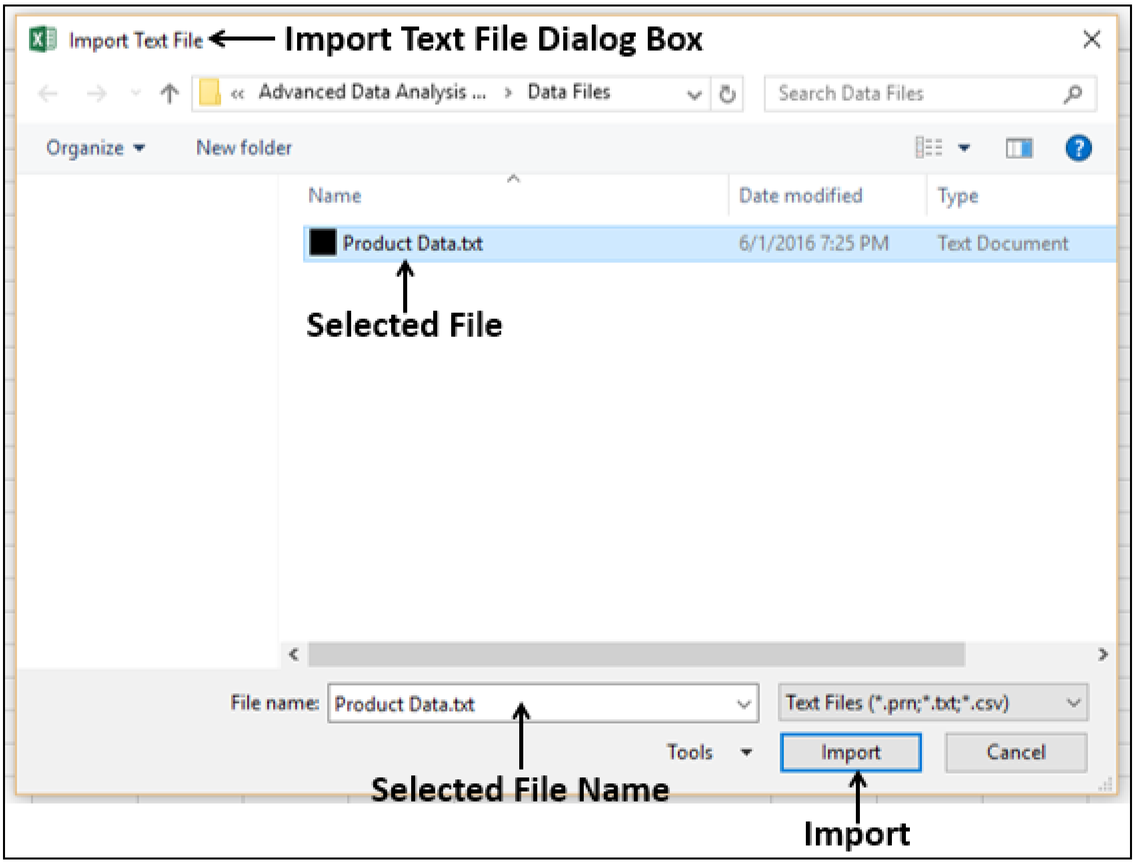
Step 2 − Click the DATA tab on the Ribbon.

Step 3 − Click From Text in the Get External Data group. The Import Text File dialog box appears.



You can see that .prn, .txt and .csv extension text files are accepted.

Step 4 − Select the file. The selected file name appears in the File name box. The Open button changes to Import button.



Import Text File ← Import Text File Dialog Box

« Advanced Data Analysis ... › Data Files

Name	Date modified	Type
Product Data.txt	6/1/2016 7:25 PM	Text Document

Selected File

File name: Product Data.txt

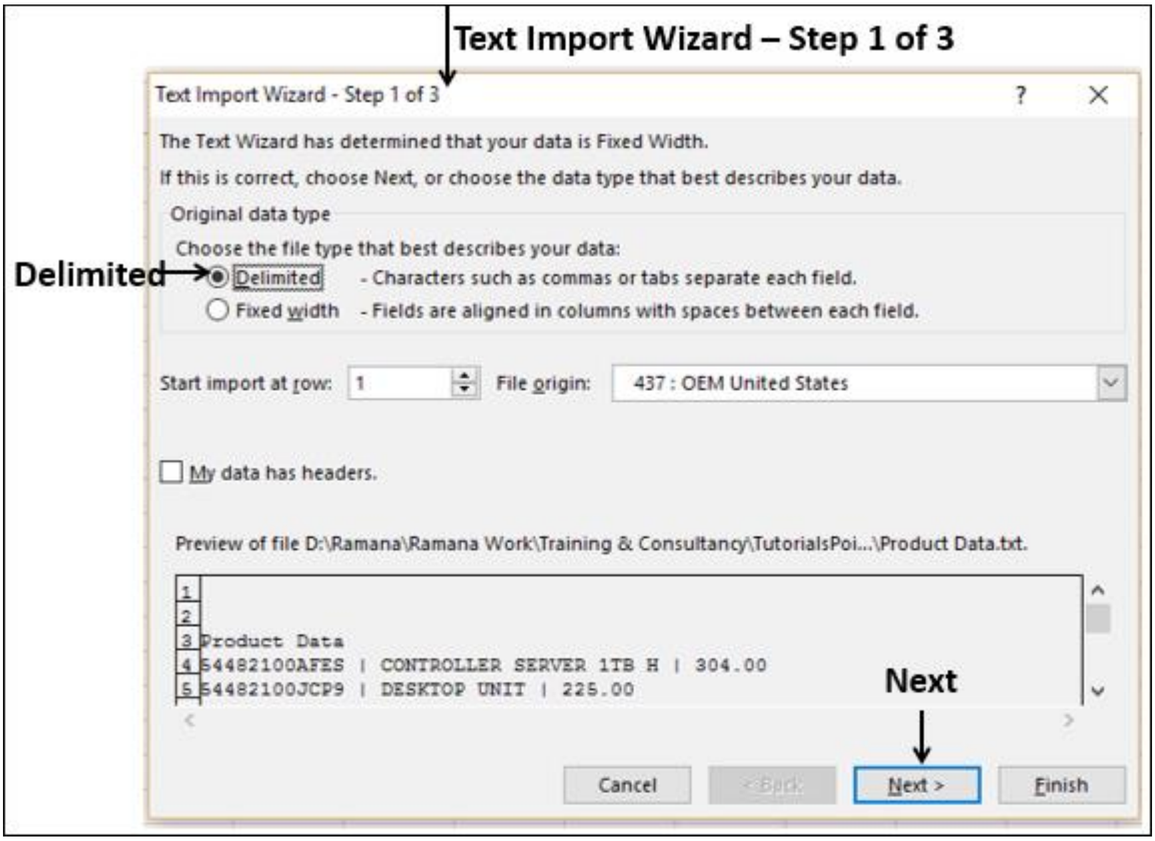
Text Files (*.prn;*.txt;*.csv)

Selected File Name

Import

Step 5 – Click the Import button. Text Import Wizard – Step 1 of 3 dialog box appears.

Step 6 – Click the option Delimited to choose the file type and click Next.



Text Import Wizard – Step 1 of 3

Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Fixed Width.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

Delimited → ● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: 437 : OEM United States

□ My data has headers.

Preview of file D:\Ramana\Ramana Work\Training & Consultancy\TutorialsPoi...\Product Data.txt.

```
1
2
3 Product Data
4 54482100AFES | CONTROLLER SERVER 1TB H | 304.00
5 54482100JCP9 | DESKTOP UNIT | 225.00
```

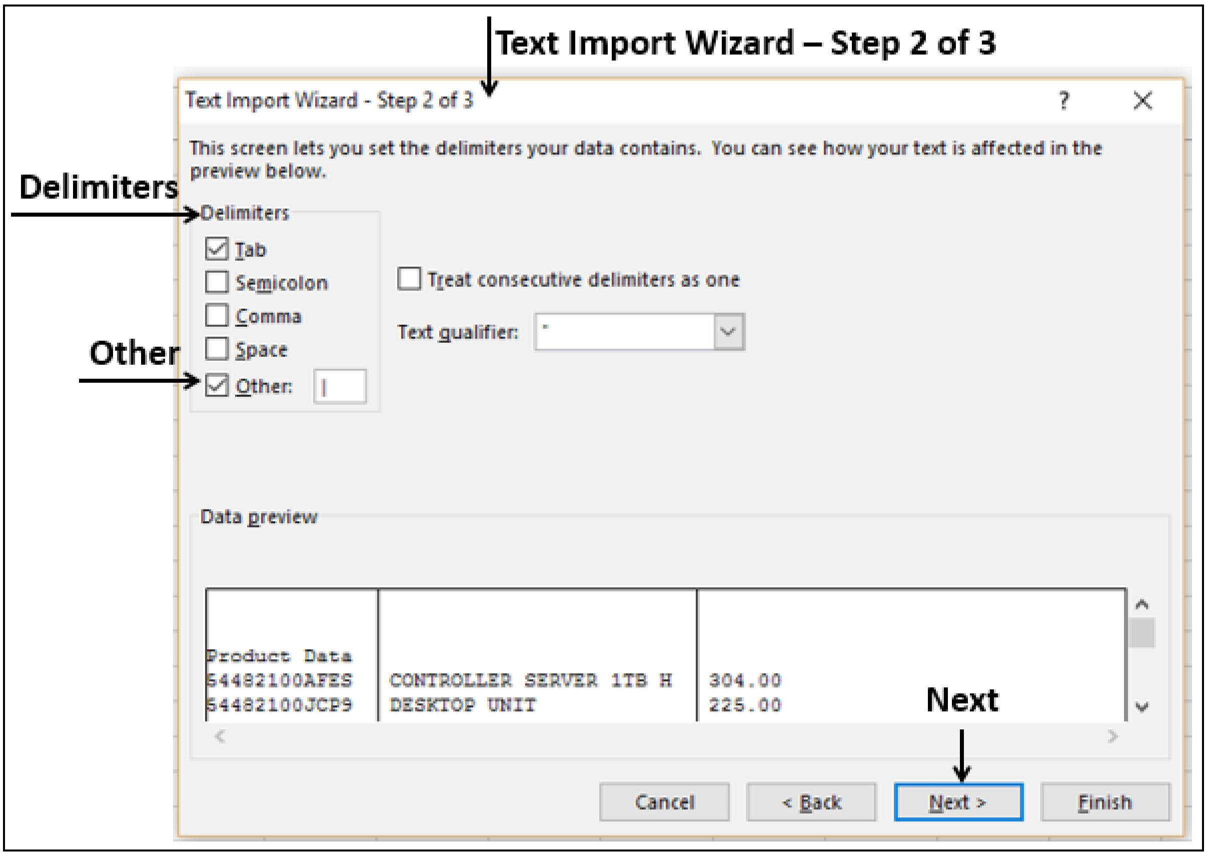
Next

The Text Import Wizard – Step 2 of 3 dialog box appears.

Step 7 − Under Delimiters, select Other.

Step 8 − In the box next to Other, type | (That is the delimiter in the text file you are importing).

Step 9 − Click Next.



Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☑ Tab
□ Semicolon
□ Comma
□ Space
☑ Other: |

□ Treat consecutive delimiters as one

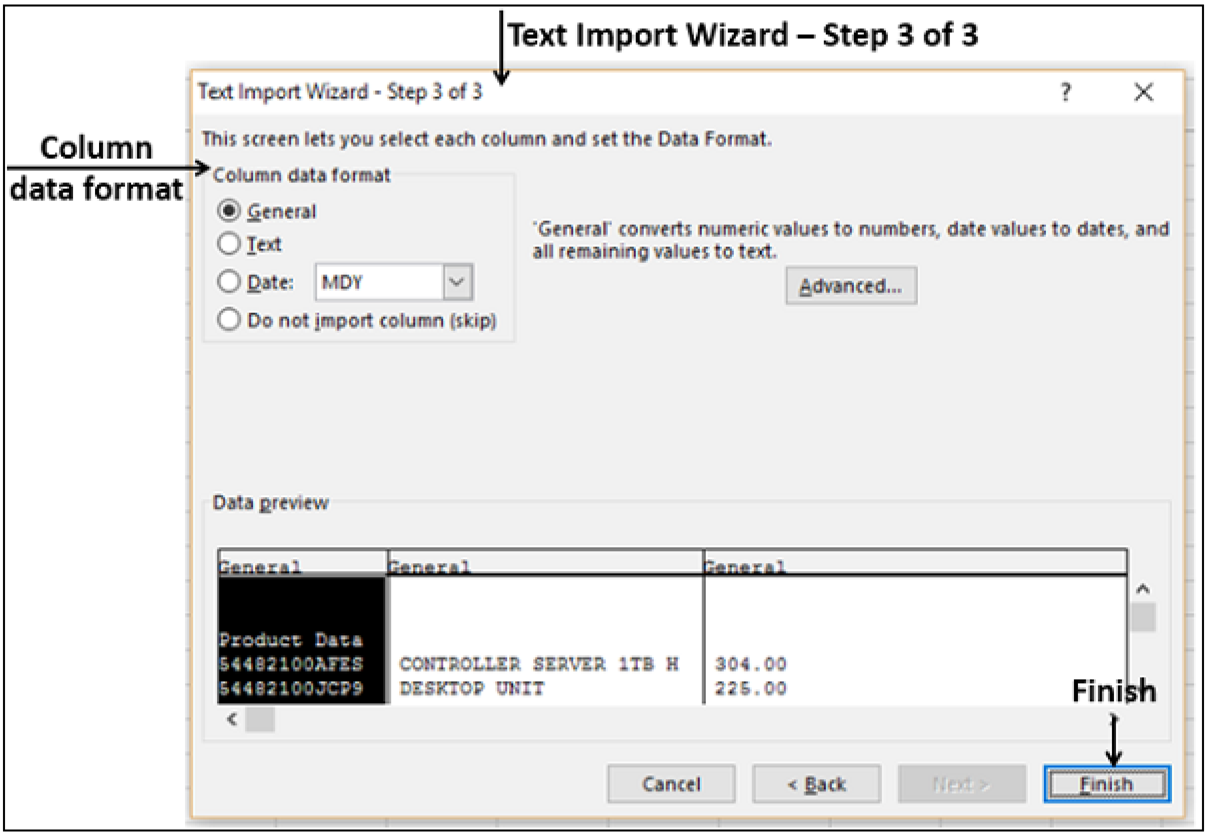
Text qualifier: "

Data preview

```
Product Data
54482100AFES   CONTROLLER SERVER 1TB H   304.00
54482100JCP9   DESKTOP UNIT              225.00
```

The Text Import Wizard – Step 3 of 3 dialog box appears.

Step 10 − In this dialog box, you can set column data format for each of the columns.


Text Import Wizard – Step 3 of 3
Column data format
Finish
Text Import Wizard - Step 3 of 3
? ×

This screen lets you select each column and set the Data Format.

Column data format

● General
○ Text
○ Date: MDY
○ Do not import column (skip)

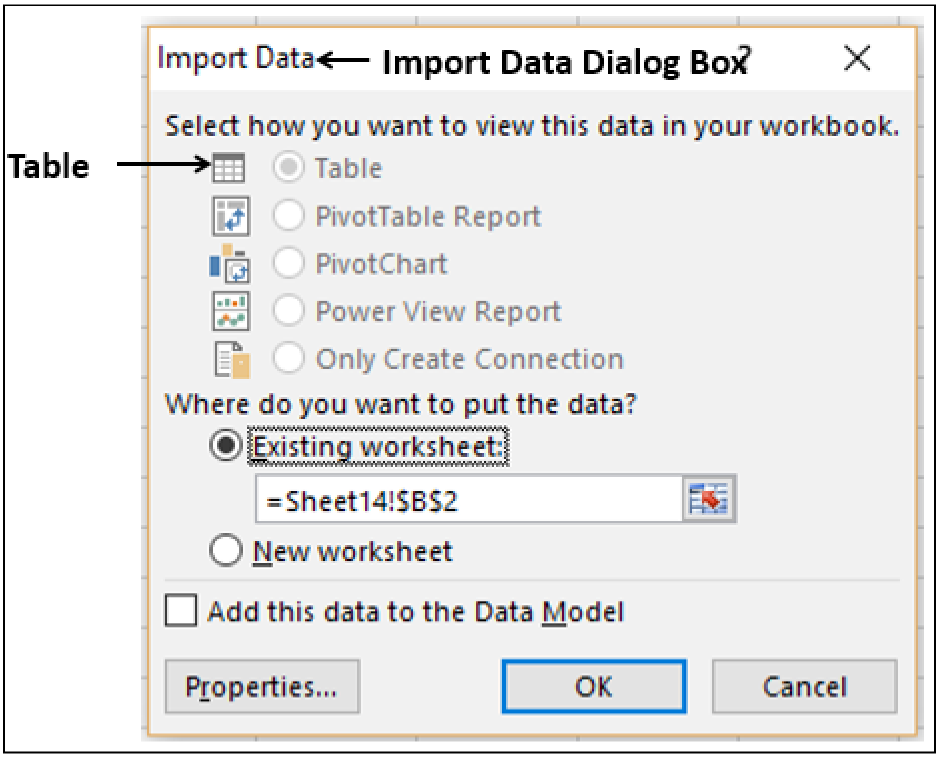
'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Data preview

```
General      General                  General
Product Data
54482100AFES  CONTROLLER SERVER 1TB H   304.00
54482100JCP9  DESKTOP UNIT              225.00
```

Cancel
< Back
Next >
Finish

Step 11 – After you complete the data formatting of columns, click Finish. The Import Data dialog box appears.



Select how you want to view this data in your workbook.

● Table
○ PivotTable Report
○ PivotChart
○ Power View Report
○ Only Create Connection

Where do you want to put the data?

● Existing worksheet:

=Sheet14!$B$2

○ New worksheet

□ Add this data to the Data Model

You will observe the following –

- Table is selected for view and is grayed. Table is the only view option you have in this case.
- You can put the data either in an existing worksheet or a New worksheet.

- You can select or not select the check box Add this data to the Data Model.
- Click OK after you have made the choices.

Data appears on the worksheet you specified. You have imported data from Text file into Excel workbook.

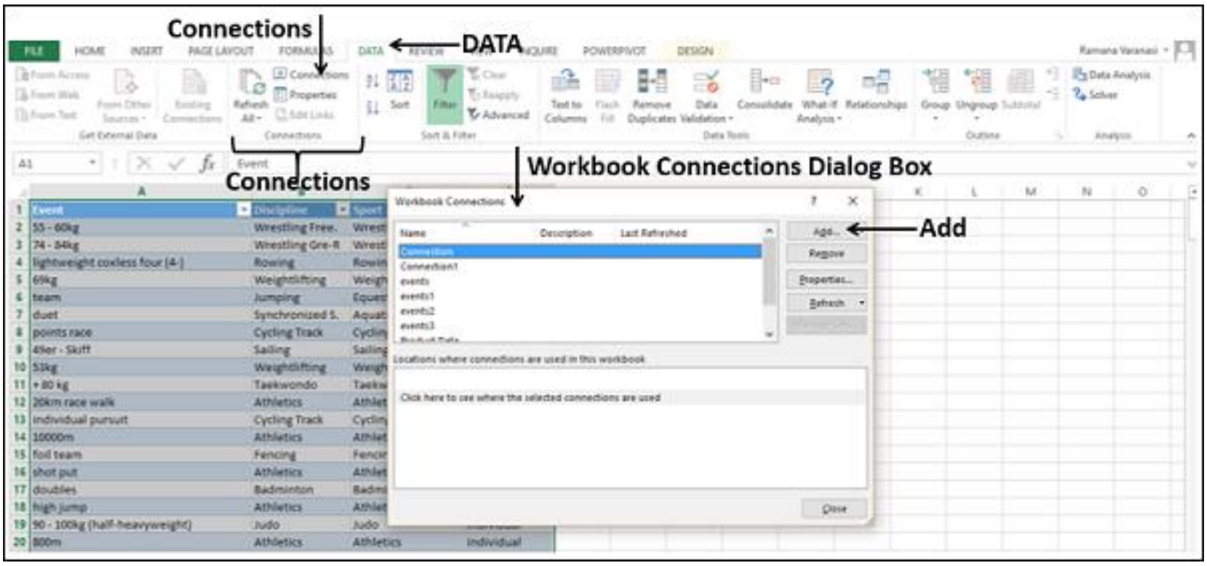
Importing Data from another Workbook

You might have to use data from another Excel workbook for your data analysis, but someone else might maintain the other workbook.

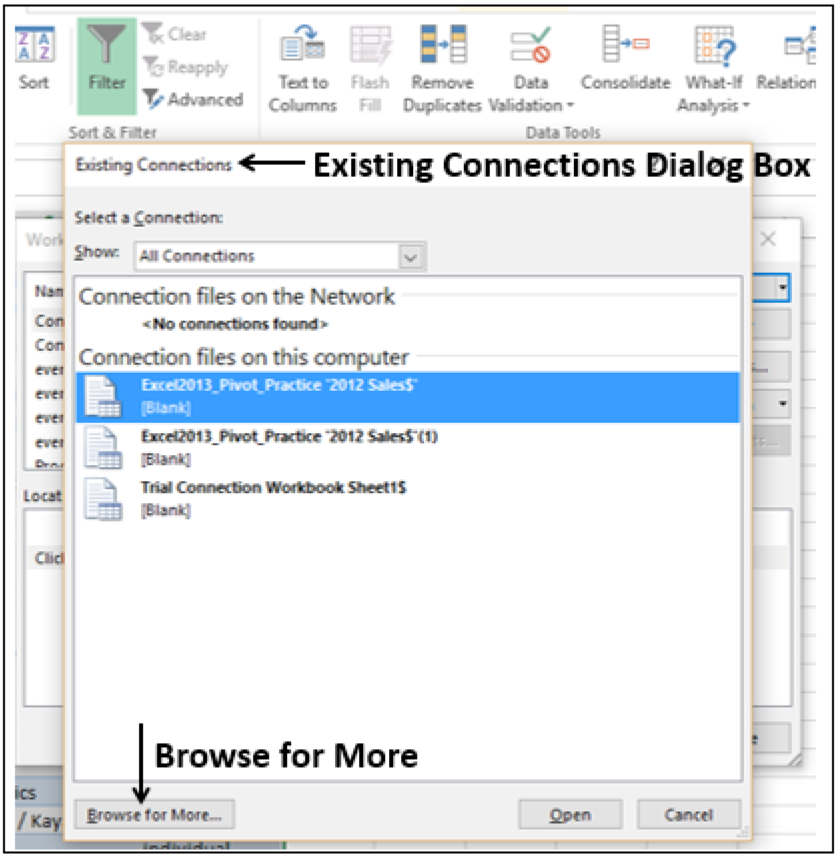
To get up to date data from another workbook, establish a data connection with that workbook.

Step 1 − Click DATA > Connections in the Connections group on the Ribbon.

The Workbook Connections dialog box appears.


Connections
DATA
Connections
Workbook Connections Dialog Box
Add
Workbook Connections
Name
Description
Last Refreshed
Add...
Remove
Properties...
Refresh
Locations where connections are used in this workbook
Click here to see where the selected connections are used
Close

Step 2 − Click the Add button in the Workbook Connections dialog box. The Existing Connections dialog box appears.



Show: All Connections

Connection files on the Network

<No connections found>

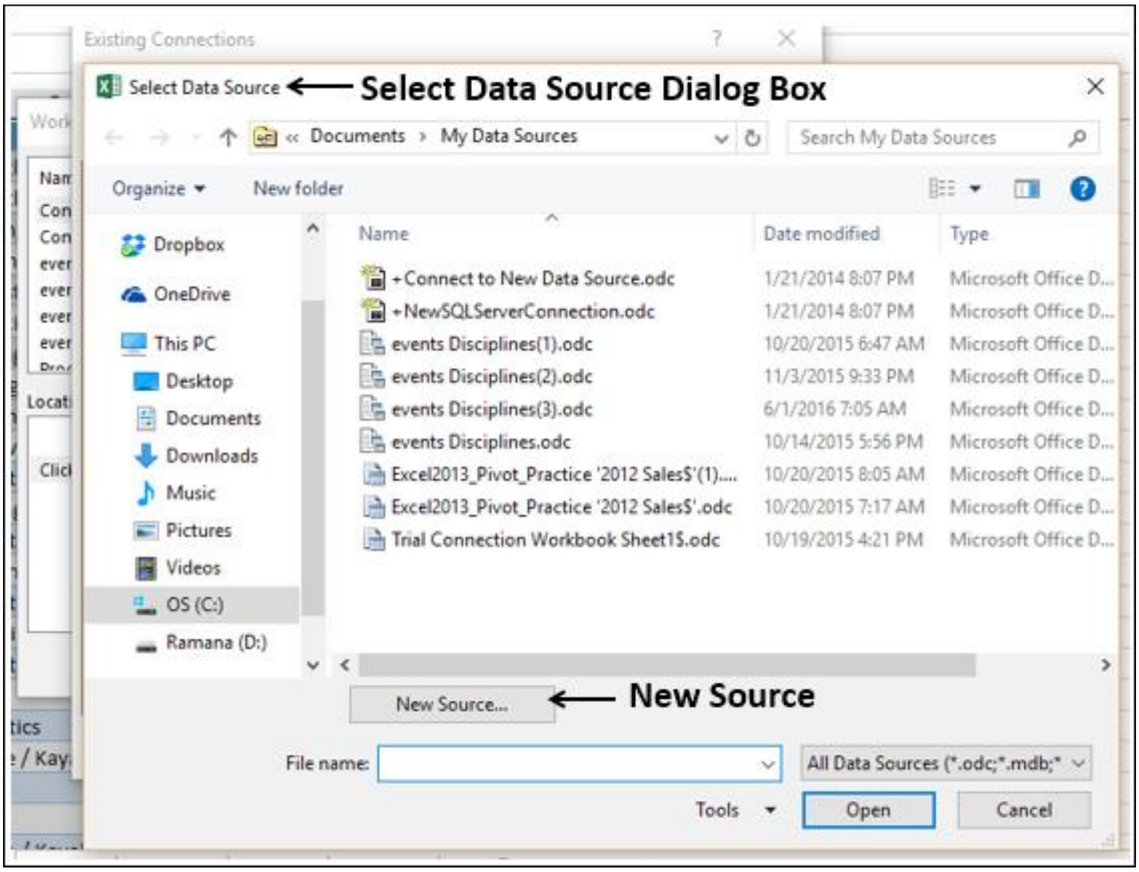
Connection files on this computer

Excel2013_Pivot_Practice '2012 Sales$'
[Blank]

Excel2013_Pivot_Practice '2012 Sales$'(1)
[Blank]

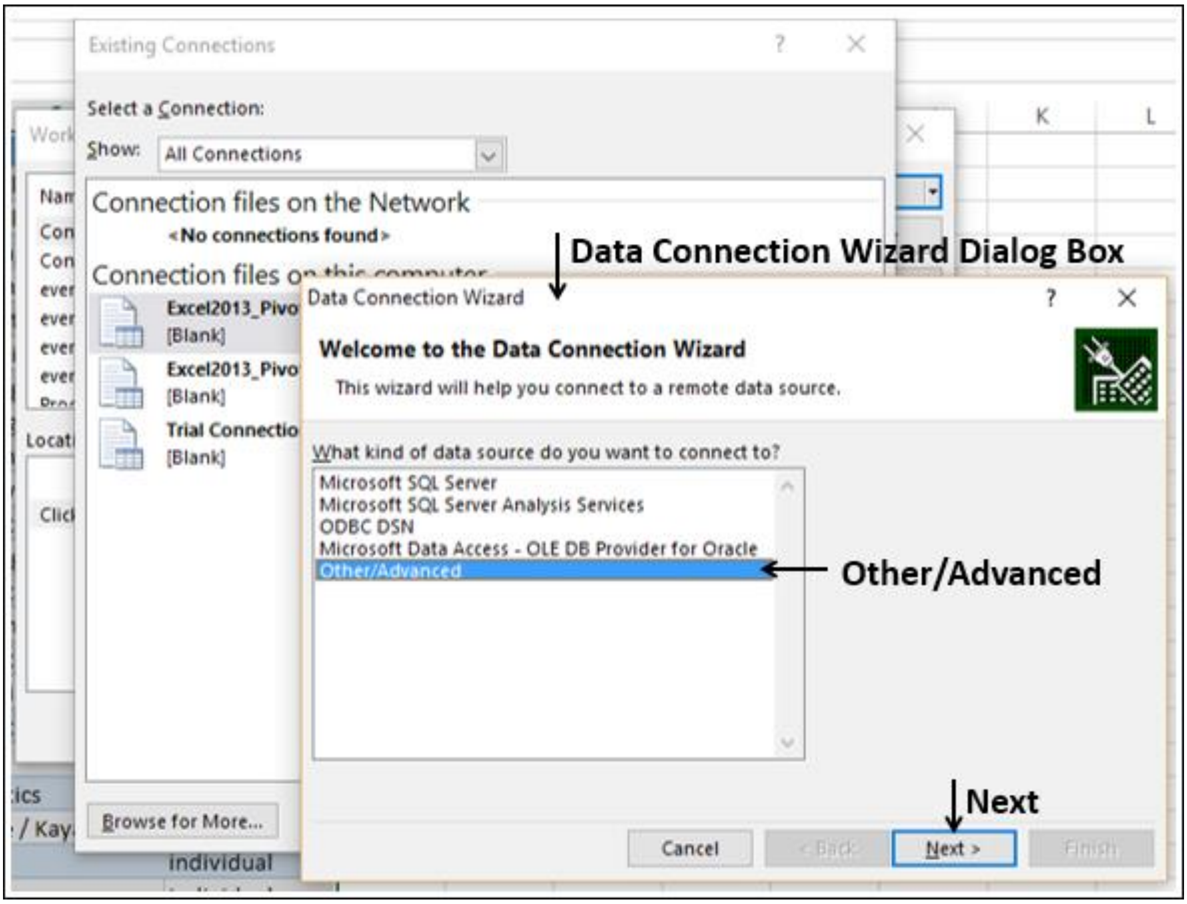
Trial Connection Workbook Sheet1$
[Blank]

Step 3 − Click Browse for More… button. The Select Data Source dialog box appears.

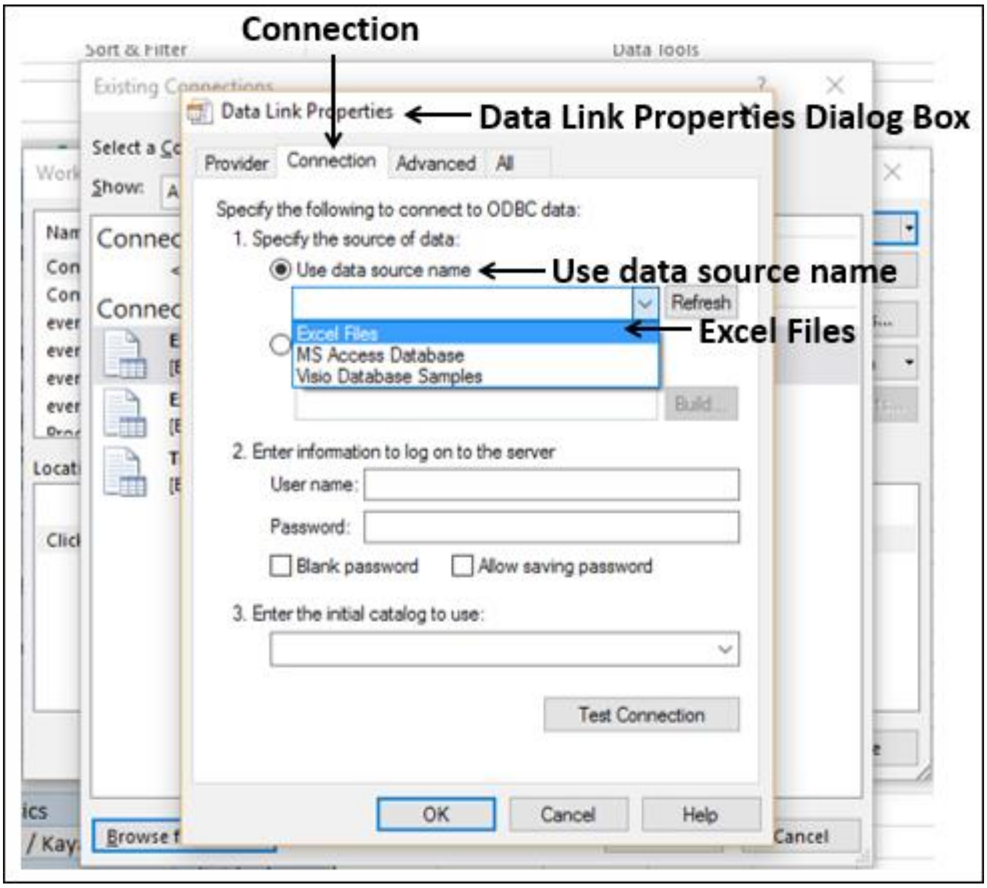


Name	Date modified	Type
+Connect to New Data Source.odc	1/21/2014 8:07 PM	Microsoft Office D...
+NewSQLServerConnection.odc	1/21/2014 8:07 PM	Microsoft Office D...
events Disciplines(1).odc	10/20/2015 6:47 AM	Microsoft Office D...
events Disciplines(2).odc	11/3/2015 9:33 PM	Microsoft Office D...
events Disciplines(3).odc	6/1/2016 7:05 AM	Microsoft Office D...
events Disciplines.odc	10/14/2015 5:56 PM	Microsoft Office D...
Excel2013_Pivot_Practice '2012 Sales$'(1)....	10/20/2015 8:05 AM	Microsoft Office D...
Excel2013_Pivot_Practice '2012 Sales$'.odc	10/20/2015 7:17 AM	Microsoft Office D...
Trial Connection Workbook Sheet1$.odc	10/19/2015 4:21 PM	Microsoft Office D...

Step 4 − Click the New Source button. The Data Connection Wizard dialog box appears.


Existing Connections
Select a Connection:
Show: All Connections
Connection files on the Network
<No connections found>
Connection files on this computer
Excel2013_Pivo
[Blank]
Excel2013_Pivo
[Blank]
Trial Connectio
[Blank]
Browse for More...
Data Connection Wizard Dialog Box
Data Connection Wizard
Welcome to the Data Connection Wizard
This wizard will help you connect to a remote data source.
What kind of data source do you want to connect to?
Microsoft SQL Server
Microsoft SQL Server Analysis Services
ODBC DSN
Microsoft Data Access - OLE DB Provider for Oracle
Other/Advanced
Other/Advanced
Next
Cancel
< Back
Next >
Finish

Step 5 − Select Other/Advanced in the data source list and click Next. The Data Link Properties dialog box appears.



Connection

Data Link Properties Dialog Box

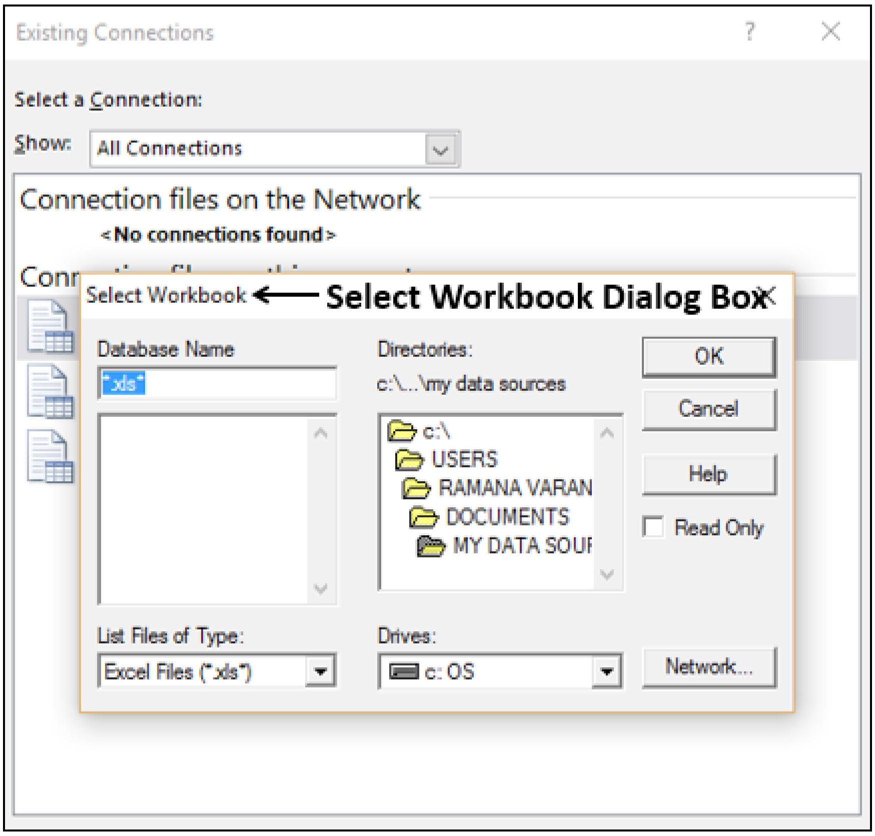
Use data source name

Excel Files

Step 6 − Set the data link properties as follows −

- Click the Connection tab.
- Click Use data source name.
- Click the down-arrow and select Excel Files from the drop-down list.
- Click OK.

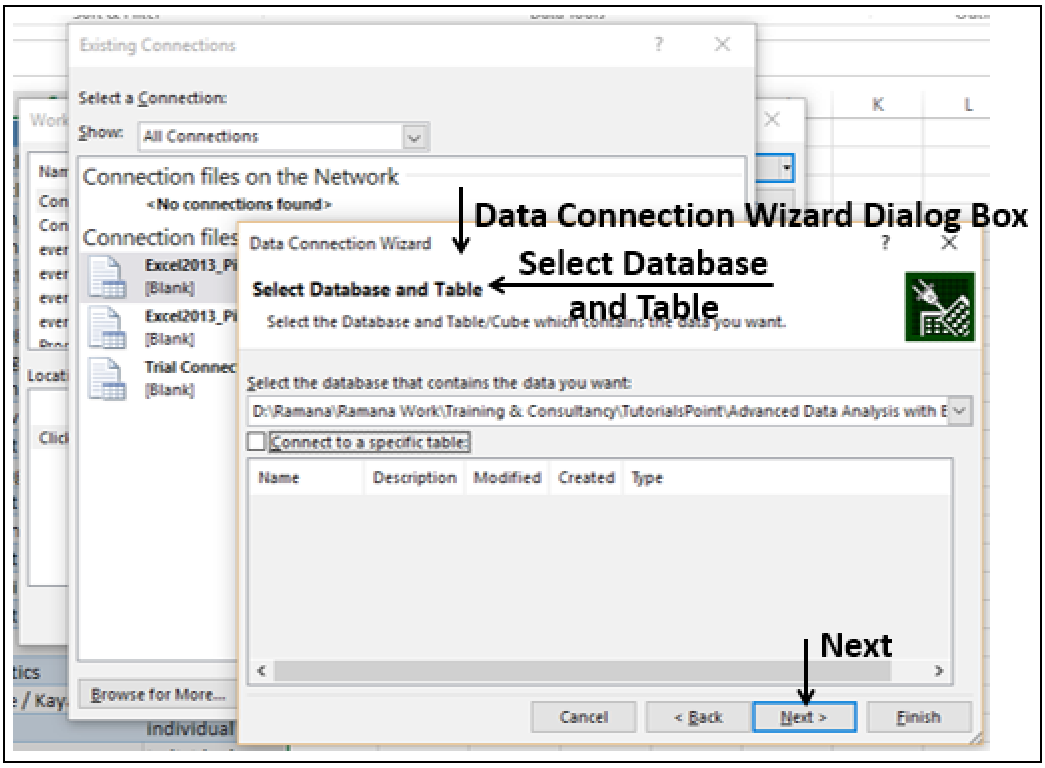
The Select Workbook dialog box appears.



Step 7 – Browse to the location where you have the workbook to be imported is located. Click OK.

The Data Connection Wizard dialog box appears with Select Database and Table.

Note – In this case, Excel treats each worksheet that is getting imported as a table. The table name will be the worksheet name. So, to have meaningful table names, name / rename the worksheets as appropriate.


Data Connection Wizard Dialog Box

Select Database and Table

Next

Data Connection Wizard

Select Database and Table

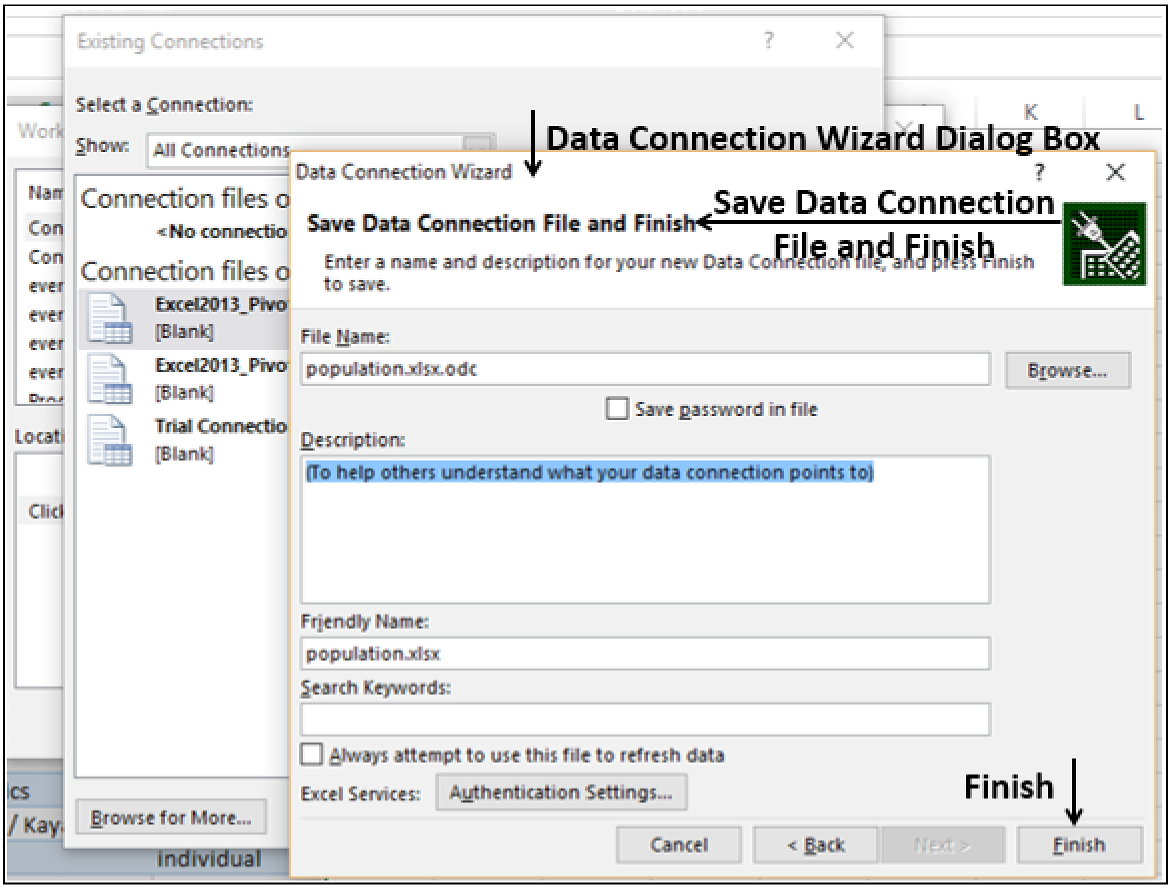
Select the Database and Table/Cube which contains the data you want.

Select the database that contains the data you want:

D:\Ramana\Ramana Work\Training & Consultancy\TutorialsPoint\Advanced Data Analysis with E

□ Connect to a specific table:

Step 8 − Click Next. The Data Connection Wizard dialog box appears with Save Data Connection File and Finish.


Data Connection Wizard Dialog Box

Save Data Connection File and Finish

Finish

Data Connection Wizard

Save Data Connection File and Finish

Enter a name and description for your new Data Connection file, and press Finish to save.

File Name: population.xlsx.odc

□ Save password in file

Description: (To help others understand what your data connection points to)

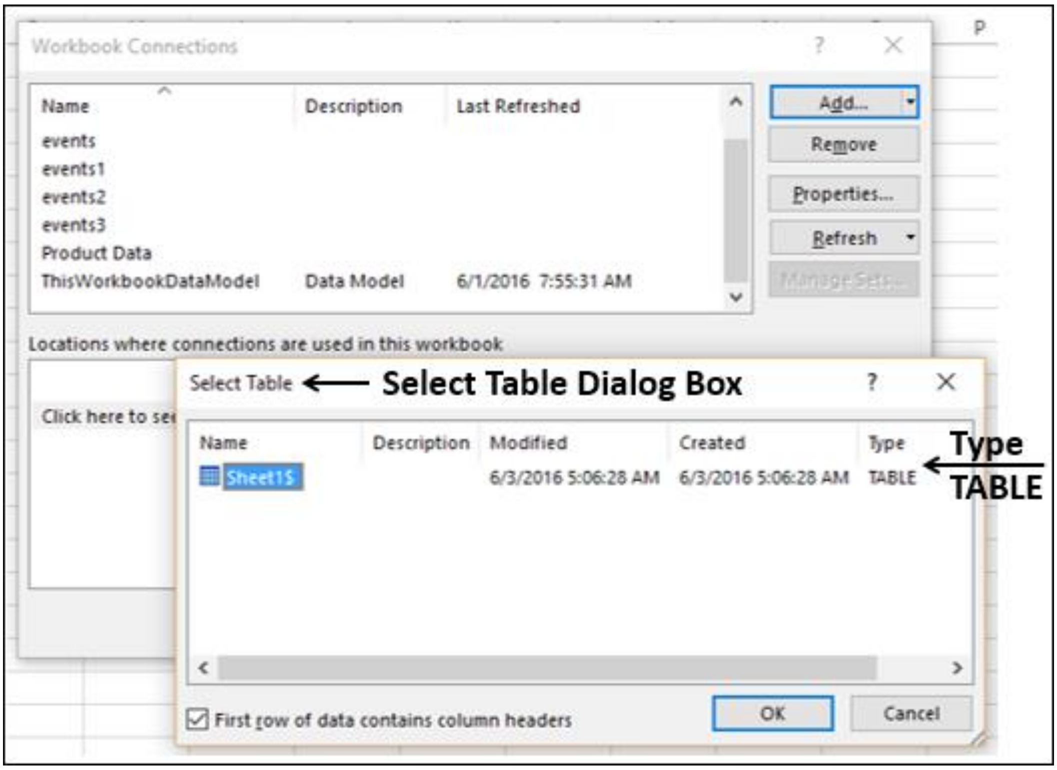
Friendly Name: population.xlsx

Search Keywords:

□ Always attempt to use this file to refresh data

Excel Services: Authentication Settings...

Step 9 − Click the Finish button. The Select Table dialog box appears.



Workbook Connections

Name	Description	Last Refreshed
events		
events1		
events2		
events3		
Product Data		
ThisWorkbookDataModel	Data Model	6/1/2016 7:55:31 AM

Locations where connections are used in this workbook

Select Table

Name	Description	Modified	Created	Type
Sheet1$		6/3/2016 5:06:28 AM	6/3/2016 5:06:28 AM	TABLE

☑ First row of data contains column headers

As you observe, Name is the worksheet name that is imported as type TABLE. Click OK.

The Data connection with the workbook you have chosen will be established.

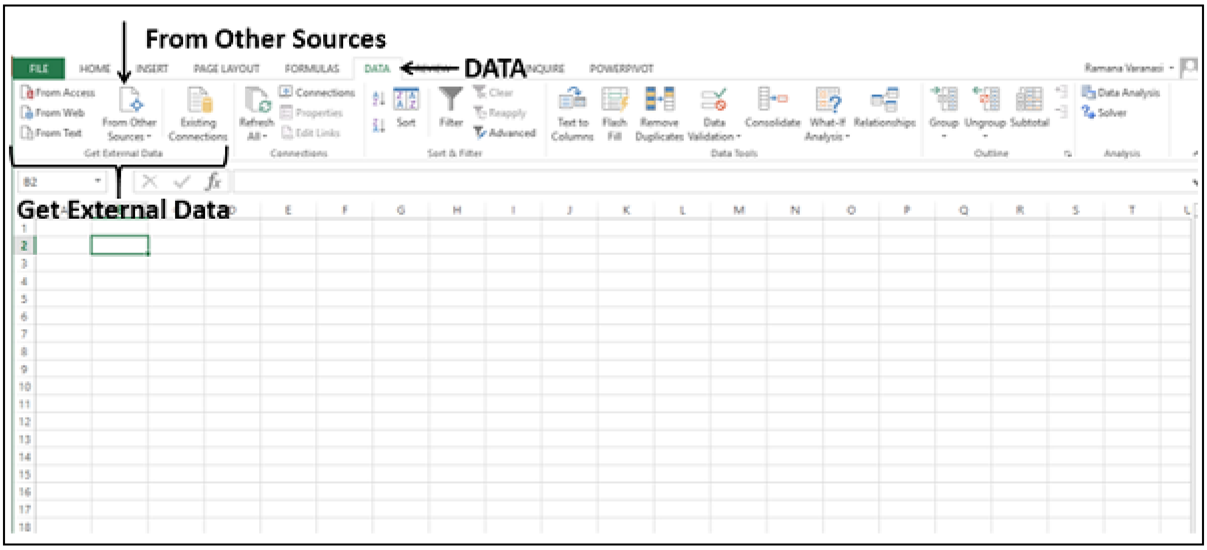
Importing Data from Other Sources

Excel provides you options to choose various other data sources. You can import data from these in few steps.

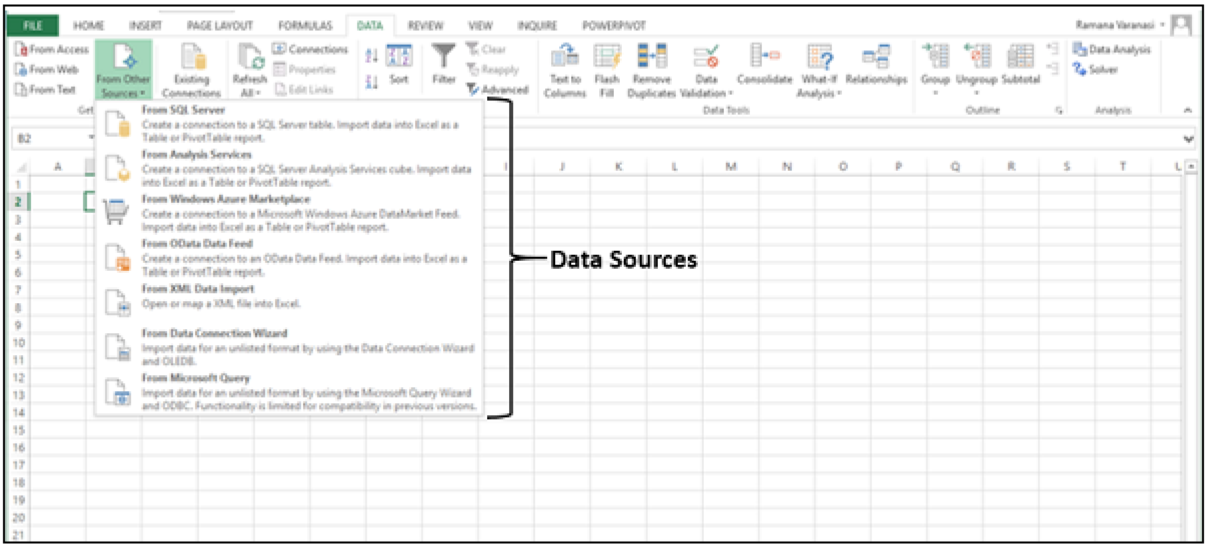
Step 1 – Open a new blank workbook in Excel.

Step 2 – Click the DATA tab on the Ribbon.

Step 3 – Click From Other Sources in the Get External Data group.



Dropdown with various data sources appears.



From SQL Server

Create a connection to a SQL Server table. Import data into Excel as a Table or PivotTable report.

From Analysis Services

Create a connection to a SQL Server Analysis Services cube. Import data into Excel as a Table or PivotTable report.

From Windows Azure Marketplace

Create a connection to a Microsoft Windows Azure DataMarket Feed. Import data into Excel as a Table or PivotTable report.

From OData Data Feed

Create a connection to an OData Data Feed. Import data into Excel as a Table or PivotTable report.

From XML Data Import

Open or map a XML file into Excel.

From Data Connection Wizard

Import data for an unlisted format by using the Data Connection Wizard and OLEDB.

From Microsoft Query

Import data for an unlisted format by using the Microsoft Query Wizard and ODBC. Functionality is limited for compatibility in previous versions.

You can import data from any of these data sources into Excel.

Importing Data using an Existing Connection

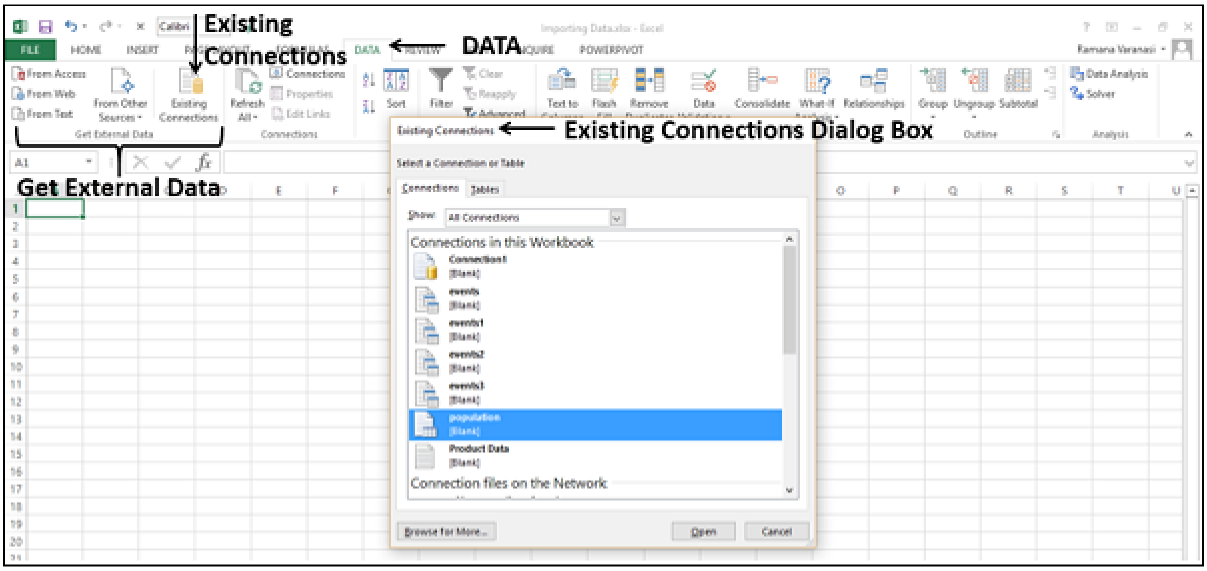
In an earlier section, you have established a data connection with a workbook.

Now, you can import data using that existing connection.

Step 1 – Click the DATA tab on the Ribbon.

Step 2 – Click Existing Connections in the Get External Data group. The Existing Connections dialog box appears.

Step 3 – Select the connection from where you want to import data and click Open.

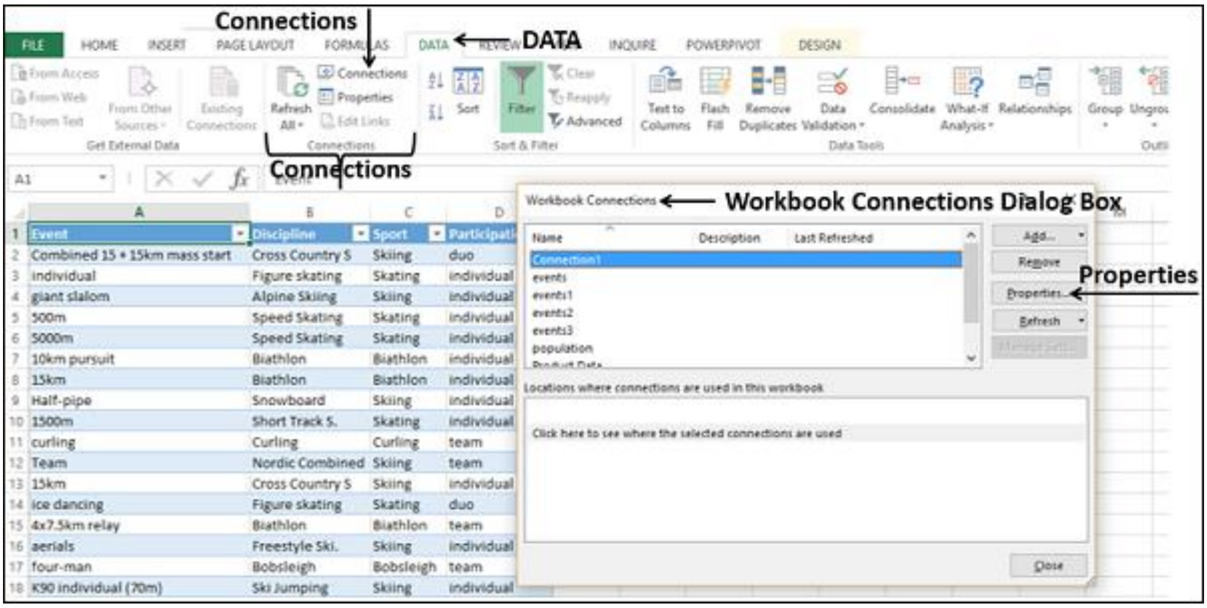


Renaming the Data Connections

It will be useful if the data connections you have in your workbook have meaningful names for the ease of understanding and locating.

Step 1 − Go to DATA > Connections on the Ribbon. The Workbook Connections dialog box appears.

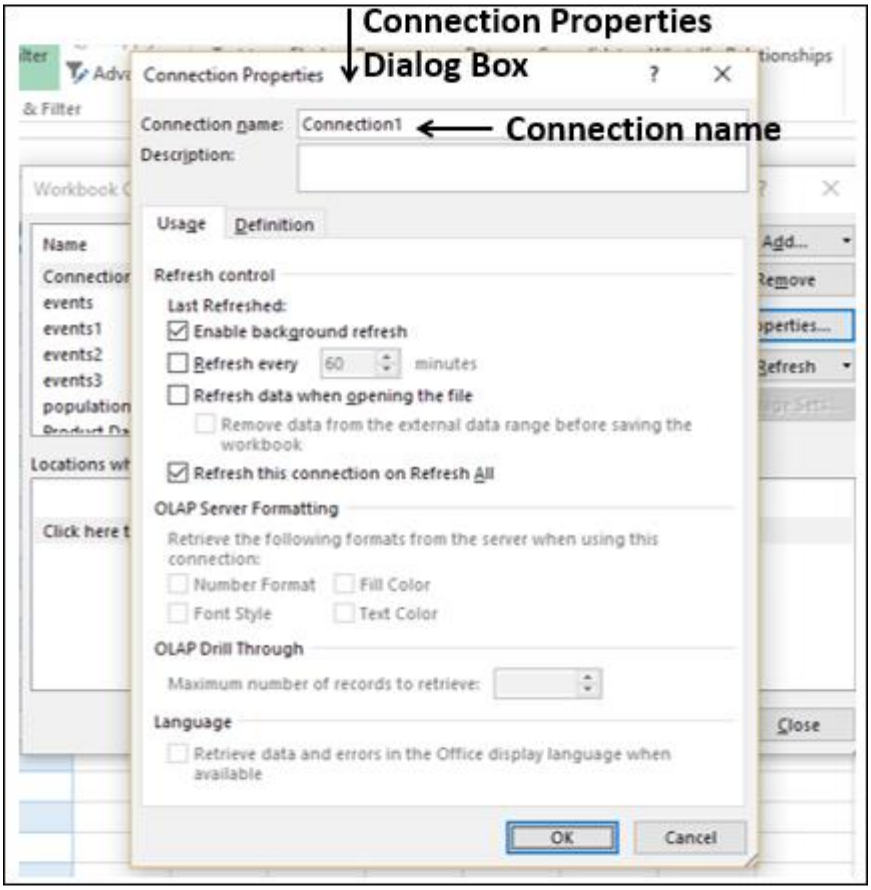
Step 2 − Select the connection that you want to rename and click Properties.



	A	B	C	D
1	Event	Discipline	Sport	Participat
2	Combined 15 + 15km mass start	Cross Country S	Skiing	duo
3	individual	Figure skating	Skating	individual
4	giant slalom	Alpine Skiing	Skiing	individual
5	500m	Speed Skating	Skating	individual
6	5000m	Speed Skating	Skating	individual
7	10km pursuit	Biathlon	Biathlon	individual
8	15km	Biathlon	Biathlon	individual
9	Half-pipe	Snowboard	Skiing	individual
10	1500m	Short Track S.	Skating	individual
11	curling	Curling	Curling	team
12	Team	Nordic Combined	Skiing	team
13	15km	Cross Country S	Skiing	individual
14	ice dancing	Figure skating	Skating	duo
15	4x7.5km relay	Biathlon	Biathlon	team
16	aerials	Freestyle Ski.	Skiing	individual
17	four-man	Bobsleigh	Bobsleigh	team
18	K90 individual (70m)	Ski Jumping	Skiing	individual

Connection1
events
events1
events2
events3
population
Product Data

The Connection Properties dialog box appears. The present name appears in the Connection name box −



Connection Properties Dialog Box

Connection name

Connection Properties

Connection name: Connection1

Description:

Refresh control

Last Refreshed:

☑ Enable background refresh

□ Refresh every 60 minutes

□ Refresh data when opening the file

□ Remove data from the external data range before saving the workbook

☑ Refresh this connection on Refresh All

OLAP Server Formatting

Retrieve the following formats from the server when using this connection:

□ Number Format □ Fill Color

□ Font Style □ Text Color

OLAP Drill Through

Maximum number of records to retrieve:

Language

□ Retrieve data and errors in the Office display language when available

Step 3 − Edit the Connection name and click OK. The data connection will have the new name that you have given.

Refreshing an External Data Connection

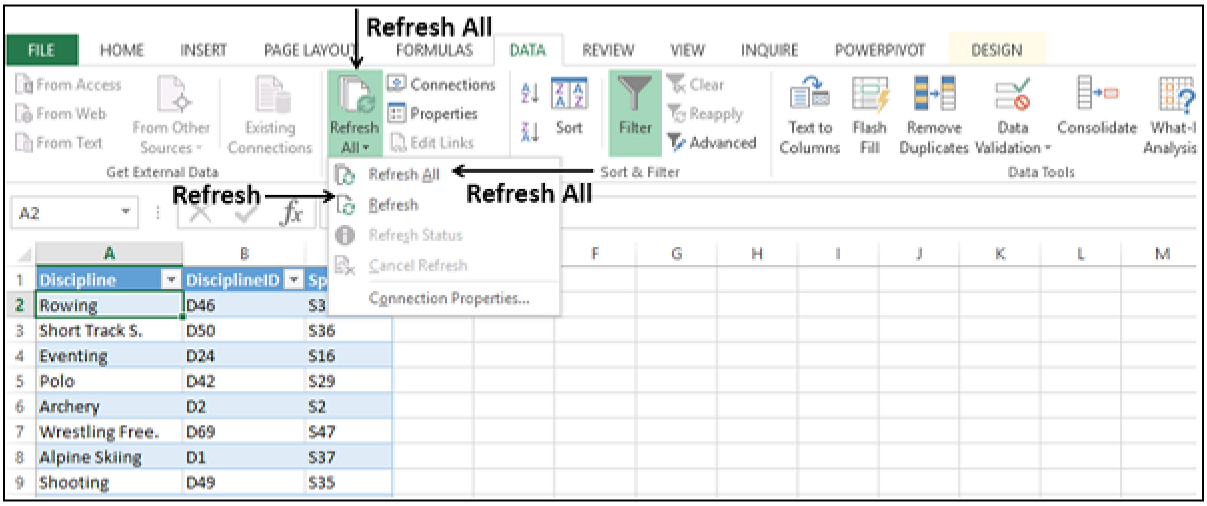
When you connect your Excel workbook to an external data source, as you have seen in the above sections, you would like to keep the data in your workbook up to date reflecting the changes made to the external data source time to time.

You can do this by refreshing the data connections you have made to those data sources. Whenever you refresh the data connection, you see the most recent data changes from that data source, including anything that is new or that is modified or that has been deleted.

You can either refresh only the selected data or all the data connections in the workbook at once.

Step 1 − Click the DATA tab on the Ribbon.

Step 2 − Click Refresh All in the Connections group.



	A	B	
1	Discipline	DisciplineID	Sp
2	Rowing	D46	S3
3	Short Track S.	D50	S36
4	Eventing	D24	S16
5	Polo	D42	S29
6	Archery	D2	S2
7	Wrestling Free.	D69	S47
8	Alpine Skiing	D1	S37
9	Shooting	D49	S35

As you observe, there are two commands in the dropdown list – Refresh and Refresh All.

- If you click Refresh, the selected data in your workbook is updated.
- If you click Refresh All, all the data connections to your workbook are updated.

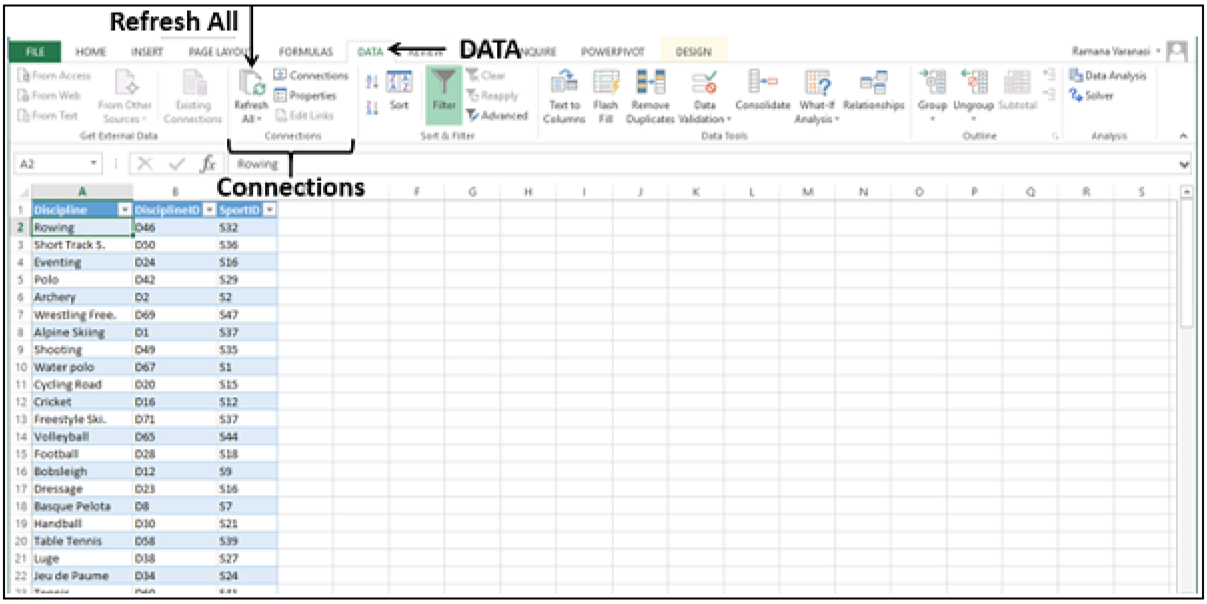
Updating all the Data Connections in the Workbook

You might have several data connections to your workbook. You need to update them from time to time so that your workbook will have access to the most recent data.

Step 1 − Click any cell in the table that contains the link to the imported data file.

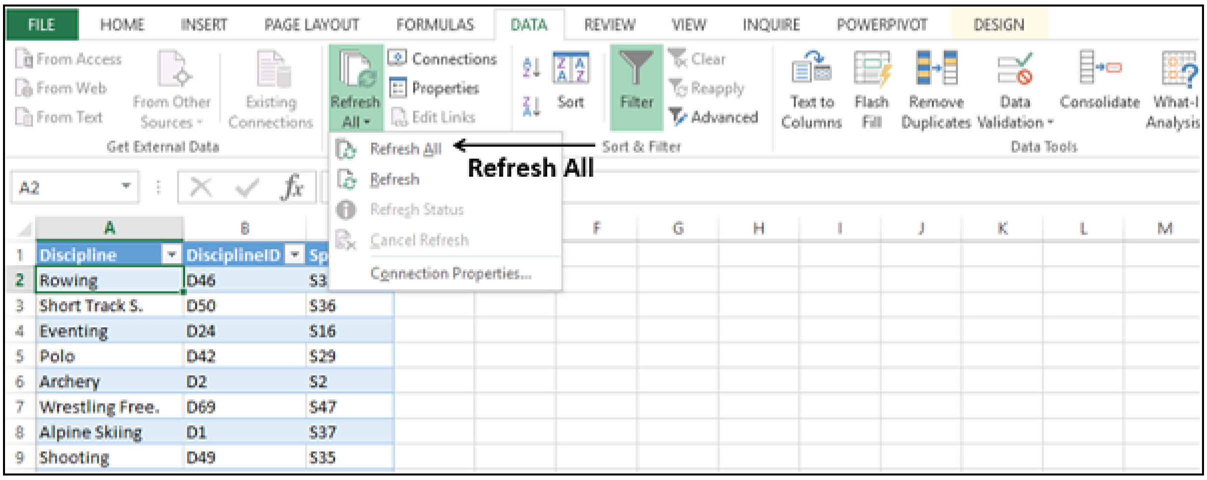
Step 2 − Click the Data tab on the Ribbon.

Step 3 − Click Refresh All in the Connections group.



	A	B	
1	Discipline	DisciplineID	SportID
2	Rowing	D46	S32
3	Short Track S.	D50	S36
4	Eventing	D24	S16
5	Polo	D42	S29
6	Archery	D2	S2
7	Wrestling Free.	D69	S47
8	Alpine Skiing	D1	S37
9	Shooting	D49	S35
10	Water polo	D67	S1
11	Cycling Road	D20	S15
12	Cricket	D16	S12
13	Freestyle Ski.	D71	S37
14	Volleyball	D65	S44
15	Football	D28	S18
16	Bobsleigh	D12	S9
17	Dressage	D23	S16
18	Basque Pelota	D8	S7
19	Handball	D30	S21
20	Table Tennis	D58	S39
21	Luge	D38	S27
22	Jeu de Paume	D34	S24

Step 4 − Select Refresh All from the dropdown list. All the data connections in the workbook will be updated.



A2		
Discipline	DisciplineID	Sp
Rowing	D46	S3
Short Track S.	D50	S36
Eventing	D24	S16
Polo	D42	S29
Archery	D2	S2
Wrestling Free.	D69	S47
Alpine Skiing	D1	S37
Shooting	D49	S35

Automatically Refresh Data when a Workbook is opened

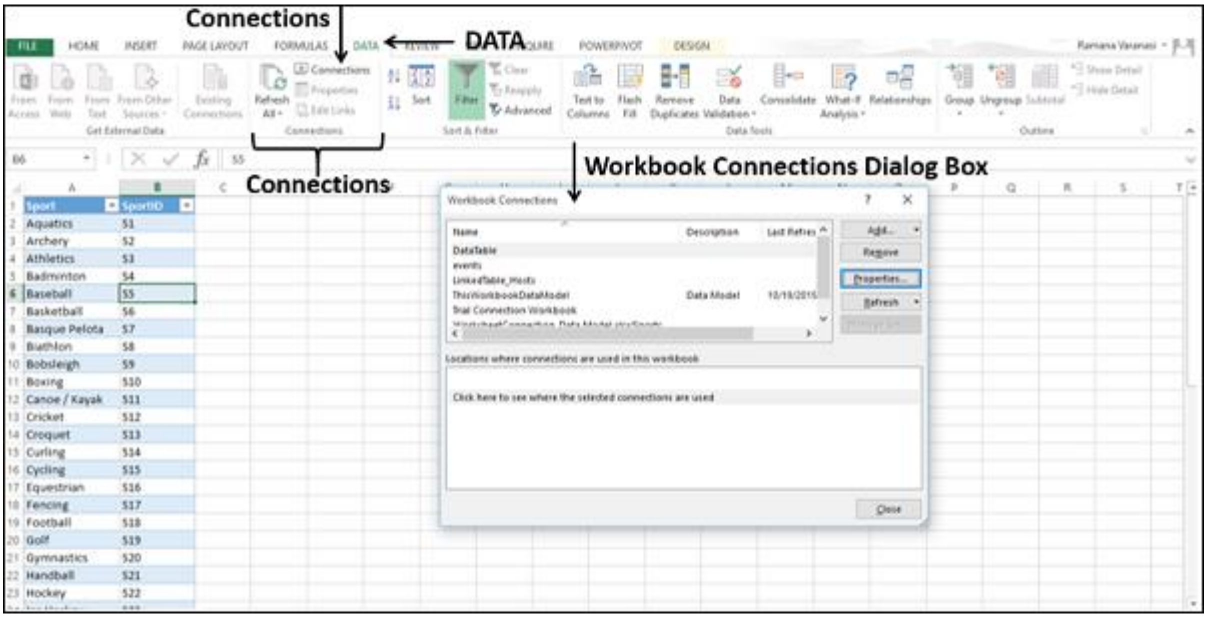
You might want to have access to the recent data from the data connections to your workbook whenever your workbook is opened.

Step 1 – Click any cell in the table that contains the link to the imported data file.

Step 2 – Click the Data tab.

Step 3 – Click Connections in the Connections group.

The Workbook Connections dialog box appears.



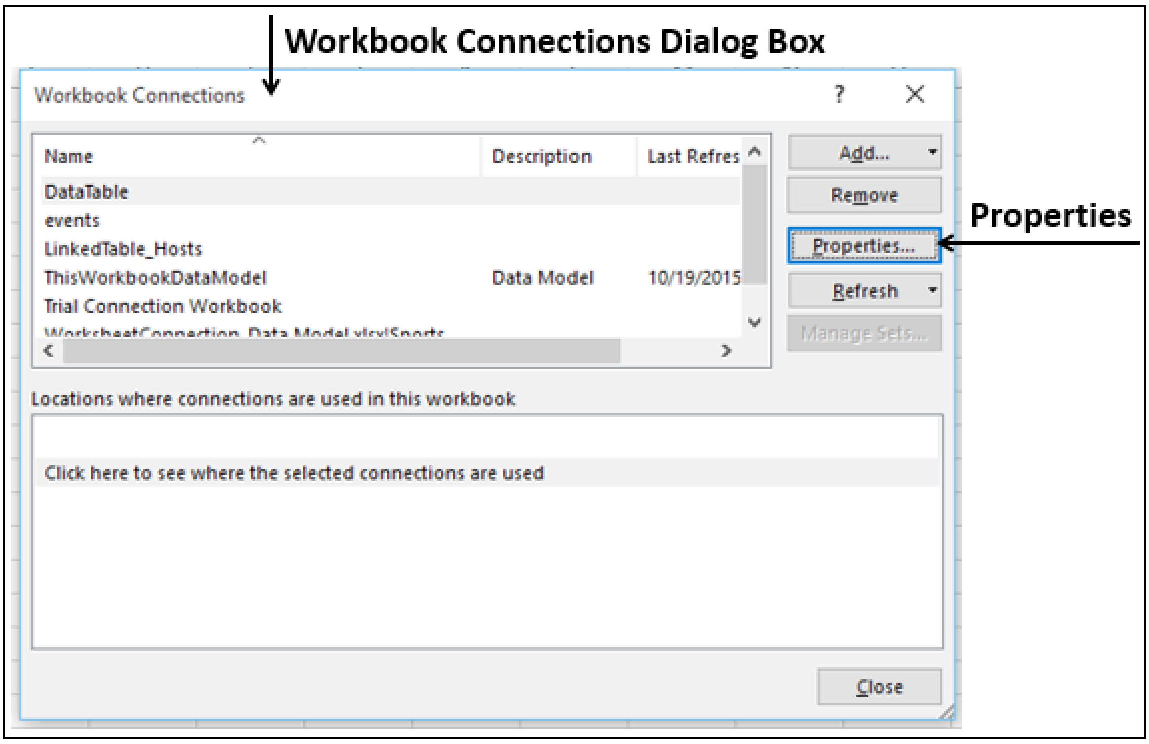
Sport	SportID
Aquatics	S1
Archery	S2
Athletics	S3
Badminton	S4
Baseball	S5
Basketball	S6
Basque Pelota	S7
Biathlon	S8
Bobsleigh	S9
Boxing	S10
Canoe / Kayak	S11
Cricket	S12
Croquet	S13
Curling	S14
Cycling	S15
Equestrian	S16
Fencing	S17
Football	S18
Golf	S19
Gymnastics	S20
Handball	S21
Hockey	S22

Name	Description	Last Refres
DataTable		
events		
LinkedTable_Hosts		
ThisWorkbookDataModel	Data Model	10/19/2015
Trial Connection Workbook		

Locations where connections are used in this workbook

Click here to see where the selected connections are used

Step 4 – Click the Properties button. The Connection Properties dialog box appears.



Workbook Connections Dialog Box

Properties

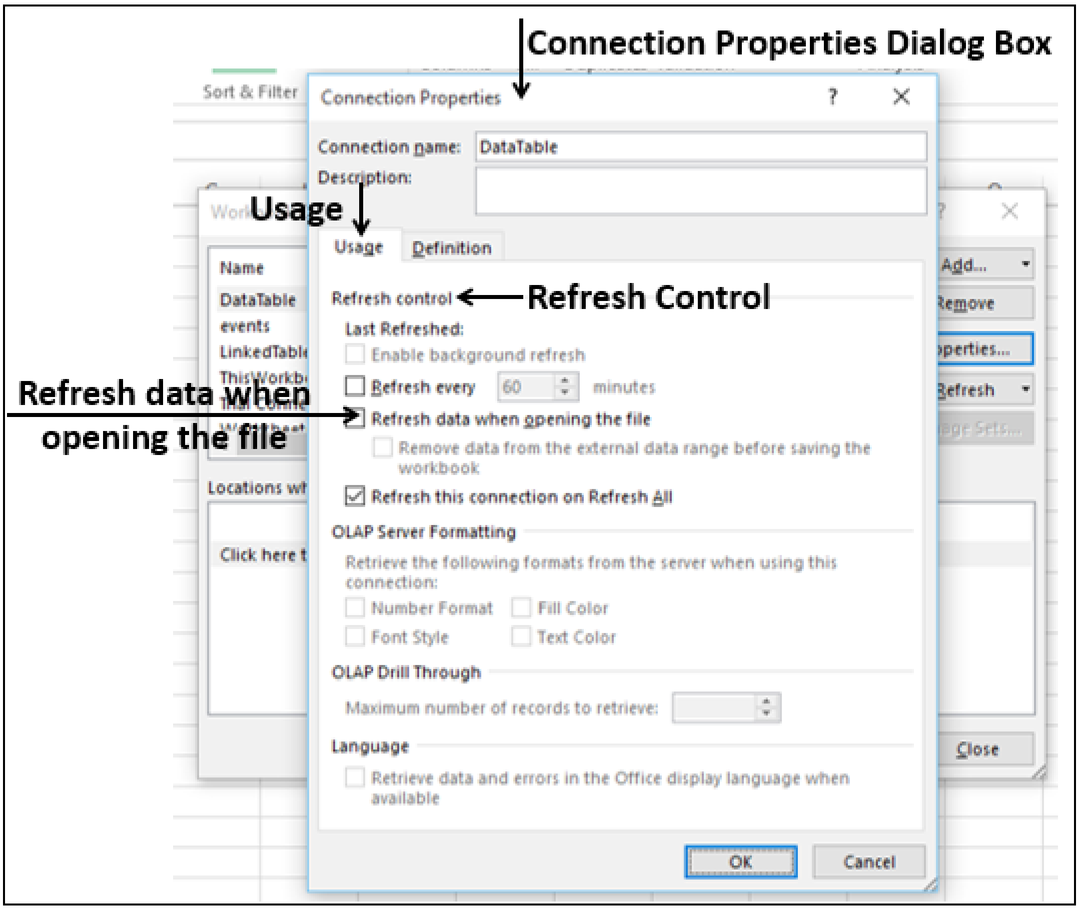
Workbook Connections

Name	Description	Last Refres
DataTable		
events		
LinkedTable_Hosts		
ThisWorkbookDataModel	Data Model	10/19/2015
Trial Connection Workbook		
WorksheetConnection_Data Model.xlsx!Sports		

Locations where connections are used in this workbook

Click here to see where the selected connections are used

Step 5 – Click the Usage tab.



Connection Properties Dialog Box

Usage

Refresh Control

Refresh data when opening the file

Connection Properties

Connection name: DataTable

Description:

Usage | Definition

Refresh control

Last Refreshed:

□ Enable background refresh

□ Refresh every 60 minutes

☑ Refresh data when opening the file

□ Remove data from the external data range before saving the workbook

☑ Refresh this connection on Refresh All

OLAP Server Formatting

Retrieve the following formats from the server when using this connection:

□ Number Format □ Fill Color

□ Font Style □ Text Color

OLAP Drill Through

Maximum number of records to retrieve:

Language

□ Retrieve data and errors in the Office display language when available

Step 6 – Check the option - Refresh data when opening the file.

You have another option also - Remove data from the external data range before saving the workbook. You can use this option to save the workbook with the query definition but without the external data.

Step 7 − Click OK. Whenever you open your workbook, the up to date data will be loaded into your workbook.

Automatically Refresh Data at regular Intervals

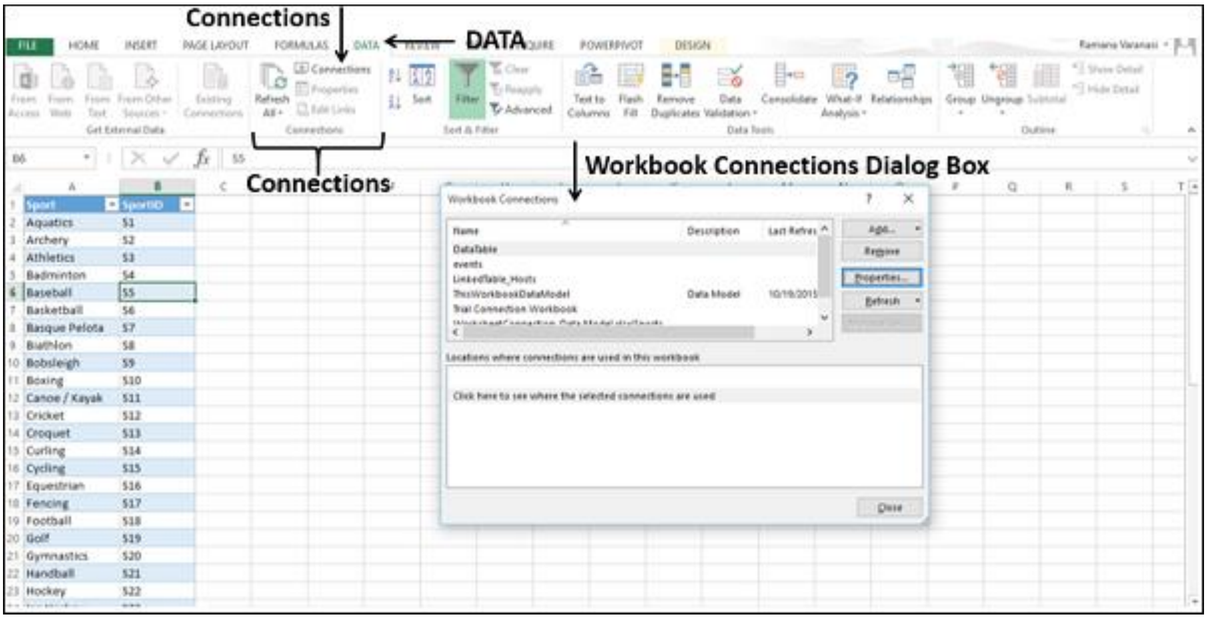
You might be using your workbook keeping it open for longer durations. In such a case, you might want to have the data refreshed periodically without any intervention from you.

Step 1 − Click any cell in the table that contains the link to the imported data file.

Step 2 − Click the Data tab on the Ribbon.

Step 3 − Click Connections in the Connections group.

The Workbook Connections dialog box appears.

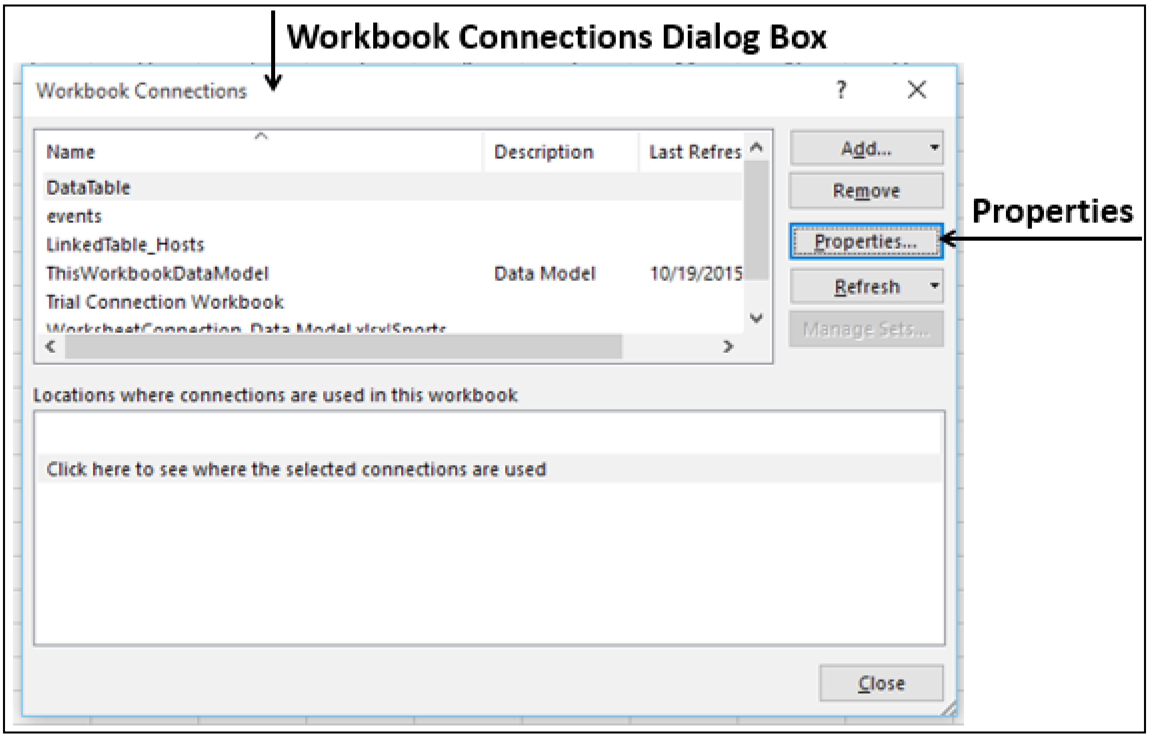


Name	Description	Last Refres
DataTable		
events		
LinkedTable_Hosts		
ThisWorkbookDataModel	Data Model	10/19/2015
Trial Connection Workbook		

Locations where connections are used in this workbook

Click here to see where the selected connections are used

Step 4 − Click the Properties button.



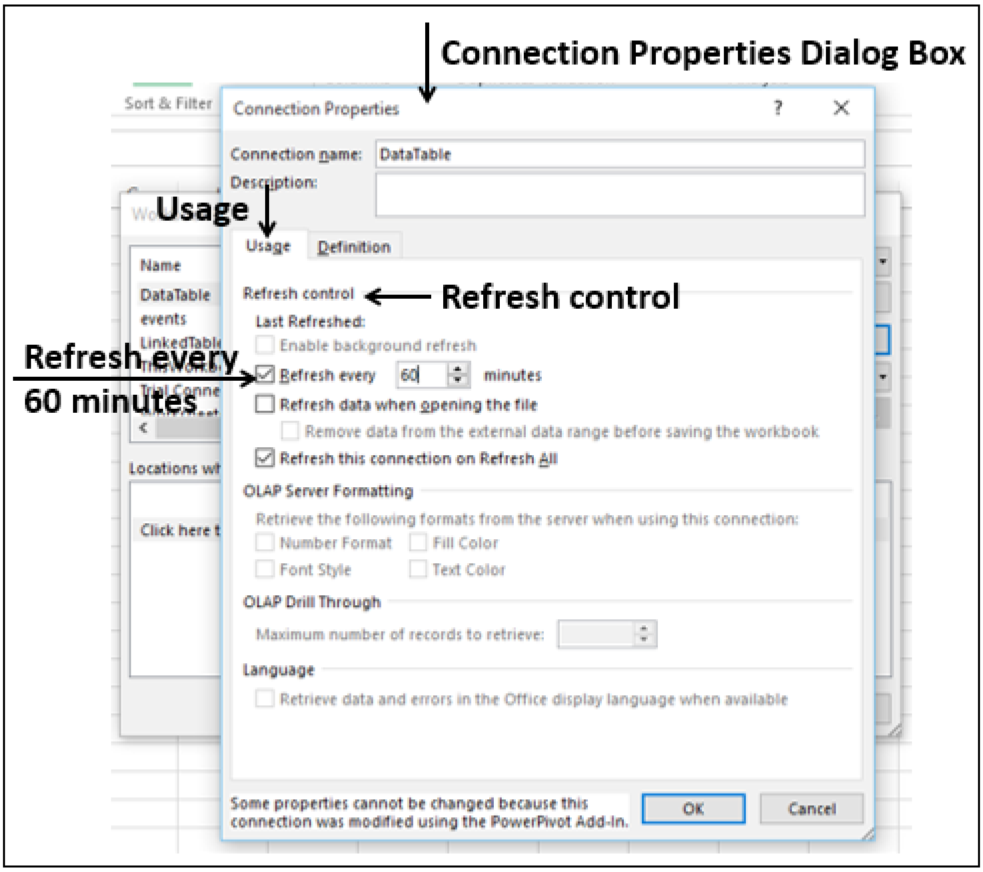
Name	Description	Last Refres
DataTable		
events		
LinkedTable_Hosts		
ThisWorkbookDataModel	Data Model	10/19/2015
Trial Connection Workbook		
WorksheetConnection_Data Model.xlsx!Sports		

Locations where connections are used in this workbook

Click here to see where the selected connections are used

The Connection Properties dialog box appears. Set the properties as follows −

- Click the Usage tab.
- Check the option Refresh every.
- Enter 60 as the number of minutes between each refresh operation and click Ok.



Connection name: DataTable

Description:

Refresh control

Last Refreshed:

□ Enable background refresh

☑ Refresh every 60 minutes

□ Refresh data when opening the file

□ Remove data from the external data range before saving the workbook

☑ Refresh this connection on Refresh All

OLAP Server Formatting

Retrieve the following formats from the server when using this connection:

□ Number Format □ Fill Color

□ Font Style □ Text Color

OLAP Drill Through

Maximum number of records to retrieve:

Language

□ Retrieve data and errors in the Office display language when available

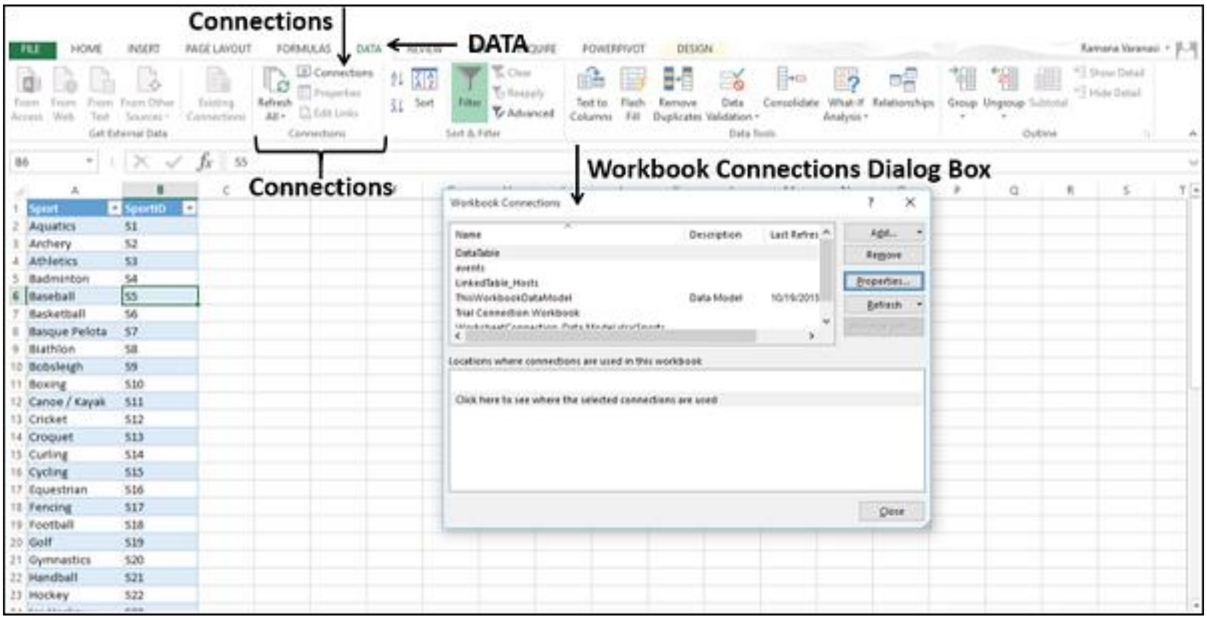
Some properties cannot be changed because this connection was modified using the PowerPivot Add-In.

Your Data will be automatically refreshed every 60 min. (i.e. every one hour).

Enabling Background Refresh

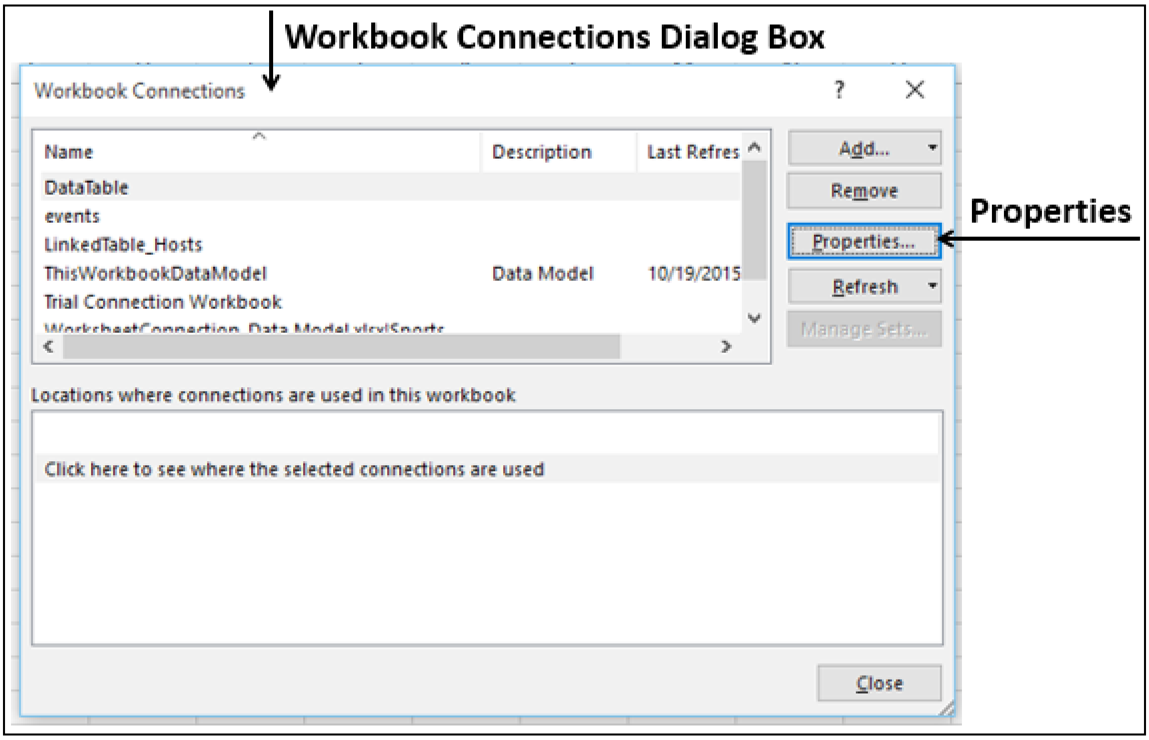
For very large data sets, consider running a background refresh. This returns control of Excel to you instead of making you wait several minutes or more for the refresh to finish. You can use this option when you are running a query in the background. However, during this time, you cannot run a query for any connection type that retrieves data for the Data Model.

- Click in any cell in the table that contains the link to the imported data file.
- Click the Data tab.
- Click Connections in the Connections group. The Workbook Connections dialog box appears.


Connections
DATA
Connections
Workbook Connections Dialog Box
Workbook Connections
Add...
Remove
Properties...
Refresh
Close
Locations where connections are used in this workbook
Click here to see where the selected connections are used

Name	Description	Last Refres
DataTable		
events		
LinkedTable_Hosts		
ThisWorkbookDataModel	Data Model	10/19/2015
Trial Connection Workbook		

Click the Properties button.

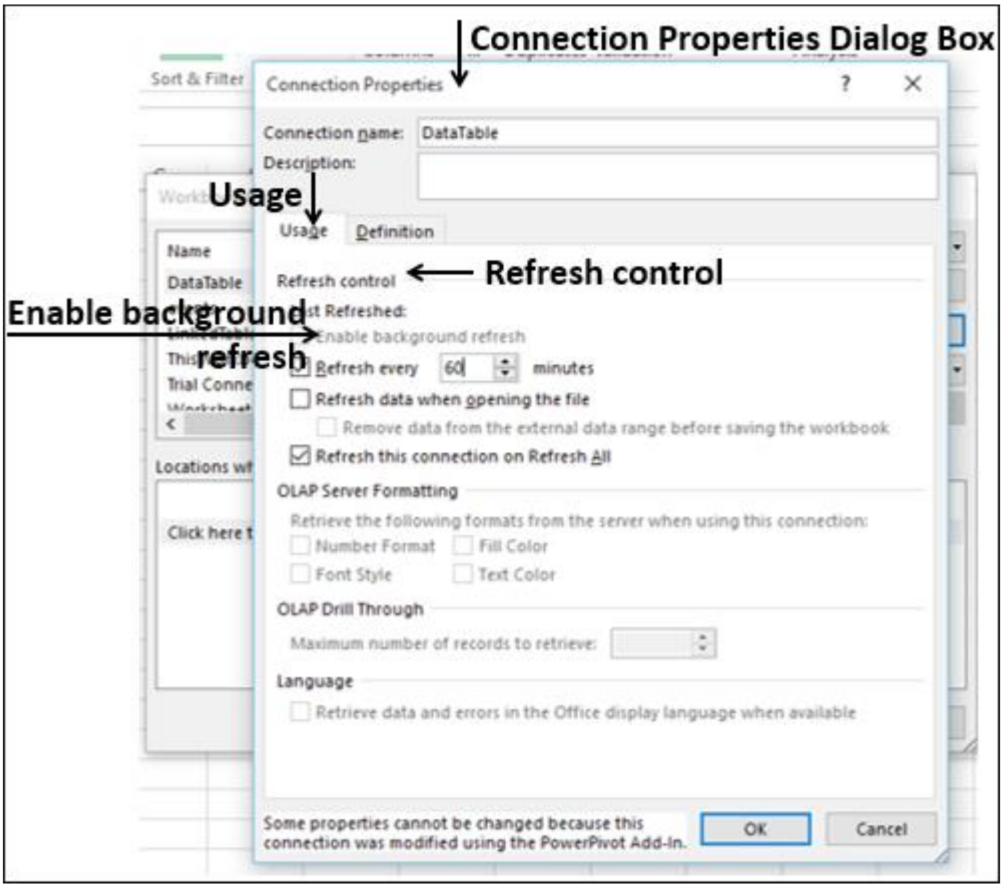


Name	Description	Last Refres
DataTable		
events		
LinkedTable_Hosts		
ThisWorkbookDataModel	Data Model	10/19/2015
Trial Connection Workbook		
WorksheetConnection_Data Model.xlsx!Sports		

Locations where connections are used in this workbook

Click here to see where the selected connections are used

The Connection Properties dialog box appears. Click the Usage tab. The Refresh Control options appear.



Connection name: DataTable

Description:

Refresh control

Last Refreshed:

□ Enable background refresh

□ Refresh every 60 minutes

□ Refresh data when opening the file

□ Remove data from the external data range before saving the workbook

☑ Refresh this connection on Refresh All

OLAP Server Formatting

Retrieve the following formats from the server when using this connection:

□ Number Format □ Fill Color

□ Font Style □ Text Color

OLAP Drill Through

Maximum number of records to retrieve:

Language

□ Retrieve data and errors in the Office display language when available

Some properties cannot be changed because this connection was modified using the PowerPivot Add-In.

- Click Enable background refresh.
- Click OK. The Background refresh is enabled for your workbook.

Experimental Designs and Their Analysis

Design of experiment means how to design an experiment in the sense that how the observations ormeasurements should be obtained to answer a query in a valid, efficient and economical way. The designing of the experiment and the analysis of obtained data are inseparable. If the experiment is designed properly keeping in mind the question, then the data generated is valid and proper analysis of data provides the valid statistical inferences. If the experiment is not well designed, the validity of the statistical inferences is questionable and may be invalid.

It is important to understand first the basic terminologies used in the experimental design.

Experimental unit:

For conducting an experiment, the experimental material is divided into smaller parts and each part isreferred to as an experimental unit. The experimental unit is randomly assigned to treatment is the experimental unit. The phrase “randomly assigned” is very important in this definition.

Experiment:

A way of getting an answer to a question which the experimenter wants to know.

Treatment

Different objects or procedures which are to be compared in an experiment are called treatments.

Sampling unit:

The object that is measured in an experiment is called the sampling unit. This may be different from the experimental unit.

Factor:

A factor is a variable defining a categorization. A factor can be fixed or random in nature. A factor is termedas a fixed factor if all the levels of interest are included in the experiment.

A factor is termed as a random factor if all the levels of interest are not included in the experiment and thosethat are can be considered to be randomly chosen from all the levels of interest.

Design of experiment:

One of the main objectives of designing an experiment is how to verify the hypothesis in an efficient and economical way. In the contest of the null hypothesis of equality of several means of normal populations having the same variances, the analysis of variance technique can be used. Note that such techniques are based on certain statistical assumptions. If these assumptions are violated, the outcome of the test of a hypothesis then may also be faulty and the analysis of data may be meaningless. So the main question is how to obtain the data such that the assumptions are met and the data is readily available for the application of tools like analysis of variance. The designing of such a mechanism to obtain such data is achieved by the design of the experiment. After obtaining the sufficient experimental unit, the treatments are allocated to the experimental units in a random fashion. Design of experiment provides a method by which the treatments are placed at random on the experimental units in such a way that the responses are estimated with theutmost precision possible.

Principles of experimental design:

There are three basic principles of design which were developed by Sir Ronald A. Fisher.

(i) Randomization

(ii) Replication

(iii) Local control

(i) Randomization

The principle of randomization involves the allocation of treatment to experimental units at random to avoid any bias in the experiment resulting from the influence of some extraneous unknown factor that may affect the experiment. In the development of analysis of variance, we assume that the errors are random and independent. In turn, the observations also become random. The principle of randomization ensures this.
The random assignment of experimental units to treatments results in the following outcomes.

a) It eliminates systematic bias.

b) It is needed to obtain a representative sample from the population.

c) It helps in distributing the unknown variation due to confounded variables throughout the experiment and breaks the confounding influence.

Randomization forms a basis of a valid experiment but replication is also needed for the validity of the experiment.
If the randomization process is such that every experimental unit has an equal chance of receiving each treatment, it is called complete randomization.

(ii) Replication:

In the replication principle, any treatment is repeated a number of times to obtain a valid and more reliable estimate than which is possible with one observation only. Replication provides an efficient way of increasing the precision of an experiment. The precision increases with the increase in the number of observations. Replication provides more observations when the same treatment is used, so it increases
Precision. For example, if the variance of x is σ^2 than variance of the sample mean $\bar{x}$ based on n Observation.

(ii) Local control (error control)

The replication is used with local control to reduce the experimental error. For example, if the experimental units are divided into different groups such that they are homogeneous within the blocks, then the variation among the blocks is eliminated and ideally, the error component will contain the variation due to the treatments only. This will, in turn, increase the efficiency.

Experimental Design for ANOVA

There is a close relationship between experimental design and statistical analysis. The way that an experiment is designed determines the types of analyses that can be appropriately conducted.

In this lesson, we review aspects of experimental design that a researcher must understand in order to properly interpret experimental data with analysis of variance.

What Is an Experiment?

An experiment is a procedure carried out to investigate cause-and-effect relationships. For example, the experimenter may manipulate one or more variables (independent variables) to assess the effect on another variable (the dependent variable).

Conclusions are reached on the basis of data. If the dependent variable is unaffected by changes in independent variables, we conclude that there is no causal relationship between the dependent variable and the independent variables. On the other hand, if the dependent variable is affected, we conclude that a causal relationship exists.

Experimenter Control

One of the features that distinguish a true experiment from other types of studies is experimenter control of the independent variable(s).

In a true experiment, an experimenter controls the level of the independent variable administered to each subject. For example, dosage level could be an independent variable in a true experiment; because an experimenter can manipulate the dosage administered to any subject.

What is a Quasi-Experiment?

A quasi-experiment is a study that lacks a critical feature of a true experiment. Quasi-experiments can provide insights into cause-and-effect relationships; but evidence from a quasi-experiment is not as persuasive as evidence from a true experiment. True experiments are the gold standard for causal analysis.

A study that used gender or IQ as an independent variable would be an example of a quasi-experiment, because the study lacks experimenter control over the independent variable; that is, an experimenter cannot manipulate the gender or IQ of a subject.

As we discuss experimental design in the context of a tutorial on analysis of variance, it is important to point out that experimenter control is a requirement for a true experiment; but it is not a requirement for analysis of variance. Analysis of variance can be used with true experiments and with quasi-experiments that lack only experimenter control over the independent variable.

What Is Experimental Design?

The term experimental design refers to a plan for conducting an experiment in such a way that research results will be valid and easy to interpret. This plan includes three interrelated activities:

- Write statistical hypotheses.
- Collect data.
- Analyze data.

Statistical Hypotheses

A statistical hypothesis is an assumption about the value of a population parameter. There are two types of statistical hypotheses:

- Null hypothesis: The null hypothesis is the statement subjected to a statistical test in an experiment. It is denoted by H_0. For example, consider the following null hypothesis:

  $H_{0:}\ \mu_i = \mu_j$

  Here, μ_i is the population mean for group *i*, and μ_j is the population mean for group *j*. This hypothesis makes the assumption that population means in groups *i* and *j* are equal.

- Alternative hypothesis: The alternative hypothesis is the hypothesis that is tenable if the null hypothesis is rejected. It is denoted by H_1 or H_a. For example, consider the following alternative hypothesis:

  $H_{1:}\ \mu_i \neq \mu_j$

  This hypothesis makes the assumption that population means in groups *i* and *j* are not equal.

The null hypothesis and the alternative hypothesis are written to be mutually exclusive. If one is true, the other is not.

Experiments rely on sample data to test the null hypothesis. If experimental results, based on sample statistics, are consistent with the null hypothesis, the null hypothesis cannot be rejected; otherwise, the null hypothesis is rejected in favor of the alternative hypothesis.

Data Collection

The data collection phase of experimental design is all about methodology - how to run the experiment to produce valid, relevant statistics that can be used to test a null hypothesis.

Identify Variables

Every experiment exists to examine a cause-and-effect relationship. With respect to the relationship under investigation, an experimental design needs to account for three types of variables:

- Dependent variable. The dependent variable is the outcome being measured, the effect in a cause-and-effect relationship.
- Independent variables. An independent variable is a variable that is thought to be a possible cause in a cause-and-effect relationship.
- Extraneous variables. An extraneous variable is any other variable that could affect the dependent variable, but is not explicitly included in the experiment.

Note: The independent variables that are explicitly included in an experiment are also called factors.

Define Treatment Groups

In an experiment, treatment groups are built around factors, each group defined by a unique combination of factor levels.

For example, suppose that a drug company wants to test a new cholesterol medication. The dependent variable is total cholesterol level. One independent variable is dosage. And, since some drugs affect men and women differently, the researchers include an second independent variable - gender.

This experiment has two factors - dosage and gender. The dosage factor has three levels (0 mg, 50 mg, and 100 mg), and the gender factor has two levels (male and female). Given this combination of factors and levels, we can define six unique treatment groups, as shown below:

Gender	Dose		
	0 mg	50 mg	100 mg
Male	Group 1	Group 2	Group 3
Female	Group 4	Group 5	Group 6

Note: The experiment described above is an example of a quasi-experiment, because the gender factor cannot be manipulated by the experimenter.

Select Factor Levels

A factor in an experiment can be described by the way in which factor levels are chosen for inclusion in the experiment:

- Fixed factor. The experiment includes all factor levels about which inferences are to be made.
- Random factor. The experiment includes a random sample of levels from a much bigger population of factor levels.

Experiments can be described by the presence or absence of fixed or random factors:

- Fixed-effects model. All of the factors in the experiment are fixed.
- Random-effects model. All of the factors in the experiment are random.
- Mixed model. At least one factor in the experiment is fixed, and at least one factor is random.

The use of fixed factors versus random factors has implications for how experimental results are interpreted. With a fixed factor, results apply only to factor levels that are explicitly included in the experiment. With a random factor, results apply to every factor level from the population.

For example, consider the blood pressure experiment described above. Suppose the experimenter only wanted to test the effect of three particular dosage levels - 0 mg, 50 mg, and 100 mg. He would include those dosage levels in the experiment, and any research conclusions would apply to only those particular dosage levels. This would be an example of a fixed-effects model.

On the other hand, suppose the experimenter wanted to test the effect of any dosage level. Since it is not practical to test *every* dosage level, the experimenter might choose three dosage levels at random from the population of possible dosage levels. Any research conclusions would apply not only to the selected dosage levels, but also to other dosage levels that were not included explicitly in the experiment. This would be an example of a random-effects model.

Select Experimental Units

The experimental unit is the entity that provides values for the dependent variable. Depending on the needs of the study, an experimental unit may be a person, animal, plant, product - anything. For example, in the cholesterol study described above, researchers measured cholesterol level (the dependent variable) of people; so the experimental units were people.

Note: When the experimental units are people, they are often referred to as subjects. Some researchers prefer the term participant, because subject has a connotation that the person is subservient.

If time and money were no object, you would include the entire population of experimental units in your experiment. In the real world, where there is never enough time or money, you will usually select a sample of experimental units from the population.

Ultimately, you want to use sample data to make inferences about population parameters. With that in mind, it is best practice to draw a random sample of experimental units from the population. This provides a defensible, statistical basis for generalizing from sample findings to the larger population.

Finally, it is important to consider sample size. The larger the sample, the greater the statistical power; and the more confidence you can have in your results.

Assign Experimental Units to Treatments

Having selected a sample of experimental units, we need to assign each unit to one or more treatment groups. Here are two ways that you might assign experimental units to groups:

- Independent groups design. Each experimental unit is randomly assigned to one, and only one, treatment group. This is also known as a between-subjects design.
- Repeated measures design. Experimental units are assigned to more than one treatment group. This is also known as a within-subjects design.

Control for Extraneous Variables

Extraneous variables can mask effects of independent variables. Therefore, a good experimental design controls potential effects of extraneous variables. Here are a few strategies for controlling extraneous variables:

- Randomization Assign subjects randomly to treatment groups. This tends to distribute effects of extraneous variables evenly across groups.
- Repeated measures design. To control for individual differences between subjects (age, attitude, religion, etc.), assign each subject to multiple treatments. This strategy is called using subjects as their own control.
- Counterbalancing. In repeated measures designs, randomize or reverse the order of treatments among subjects to control for order effects (e.g., fatigue, practice).

As we describe specific experimental designs in upcoming lessons, we will point out the strategies that are used with each design to control the confounding effects of extraneous variables.

Data Analysis

Researchers follow a formal process to determine whether to reject a null hypothesis, based on sample data. This process, called hypothesis testing, consists of five steps:

- Formulate hypotheses. This involves stating the null and alternative hypotheses. Because the hypotheses are mutually exclusive, if one is true, the other must be false.
- Choose the test statistic. This involves specifying the statistic that will be used to assess the validity of the null hypothesis. Typically, in analysis of variance studies, researchers compute a F ratio to test hypotheses.
- Compute a P-value, based on sample data. Suppose the observed test statistic is equal to *S*. The P-value is the probability that the experiment would yield a test statistic as extreme as *S*, assuming the null hypothesis is true.

- Choose a significance level. The significance level, denoted by α, is the probability of rejecting the null hypothesis when it is really true. Researchers often choose a significance level of 0.05 or 0.01.
- Test the null hypothesis. If the P-value is smaller than the significance level, we reject the null hypothesis; if it is larger, we fail to reject.

A good experimental design includes a precise plan for data analysis. Before the first data point is collected, a researcher should know how experimental data will be processed to accept or reject the null hypotheses.

STATISTICAL PROCESS CONTROL

Definition:

Statistical process control (SPC) is defined as the use of statistical techniques to control a process or production method. SPC tools and procedures can help you monitor process behavior, discover issues in internal systems, and find solutions for production issues. Statistical process control is often used interchangeably with statistical quality control (SQC).

- SPC tools
- SQC vs. SPC
- The 7 quality control tools
- The 7 supplemental tools
- History of SPC
- SPC resources

SPC TOOLS

A popular SPC tool is the control chart, originally developed by Walter Shewhart in the early 1920s. A control chart helps one record data and lets you see when an unusual event, such as a very high or low observation compared with "typical" process performance, occurs.

Control charts attempt to distinguish between two types of process variation:

1. Common cause variation, which is intrinsic to the process and will always be present
2. Special cause variation, which stems from external sources and indicates that the process is out of statistical control

Various tests can help determine when an out-of-control event has occurred. However, as more tests are employed, the probability of a false alarm also increases.

SQC VERSUS SPC

Statistical quality control (SQC) is defined as the application of the 14 statistical and analytical tools (7-QC and 7-SUPP) to monitor process *outputs* (dependent variables). Statistical process control (SPC) is the application of the same 14 tools to control process *inputs* (independent variables). Although both terms are often used interchangeably, SQC includes acceptance sampling where SPC does not.

THE 7 QUALITY CONTROL (7-QC) TOOLS

In 1974, Dr. Kaoru Ishikawa brought together a collection of process improvement tools in his text *Guide to Quality Control*. Known around the world as the seven quality control (7-QC) tools, they are:

1. Cause-and-effect diagram (also called Ishikawa diagram or fishbone diagram)
2. Check sheet
3. Control chart
4. Histogram
5. Pareto chart
6. Scatter diagram
7. Stratification

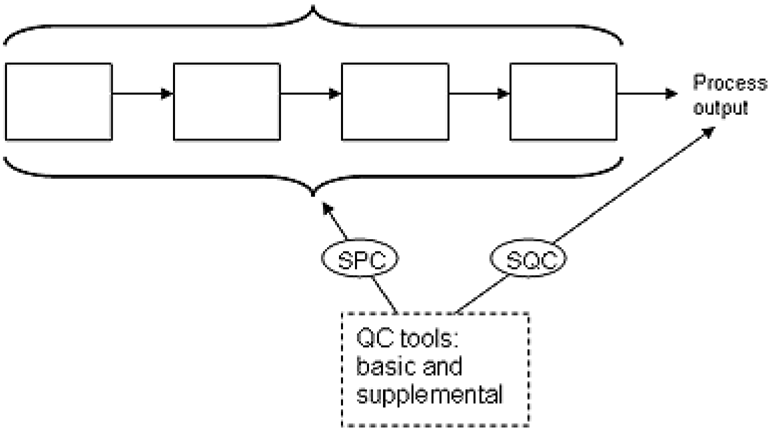
THE 7 SUPPLEMENTAL (7-SUPP) TOOLS

In addition to the basic 7-QC tools, there are also some additional statistical quality tools known as the seven supplemental (7-SUPP) tools:

1. Data stratification
2. Defect maps
3. Events logs
4. Process flowcharts
5. Progress centers
6. Randomization
7. Sample size determination

The Relationship Between Statistical Quality Control and Statistical Process Control

Design of experiments (DOE) and analysis of variance (AOV or ANOVA)



Why use Statistical Process Control?

Today companies are facing increasing competition and also operational costs, including raw material continuously increasing. So for the organizations, it is beneficial if they have control over their operation.

Organizations must make an effort for continuous improvement in quality, efficiency, and cost reduction. Many organizations still follow inspection after the production for quality related issues.

SPC helps companies to move towards prevention-based quality control instead of detection based quality controls. By monitoring SPC graphs, organizations can easily predict the behavior of the process.

Statistical Process Control Benefits

- Reduce scrap and rework
- Increase productivity
- Improve overall quality
- Match process capability to product requirement
- Continuously monitor process to maintaining control
- Provide data to support decision making
- Streamline the process
- Increase in product reliability
- Opportunity for company-wide improvements

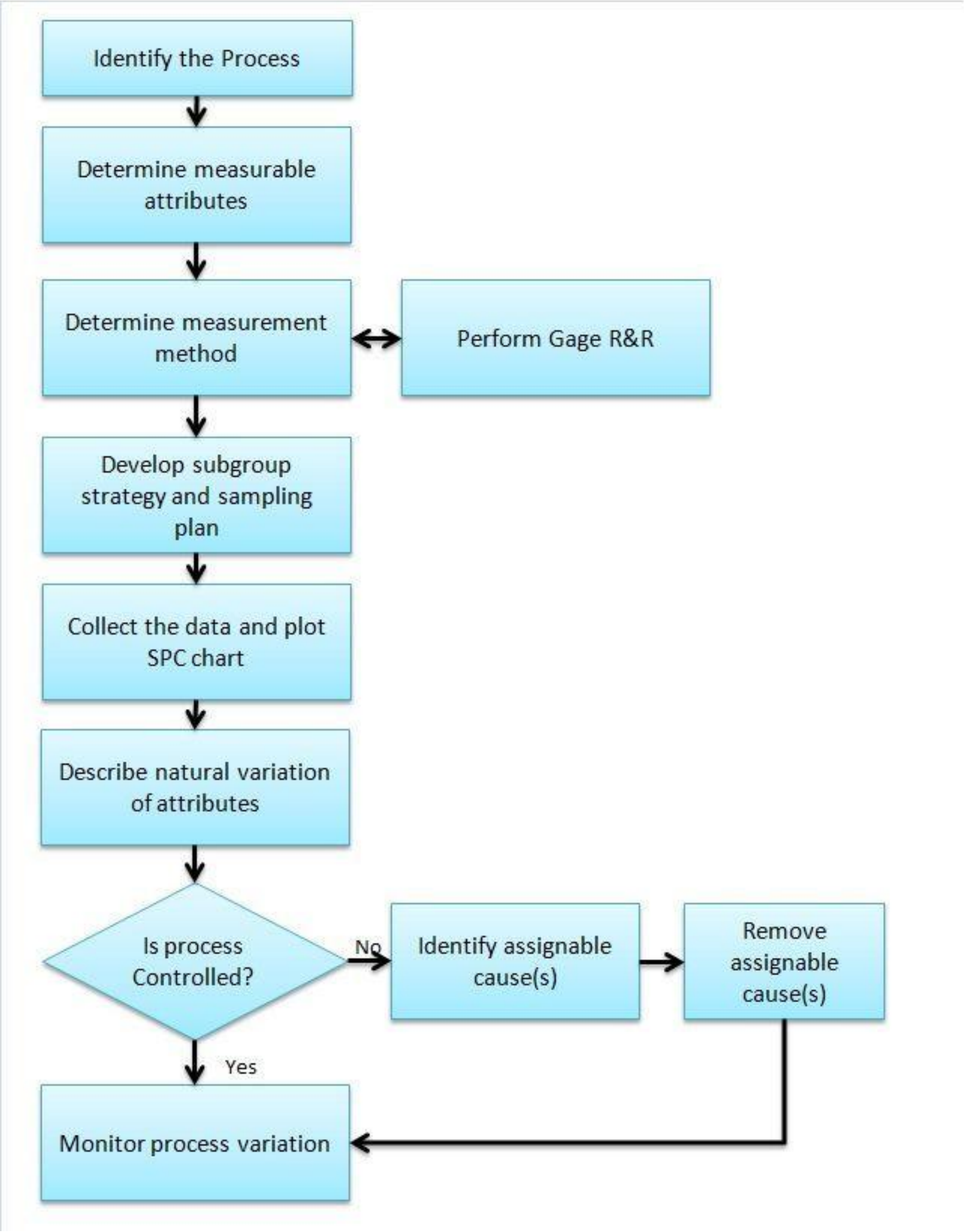
Statistical Process Control Objective

SPC focuses on optimizing continuous improvement by using statistical tools to analyze data, make inferences about process behavior, and then make appropriate decisions.

The basic assumption of SPC is that all processes are subject to variation. Variation measures how data are spread around the central tendency. Moreover, variation may be classified as one of two types, random or chance cause variation and assignable cause variation.

Common Cause: A cause of variation in the process is due to chance, but not assignable to any factor. It is the variation that is inherent in the process. Process under the influence of common cause will always be stable and predictable.

Assignable Cause: It is also known as “special cause”. The variation in a process that is not due to chance therefore can be identified and eliminated. Process under influence of special cause will not be stable and predictable.


Identify the Process
Determine measurable attributes
Determine measurement method
Perform Gage R&R
Develop subgroup strategy and sampling plan
Collect the data and plot SPC chart
Describe natural variation of attributes
Is process Controlled?
No
Identify assignable cause(s)
Remove assignable cause(s)
Yes
Monitor process variation

How to Perform SPC

1.Identify the processes: Identify the key process that impacts the output of the product or the process that is very critical to the customer. For example, plate thickness impacts the product's performance in a manufacturing company, then consider the plate manufacturing process.

2. Determine measurable attributes of the process: Identify the attributes that need to measure during the production. From the above example, consider the plate thickness as a measurable attribute.

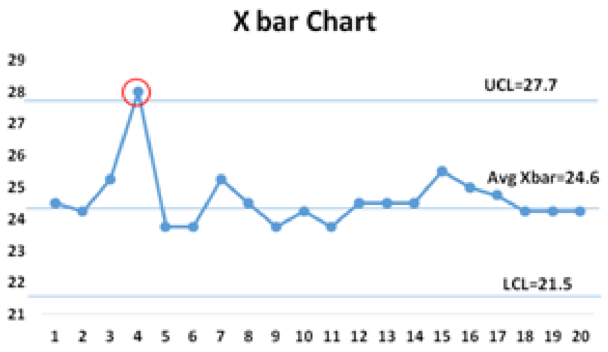
3. Determine the measurement method and also perform Gage R&R: Create a measurement method work instructions or procedure including the measuring instrument. For example, consider thickness gage to measure the thickness and create an appropriate measuring procedure. Perform Gage Repeatability and Reproducibility (Gage R & R) to define the amount of variation in the measurement data due to the measurement system.

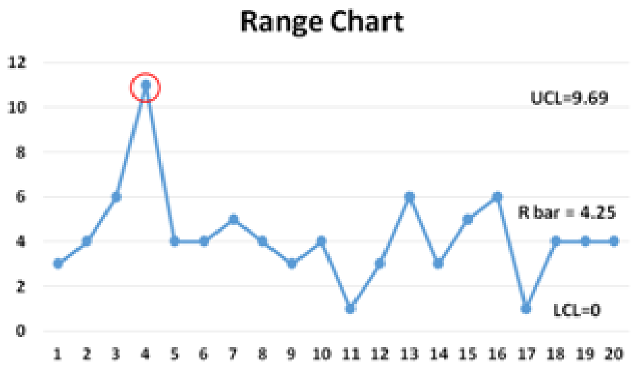
4. Develop a subgroup strategy and sampling plan: Determine the subgroup size based on the product's criticality and determine the sampling size and frequency. For example collect 20 sets of plate thicknesses in a time sequence with a subgroup size of 4.

5. Collect the data and plot SPC chart: Collect the data as per sample size and select an appropriate SPC chart based on data type (Continuous or Discrete) and also subgroup size. For Example, plate thicknesses with a subgroup size of 4, select Xbar -R chart.

6. Describe natural variation of attributes: Calculate the control limits. From the above example, calculate the upper control limit (UCL) and lower control limit (LCL) for both Xbar Range.

7. Monitor process variation: Interpret the control chart and check whether any point is out of control and the pattern. Example: check Xbar R chart If the process is not in control, then identify the assignable cause(s) and address the issue. This is an ongoing process to monitor the process variation.



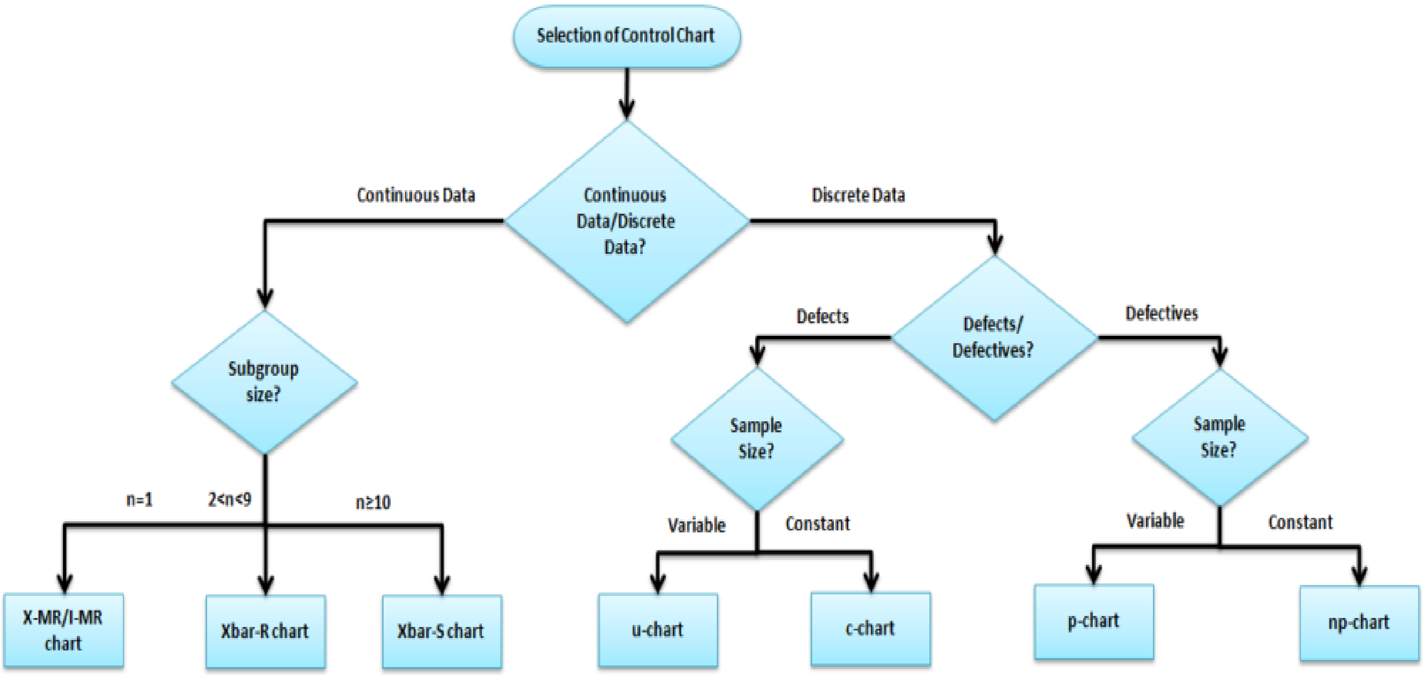


Additional Statistical Process Control Resources

Control limits are the voice of the process (different from specification limits, which are the customer's voice.) They show what the process is doing and act as a guide for what it should be doing. Control limits also indicate that a process event or measurement is likely to fall within that limit.

Control charts : A Control chart is one of the primary techniques of statistical process control (SPC). The control chart is a graphical display of quality characteristics that have been measured or computed from a sample versus the sample number or time. Furthermore, control chart contains a center line represents the average value of the quality characteristics and two other horizontal lines known as upper control limit (UCL) and lower control limit (LCL)

Selection of an appropriate control chart is very important in control charts mapping. Otherwise, it ended up with inaccurate control limits for the data. The selection of control chart depends on the data type: Continuous or Discrete?



Variable (Continuous) Control Charts

Measure the output on a continuous scale. It is possible to measure the quality characteristics of a product.

- X bar – R Charts (when data is readily available) X bar R chart is used to monitor the process performance of a continuous data and the data to be collected in subgroups at a set time periods. It is actually a two plots to monitor the process mean and the process variation over the time.
- Run Charts (limited single-point data) A run chart displays observed data as they evolve over time. Just a basic graph that displays data values in a time order. Can be useful for identifying trends or shifts in process but also allows you to test for randomness in the process.
- X – MR Charts (I – MR, individual moving range) I-MR chart also known as X-MR chart. An Individual moving range (I-MR) chart is used when data is continuous and not collected in

subgroups. In other words collect the single observation at a time. An I-MR chart provides process variation over time in graphical method.

- X bar – S Charts (when sigma is readily available) X Bar S charts often used control chart to examine the process mean and standard deviation over the time. These charts are used when the subgroups have large sample size and S chart provides better understanding of the spread of subgroup data than range.
- EWMA Chart: The EWMA – Exponentially Weighted Moving Average chart is used in statistical process control to monitor variables (or attributes that act like variables) that make use of the entire history of a given output. This is different from other control charts that tend to treat each data point individually.

Attribute (Discrete) Control Charts:

The output is a decision or counting. It is not possible to measure the quality characteristics of a product. In other words, it is based on the visual inspection like good or bad, fail or pass, accept or reject.

- p Charts: (for defectives – sample size varies) – Use P chart when the data are the fraction defective of some set of process output. It may also shown as percentage of defective. The points plotted on p-chart are the fraction of non confirming units or defective pieces found in n samples.
- np Charts (for defectives – sample size fixed) – Use an np-chart when the data collected in subgroups that are the same size. np-charts show how the process, measured by the number of nonconforming items (defectives) it produces, changes over time. In other words, the process describes pass or fail, yes or no.
- c Charts (for defects – sample size fixed) – Use c charts when the data are concerned with the number of defects in a product. The number of defects collected for the area of opportunity in each subgroup.
- u Charts (for defects – sample size varies) – A u chart is an attribute control chart that displays how the frequency of defects, or nonconformities, is changing over time for a process or system. The number of defects collected for the area of opportunity in each subgroup.

Statistical Reporting

Through statistics, the collected data can be abridged and presented in such a way that it can be easily understood. When building a statistical reporting strategy there are three main classification of statistical tests used in surveys: descriptive statistics, inferential statistics and psychometric tests.

Descriptive Statistics

When using statistical reporting descriptive statistics aims to illustrate a huge portion of the collected data through charts and tables. However, it does not seek to derive a conclusion based from this information or from the sampling population. What it intends to accomplish is to give a rundown of the gathered information by using descriptive charts and tables. This type classification of statistical test is typified with uni-variate analysis and its corresponding survey sample. In addition, it is associated with measures of tendency such as the mean, median, and mode, and measures of dispersion like variance and standard deviation. Descriptive statistics are a very important variable when running statistical reports.

Inferential Statistics

Inferential statistics provide a more compelling and effective statistical data analysis. As the name implies, inferential statistics is involved with making broader and deeper deductions and interpretations usually on the interaction between variables, cause and effect relationship, and identifying the scope of the sample's representation in the population. Based on the sample, the surveyor will verify the hypothesis and then come up with a conclusion. Commonly used inferential statistics in data analysis are Analysis of Variance (ANOVA), T-test, linear regression, and multiple regression.

Psychometric Tests

Psychometric tests analyze the attributes and performance of the employed survey to ensure that the survey data is reliable and valid. Example of a psychometric test is Cronbach's Alpha.

Statistical Reporting Tools

Statistical reporting tools are also used in further understanding the survey data, which is a key factor in making business decisions. Among these are factor analysis, cluster analysis, gap analysis, Z-test, and U-test. In a factor analysis, the obtained data are classified into recognizable clusters. On the other hand, a cluster analysis specifies data clusters that have unique And traceable attributes. Moreover, a gap analysis correlates data and determines if the data disparities are statistically important. A Z-test matches two percentile scores and determines if they are statistically important while a U-test equates median scores of axis-define groups and identifies if their differences are statistically relevant.

These statistical reporting tools can also be supplemented with tables such as frequency tables and cross tabs. Frequency tables depict all the response choices, how many times it has been answered, and the percentage of participants who chose those responses. These are beneficial especially when there are various response choices present or if there is a little disparity between the responses. When two different subgroups or subsets of data will be compared, cross tabulation or cross tab is the best way to go. It is commonly used for questions relating to demographics. Cross tab lets you correlate data from two queries so as to establish the relationship that exists between them.

Types of Statistical Reporting Data

There are four types of data normally encountered during statistical analysis and are presented in statistical reporting data: categorical data, ordinal data, interval data, and ratio data.

- Categorical Data – This type of data is a result of relative frequency statistics. Example is dividing the sum of a certain response with the total number of responses. Let us say that for a brand survey, the brand quality choice accumulated 25 responses out of 100. Therefore, it can be inferred that 25 percent of the participants prefer brand quality in choosing a brand.
- Ordinal Data – Ordinal data are best presented using frequency tables. These are data that have scales and ordered according to preference. For example, out of 100 respondents, 45 of them agree that the brand needs to improve its packaging. In percent, that is equivalent to 45%.
- Interval Data – Encapsulating interval data can be best be done when treated as an ordinal data. Averaging and standard deviation are the ideal techniques in evaluating this type of data.
- Ratio Data – Ratio data can be converted to a normal data using logarithms and square roots. A distinguishing characteristic of this set of data is that it has a defined zero point. Decimals and fractions are also available in a ratio data.

Interpretation in Excel

Microsoft Excel is one of the most widely used tools in any industry. While some enjoy playing with pivotal tables and histograms, others limit themselves to simple pie-charts and conditional formatting.

Some may create an artwork out of the dull monochrome Excel, while others may be satisfied with its data analysis. In this discussion, we will make a deep delving analysis of Microsoft Excel and its utility. We will focus on how to analyze data in Excel Analytics, the various tricks, and techniques for it. The discussion will also explore the various ways to analyze data in Excel.

We will discuss the different features of Excel analytics to know how to analyze data in excel (much of which are unexplored to the mass), functions, and best practices.

Our discussion will include, but not be limited to:

(i) Best Way to Analyze Data in Excel

(ii) How to Analyze Sales Data in Excel

(iii) Analyzing Data Sets with Excel

(iv) Data Segregation with Excel

(v) The Importance of Data Reporting

(vi) How to Analyze Data in Excel

A pivot tool helps us summarize huge amounts of data. One of the best ways to analyze data in excel, it is mostly used to understand and recognize patterns in the data set.

Recognizing patterns in a small dataset is pretty simple. But the enormity of the datasets often calls for additional efforts to find the patterns. In such cases, a pivot table can be a huge advantage as it takes only a few minutes to summarize groups of data using a pivot table.

A data analysis example can be, you have a dataset consisting of regions and number of sales. You may want to know the number of sales based on the regions, which can be used to determine why a region is lacking and how to possibly improve in that area. Using a pivot table, you can create a report in excel within a few minutes and save it for future analysis.

A Pivot Table allows you to summarize data as averages, sums, or counts in Excel from data that is stored in another Spread sheet, or table. It is great for quickly building reports because you can sort and visualize the data quickly.

Taking a data analysis example like, you may have put together a spreadsheet, which you can copy, and paste into Excel, or use in Google Docs if you would prefer (just click File > Make a Copy).

The spreadsheet contains data with a mock company's customer purchase information. Since companies purchase at different dates, a pivot table will help us to consolidate this data to allow us to see total buys per company, as well as to compare purchases across companies, for quick analysis.

The Pivot table allows you to take a table with a lot of data in it and rearrange the table so that you only look at only what matters to you.

a) Whether you are using a Mac or a PC, you can select the whole dataset that you want to look at and select: “Data” -> “Pivot Table”. When you hit that, a new tab should be opened with a table.

Data Set

b) Once you have your table in front of you, you can drag and drop the Column Labels, Row Labels, and Report Filter

Column Labels go across the top row of your table (for example Date, Month, Company Name)

Row Labels go across the left-hand side of your table [for example Date, Month, Company Name (same as with column labels, it depends on how you would prefer to look at the data, vertically or horizontally)]

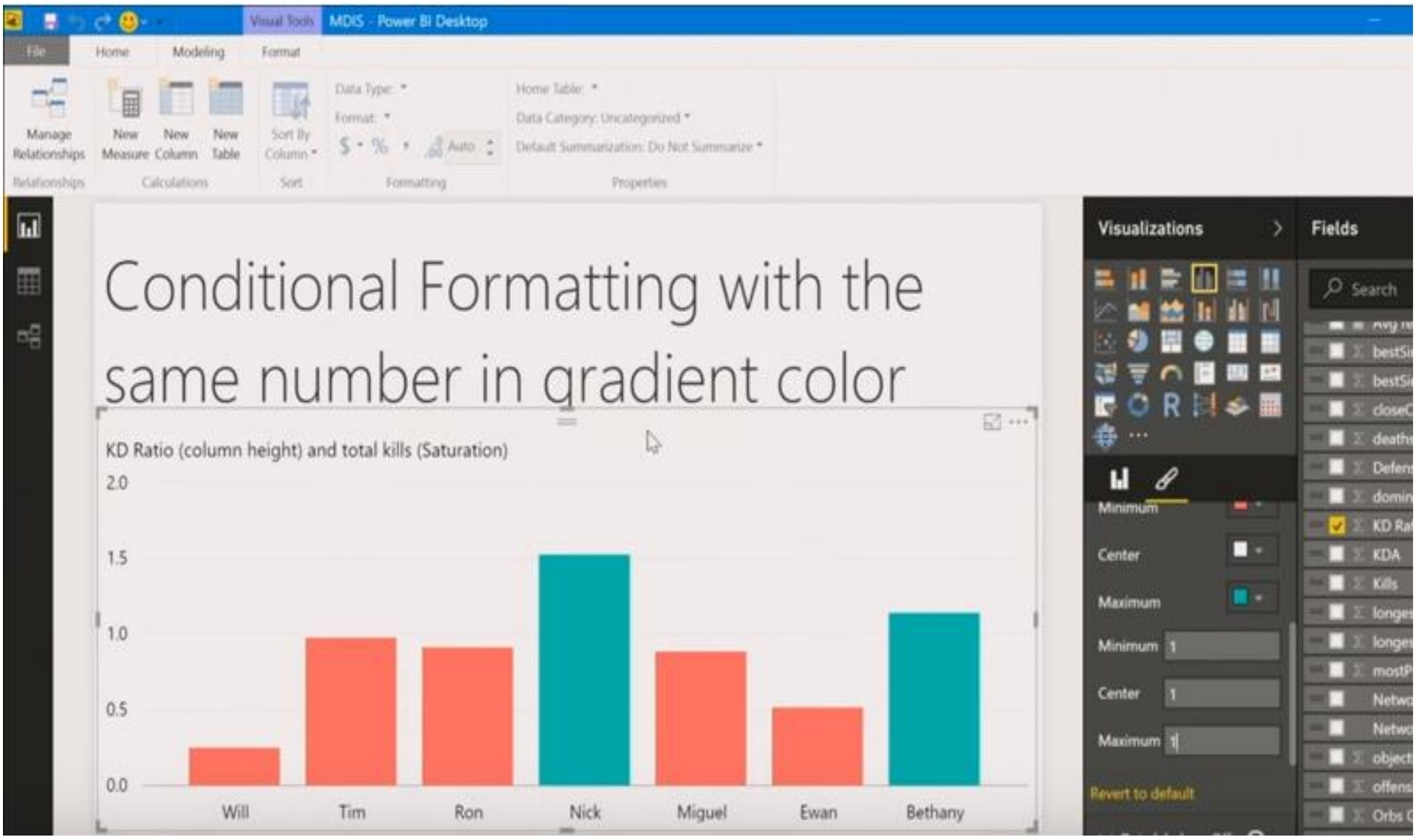
The Values section is where you put the data you would like calculated (for example Purchases, Revenue)

Report Filter helps you refine your results. Add anything you would like to Filter by (for example you want to look at Lead Referral Sources, but exclude Google and Direct)

Pivot tables are a great way to manage the data from your reports. You can copy and paste the data into your own Excel file, or create a copy in Google Apps (File > Make a Copy).

To know how to analyze data in excel, you can instantly create different types of charts, including line and column charts, or add miniature graphs. You can also apply a table style, create PivotTables, quickly insert totals, and apply conditional formatting. Analyzing large data sets with Excel makes work easier if you follow a few simple rules:

1. Select the cells that contain the data you want to analyze.
2. Click the Quick Analysis button image button that appears to the bottom right of your selected data (or press CRTL + Q).
3. Selected data with Quick Analysis Lens button visible
4. In the Quick Analysis gallery, select a tab you want. For example, choose Charts to see your data in a chart.
5. Pick an option, or just point to each one to see a preview.
6. You might notice that the options you can choose are not always the same. That is often because the options change based on the type of data you have selected in your workbook.



Data Type:
Format:
Auto
Home Table:
Data Category: Uncategorized
Default Summarization: Do Not Summarize

Conditional Formatting with the same number in gradient color

Minimum
Center
Maximum
Minimum 1
Center 1
Maximum 1

To understand the best way to analyze data in excel, you might want to know which analysis option is suitable for you. Here we offer you a basic overview of some of the best options to choose from.

1. Formatting: Formatting lets you highlight parts of your data by adding things like data bars and colors. This lets you quickly see high and low values, among other things.
2. Charts: Charts Excel recommends different charts, based on the type of data you have selected. If you do not see the chart you want, click More Charts.
3. Totals: Totals let you calculate the numbers in columns and rows. For example, Running Total inserts a total that grows as you add items to your data. Click the little black arrows on the right and left to see additional options.
4. Tables: Tables make it easy to filter and sort your data. If you do not see the table style you want, click More.
5. Sparklines: Sparklines are like tiny graphs that you can show alongside your data. They provide a quick way to see trends.

Ways to Analyze Data in Excel: Tips and Tricks

It is fun to analyze data in MS Excel if you play it right. Here, we offer some quick hacks so that you know how to analyze data in excel.

1. How to Analyze Data in Excel: Data Cleaning

Data Cleaning, one of the very basic excel functions, becomes simpler with a few tips and tricks. You may learn how to use a native Excel feature and how to accomplish the same goal with Power Query. Power Query is a built-in feature in Excel 2016 and an Add-in for Excel 2010/2013. It helps you to extract, transform, and load your data with just a few clicks.

1. Change the format of numbers from text to numeric

Sometimes when you import data from an external source other than Excel, numbers are imported as text. Excel will alert you by showing a green tooltip in the top-left corner of the cell.

Depending on the number of values in the range, you can quickly convert the values to numbers by clicking on 'Convert to a number' within the tooltip options.

However, if you have more than 1000 values, you will have to wait a couple of seconds while Excel finishes the conversion.

You may also convert the values to number format is to use Text-to-Columns using the following steps:

1. Select the range with the values to be converted.
2. Go to Data > Text to Columns.
3. Select Delimited and click Next.
4. Uncheck all the checkboxes for delimiters (see below) and click Next.
5. Text-Columns-Checkboxes

2. Select General and click on Finish

When you have lots of numbers to convert this tip will be much faster than waiting for all the numbers to be converted. In Power Query, you just have to right-click on the column header of the column you want to convert.

1. Then go to Change Type.
2. Then select the type of number you want (such as Decimal or Whole Number)
3. Power-Query-Data-Type

Data Analysis is simpler and faster with Excel analytics. Here, we offer some tips for work:

1. Create auto expandable ranges with Excel tables: One of the most underused features of MS Excel is Excel Tables. Excel Tables have wonderful properties that allow you to work more efficiently. Some of these features include:
2. Formula Auto Fill: Once you enter a formula in a table it will be automatically be copied to the rest of the table.
3. Auto Expansion: New items typed below or at the right of the table become part of the table.
4. Visible headers: Regardless of your position within the table, your headers will always be visible.
5. Automatic Total Row: To calculate the total of a row, you just have to select the desired formula.
6. Use Excel Tables as part of a formula: Like in dropdown lists, if you have a formula that depends on a Table, when you add new items to the Table, the reference in the formula will be automatically updated.

7. Use Excel Tables as a source for a chart: Charts will be updated automatically as well if you use an Excel Table as a source. As you can see, Excel Tables allow you to create data sources that do not have to be updated when new data is included.

How to Analyze Data in Excel: Data Visualization

Quickly visualize trends with sparklines: Sparklines are a visualization feature of MS Excel that allows you to quickly visualize the overall trend of a set of values. Sparklines are mini-graphs located inside of cells. You may want to visualize the overall trend of monthly sales by a group of salesmen.

To create the sparklines, follow these steps below:

1. Select the range that contains the data that you will plot (This step is recommended but not required, you can select the data range later).
2. Go to Insert > Sparklines > Select the type of sparkline you want (Line, Column, or Win/Loss). For this specific example, I will choose Lines.
3. Click on the range selection button Select Range Excel Button to browse for the location of the sparklines, press Enter and click OK. Make sure you select a location that is proportional to the data source. For example, if the data source range contains 6 rows then the location of the sparkline must contain 6 rows.

To format the sparkline you may try the following:

To change the colour of markers:

1. Click on any cell within the sparkline to show the Sparkline Tools menu.
2. In the Sparkline tools menu, go to Marker Color and change the colour for the specific markers you want.

For example High points on the green, Low points on red, and the remaining in blue.

To change the width of the lines:

1. Click on any cell within the sparkline to show the Sparkline Tools menu.
2. In the Sparkline tools contextual menu, go to Sparkline Color > Weight and change the width of the line as you desire.

Save Time with Quick Analysis: One of the major improvements introduced back in Excel 2013 was the Quick Analysis feature. This feature allows you to quickly create graphs, sparklines, PivotTables, PivotCharts, and summary functions by just clicking on a button.

When you select data in Excel 2013 or later, you will see the Quick Analysis button Quick Analysis Excel Button in the bottom-right corner of the range selected. If you click on the Quick Analysis button you will see the following options:

1. Formatting
2. Charts

3. Totals
4. Tables
5. Sparklines

When you click on any of the options, Excel will show a preview of the possible results you could obtain given the data you selected.

1. If you click on the Quick Analysis button and go to charts, you could quickly create the graph below just by clicking a button.
2. If you go to Totals, you can quickly insert a row with the average for each column:
3. If you click on Sparklines, you can quickly insert Sparklines:
4. As you can see, the Quick Analysis feature really allows you to quickly perform different visualizations and analysis with almost no effort.

How to Analyze Data in Excel: Data reporting

Data reporting in Excel analytics requires more than just accounting skills, it also requires a thorough knowledge of excel functionalities and the ability to add beauty to your report.

1. Turn Auto Refresh off before editing the Excel workbook. This will stop the table from refreshing when you are making changes on the worksheet. To do this, click on the Refresh icon at the bottom of the Excel Report Designer Task Pane and then select “Switch auto-refresh off”.
2. When adding a new row to the layout, select a cell in the table area below where you want to insert the new row. Then right-click and from the context menu select Insert > Table Rows Above.
3. When deleting columns or rows make sure you use the Table Delete functions similar to the Insert functions above. To remove a column or row, select a cell in the table area of the row or column you want to remove. Then right-click and from the context menu select Delete and then either Table Columns or Table Rows.
4. Remove unneeded rows or columns from the table. Having fewer cells in the layout makes it easier for the table to refresh, so removing any unneeded ones will improve performance.

Excel has several other uses that may not have been covered here. Play around with visualize complex data or organize disparate numbers, to discover the infinite variety of functions in Excel analytics. Strong knowledge of excel is a boon if you are looking forward to a career in data analytics.

This was all about how to analyze data in excel, how to analyze sales data in excel along with a few data analysis example.

Hopefully, you must have understood the best way to analyze data in excel.

We offer one of the best-known courses in Data Analytics Using Excel Course. The course enables you to learn tools such as Advanced Excel, PowerBI, and SQL. The live projects and intensive training program also empower you to come up with solutions for real-life problems.

R Programming

R Programming Tutorial is designed for both beginners and professionals. Our tutorial provides all the basic and advanced concepts of data analysis and visualization.

R is a software environment which is used to analyze statistical information and graphical representation. R allows us to do modular programming using functions.

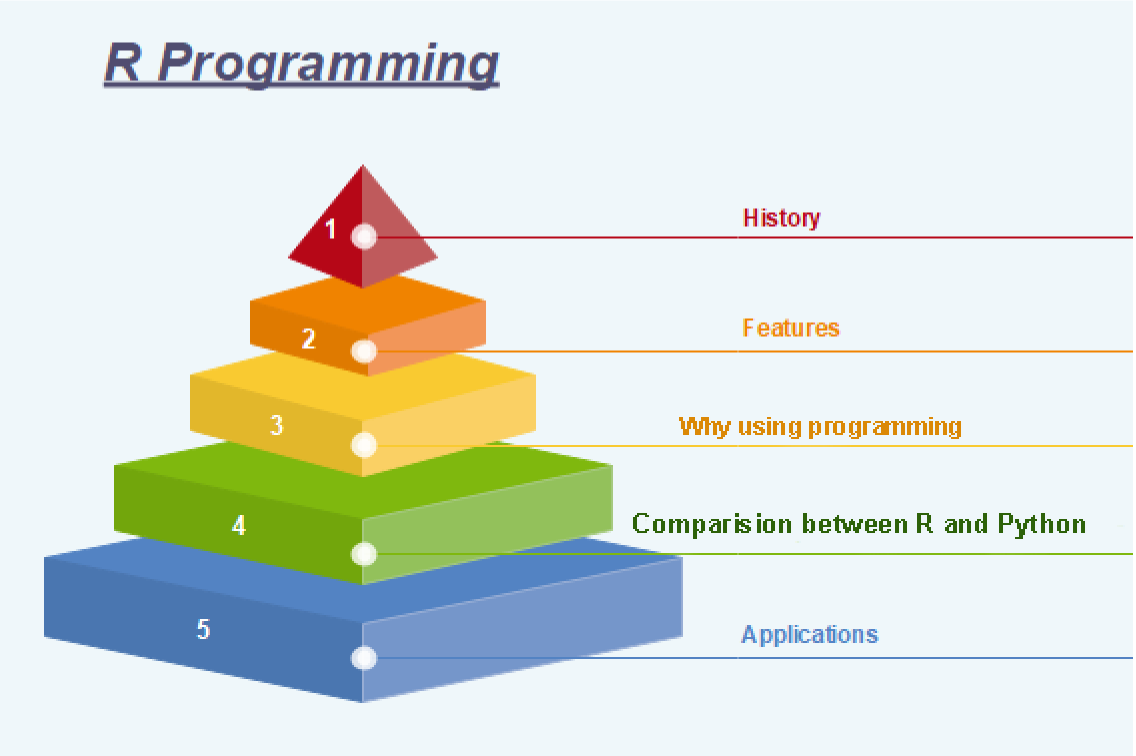
Our R tutorial includes all topics of R such as introduction, features, installation, rstudio ide, variables, datatypes, operators, if statement, vector, data handing, graphics, statistical modelling, etc. This programming language was named R, based on the first name letter of the two authors (Robert Gentleman and Ross Ihaka).

What is R Programming?

"R is an interpreted computer programming language which was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand." The *R Development Core Team* currently develops R. It is also a software environment used to analyze statistical information, graphical representation, reporting, and data modeling. R is the implementation of the S programming language, which is combined with lexical scoping semantics.

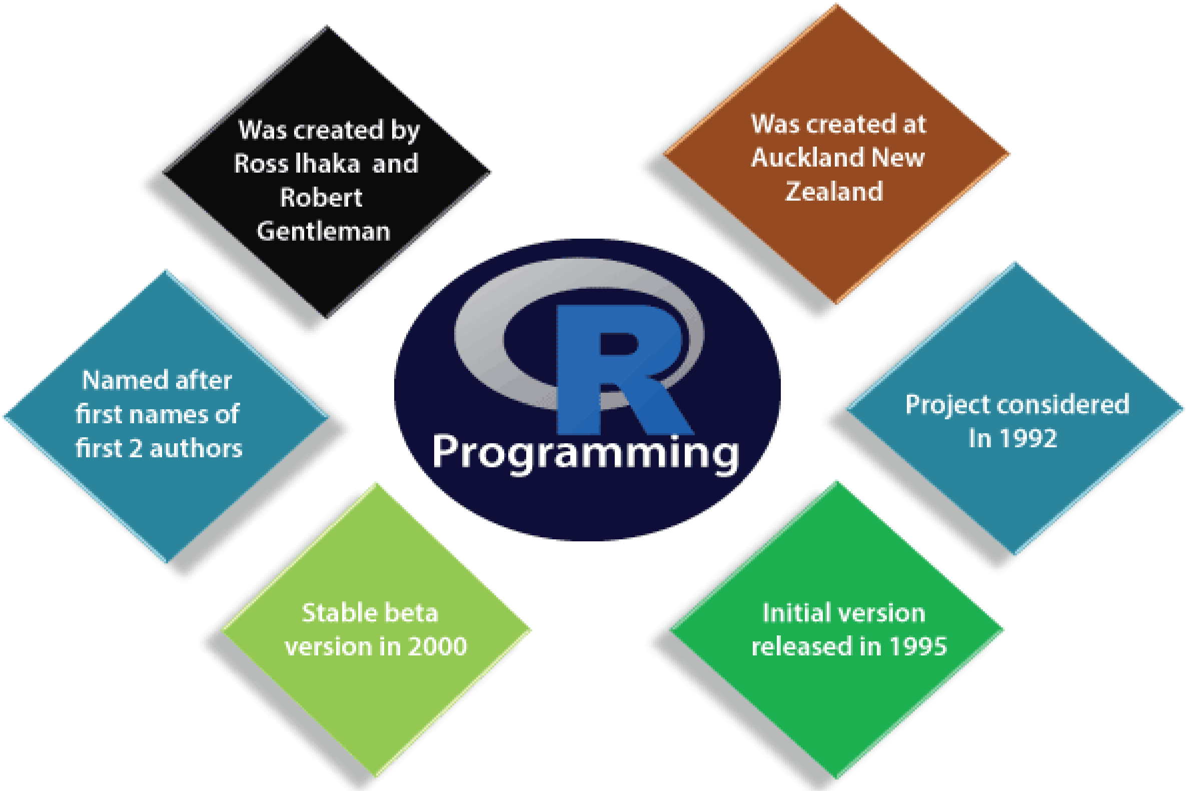
R not only allows us to do branching and looping but also allows to do modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python, and FORTRAN languages to improve efficiency.

In the present era, R is one of the most important tool which is used by researchers, data analyst, statisticians, and marketers for retrieving, cleaning, analyzing, visualizing, and presenting data.



History of R Programming

The history of R goes back about 20-30 years ago. R was developed by Ross lhaka and Robert Gentleman in the University of Auckland, New Zealand, and the R Development Core Team currently develops it. This programming language name is taken from the name of both the developers. The first project was considered in 1992. The initial version was released in 1995, and in 2000, a stable beta version was released.


Was created by Ross Ihaka and Robert Gentleman
Was created at Auckland New Zealand
Named after first names of first 2 authors
Programming
Project considered In 1992
Stable beta version in 2000
Initial version released in 1995

The following table shows the release date, version, and description of R language:

Version-Release	Date	Description
0.49	1997-04-23	First time R's source was released, and CRAN (Comprehensive R Archive Network) was started.
0.60	1997-12-05	R officially gets the GNU license.
0.65.1	1999-10-07	update.packages and install.packages both are included.
1.0	2000-02-29	The first production-ready version was released.
1.4	2001-12-19	First version for Mac OS is made available.
2.0	2004-10-04	The first version for Mac OS is made available.

2.1	2005-04-18	Add support for UTF-8encoding, internationalization, localization etc.
2.11	2010-04-22	Add support for Windows 64-bit systems.
2.13	2011-04-14	Added a function that rapidly converts code to byte code.
2.14	2011-10-31	Added some new packages.
2.15	2012-03-30	Improved serialization speed for long vectors.
3.0	2013-04-03	Support for larger numeric values on 64-bit systems.
3.4	2017-04-21	The just-in-time compilation (JIT) is enabled by default.
3.5	2018-04-23	Added new features such as compact internal representation of integer sequences, serialization format etc.

Features of R programming

R is a domain-specific programming language which aims to do data analysis. It has some unique features which make it very powerful. The most important arguably being the notation of vectors. These vectors allow us to perform a complex operation on a set of values in a single command. There are the following features of R programming:

1. It is a simple and effective programming language which has been well developed.
2. It is data analysis software.
3. It is a well-designed, easy, and effective language which has the concepts of user-defined, looping, conditional, and various I/O facilities.
4. It has a consistent and incorporated set of tools which are used for data analysis.
5. For different types of calculation on arrays, lists and vectors, R contains a suite of operators.
6. It provides effective data handling and storage facility.
7. It is an open-source, powerful, and highly extensible software.
8. It provides highly extensible graphical techniques.

9. It allows us to perform multiple calculations using vectors.
10. R is an interpreted language.

Why use R Programming?

There are several tools available in the market to perform data analysis. Learning new languages is time taken. The data scientist can use two excellent tools, i.e., R and Python. We may not have time to learn them both at the time when we get started to learn data science. Learning statistical modeling and algorithm is more important than to learn a programming language. A programming language is used to compute and communicate our discovery.

The important task in data science is the way we deal with the data: clean, feature engineering, feature selection, and import. It should be our primary focus. Data scientist job is to understand the data, manipulate it, and expose the best approach. For machine learning, the best algorithms can be implemented with R. Keras and Tensor Flow allow us to create high-end machine learning techniques. R has a package to perform Boost. Boost is one of the best algorithms for Kaggle competition.

R communicates with the other languages and possibly calls Python, Java, C++. The big data world is also accessible to R. We can connect R with different databases like Spark or Hadoop.

In brief, R is a great tool to investigate and explore the data. The elaborate analysis such as clustering, correlation, and data reduction are done with R.

Applications of R

There are several-applications available in real-time. Some of the popular applications are as follows:

- Facebook
- Google
- Twitter
- HRDAG
- Sunlight Foundation
- Real Climate
- NDAA
- XBOX ONE
- ANZ

- FDA

Prerequisite

R programming is used for statistical information and data representation. So it is required that we should have the knowledge of statistical theory in mathematics. Understanding of different types of graphs for data representation and most important is that we should have prior knowledge of any programming.

Audience

This tutorial is helpful for those students who are interested in gaining the knowledge of how data analysis projects are implemented. This tutorial covers all the basics of R and how data analysis is done using R.

Introduction of STATA

Stata is a full-featured statistical programming language for Windows, Mac OS X, Unix and Linux. It can be considered a “stat package,” like SAS, SPSS, RATS, or eViews. Stata is available in several versions: Stata/IC (the standard version), Stata/SE (an extended version) and Stata/MP (for multiprocessing). The major difference between the versions is the number of variables allowed in memory, which is limited to 2,047 in standard Stata/IC, but can be much larger in Stata/SE or Stata/MP. The number of observations in any version is limited only by memory. Christopher F Baum (Boston College FMRC) Introduction to Stata August 2011 2 / 157 Strengths of Stata Stata/SE relaxes the Stata/IC constraint on the number of variables, while Stata/MP is the multiprocessor version, capable of utilizing 2, 4, 8... processors available on a single computer. Stata/IC will meet most users’ needs; if you have access to Stata/SE or Stata/MP, you can use that program to create a subset of a large survey dataset with fewer than 2,047 variables. Stata runs on all 64-bit operating systems, and can access larger datasets on a 64-bit OS, which can address a larger memory space. All versions of Stata provide the full set of features and commands: there are no special add-ons or ‘toolboxes’. Each copy of Stata includes a complete set of manuals (over 6,000 pages) in PDF format, hyperlinked to the on-line help.

A Stata license may be used on any machine which supports Stata (Mac OS X, Windows, Linux): there are no machine-specific licenses for Stata versions 11 or 12. You may install Stata on a home and office machine, as long as they are not used concurrently. Licenses can be either annual or perpetual. Stata works differently than some other packages in requiring that the entire dataset to be analyzed must reside in memory. This brings a considerable speed advantage, but implies that you may need more

RAM (memory) on your computer. There are 32-bit and 64-bit versions of Stata, with the major difference being the amount of memory that the operating system can allocate to Stata (or any other application) In some cases, the memory requirement may be of little concern. Stata is capable of holding data very efficiently, and even a quite sizable dataset (e.g., more than one million observations on 20–30 variables) may only require 500 Mb or so. You should take advantage of the compress command, which will check to see whether each variable may be held in fewer bytes than its current allocation. For instance, indicator (dummy) variables and categorical variables with fewer than 100 levels can be held in a single byte, and integers less than 32,000 can be held in two bytes: see help datatypes for details. By default, floating-point numbers are held in four bytes, providing about seven digits of accuracy. Some other statistical programs routinely use eight bytes to store all numeric variables.

The memory available to Stata may be considerably less than the amount of RAM installed on your computer. If you have a 32-bit operating system, it does not matter that you might have 4 Gb or more of RAM installed; Stata will only be able to access about 1 Gb, depending on other processes' demands. To make most effective use of Stata with large datasets, use a computer with a 64-bit operating system. Stata will automatically install a 64-bit version of the program if it is supported by the operating system. All Linux, Unix and Mac OS X computers today come with 64-bit operating systems Stata is eminently portable, and its developers are committed to cross-platform compatibility. Stata runs the same way on Windows, Mac OS X, Unix, and Linux systems. The only platform-specific aspects of using Stata are those related to native operating system commands: e.g. is the file to be accessed C:\Stata\StataData\myfile.dta or /users/baum/statadata/myfile.dta Perhaps unique among statistical packages, Stata's binary data files may be freely copied from one platform to any other, or even accessed over the Internet from any machine that runs Stata. You may store Stata's binary datafiles on a webserver (HTTP server) and open them on any machine with access to that server.

Stata's user interface

Stata has traditionally been a command-line-driven package that operates in a graphical (windowed) environment. Stata version 11 (released June 2009) and version 12 (released July 2011) contains a graphical user interface (GUI) for command entry via menus and dialogs. Stata may also be used in a command-line environment on a shared system (e.g., a Unix server) if you do not have a graphical interface to that system. A major advantage of Stata's GUI system is that you always have the option of reviewing the command that has been entered in Stata's Review window. Thus, you may examine the syntax, revise it in the Command window and resubmit it. You may find that this is a more efficient way of using the program than relying wholly on dialogs.

PSPP

Introduction to PSPP

PSPP is a statistical analysis tool developed to be a free, open-source alternative to SPSS (which is now developed by IBM). Although not identical, it is similar in many respects, and allows one to work with file formats common to SPSS. Whereas SPSS was originally an acronym for Statistical Package for the Social Sciences, "PSPP" has no corresponding meaning (although the letters obviously relate to the SPSS name).

Unlike SPSS, PSPP does not limit the number of cases or variables which you are able to use, nor will it require you to purchase add-ons to gain access to more advanced functions. As a fully functional statistical analysis program, it is capable of performing descriptive statistics, y-tests, linear regression, as well as non-parametric tests. Included in its basic design is the ability to perform analyses as quickly as possible, regardless of the volume of data entered. In addition, you have the option of using the program through its graphical user interface (GUI), or through the more traditional method of utilizing syntax commands.

PSPP is a tool for statistical analysis of sampled data. It reads the data, analyzes the data according to commands provided, and writes the results to a listing file, to the standard output or to a window of the graphical display.

The language accepted by PSPP is similar to those accepted by SPSS statistical products. The details of PSPP’s language are given later in this manual.

PSPP produces tables and charts as output, which it can produce in several formats; currently, ASCII, PostScript, PDF, HTML, DocBook and TeX are supported.

The current version of PSPP, 1.6.2, is incomplete in terms of its statistical procedure support. PSPP is a work in progress. The authors hope to fully support all features in the products that PSPP replaces, eventually. The authors welcome questions, comments, donations, and code submissions. See Submitting Bug Reports, for instructions on contacting the authors.

What is SPSS?

SPSS stands for Statistical Package for the Social Sciences. It was created in 1968 to allow social scientists (and others) to perform statistical tests. IBM purchased it in 2009 and renamed it IBM SPSS. It and SAS are used widely for data analysis.

Why PSPP?

SPSS is excellent. It also costs more than $1,000. So the author of PSPP (the acronym does not stand for anything) created a free program that follows the nomenclature and style of SPSS. Unlike the statistics program R, which is also free, PSPP does not require that you learn a new language. And its graphical-user interface is an advantage for those less comfortable with a command-line environment.

PSPP is not as competent as its paid rival For example, PSPP misinterpreted a CSV file I created that SPSS handled just fine. And PSPP is missing some advanced statistical tests.However, PSPP performs enough statistics to be useful for journalists.

Why would journalists use statistics?

Life is random. What appears to be a pattern may just be chance. Statistics let us distinguish between a meaningful pattern and random chance. In turn, that lets journalists avoid mistaking coincidence for significance.

For example, does that cancer cluster have an environmental cause or is it just bad luck? Are improvements in school test scores the product of a new curriculum or good luck? Is a change in the local crime rate substantially different from the state or national average? In other words, are these differences newsworthy or the product of chance? Statistics can tell us.

SCALES OF MEASUREMENT

Before we can analyze the data, we have to understand what kind it is. This involves *scales of measurement*, which come in four types.

	Scale	Description	Examples
	Nominal	Categorizes data without ranking; group 1 is not “better” than group 2	Religion, political party, sex, color, national origin
	Ordinal	Categorizes and ranks; placing first is better than placing second but intervals between places are unequal	Finish order, grades, rank in class, Likert scale
	Interval	Categorizes and ranks with equal distance but without a true zero; 0 degrees Celsius is not zero energy	Fahrenheit and Celsius temperature scales, time of day
	Ratio	Categorizes and ranks with equal distance, and has a true zero	Kelvin temperature scale, age, weight, distance, money

For now, we can group these four into two: *categorical* and *continuous*. Then we can parsethe HMDA variables into these two groups and consider potential statistics.

	Scale	Group	DenverHMDA variables	Potential stats
	Nominal	Categorical	LoanType Action CenTract Ethnicity Race Sex DenialReason	Sum, percent, mode, chi-square
	Ordinal			
	Interval			

		Continuous	Loan Income	Sum, percent, mean, median, standard deviation, t-test, ANOVA, regression
	Ratio			

Introduction to EViews

EViews is a software package that provides tools for data analysis, regression, and forecasting. It is a “canned” regression package for econometric analysis. EViews has an object-oriented design. Each type of object has specific ‘views’ and procedures that are used in Eviews.

Eviews has been a command-line-only program up until recently, and all the advanced features, such as Kronecker products, eigenvector solution, and singular value decomposition are still processed using the command line. As well, the command line mode records all of the steps in an analysis. The GUI (Graphical User Interface) is convenient, but you will have no record of your session and will have to go back to command line mode to use advanced features.

CREATING A WORKFILE AND IMPORTING DATA

Open Eviews.

Note the date range of the csv raw file.

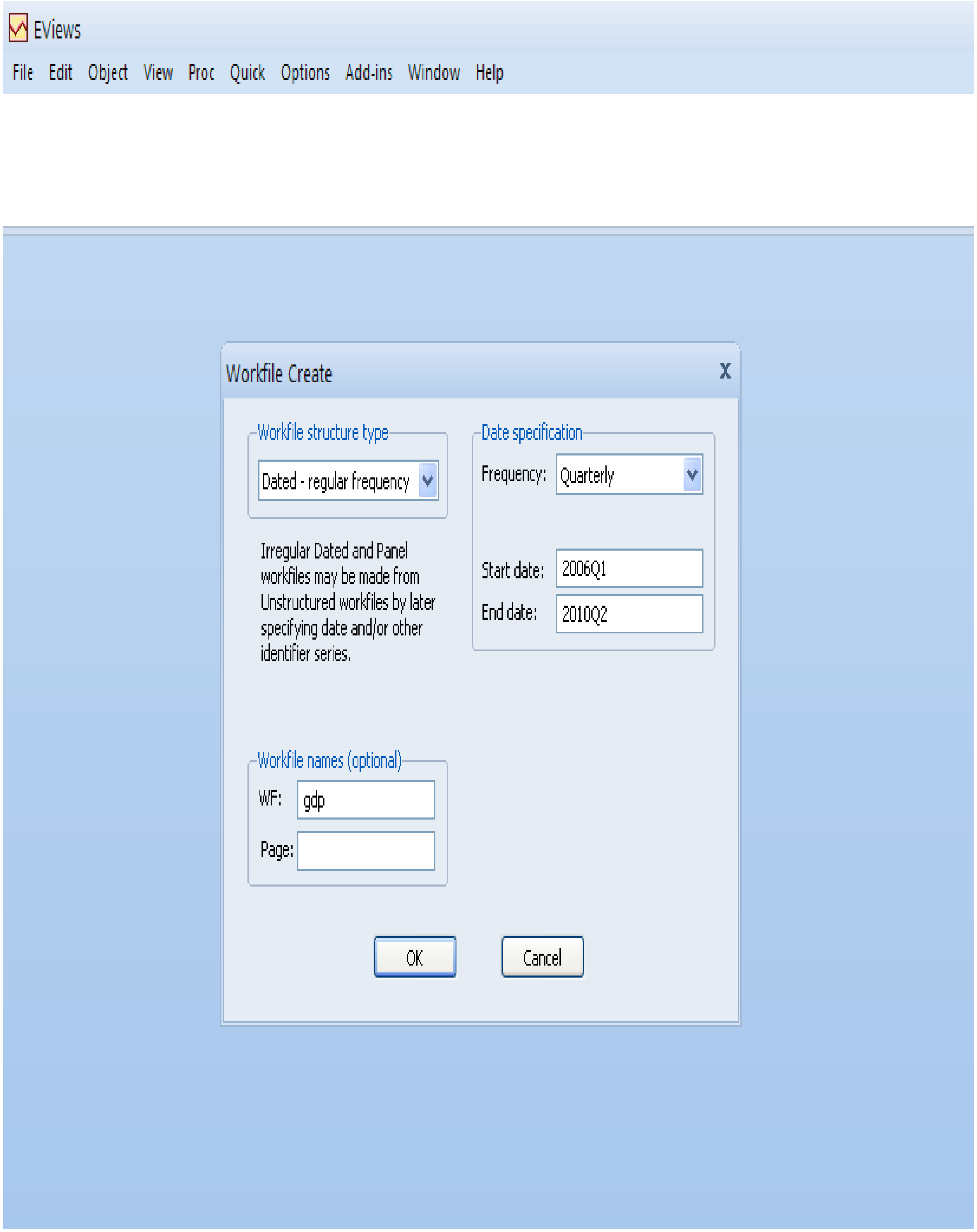
To create a new workfile, select File/New/Workfile… from the main menu.

The Workfile Range dialog box opens.

Under Frequency, select Quarterly.

Under Range enter 2006Q1 as the Start date and 2010Q2 as the End date.

Give the newly created workfile a name. Ours is named gdp. Notice that Eviews is not case-sensitive.


EViews
File Edit Object View Proc Quick Options Add-ins Window Help
Workfile Create
X
OK
Cancel

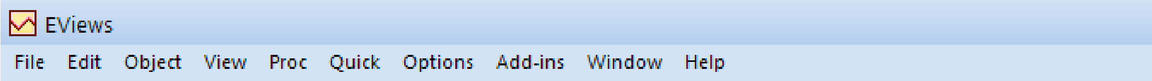
Workfile structure type: Dated - regular frequency

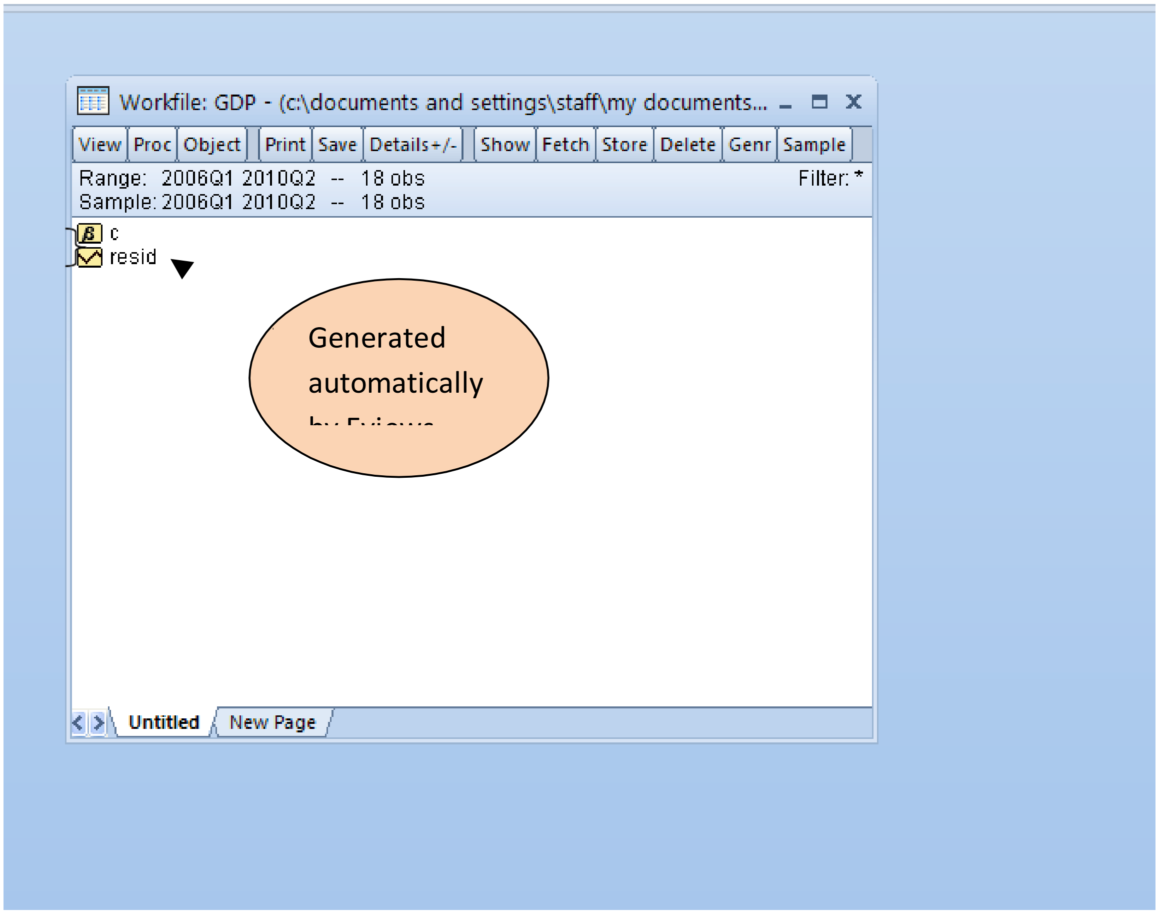
Irregular Dated and Panel workfiles may be made from Unstructured workfiles by later specifying date and/or other identifier series.

Date specification
Frequency: Quarterly
Start date: 2006Q1
End date: 2010Q2

Workfile names (optional)
WF: gdp
Page:

A new workfile called gdp is now created.





Range: 2006Q1 2010Q2 -- 18 obs
Sample: 2006Q1 2010Q2 -- 18 obs
Filter: *

Generated automatically by Eviews

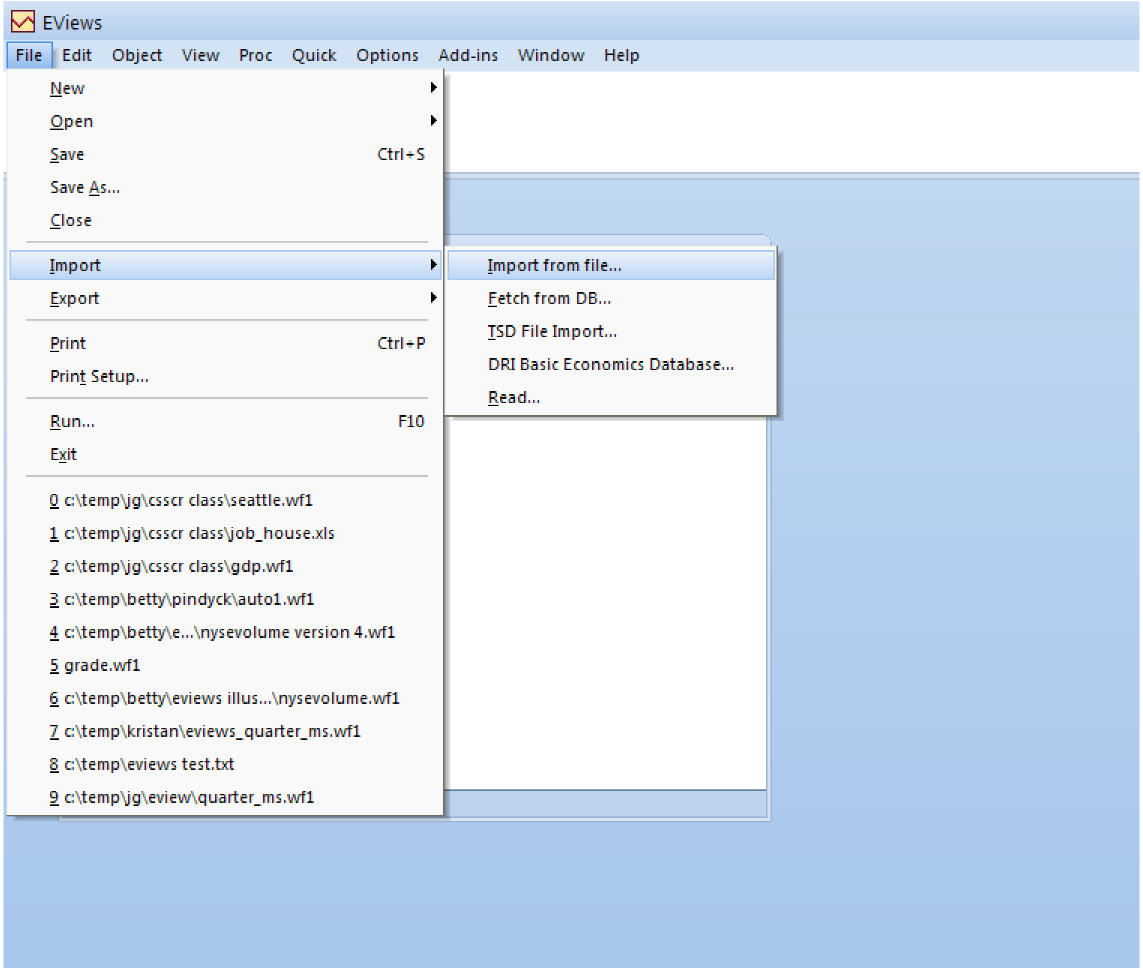
In the workfile, C represents the constant term for any regression analysis in Eviews. By default, residrepresents the residuals from the most recent regression. You haven’t run a regression yet, so they are empty.

Import your data

Select File/Import/Import from file …from the main menu.

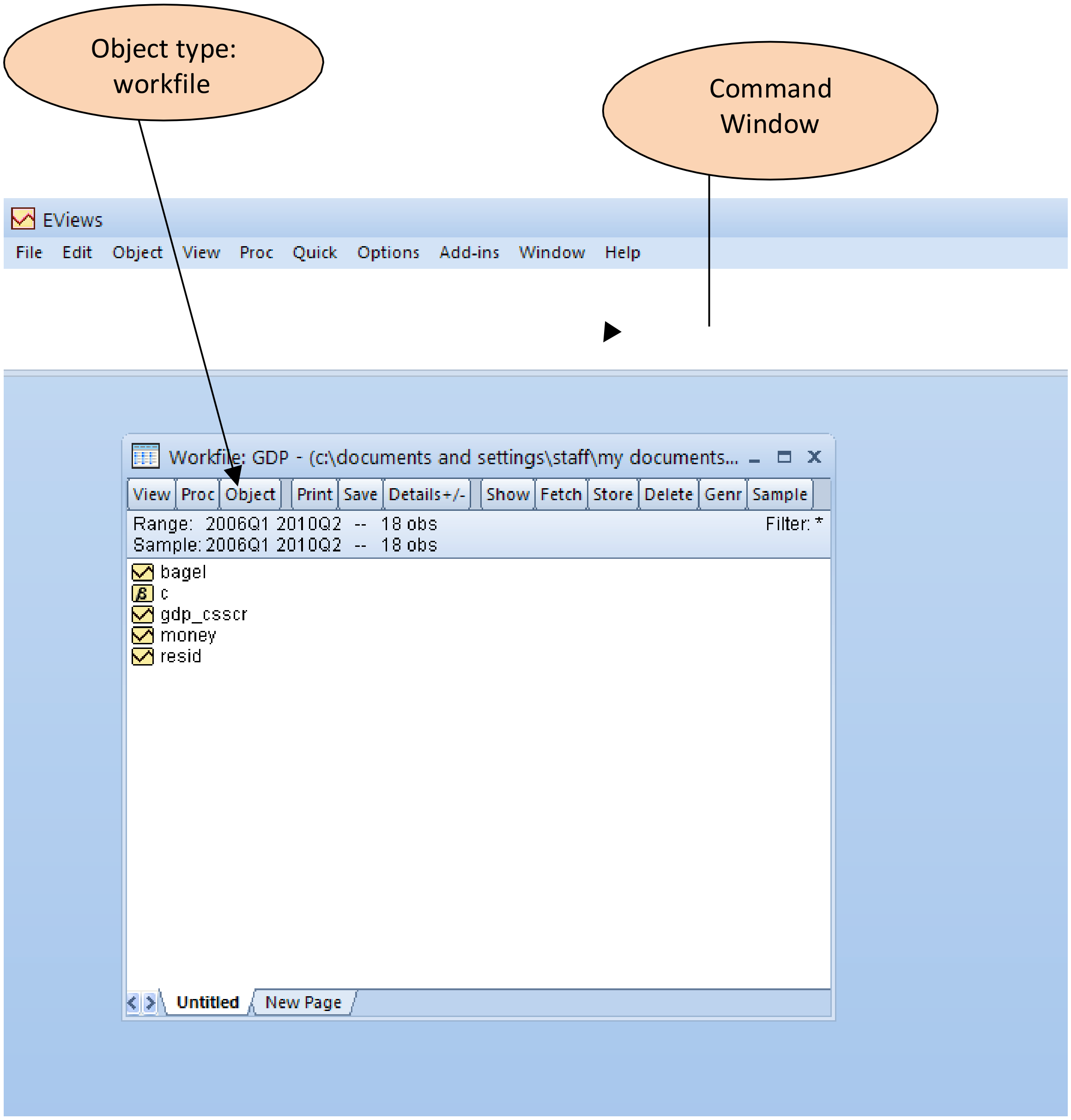
Change the type of opening files to all.

Find and open the csscr_gdp.csv file in the Open dialog box.


EViews
File
Edit
Object
View
Proc
Quick
Options
Add-ins
Window
Help
New
Open
Save Ctrl+S
Save As...
Close
Import
Export
Print Ctrl+P
Print Setup...
Run... F10
Exit
0 c:\temp\jg\csscr class\seattle.wf1
1 c:\temp\jg\csscr class\job_house.xls
2 c:\temp\jg\csscr class\gdp.wf1
3 c:\temp\betty\pindyck\auto1.wf1
4 c:\temp\betty\e...\nysevolume version 4.wf1
5 grade.wf1
6 c:\temp\betty\eviews illus...\nysevolume.wf1
7 c:\temp\kristan\eviews_quarter_ms.wf1
8 c:\temp\eviews test.txt
9 c:\temp\jg\eview\quarter_ms.wf1
Import from file...
Fetch from DB...
TSD File Import...
DRI Basic Economics Database...
Read...

A couple of steps will lead you to the final workfile which shows all the variables. The illustration below shows a workfile with three variables imported. The white rectangle under the menu bar is the command window, mainly used for writing and executing commands. EViews also has a powerful point-and-click GUI(graphical user interface). Try both.

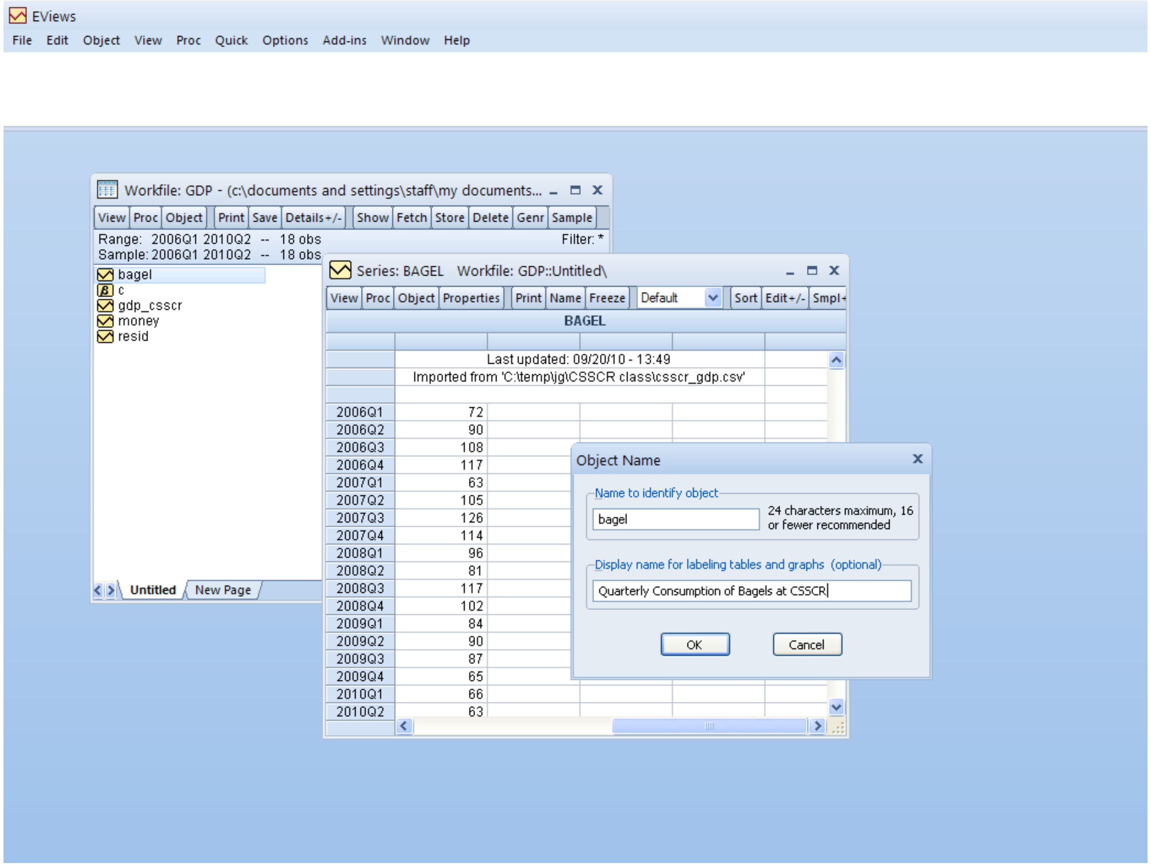
Another way of importing is to drag the external file into the area under the command window. Notice this way of importing does not require you to create a new workfile first. But as discussed previously, using drag-and-drop doesn't give you a record of your actions.


Object type: workfile
Command Window
EViews
File Edit Object View Proc Quick Options Add-ins Window Help
Workfile: GDP - (c:\documents and settings\staff\my documents...
View Proc Object Print Save Details+/- Show Fetch Store Delete Genr Sample
Untitled
New Page

```
Range:  2006Q1 2010Q2  --  18 obs                Filter: *
Sample: 2006Q1 2010Q2  --  18 obs
bagel
c
gdp_csscr
money
resid
```

LABEL THE VARIABLES

Suppose we want to label the variable bagel as ‘quarterly consumption of bagels at CSSCR.’ Bring up the variable view window by double clicking on bagel. In the variable view window, click on ‘Name.’ The labelcan then be added.


EViews
File Edit Object View Proc Quick Options Add-ins Window Help
View Proc Object Print Save Details+/- Show Fetch Store Delete Genr Sample
View Proc Object Properties Print Name Freeze Default Sort Edit+/- Smpl+
Untitled New Page
OK Cancel

Workfile: GDP - (c:\documents and settings\staff\my documents...

Range: 2006Q1 2010Q2 -- 18 obs

Sample: 2006Q1 2010Q2 -- 18 obs

Filter: *

bagel
c
gdp_csscr
money
resid

Series: BAGEL Workfile: GDP::Untitled\

BAGEL

Last updated: 09/20/10 - 13:49

Imported from 'C:\temp\jg\CSSCR class\csscr_gdp.csv'

	BAGEL
2006Q1	72
2006Q2	90
2006Q3	108
2006Q4	117
2007Q1	63
2007Q2	105
2007Q3	126
2007Q4	114
2008Q1	96
2008Q2	81
2008Q3	117
2008Q4	102
2009Q1	84
2009Q2	90
2009Q3	87
2009Q4	65
2010Q1	66
2010Q2	63

Object Name

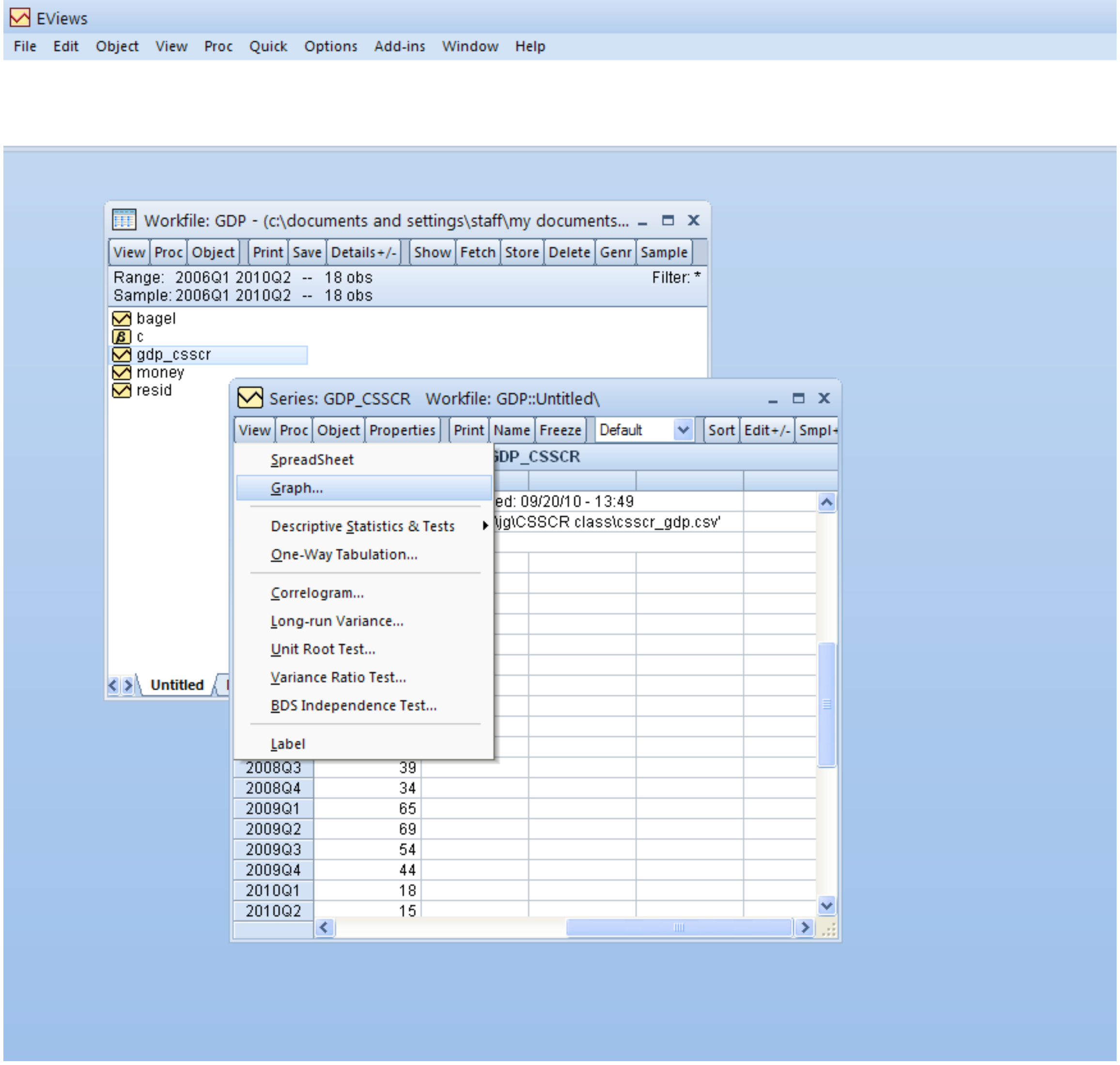
Name to identify object: bagel

24 characters maximum, 16 or fewer recommended

Display name for labeling tables and graphs (optional): Quarterly Consumption of Bagels at CSSCR

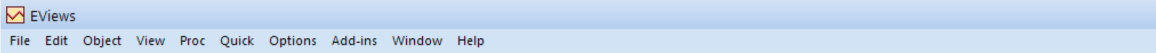
PLOT YOUR DATA.

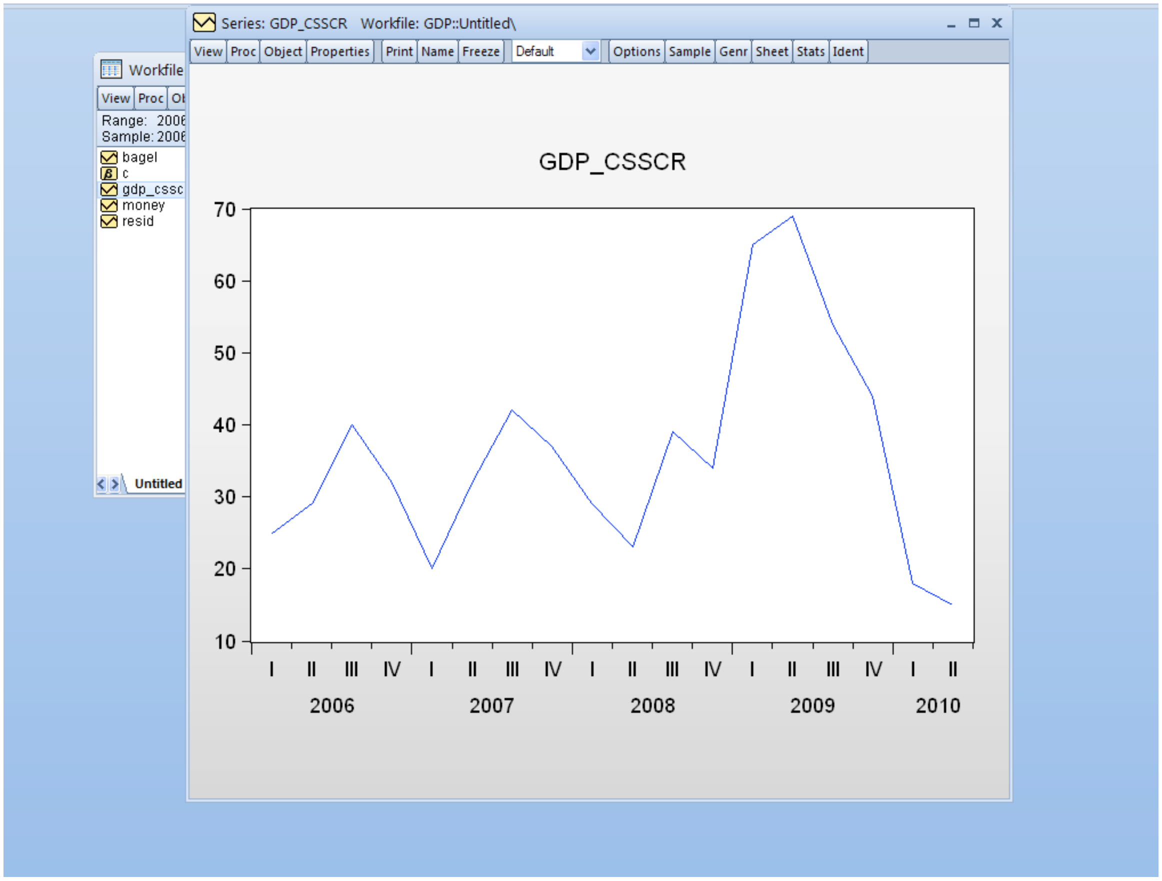
Plotting your data is an often overlooked but very good practice. Plotting can help you catch errors in yourdata before you’ve run lots of regressions that make no sense. To create a time series plot of gdp_csscr (Quarterly GDP created at CSSCR), double-click on the workfile to open up the series gdp_ccsscr. In this window select View/Graph/Line to plot a line graph.



```
Range:  2006Q1 2010Q2  --  18 obs                Filter: *
Sample: 2006Q1 2010Q2  --  18 obs
```

```
DP_CSSCR
ed: 09/20/10 - 13:49
\jg\CSSCR class\csscr_gdp.csv'
2008Q3    39
2008Q4    34
2009Q1    65
2009Q2    69
2009Q3    54
2009Q4    44
2010Q1    18
2010Q2    15
```


EViews
File Edit Object View Proc Quick Options Add-ins Window Help



It looks like starting from the second quarter of 2009, CSSCR was heading to a moderate recession.

SIMPLE DATA MANIPULATION

To explore the statistical characteristics of a variable, double-click on it in the workfile, selectView/Descriptive Stats/Histogram and Stats. This generates the summary statistics.

If we know CSSCR GDP is actually measured in thousands of dollars, we can transform it to a new variable, gdp_new, for later use. This can be achieved by simply typing genr gdp_new=1000*gdp_csscr in the command window.

Machine Learning

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes.

In this tutorial, we’ll look into the common machine learning methods of supervised and unsupervised learning, and common algorithmic approaches in machine learning, including the k-nearest neighbor algorithm, decision tree learning, and deep learning. We’ll explore which programming languages are most used in machine learning, providing you with some of the positive and negative attributes of each. Additionally, we’ll discuss biases that are perpetuated by machine learning algorithms, and consider what can be kept in mind to prevent these biases when building algorithms.

Machine Learning Methods

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data. Let’s explore these methods in more detail.

Supervised Learning

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to “learn” by comparing its actual output with the “taught” outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data.

For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

Unsupervised Learning

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Without being told a “correct” answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

Approaches

As a field, machine learning is closely related to computational statistics, so having a background knowledge in statistics is useful for understanding and leveraging machine learning algorithms.

For those who may not have studied statistics, it can be helpful to first define correlation and regression, as they are commonly used techniques for investigating the relationship among quantitative variables. Correlation is a measure of association between two variables that are not designated as either dependent or independent. Regression at a basic level is used to examine the relationship between one dependent and one independent variable. Because regression statistics can be used to anticipate the dependent variable when the independent variable is known, regression enables prediction capabilities.

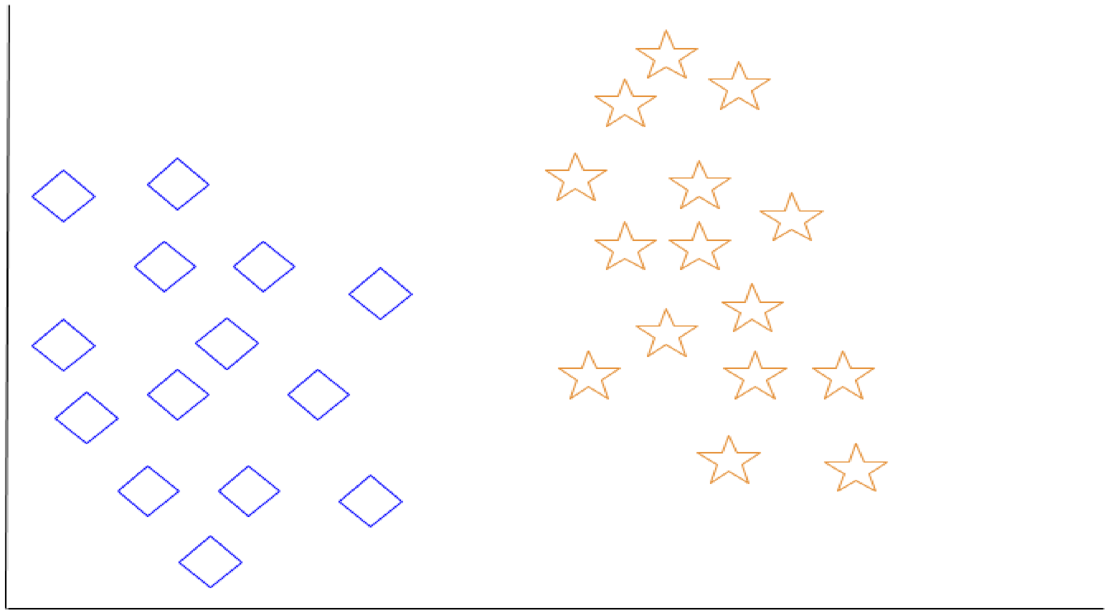
Approaches to machine learning are continuously being developed. For our purposes, we'll go through a few of the popular approaches that are being used in machine learning at the time of writing.

k-nearest neighbor

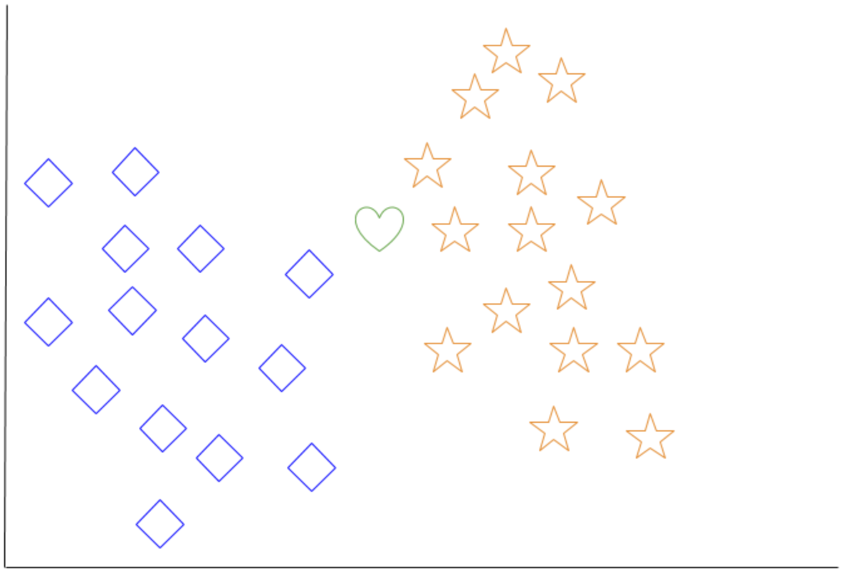
The k-nearest neighbor algorithm is a pattern recognition model that can be used for classification as well as regression. Often abbreviated as k-NN, the k in k-nearest neighbor is a positive integer, which is typically small. In either classification or regression, the input will consist of the k closest training examples within a space.

We will focus on k-NN classification. In this method, the output is class membership. This will assign a new object to the class most common among its k nearest neighbors. In the case of k = 1, the object is assigned to the class of the single nearest neighbor.

Let's look at an example of k-nearest neighbor. In the diagram below, there are blue diamond objects and orange star objects. These belong to two separate classes: the diamond class and the star class.

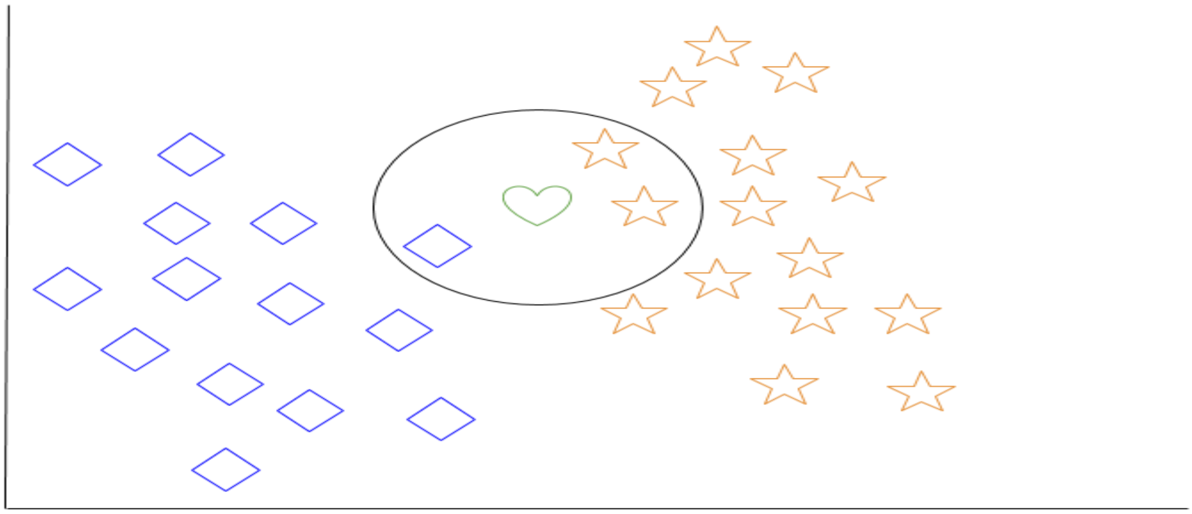


When a new object is added to the space — in this case a green heart — we will want the machine learning algorithm to classify the heart to a certain class.



When we choose k = 3, the algorithm will find the three nearest neighbors of the green heart in order to classify it to either the diamond class or the star class.

In our diagram, the three nearest neighbors of the green heart are one diamond and two stars. Therefore, the algorithm will classify the heart with the star class.



Among the most basic of machine learning algorithms, k-nearest neighbor is considered to be a type of "lazy learning" as generalization beyond the training data does not occur until a query is made to the system.

Decision Tree Learning

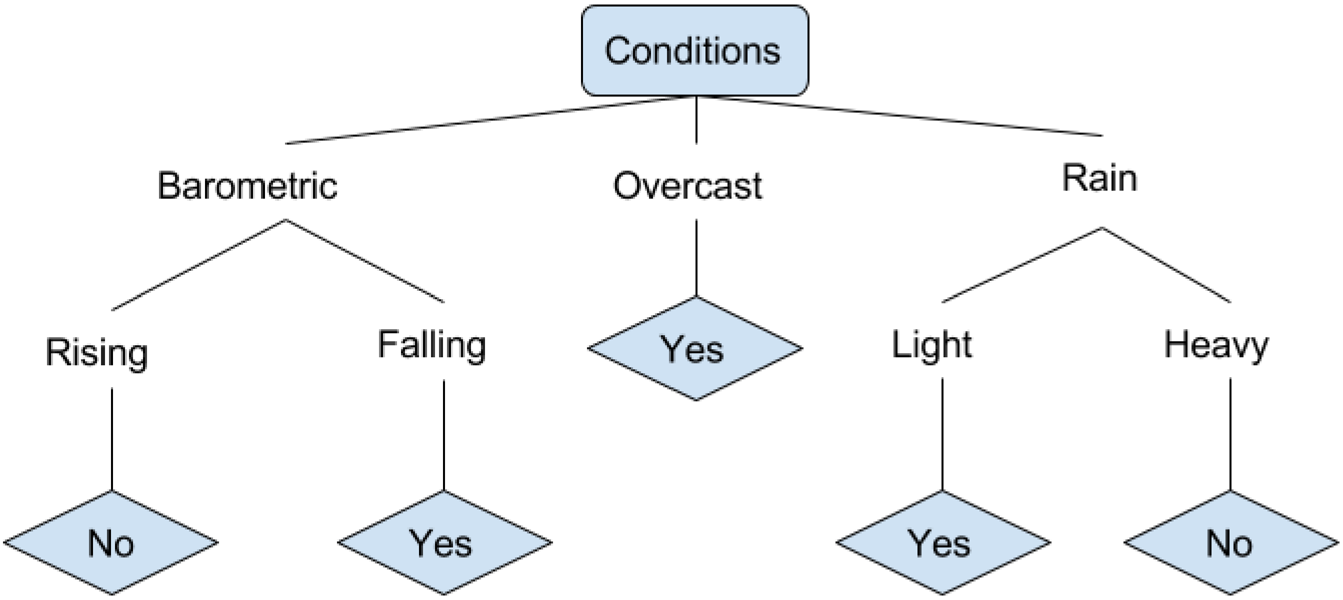
For general use, decision trees are employed to visually represent decisions and show or inform decision making. When working with machine learning and data mining, decision trees are used as a predictive model. These models map observations about data to conclusions about the data's target value.

The goal of decision tree learning is to create a model that will predict the value of a target based on input variables.

In the predictive model, the data's attributes that are determined through observation are represented by the branches, while the conclusions about the data's target value are represented in the leaves.

When "learning" a tree, the source data is divided into subsets based on an attribute value test, which is repeated on each of the derived subsets recursively. Once the subset at a node has the equivalent value as its target value has, the recursion process will be complete.

Let's look at an example of various conditions that can determine whether or not someone should go fishing. This includes weather conditions as well as barometric pressure conditions.



In the simplified decision tree above, an example is classified by sorting it through the tree to the appropriate leaf node. This then returns the classification associated with the particular leaf, which in this case is either a `Yes` or a `No`. The tree classifies a day's conditions based on whether or not it is suitable for going fishing.

A true classification tree data set would have a lot more features than what is outlined above, but relationships should be straightforward to determine. When working with decision tree learning, several determinations need to be made, including what features to choose, what conditions to use for splitting, and understanding when the decision tree has reached a clear ending.

Deep Learning

Deep learning attempts to imitate how the human brain can process light and sound stimuli into vision and hearing. A deep learning architecture is inspired by biological neural networks and consists of multiple layers in an artificial neural network made up of hardware and GPUs.

Deep learning uses a cascade of nonlinear processing unit layers in order to extract or transform features (or representations) of the data. The output of one layer serves as the input of the successive layer. In deep learning, algorithms can be either supervised and serve to classify data, or unsupervised and perform pattern analysis.

Among the machine learning algorithms that are currently being used and developed, deep learning absorbs the most data and has been able to beat humans in some cognitive tasks. Because of these attributes, deep learning has become an approach with significant potential in the artificial intelligence space

Computer vision and speech recognition have both realized significant advances from deep learning approaches. IBM Watson is a well-known example of a system that leverages deep learning.

Programming Languages

When choosing a language to specialize in with machine learning, you may want to consider the skills listed on current job advertisements as well as libraries available in various languages that can be used for machine learning processes.

Python's is one of the most popular languages for working with machine learning due to the many available frameworks, including Tensor Flow, PyTorch, and Keras. As a language that has readable syntax and the ability to be used as a scripting language, Python proves to be powerful and straightforward both for pre-processing data and working with data directly. The scikit-learn machine learning library is built on top of several existing Python packages that Python developers may already be familiar with, namely NumPy, SciPy, and Matplotlib.

Java is widely used in enterprise programming, and is generally used by front-end desktop application developers who are also working on machine learning at the enterprise level. Usually it is not the first choice for those new to programming who wants to learn about machine learning, but is favored by those with a background in Java development to apply to machine learning. In terms of machine learning applications in industry, Java tends to be used more than Python for network security, including in cyber-attack and fraud detection use cases.

Among machine learning libraries for Java are Deeplearning4j, an open-source and distributed deep-learning library written for both Java and Scala; MALLET (MAchine Learning for Languag E Toolkit) allows for machine learning applications on text, including natural language processing, topic modeling, document classification, and clustering; and Weka, a collection of machine learning algorithms to use for data mining tasks.

C++ is the language of choice for machine learning and artificial intelligence in game or robot applications (including robot locomotion). Embedded computing hardware developers and electronics engineers are more likely to favor C++ or C in machine learning applications due to their proficiency and level of control in the language. Some machine learning libraries you can use with C++ include the scalable mlpack, Dlib offering wide-ranging machine learning algorithms, and the modular and open-source Shark.

Human Biases

Although data and computational analysis may make us think that we are receiving objective information, this is not the case; being based on data does not mean that machine learning outputs are neutral. Human bias plays a role in how data is collected, organized, and ultimately in the algorithms that determine how machine learning will interact with that data.

If, for example, people are providing images for “fish” as data to train an algorithm, and these people overwhelmingly select images of goldfish, a computer may not classify a shark as a fish. This would create a bias against sharks as fish, and sharks would not be counted as fish.

When using historical photographs of scientists as training data, a computer may not properly classify scientists who are also people of color or women. In fact, recent peer-reviewed research has indicated that AI and machine learning programs exhibit human-like biases that include race and gender prejudices. See, for example “Semantics derived automatically from language corpora contain human-like biases” and “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints” .

As machine learning is increasingly leveraged in business, uncaught biases can perpetuate systemic issues that may prevent people from qualifying for loans, from being shown ads for high-paying job opportunities, or from receiving same-day delivery options.

Because human bias can negatively impact others, it is extremely important to be aware of it, and to also work towards eliminating it as much as possible. One way to work towards achieving this is by ensuring that there are diverse people working on a project and that diverse people are testing and reviewing it. Others have called for regulatory third parties to monitor and audit algorithms, building alternative systems that can detect biases, and ethics reviews as part of data science project planning. Raising awareness about biases, being mindful of our own unconscious biases, and structuring equity in our machine learning projects and pipelines can work to combat bias in this field.

UNIT – V - REGRESSION ANALYSIS

Introduction

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Also called simple regression or ordinary least squares (OLS), linear regression is the most common form of this technique. Linear regression establishes the linear relationship between two variables based on a line of best fit. Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is zero. Non-linear regression models also exist, but are far more complex.

Regression analysis is a powerful tool for uncovering the associations between variables observed in data, but cannot easily indicate causation. It is used in several contexts in business, finance, and economics. For instance, it is used to help investment managers value assets and understand the relationships between factors such as commodity prices and the stocks of businesses dealing in those commodities.

Regression as a statistical technique should not be confused with the concept of regression to the mean (mean reversion).

- A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables.
- A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.
- It does this by essentially fitting a best-fit line and seeing how the data is dispersed around this line.
- Regression helps economists and financial analysts in things ranging from asset valuation to making predictions.

- In order for regression results to be properly interpreted, several assumptions about the data and the model itself must hold.

Understanding Regression

Regression captures the correlation between variables observed in a data set, and quantifies whether those correlations are statistically significant or not.

The two basic types of regression are simple linear regression and multiple linear regression, although there are non-linear regression methods for more complicated data and analysis. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome (while holding all others constant).

Regression can help finance and investment professionals as well as professionals in other businesses. Regression can also help predict sales for a company based on weather, previous sales, GDP growth, or other types of conditions. The capital asset pricing model (CAPM) is an often-used regression model in finance for pricing assets and discovering costs of capital.

Regression and Econometrics

Econometrics is a set of statistical techniques used to analyze data in finance and economics. An example of the application of econometrics is to study the income effect using observable data. An economist may, for example, hypothesize that as a person increases their income their spending will also increase.

If the data show that such an association is present, a regression analysis can then be conducted to understand the strength of the relationship between income and consumption and whether or not that relationship is statistically significant—that is, it appears to be unlikely that it is due to chance alone.

Note that you can have several explanatory variables in your analysis—for example, changes to GDP and inflation in addition to unemployment in explaining stock market prices. When more than one explanatory variable is used, it is referred to as multiple linear regression. This is the most commonly used tool in econometrics.

Econometrics is sometimes criticized for relying too heavily on the interpretation of regression output without linking it to economic theory or looking for causal mechanisms. It is crucial that the findings revealed in the data are able to be adequately explained by a theory, even if that means developing your own theory of the underlying processes.

Calculating Regression

Linear regression models often use a least-squares approach to determine the line of best fit. The least-squares technique is determined by minimizing the sum of squares created by a mathematical function. A square is, in turn, determined by squaring the distance between a data point and the regression line or mean value of the data set.

Once this process has been completed (usually done today with software), a regression model is constructed. The general form of each type of regression model is:

Simple linear regression:

\begin{aligned}&Y = a + bX + u \\\end{aligned}Y=a+bX+u

Multiple linear regression:

\begin{aligned}&Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u \\&\textbf{where:} \\&Y = \text{The dependent variable you are trying to predict} \\&\text{or explain} \\&X = \text{The explanatory (independent) variable(s) you are } \\&\text{using to predict or associate with Y} \\&a = \text{The y-intercept} \\&b = \text{(beta coefficient) is the slope of the explanatory} \\&\text{variable(s)} \\&u = \text{The regression residual or error term} \\\end{aligned}*Y*=*a*+*b*1*X*1+*b*2*X*2+*b*3*X*3+...+*btXt*
+*u***where:***Y*=The dependent variable you are trying to predictor explain*X*=The explanatory (independent) variable(s) you are using to predict or associate with Y*a*=The y-interceptb=(beta coefficient) is the slope of the explanatoryvariable(s)*u*=The regression residual or error term

Example of How Regression Analysis Is Used in Finance

Regression is often used to determine how many specific factors such as the price of a commodity, interest rates, particular industries, or sectors influence the price movement of an asset. The aforementioned CAPM is based on regression, and it is utilized to project the expected returns for stocks and to generate costs of capital. A stock's returns are regressed against the returns of a broader index, such as the S&P 500, to generate a beta for the particular stock.

Beta is the stock's risk in relation to the market or index and is reflected as the slope in the CAPM model. The return for the stock in question would be the dependent variable Y, while the independent variable X would be the market risk premium.

Additional variables such as the market capitalization of a stock, valuation ratios, and recent returns can be added to the CAPM model to get better estimates for returns. These additional factors are known as the Fama-French factors, named after the professors who developed the multiple linear regression model to better explain asset returns.1

Why Is It Called Regression?

Although there is some debate about the origins of the name, the statistical technique described above most likely was termed "regression" by Sir Francis Galton in the 19th century to describe the statistical feature of biological data (such as heights of people in a population) to regress to some mean level. In other words, while there are shorter and taller people, only outliers are very tall or short, and most people cluster somewhere around (or "regress" to) the average.2

What Is the Purpose of Regression?

In statistical analysis, regression is used to identify the associations between variables occurring in some data. It can show both the magnitude of such an association and also determine its statistical significance (i.e., whether or not the association is likely due to chance). Regression is a powerful tool for statistical inference and has also been used to try to predict future outcomes based on past observations.

How Do You Interpret a Regression Model?

A regression model output may be in the form of Y = 1.0 + (3.2)X_1 - 2.0(X_2) + 0.21.

Here we have a multiple linear regression that relates some variable Y with two explanatory variables X_1 and X_2. We would interpret the model as the value of Y changes by 3.2x for every one-unit change in X_1 (if X_1 goes up by 2, Y goes up by 6.4, etc.) *holding all else constant* (all else equal). That means controlling for X_2, X_1 has this observed relationship. Likewise, holding X1 constant, every one unit increase in X_2 is associated with a 2x *decrease* in Y. We can also note the y-intercept of 1.0, meaning that Y = 1 when X_1 and X_2 are both zero. The error term (residual) is 0.21.

What Are the Assumptions That Must Hold for Regression Models?

In order to properly interpret the output of a regression model, the following main assumptions about the underlying data process of what you analyzing must hold:

- The relationship between variables is linear
- Homoskedasticity, or that the variance of the variables and error term must remain constant
- All explanatory variables are independent of one another
- All variables are normally-distributed

Definition of Linear and Non-Linear Equation

Linear means something related to a line. All the linear equations are used to construct a line. A non-linear equation is such which does not form a straight line. It looks like a curve in a graph and has a variable slope value.

The major difference between linear and nonlinear equations is given here for the students to understand it in a more natural way. The differences are provided in a tabular form with examples.

What is the difference between Linear and Nonlinear Equations?

To find the difference between the two equations, i.e. linear and nonlinear, one should know the definitions for them. So, let us define and see the difference between them.

Linear Equations	**Non-Linear Equations**
It forms a straight line or represents the equation for the straight line	It does not form a straight line but forms a curve.
It **has only one degree**. Or we can also define it as an equation having the maximum degree 1.	A nonlinear equation **has the degree as 2 or more than 2**, but not less than 2.
All these equations form a straight line in XY plane. These lines can be extended to any direction but in a straight form.	It forms a curve and if we increase the value of the degree, the curvature of the graph increases.
The general representation of linear equation is; **$y = mx + c$** Where x and y are the variables, m is the slope of the line and c is a constant value.	The general representation of nonlinear equations is; **$ax^2 + by^2 = c$** Where x and y are the variables and a,b and c are the constant values
Examples: • $10x = 1$ • $9y + x + 2 = 0$	**Examples:** • $x^2+y^2 = 1$ • $x^2 + 12xy + y^2 = 0$

• $4y = 3x$ • $99x + 12 = 23\ y$	• $x^2+x+2 = 25$

Note:

The linear equation has only one variable usually and if any equation has two variables in it, then the equation is defined as a Linear equation in two variables. For example, $5x + 2 = 1$ is Linear equation in one variable. But $5x + 2y = 1$ is a Linear equation in two variables.

Let us see some examples based on these concepts.

Solved Examples

Example: Solve the linear equation $3x+9 = 2x + 18$.

Solution: Given, $3x+9 = 2x + 18$

$\Rightarrow 3x - 2x = 18 - 9$

$\Rightarrow x = 9$

Example: Solve the nonlinear equation $x+2y = 1$ and $x = y$.

Solution: Given, $x+2y = 1$

$x = y$

By putting the value of x in the first equation we get,

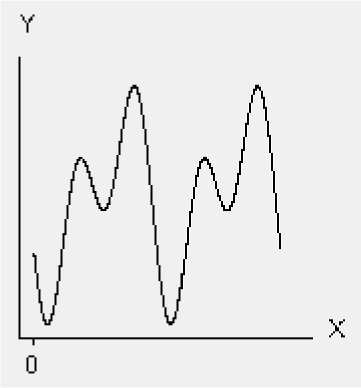
$\Rightarrow y + 2y = 1$

$\Rightarrow 3y = 1$

$\Rightarrow y = \frac{1}{3}$

$\therefore x = y = \frac{1}{3}$

The difference between linear and nonlinear regression models isn't as straightforward as it sounds. You'd think that linear equations produce straight lines and nonlinear equations model curvature. Unfortunately, that's *not* correct. Both types of models can fit curves to your data—so that's not the defining characteristic. In this post, I'll teach you how to identify linear and nonlinear regression models.



The difference between nonlinear and linear is the "non." OK, that sounds like a joke, but, honestly, that's the easiest way to understand the difference. First, I'll define what linear regression is, and then everything else must be nonlinear regression. I'll include examples of both linear and nonlinear regression models.

Linear Regression Equations

A linear regression model follows a very particular form. In statistics, a regression model is linear when all terms in the model are one of the following:

- The constant
- A parameter multiplied by an independent variable (IV)

Then, you build the equation by only adding the terms together. These rules limit the form to just one type:

Dependent variable = constant + parameter * IV + … + parameter * IV

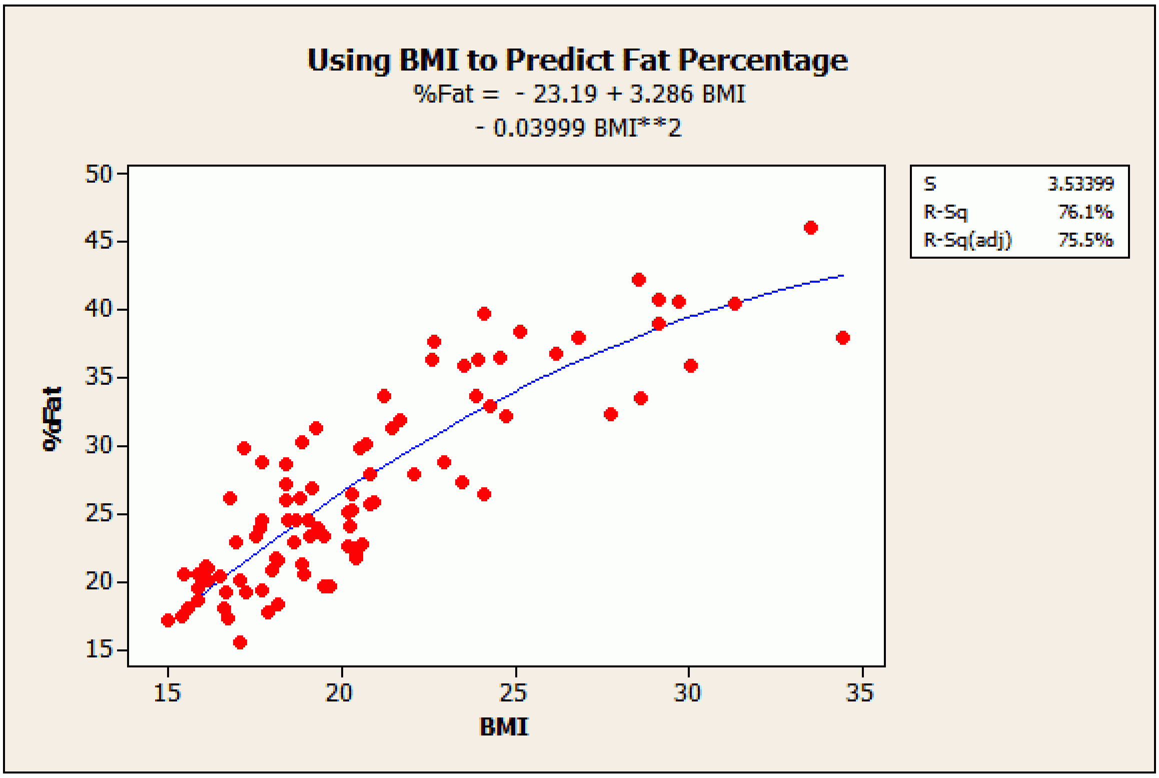
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Statisticians say that this type of regression equation is linear in the parameters. However, it is possible to model curvature with this type of model. While the function must be linear in the parameters, you can raise an independent variable by an exponent to fit a curve. For example, if you square an independent variable, the model can follow a U-shaped curve.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

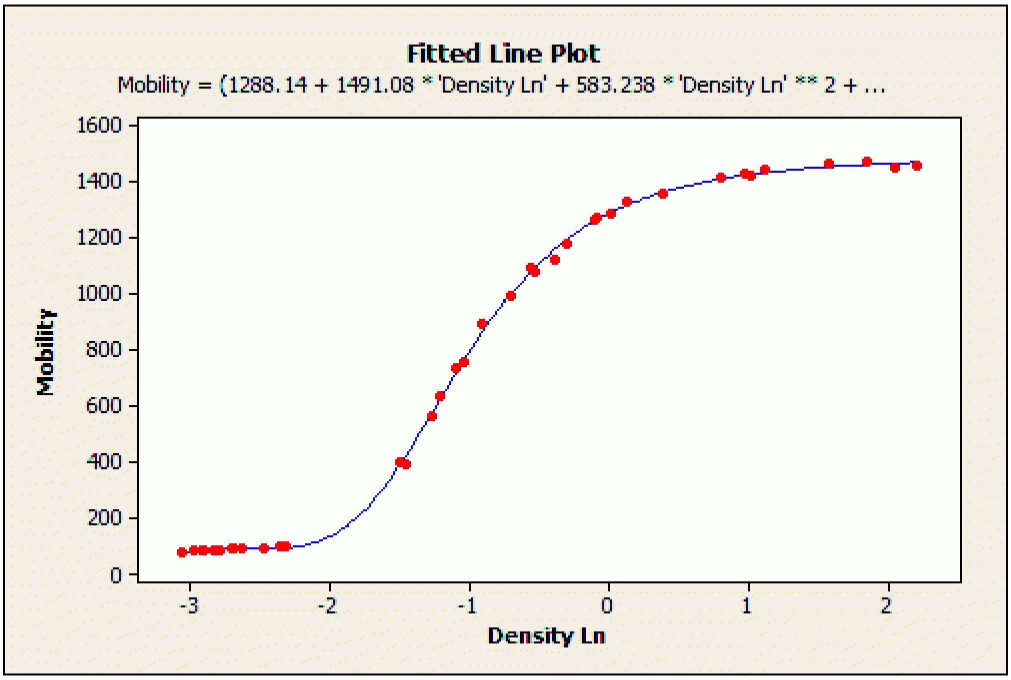
While the independent variable is squared, the model is still linear in the parameters. Linear models can also contain log terms and inverse terms to follow different kinds of curves and yet continue to be linear in the parameters.

The regression example below models the relationship between body mass index (BMI) and body fat percent. In a different blog post, I use this model to show how to make predictions with regression analysis. It is a linear model that uses a quadratic (squared) term to model the curved relationship.



Nonlinear Regression Equations

Nonlinear regression example below models the relationship between density and electron mobility.



The equation for the nonlinear regression analysis is too long for the fitted line plot:

Statistics is a branch of Mathematics, that deals with the collection, analysis, interpretation, and the presentation of the numerical data. In other words, it is defined as the collection of quantitative data. The main purpose of Statistics is to make an accurate conclusion using a limited sample about a greater population.

Types of Statistics

Statistics can be classified into two different categories. The two different types of Statistics are:

- Descriptive Statistics
- Inferential Statistics

In Statistics, **descriptive statistics** describe the data, whereas **inferential statistics** help you make predictions from the data. In inferential statistics, the data are taken from the sample and allows you to generalize the population. In general, inference means "guess", which means making inference about something. So, statistical inference means, making inference about the population. To take a conclusion about the population, it uses various statistical analysis techniques. In this article, one of the types of statistics called inferential statistics is explained in detail. Now, you are going to learn the proper definition of statistical inference, types, solutions, and examples.

Statistical Inference Definition

Statistical inference is the process of analysing the result and making conclusions from data subject to random variation. It is also called inferential statistics. Hypothesis testing and confidence intervals are the applications of the statistical inference. Statistical inference is a method of making decisions about the parameters of a population, based on random sampling. It helps to assess the relationship between the dependent and independent variables. The purpose of statistical inference to estimate the uncertainty or sample to sample variation. It allows us to provide a probable range of values for the true values of something in the population. The components used for making statistical inference are:

- Sample Size
- Variability in the sample
- Size of the observed differences

Types of Statistical Inference

There are different types of statistical inferences that are extensively used for making conclusions. They are:

- One sample hypothesis testing
- Confidence Interval
- Pearson Correlation
- Bi-variate regression
- Multi-variate regression
- Chi-square statistics and contingency table

- ANOVA or T-test

Statistical Inference Procedure

The procedure involved in inferential statistics are:

- Begin with a theory
- Create a research hypothesis
- Operationalize the variables
- Recognize the population to which the study results should apply
- Formulate a null hypothesis for this population
- Accumulate a sample from the population and continue the study
- Conduct statistical tests to see if the collected sample properties are adequately different from what would be expected under the null hypothesis to be able to reject the null hypothesis

Statistical Inference Solution

Statistical inference solutions produce efficient use of statistical data relating to groups of individuals or trials. It deals with all characters, including the collection, investigation and analysis of data and organizing the collected data. By statistical inference solution, people can acquire knowledge after starting their work in diverse fields. Some statistical inference solution facts are:

- It is a common way to assume that the observed sample is of independent observations from a population type like Poisson or normal
- Statistical inference solution is used to evaluate the parameter(s) of the expected model like normal mean or binomial proportion

Importance of Statistical Inference

Inferential Statistics is important to examine the data properly. To make an accurate conclusion, proper data analysis is important to interpret the research results. It is majorly used in the future prediction for various observations in different fields. It helps us to make inference about the data. The statistical inference has a wide range of application in different fields, such as:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector

Statistical Inference Examples

An example of statistical inference is given below.

Problem 1: From the shuffled pack of cards, a card is drawn. This trial is repeated for 400 times, and the suits are given below:

Suit	Spade	Clubs	Hearts	Diamonds
No.of times drawn	90	100	120	90

While a card is tried at random, then what is the probability of getting a

1. Diamond cards
2. Black cards
3. Except for spade

Solution:

By statistical inference solution,

Total number of events = 400

i.e.,90+100+120+90=400

(1) The probability of getting diamond cards:

Number of trials in which diamond card is drawn = 90

Therefore, P(diamond card) = 90/400 = 0.225

(2) The probability of getting black cards:

Number of trials in which black card showed up = 90+100 =190

Therefore, P(black card) = 190/400 = 0.475

(3) Except for spade

Number of trials other than spade showed up = 90+100+120 =310

Therefore, P(except spade) = 310/400 = 0.775

Time – Series Forecasting

In the simplest terms, **time-series forecasting** is a technique that utilizes historical and current data to predict future values over a period of time or a specific point in the future. By analyzing data that we stored in the past, we can make informed decisions that can guide our business strategy and help us understand future trends.

Some of you may be asking yourselves what the difference is between time-series forecasting and algorithmic predictions using, for example, machine learning. Well, machine learning techniques such as random forest, gradient boosting regressor, and time delay neural networks can be used to extrapolate time-series data, but they are far from the only available options or the best ones (as you will see in this article). The most important property of a time-series algorithm is the ability to extrapolate patterns outside of the domain of training data, which most machine learning techniques cannot do by default. This is where specialized time-series forecasting techniques come in.There are plenty of forecasting techniques to choose from, and this article will help you acquire a basic understanding of the most popular ones. From simple linear regression models to complex and vast neural networks, each forecasting method has its own benefits and drawbacks

Application of Time Series Forecasting

Quite a few industries and scientific fields are utilizing time-series forecasting. Some of the most relevant ones include:

- Business planning
- Control engineering
- Cryptocurrency trends
- Financial markets
- Modeling disease spreading
- Pattern recognition
- Resources allocation
- Signal processing
- Sports analytics
- Statistics
- Weather

Time – Series Forecasting Usefulness

- Even though time-series forecasting may seem like a universally applicable technique, there are some limitations that developers need to be aware of. Because forecasting isn't a strictly defined method but rather a combination of data analysis techniques, analysts, and data scientists need to consider the limitations that the prediction models hold, as well as the data itself.
- The most crucial step when considering time-series forecasting is **understanding your data** model and knowing **which business questions** need to be answered using this data. By diving into the problem domain, a developer can more easily distinguish random fluctuations from stable and constant trends in historical data. This will come in handy

when tuning the prediction model to generate the best forecasts and even considering which forecasting method to use.

- When using time-series analysis, some data limitations need to be considered. Common problems include generalizing from a single data source and difficulty in obtaining appropriate measurements and accurately identifying the correct model to represent the data.

There are quite a few factors associated with time-series forecasting, but the most important ones include the following:

- **Amount of data**
- **Data quality**
- **Seasonality**
- **Trends**
- **Unexpected events**

The **amount of data** is probably the most important factor (assuming that the data is accurate). A good rule of thumb would be *the more data we have, the better our model will generate forecasts*. This also makes it much easier for our model to distinguish between trends and noise in the data.

Time series forecasting is hardly a new problem in data science and statistics. The term is self-explanatory and has been on business analysts' agenda for decades now: The very first instances of time series analysis and forecasting trace back to the early 1920s.

Although an intern analyst today can work with time series in Excel, the growth of computing power and data tools allows for leveraging time series for much more complex problems than before to achieve higher prediction accuracy.

In this article, we'll have a broad look at how time series forecasting is used, discuss several methods, and name the key tools that help data scientists put this technique to use.

What are time series forecasting and analysis?

Time series forecasting is a set of methods in statistics and data science to predict some variables that develop and change over time. The underlying intention of time series forecasting is determining how target variables will change in the future by observing historical data from the time perspective, defining the patterns, and yielding short or long-term predictions on how change occurs – considering the captured patterns.

The use cases for this approach are numerous, ranging from sales and demand predictions to highly specialized scientific works on bacterial ecosystems.

What is time series analysis?

Time series analysis focuses on understanding the dependencies in data as it changes over time. Unlike forecasting, it tries to answer the questions *what happens?* and *why does that happen?* Forecasting, on the other hand, corresponds to finding out *what will happen*.

That said, analysis precedes forecasting and helps data scientists prepare data for training machine learning models.

Trends, seasons, cycles, and irregularities

Many machine learning and data mining tasks operate with datasets that have a single slice of time or don't consider the time aspect at all. Natural language processing, image or sound recognition, and numerous classification and regression problems can be solved without time variables at all. For example, the sound recognition solution that we worked with entailed capturing specific teeth grinding sounds of patients as they slept. So, we weren't interested in how these sounds change over time, but rather how to distinguish them from ambient sounds.

Time series problems, on the other hand, are always time-dependent and we usually look at four main components: seasonality, trends, cycles, and irregular components.

Forecasting pandemic spread, diagnosis, and medication planning in healthcare

Time series analysis and forecasting became the key technique applied in healthcare to predict the spread of Covid-19. It was used for transmission predictions, mortality ratios, the spread of the epidemic, and more.

Besides this obviously pertinent set of use cases, time series analysis and forecasting found their applications across almost all healthcare fields, from genetics to diagnosis and treatment.

There’s significant progress in both short- and long-term prognosis of medical conditions. For instance, times series forecasting is used to predict possible death after heart attacks to apply preventive care.

Revenue cycle management and medical billing are at the other end of the treatment spectrum. One of the use cases for time series forecasts is predicting expenditures on prescribed medications.

Anomaly detection for fraud detection, cyber security, and predictive maintenance

Anomaly detection is one of the common machine learning tasks that looks for outliers in the way data points are normally distributed. While it doesn’t necessarily has to be time series data, anomaly detection often goes hand in hand with it. Detecting anomalies in time series entails finding irregular spikes or valleys that significantly deviate from the way seasons and trends look.

Fraud detection is a business-critical activity for any industry dealing with payments and other financial operations. PayPal — an ML powerhouse when it comes to fraud detection — applies time series analysis to track the normal frequency of operations of each account to find irregular spikes in the number of transactions. These findings are also checked against other suspicious activities, like recent changes in the shipping address or massive funds withdrawal, to highlight a given transaction as likely being fraudulent.

Trends. The trend component describes how the variable – drug sales in this case – changes over long periods of time. We see that the sales revenues of antidiabetic drugs have substantially increased during the period from the 1990s to 2010s.

Seasons. The seasonal component showcases each year's wave-like changes in sales patterns. Sales were increasing and decreasing seasonally. Seasonal series can be tied to any time measurement, but these periods of time always have a fixed length and succession. We can consider monthly or quarterly patterns for sales in midsize or small eCommerce, or track micro-interactions across a day.

Cycles. Cycles are long-term patterns that have a waveform and recurring nature similar to seasonal patterns but with variable length, they don't have a fixed time period. For example, business cycles have recognizable elements of growth, recession, and recovery. But the cycles themselves stretch in time differently for a given country throughout its history.

Irregularities. Irregular components appear due to unexpected events, like cataclysms, or are simply representative of noise in the data.

Time series forecasting and analysis: examples and uses cases

Time series forecasting is arguably one of the most common areas where machine learning is applied in business. Let's discuss some use cases from our experience and other businesses.

Demand forecasting for retail, procurement, and dynamic pricing

Predicting customer demand is a cornerstone task for businesses that manage supplies and procurement. Another common application is predicting prices and rates for products and services that dynamically adjust prices depending on demand and revenue targets.

Introduction to Simulation

In simulation, an analyst runs multiple scenarios to predict how a system or process performs under different conditions, and it's the basis for predictive analytics.

Modeling, also known as optimization modeling, differs in that it can determine a specific, optimal or best outcome of a specific scenario. This is known as prescribing an outcome, hence the name prescriptive analytics.

Let's learn more about the difference between modeling and simulation.

Definition of Modeling

A model is a representation of a physical object or process. The model may be physical or mathematical and closely replicates the object or process. It may be used to determine answers to questions such as which is the best business decision to make given a particular set of circumstances.

A key aspect of optimization modeling (mathematical) is the use of mathematical equations and techniques to create models that perform in much the same way as the process being modeled. This applies as much to a business model as it does to a CAD model that incorporates formulae representing the physical strength and rigidity of structural components.

Modeling Examples

As mentioned, modeling can be performed with both physical and mathematical models. For example, Formula 1 racing car teams create exact scale models of their cars to determine the best arrangement of various aero components to minimize drag at high speed or to increase downforce.

In the business environment, supply chain managers may use optimization modeling for supply chain network design to determine the best mix of suppliers, locations, logistics and production for optimal supply. Such an approach models the supply chain, then populates the model with organizational and other data, using solver software to determine, in specific financial terms, financial returns to facilitate correct decisions.

What Is Simulation?

The dictionary definition of simulation refers to two specific processes: that of representing how a system functions by creating a representation of the system or the examination of a problem by simulating a real-life situation.

In other words, it relates to a study of the behavior and performance of a real-world system through testing a theoretical, mathematical or physical representation of the system. Through the application of different conditions, it's possible to discover and understand how a real-life system performs. Simulation is used to gain insight into how the system performs under different scenarios.

Simulation examples

In the Second World War, Britain created an exact scale model of the Mohne Dam in the Ruhr Valley in Germany. They used it to study how to bounce massive spherical bombs across the surface of the water so they fell to the bottom of the dam wall before exploding. They also simulated dropping full-sized practice bombs to establish the correct speed, altitude and position needed to drop the bombs correctly.

Weather forecasters use simulation models to predict future weather. These predictions allow organizers of outdoor events to prepare for likely weather as well as letting forecasters give warning of possible abnormal weather. While weather forecasting is often remarkably accurate, based on multiple simulations, it's nevertheless incapable of providing specific answers except possibly very close to the time.

Difference between Optimization Modeling and Simulation

It's important to understand that each approach is valid, offering differing insights and information.

The key difference between modeling and simulation is that optimization modeling provides a definite recommendation for action in a specific situation, while simulation allows users to determine how a system responds to different inputs so as to better understand how it operates.

In optimization modeling, mathematical techniques are used to represent a real-world situation. This approach takes into account key variables, constraints and trade-offs so as to determine an optimal solution that's feasible and achieves certain objectives.

Simulation, on the other hand, evaluates multiple different scenarios in an attempt to understand how a system would work. A good example is the Monte Carlo simulation that uses randomly-generated inputs based on probability to determine the outcome of various scenarios. Limitations of this approach are that outcomes may not be feasible or even workable. Additionally, it isn't possible to accurately determine the single best optimal solution, as answers are dependent upon the randomly generated variables, although in some instances answers can come close.

Using Optimization Modeling and Simulation Together

The differences between optimization modeling and simulation can come in handy at times. While optimization modeling will recommend specific strategies, simulation can be used for optimizing scenario planning and help answer what-if questions.

Another way to exploit these differences is to use simulation to broadly understand how a system works before using optimization modeling to determine specific answers. For example, an organization may use simulation modeling to determine whether there's a need for an additional distribution center and optimization modeling to determine its optimal location and size.

Gaining Greater Certainty by Exploiting the Difference Between Modeling and Simulation.

Simulation analysis is useful for predicting future performance. Sales executives use predictive analytics to simulate future market trends and changes, while stockbrokers use it to anticipate changes in stock and bond indices. Optimization modeling takes this capability a step further with its ability to provide specific, data-driven answers to specific business problems.

True optimization modeling, the basis for prescriptive analytics software such as River Logic's Enterprise Optimizer, would always use data derived from descriptive and prescriptive analytics to determine optimal solutions to specific business problems and, as a tool, to analyze different scenarios.

What's the Difference Between Simulation and Optimization?

Although simulation and optimization are similar and leverage many of the same computational techniques and algorithms, they are different activities. Each has its advantages and disadvantages, and each is better suited for certain types of problems. Here are some key differences between them:

- **"What-if" analysis**: Simulation is better suited to observing the performance of the simulated system by tweaking the initial conditions (that is, the values of the input variables). Optimization is used more often to determine an optimal system design.
- **Constraints**: Successful optimization depends on properly identifying the constraints placed on various parameters—for example, a business might have a maximum number of employees it can hire to work on production lines. With simulation, the analyst starts with realistic values for inputs and modifies them within reasonable ranges to determine what happens with the outputs.
- **Influence of randomness**: Simulations can account for random variation in the parameters—in the barbershop example, each barber's hair cutting speed can be expressed as a normal distribution around an average. This variability can make a large difference in the accuracy of the results. Optimization works better clearly defined mathematical relationships that don't have variability.
- **Planning and decision support:** Optimization methods can be used to support both tactical and strategic planning decisions, because they provide a single "best" answer to a given problem. This is one of the advantages of optimization. Simulation, by contrast, is considered more exploratory.

- **Modeling difficulty:** Simulations are generally easier to model, because fewer assumptions need to be made. An optimization solution requires either more assumptions about the inputs or more computing power to deal with all the different variables to calculate the optimized result.

How Simulation and Optimization Can Work Together

Although simulation and optimization help solve different problems, they can work together to drive business results. How?

By application of simulation techniques for a system—say, the receiving dock at a warehouse—on the basis of observed factors that influence the efficiency, throughput, or other metrics of the receiving team, a business can get a feel for the factors that have the most pronounced effect on the outputs.

Armed with these insights and data simulation techniques, the business can then make better assumptions about the mathematical relationships between the parameters, which drives better optimization—that is, better decisions about what to change and by how much to improve the team's performance.

Examples of Application of Simulation in Business

Data simulation tools are used in businesses of all sizes and in all industries to analyze current processes and determine where to focus on improvements. Here are some simulation system examples:

- **Royal Dutch Shell:** Used a simulation based model to support vessel servicing of offshore oil platforms, including factors such as vessel capacity, storage at port facilities, and more. The simulation showed Shell where best to invest in improvements.
- **Cancer center:** A major medical center in the Midwest modeled internal patient care processes. This helped them determine the best arrangement of different types of patient populations, thereby minimizing patient, doctor, and caregiver travel times and maximizing operating room utilization.
- **Agricultural logistics:** A sugarcane producer in Brazil used simulation to improve operational capacity and reduced capital costs, while increasing the efficiency of the vehicle fleet carrying sugarcane from plantation to mill.
- **Walmart:** Before investing millions into a robotic based system that picks groceries for its online grocery pickup system, Walmart ran a simulation to test viability before making the change – you can read the case study here.

In all simulation exercises, it’s important to account for all the factors that influence the system’s performance and to accurately characterize each one. Simulation modeling requires keen observation and data analysis—in some cases, large amounts of data—to get an accurate picture of the system. It’s also important to validate the model by comparing model data with real system data.

The better the model, the better the simulation’s response to different inputs; good data modeling and simulation can result in better optimization.

Simulation and optimization can therefore be seen as two complementary approaches to solving business problems. With advances in big data simulation software and computing power, simulation and optimization will become increasingly important tools in every company’s decision-making toolkit, enabling better insights and better business decisions.

Decision Support System (DSS)

Introduction

A decision support system (DSS) is a computerized program used to support determinations, judgments, and courses of action in an organization or a business. A DSS sifts through and analyzes massive amounts of data, compiling comprehensive information that can be used to solve problems and in decision-making.

Typical information used by a DSS includes target or projected revenue, sales figures or past ones from different time periods, and other inventory- or operations-related data.

Understanding a Decision Support System (DSS)

A decision support system gathers and analyzes data, synthesizing it to produce comprehensive information reports. In this way, as an informational application, a DSS differs from an ordinary operations application, whose function is just to collect data.

The DSS can either be completely computerized or powered by humans. In some cases, it may combine both. The ideal systems analyze information and actually make decisions for the user. At the very least, they allow human users to make more informed decisions at a quicker pace.

Using a DSS

The DSS can be employed by operations management and other planning departments in an organization to compile information and data and synthesize it into actionable intelligence. In fact, these systems are primarily used by mid- to upper-level management.

For example, a DSS may be used to project a company's revenue over the upcoming six months based on new assumptions about product sales. Due to a large number of factors that surround projected revenue figures, this is not a straightforward calculation that can be done manually. However, a DSS can integrate all the multiple variables and generate an outcome and alternate outcomes, all based on the company's past product sales data and current variables.

Characteristics of a DSS

The primary purpose of using a DSS is to present information to the customer in an easy-to-understand way. A DSS system is beneficial because it can be programmed to generate many types of reports, all based on user specifications. For example, the DSS can generate information and output its information graphically, as in a bar chart that represents projected revenue or as a written report.

As technology continues to advance, data analysis is no longer limited to large, bulky mainframe computers. Since a DSS is essentially an application, it can be loaded on most computer systems, whether on desktops or laptops. Certain DSS applications are also available through mobile devices.

The flexibility of the DSS is extremely beneficial for users who travel frequently. This gives them the opportunity to be well-informed at all times, providing them the ability to make the best decisions for their company and customers on the go or even on the spot.

Purpose of a Decision Support System

A decision support system produces detailed information reports by gathering and analyzing data. Hence, a DSS is different from a normal operations application, whose goal is to collect data and not analyze it.

In an organization, a DSS is used by the planning departments – such as the operations department – which collects data and creates a report that can be used by managers for decision-making. Mainly, a DSS is used in sales projection, for inventory and operations-related data, and to present information to customers in an easy-to-understand manner.

Theoretically, a DSS can be employed in various knowledge domains from an organization to forest management and the medical field. One of the main applications of a DSS in an organization is real-time reporting. It can be very helpful for organizations that take part in just-in-time (JIT) inventory management.

In a JIT inventory system, the organization requires real-time data of their inventory levels to place orders "just in time" to prevent delays in production and cause a negative domino effect. Therefore, a DSS is more tailored to the individual or organization making the decision than a traditional system.

Components of a Decision Support System

The three main components of a DSS framework are:

1. Model Management System

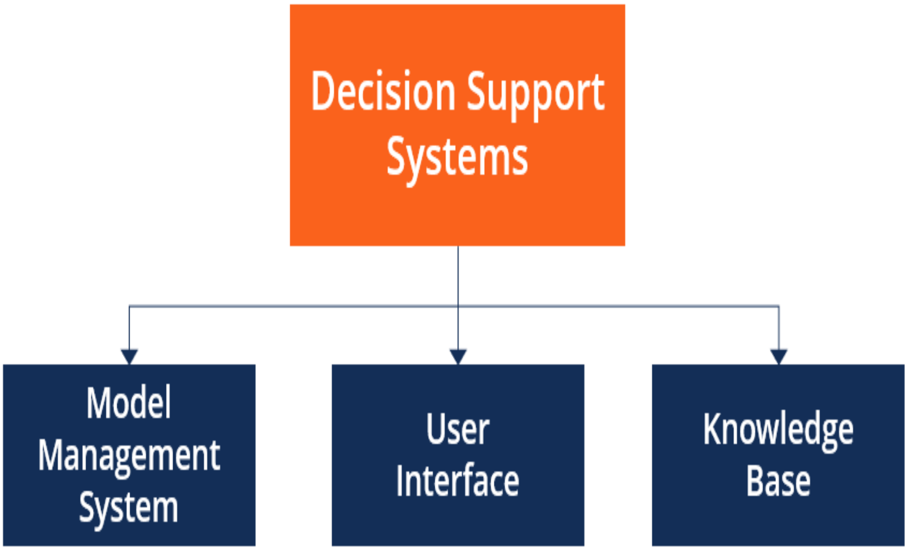
The model management system S=stores models that managers can use in their decision-making. The models are used in decision-making regarding the financial health of the organization and forecasting demand for a good or service.

2. User Interface

The user interface includes tools that help the end-user of a DSS to navigate through the system.

3. Knowledge Base

The knowledge base includes information from internal sources (information collected in a transaction process system) and external sources (newspapers and online databases).



Types of Decision Support Systems

- **Communication-driven**: Allows companies to support tasks that require more than one person to work on the task. It includes integrated tools such as Microsoft SharePoint Workspace and Google Docs.
- **Model-driven**: Allows access to and the management of financial, organizational, and statistical models. Data is collected, and parameters are determined using the information provided by users. The information is created into a decision-making model to analyze situations. An example of a model-driven DSS is Dicodess – an open-source model-driven DSS.
- **Knowledge-driven**: Provides factual and specialized solutions to situations using stored facts, procedures, rules, or interactive decision-making structures like flowcharts.
- **Document-driven**: Manages unstructured information in different electronic formats.
- **Data-driven**: Helps companies to store and analyze internal and external data.

Advantages of a Decision Support System

- A decision support system increases the speed and efficiency of decision-making activities. It is possible, as a DSS can collect and analyze real-time data.

- It promotes training within the organization, as specific skills must be developed to implement and run a DSS within an organization.
- It automates monotonous managerial processes, which means more of the manager's time can be spent on decision-making.
- It improves interpersonal communication within the organization.

Disadvantages of a Decision Support System

- The cost to develop and implement a DSS is a huge capital investment, which makes it less accessible to smaller organizations.
- A company can develop a dependence on a DSS, as it is integrated into daily decision-making processes to improve efficiency and speed. However, managers tend to rely on the system too much, which takes away the subjectivity aspect of decision-making.
- A DSS may lead to information overload because an information system tends to consider all aspects of a problem. It creates a dilemma for end-users, as they are left with multiple choices.
- Implementation of a DSS can cause fear and backlash from lower-level employees. Many of them are not comfortable with new technology and are afraid of losing their jobs to technology.

Decision support system examples

Organizations use decision support systems in several different contexts, including the following:

- **GPS routing.** GPS route planning is an example of a typical DSS. It compares different routes, taking into account factors such as distance, driving time and cost. The GPS navigating system also enables users to choose alternative routes, displaying them on a map and providing step-by-step instructions.

- **ERP dashboards.** ERP (enterprise resource planning) dashboards can use a decision support system to visualize changes in production and business processes, monitor current business performance against set goals and identify areas for improvement. ERP dashboards let business owners see a snapshot of their company's most important numbers and metrics.

- **Clinical decision support system.** A clinical decision support system (CDSS) is a software program that uses advanced decision-making algorithms to help physicians make the best medical decisions. Healthcare professionals often use these to interpret patient records and test results, and to calculate the best treatment plan. CDSS in healthcare can help providers identify abnormalities during specific tests, as well as monitor patients after certain procedures to determine if they are having any adverse reactions.