COURSE
GUIDE

PAD108
SOCIAL STATISTICS II

Course Team

Dr. Musa Zakari-NOUN
& Senusi Mohammed-Federal Polytechnic
Nasarawa (Course Writer/Developer)
Dr. Hassan Mohammed Nazif (Content
Editor)-Nasarawa State University Keffi

NATIONAL OPEN UNIVERSIT OF NIGERIA

© 2024 by NOUN Press
National Open University of Nigeria
Headquarters
University Village
Plot 91, Cadastral Zone
Nnamdi Azikiwe Expressway
Jabi, Abuja

Lagos Office
14/16 Ahmadu Bello Way
Victoria Island, Lagos

E-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

Printed 2024

ISBN: 978-978-786-229-2

CONTENTS

Introduction

Welcome to PAD108: Social Statistics II for Public Administration. This is a 2-credit unit compulsory course for students pursuing a Bachelor of Science (B.Sc) in Public Administration and other related disciplines.

Course Objectives

1. Enhance Statistical Proficiency: This course aims to enhance students' proficiency in the application of statistical methods and techniques in the context of public administration.
2. Develop Analytical Skills: Students will develop critical analytical skills to interpret and manage statistical data relevant to public policy and administration.
3. Foster Decision-Making Abilities: The course is designed to foster decision-making abilities by equipping students with the knowledge to apply statistical analysis in real-world public administration scenarios.
4. Promote Evidence-Based Practices: It promotes the use of evidence-based practices in public administration through rigorous statistical analysis.

Course Content

1. Advanced Statistical Methods: In-depth study of statistical methods including inferential statistics, hypothesis testing, and regression analysis.
2. Data Collection and Analysis: Techniques for efficient data collection, data organization, and data analysis relevant to public administration.
3. Statistical Software: Introduction to statistical software tools used in data analysis and interpretation.
4. Case Studies and Applications: Practical applications of statistical methods through case studies relevant to public policy and administration.

Learning Outcomes

By the end of this course, students should be able to:

1. Apply advanced statistical methods to analyze data in public administration.
2. Interpret statistical results and make informed decisions based on data analysis.
3. Use statistical software to manage and analyze data.
4. Critically evaluate and apply statistical information in public administration scenarios.

Teaching Methods

1. Facilitation and interactive sessions
2. Practical exercises and assignments
3. Case study analysis
4. Use of statistical software for data analysis

This course is essential for those aiming to build a career in public administration, providing the necessary tools and skills to navigate and utilize statistical data effectively in their professional roles.

Course Guide

The purpose of this course is to present an examination of quantitative methods and techniques that are used in the public sector. Areas of focus include Statistical analysis, sampling, forecasting and time-series analysis, and design issues in generating and testing research questions and hypotheses. Quasi-experimental and non-experimental designs will be used including survey research to assess public input on government service quality and applied to practice and policy issues. IBM SPSS software will be used to analyze government datasets using descriptive and inferential statistics including correlations, cross-tabulations, t-tests, ANOVA, and regression

Measurable Learning Outcomes

Upon successful completion of these modules, you will be able to:
Develop research questions and testable hypotheses linked to existing theory or research.
Develop hypotheses, choose appropriate statistics to test them, and describe the results correctly in a short research paper.
Develop research design and literature review related to dataset.
Perform descriptive and inferential statistical analysis of public administrative datasets using IBM SPSS software.
Interpret results from descriptive and inferential statistical analysis of public administrative datasets and place results in APA formatted text, tables, and figures.
Compare statistical test results to those of scholars' studies as reported in the literature review of the assignment.
The course is aimed at acquainting you with what quantitative techniques are all about and letting you understand the practical applications of quantitative techniques in business and economic decision making. To ensure that this aim is achieved, some important background information will be provided and discussed, including:
Definition of quantitative techniques
Uses of quantitative techniques
Tools and applications of quantitative techniques

The correlation theory
Forecasting and time-series analysis
Index numbers
Inventory control
Decision analysis
Network planning and analysis

Self-Assessment-Exercise (SAEs)

Two Self-assessment Exercises each are incorporated in the study material for each unit. Self-assessment Exercise helps students to be a realistic judge of their own performance and to improve their work. Promotes the skills of reflective practice and self-monitoring; Promotes academic integrity through student self-reporting of learning progress; Develops self-directed learning; Increases student motivation and Helps students develop a range of personal, transferrable skills

Summary

Each Unit contained a summary of the entire unit. A summary is a brief statement or restatement of main points, especially as a conclusion to a work: a summary of a chapter. A brief is a detailed outline, by heads and subheads, of a discourse (usually legal) to be completed: a brief for an argument.

Possible Answers to Self-Assessment Exercise(s)

The materials contained Possible Answers to Self-Assessment Exercise(s) within the content. The possible Self-assessments answers enable you to understand how well you're performing in the contents. It is a way of analysing your work performance and any areas for growth. Reflecting on your strengths, weaknesses, values and accomplishments can help you determine what goals to work toward next.

Course Material

The course material package is comprises of following Modules and unit structure:

Module 1

Module 2

Unit 1	Statistical Tools I
Unit 2	Statistical Tools II
Unit 3	Statistical Tools III
Unit 4	Statistical Tools IV
Unit 5	Basic Advance Mathematics

Module 3

Unit 1	Population vs. Sample
Unit 2	Correlation Analysis
Unit 3	Simple Linear Regression
Unit 4	Multiple Linear Regressions
Unit 5	Spearman’s Rank Correlation

Module 4

Unit 1	Pearson Correlation Coefficient
Unit 2	Analysis of Variance (ANOVA)
Unit 3	The Use of Statistical Package for Social Science (SPSS) for Analysis
Unit 4	Time Series Analysis
Unit 5	Forecasting

Module 5

Unit 1	Index numbers
Unit 2	Inventory Control
Unit 3	Economic order quantity (EOQ)
Unit 4	Decision Analysis
Unit 5	Network Analysis
Unit 6	Critical Path of a Project (CPA) and Program Evaluation Review Technique (PERT)

MAIN COURSE

CONTENTS

MODULE 1

Unit 1 Introduction to Advanced Statistics

Unit Structure

1.1 Introduction

Advanced statistics involves complex methods and techniques used to analyze data beyond basic descriptive and inferential statistics. These methods are crucial for uncovering deeper insights, making predictions, and making informed decisions in various fields such as economics, engineering, medicine, and social sciences. Therefore, this unit will be discussing.

1.2 Learning Outcome

By the end of the unit, you should be able to:

- discuss the introduction to Advanced statistics
- explain the Introduction to Regression analysis
- explore the Introduction to Time Series Analysis
- discuss the Introduction to Structural Equation Modeling (SEM)
- explain the Introduction to Statistical Package for the Social Sciences
- outline the Introduction to Popular Statistical Software (R)
- discuss the Introduction to Stata

1.3 Introduction to Advanced statistics

Definition of Advanced statistics is referred to as statistics analysis that is complex, multifaceted, and involves the use of sophisticated mathematical techniques to extract deep insights and patterns from data sets. This specialized branch of statistics goes beyond traditional methods to delve into the realm of predictive modeling, hypothesis testing, and data visualization. It requires a solid foundation in probability theory, linear algebra, and calculus, enabling practitioners to tackle real-world problems with precision and accuracy. Furthermore, advanced statistics often incorporates cutting-edge software and programming tools like SPSS, R, Python, and SAS to streamline the analysis process and facilitate the exploration of large datasets. In essence, advanced statistics empowers professionals across various fields – from finance to healthcare to marketing – to make informed decisions, gain a competitive edge, and drive innovation through data-driven strategies and solutions. As technology continues to evolve and the volume of data grows exponentially, the importance of advanced statistics will only increase, making it a pivotal tool for transforming raw information into valuable knowledge and actionable insights.

Machine learning algorithms harness sophisticated statistical methodologies to develop models capable of generating predictions or categorizing information. In this realm, several key algorithms play pivotal roles.

Linear regression emerges as a powerful tool, enabling the forecasting of continuous results by analyzing the influence of one or multiple predictor variables.

Decision trees, another prominent algorithm, facilitate data classification through a branching process, where data subsets are segregated based on distinct feature values. These algorithms significantly contribute to the scientific evolution of machine learning, enhancing its predictive capabilities and providing vital insights into complex data patterns. While linear regression focuses on predicting outcomes in a continuous manner, decision trees excel in organizing data based on specific features, thereby offering valuable clarity and structure to the analytical process. Such algorithms exemplify the dynamic nature of machine learning and its transformative impact on data analysis across diverse fields, from finance and healthcare to marketing and transportation. Deepening our understanding of these algorithms not only enriches our analytical skills but also amplifies the potential for innovation and discovery in the vast landscape of machine learning technology.

1.3.1 Regression analysis

Regression analysis is a powerful statistical technique used to study and understand the relationship between a dependent variable and one or more independent variables. In the context of multiple regressions, this method allows researchers to model the linear connection between a particular dependent variable and several independent variables simultaneously. By examining these relationships, analysts can gain valuable insights into how changes in the independent variables affect the dependent variable.

Multiple regressions enable a comprehensive exploration of potential influences on the outcome of interest, providing a more nuanced understanding of the underlying dynamics at play. This type of analysis is particularly useful in scenarios where the dependent variable may be influenced by a combination of factors, necessitating a more sophisticated approach to modeling. Through multiple regressions, researchers can identify and quantify the individual impact of each independent variable on the dependent variable, shedding light on the complex interplay between various factors. By utilizing this advanced analytical method, practitioners can uncover hidden patterns, make informed predictions, and draw meaningful conclusions based on the data. In summary, multiple regression analysis serves as a valuable tool for examining the intricate relationships between variables and extracting meaningful insights from complex datasets.

Self-Assessment Exercise 1

i. Discuss the introduction to Advanced statistics
ii. Explain the Introduction to Regression analysis

1.3.2 Time Series Analysis

Time series analysis involves studying and analyzing a sequence of data points collected over time to identify patterns, trends, and relationships within the data. By examining how variables change over time, this statistical technique helps in making predictions and understanding the underlying factors driving the observed behavior. Time series analysis plays a crucial role in various fields such as finance, economics, weather forecasting, and signal processing. It allows researchers and analysts to model time-dependent data, forecast future values, and assess the impact of certain variables on the data series. This analytical method often involves techniques like autocorrelation, spectral analysis, and regression modeling to uncover valuable insights from the time-dependent data. Furthermore, time series analysis can provide valuable information for decision-making processes, such as projecting future sales, analyzing seasonal trends, and understanding the fluctuations in stock prices.

Overall, mastering time series analysis can empower professionals to make informed decisions, anticipate future trends, and optimize strategies based on historical data patterns.

1.3.3 Structural Equation Modeling (SEM)

Structural Equation Modeling (SEM) is a statistical method that encompasses both confirmatory factor analysis and path analysis within a single framework. This methodology allows researchers to examine complex relationships between observed and latent variables by utilizing a series of simultaneous equations. Through SEM, researchers can test hypotheses about causal relationships, model latent constructs, and estimate the strength and direction of effects in a comprehensive manner. By incorporating measurement models and structural equations, SEM provides a powerful tool for investigating intricate relationships in various fields such as psychology, sociology, economics, and more.

SEM offers a systematic approach to analyze multivariate data and understand the underlying structure of a theoretical model by combining statistical techniques and theoretical assumptions. Moreover, SEM allows researchers to account for measurement error, address issues of endogeneity, and explore mediating and moderating effects in a unified statistical framework. As a versatile analytical technique, SEM is widely used in research to assess complex theoretical models, evaluate the fit of the model to the data, and derive meaningful insights into the relationships between variables.

Overall, SEM serves as a valuable tool for researchers seeking to understand the interplay of multiple variables, establish causal pathways, and validate theoretical constructs in a robust and systematic manner.

1.3.4 Statistical Package for the Social Sciences

SPSS, officially known as the Statistical Package for the Social Sciences, is a proprietary software framework meticulously crafted by IBM specifically designed for conducting comprehensive statistical analyses. This robust tool boasts an array of practical features tailored to meet the analytical needs of researchers in various fields. With its intuitive user interface, SPSS provides a seamless experience even for individuals with limited programming knowledge, making it widely accessible across different disciplines.

One of the keys defining characteristics of SPSS is its diverse set of statistical tests, allowing users to delve deep into their data with precision and accuracy. From basic descriptive statistics to advanced analytical methods, SPSS offers a wide spectrum of options to cater to the intricate requirements of data analysis. This versatility makes SPSS a go-to choice for professionals working in the social sciences, market research, health studies, and survey data analysis.

Apart from its rich feature set, SPSS has garnered popularity for its reliability and efficiency in handling complex data sets. Researchers rely on SPSS not only for its analytical capabilities but also for its ability to streamline the entire data analysis process, ultimately saving time and enhancing productivity. The software's reputation for producing insightful results has solidified its position as a trusted tool in the realm of statistical analysis.

SPSS stands as a cornerstone in the realm of statistical software, providing a robust and user-friendly platform for researchers to explore, analyze, and interpret data effectively across a myriad of research domains.

1.3.5 Popular Statistical Software (R)

Popular Statistical Software (R) is a widely renowned open-source programming language and software environment known for its powerful capabilities in statistical computing and graphics. This software boasts an impressive array of features, including extensive libraries that provide users with a wealth of tools to perform complex statistical analyses with ease. Additionally, its robust data visualization capabilities allow users to effectively present their findings in a visually appealing manner. One of the key strengths of this statistical software is the strong community support it enjoys. Users have the opportunity to tap into a vast network of like-minded individuals who are ready to offer assistance, share insights, and collaborate on projects.

Academia and industry alike have embraced this software for its versatility and reliability in performing a wide range of statistical analyses. From regression and time-series analysis to advanced machine learning algorithms, the capabilities of this software are truly impressive. The impact of Popular Statistical Software (R), first introduced by Ihaka and Gentleman in 1996, continues to be felt across various fields, shaping the way researchers and analysts approach data-driven decision-making. Its user-friendly interface, comprehensive documentation, and active online community make it a go-to choice for professionals seeking to leverage the power of statistical computing in their work.

Self-Assessment Exercises 2

i. Explore the Introduction to Time Series Analysis
ii. Discuss the Introduction to Structural Equation Modeling (SEM)
iii. Explain the Introduction to Statistical Package for the Social Sciences

1.3.6 Stata

Stata is a renowned proprietary statistical software package specializing in data analysis, data management, and graphics, catering to a broad spectrum of users across various disciplines. Known for its user-friendly interface and comprehensive documentation, Stata offers robust capabilities in advanced fields such as econometrics and biostatistics, making it a top choice for researchers and analysts alike. Its versatility extends to a range of applications, finding extensive utilization in disciplines like economics, sociology, political science, and epidemiology. StataCorp, the company behind Stata, ensures that the software remains at the forefront of statistical analysis tools, providing researchers with innovative solutions to complex data challenges. With its powerful features and dedicated user base, Stata continues to be a key player in the realm of statistical software, empowering users in academia, research, and industry to delve into data with confidence and precision.

1.4 References/Further Reading/Web Resources

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis. CRC Press.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). Multivariate Data Analysis. Cengage Learning.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

1.5 Possible Answers SAEs

Answer to SAEs 1

1. Introduction to Advanced statistic

Definition of Advanced statistics is referred to as statistics analysis that is complex, multifaceted, and involves the use of sophisticated mathematical techniques to extract deep insights and patterns from data sets. This specialized branch of statistics goes beyond traditional methods to delve into the realm of predictive modeling, hypothesis testing, and data visualization

2. Introduction to Regression analysis

Regression analysis is a powerful statistical technique used to study and understand the relationship between a dependent variable and one or more independent variables. In the context of multiple regressions, this method allows researchers to model the linear connection between a particular dependent variable and several independent variables simultaneously. By examining these relationships, analysts can gain valuable insights into how changes in the independent variables affect the dependent variable. Multiple regressions enable a comprehensive exploration of potential influences on the outcome of interest, providing a more nuanced understanding of the underlying dynamics at play.

Answer to SAEs 2

1. Introduction to Time Series Analysis

Time series analysis involves studying and analyzing a sequence of data points collected over time to identify patterns, trends, and relationships within the data. By examining how variables change over time, this statistical technique helps in making predictions and understanding the underlying factors driving the observed behavior. Time series analysis plays a crucial role in various fields such as finance, economics, weather forecasting, and signal processing. It allows researchers and analysts to model time-dependent data, forecast future values, and assess the impact of certain variables on the data series.

2. Introduction to Structural Equation Modeling (SEM)

Structural Equation Modeling (SEM) is a statistical method that encompasses both confirmatory factor analysis and path analysis within a single framework. This methodology allows researchers to examine complex relationships between observed and latent variables by utilizing a series of simultaneous equations. Through SEM, researchers can test hypotheses about causal relationships, model latent constructs, and estimate the strength and direction of effects in a comprehensive manner. By incorporating measurement models and structural equations, SEM provides a powerful tool for investigating intricate relationships in various fields such as psychology, sociology, economics, and more.

3. Introduction to Statistical Package for the Social Sciences

SPSS, officially known as the Statistical Package for the Social Sciences, is a proprietary software framework meticulously crafted by IBM specifically designed for conducting comprehensive statistical analyses. This robust tool boasts an array of practical features tailored to meet the analytical needs of researchers in various fields.

Unit 2 Data and Data Analysis

Unit Structure

1.1 Introduction

In our last class, we discussed the concept of quantitative methods and technique. We went further to explain the relevance of quantitative Methods. We concluded our last class by identifying the tools of quantitative Methods/Techniques. In today's class, we will be examining the concept of data, types of data and data classification.

.

2.2 Learning Outcomes

By the end of this unit, you will be able to:

- define the concept of data
- identify and explains types of data
- outline the data classification

2.3 Concept of Data

What is data, the term "data" refers to discrete pieces of factual information that are recorded and utilized for the purpose of analysis? The statistics that are compiled come from the raw information that is provided. The end product of data analysis is statistical information, which includes both the interpretation and presentation of the data.

What is raw data; raw data (sometimes called source data, atomic data or primary data) is data that has not been processed for use. A distinction is sometimes made between data and information to the effect that information is the end product of data processing

According to an article published online by the Star Tribune on July 27, 2015, Example of raw data was utilized to produce a graph that showed the average number of passengers that boarded the Green Line train each day, broken down by month. Readers were able to have a better understanding of the raw data thanks to these statistical explanations.

Keep in mind that the creation of statistics requires both the analysis of data and the performance of computations. Tables, charts, and graphs are the most common formats for conveying statistical information, though this is not always the case. In this particular instance, the graph was utilized by the writer from the Star Tribune to illustrate the typical quantity of passengers who boarded the LRT at each of the several stations located along the Green Line throughout the course of 2014. The raw data provided by MTC, which counted the number of riders who boarded at each station on each day, were used to calculate an average of the number of riders who boarded at each station during each month.

Self-Assessment Exercises 1

1. Define Data
2. Explain the concept of raw data in statistics

2.3.1 Types of Data

The collection and interpretation of data is the focus of the field of statistics. The values of one or more variables are measured in order to get data on those variables. Data can be classified as either quantitative data or qualitative data.

i. Quantitative data:

Quantitative data are measurements that are recorded on a numerical scale that occurs naturally in the world. The amount of time you have to wait for the next bus is an example of quantitative data. Your height and weight are also examples of quantitative data. Quantitative data are real numbers. They are not arbitrary numerical values that have been assigned in order to represent qualitative data. In order to collect qualitative data, an experiment will invariably seek out verbal or other non-numerical responses (eg, yes and no; defective and non-defective; Catholic, Protestant and other).

ii. Qualitative data:

Qualitative data are non-numeric. The political party that you support is one example of qualitative data. Other examples include: Your gender. Sometimes qualitative data are given

arbitrary numerical values, such as a 1 for male and a 2 for female. In this case, the values for gender are as follows:
male = 1, female = 2.

When presenting data, the most effective graphic method to employ is one that takes into account the particular kind of data that is being considered. When the topic of statistical inference is discussed later in the guide, the data type will play an important role in determining which statistical approach is most suitable for the particular problem that has to be solved.

Self-Assessment Exercises 2

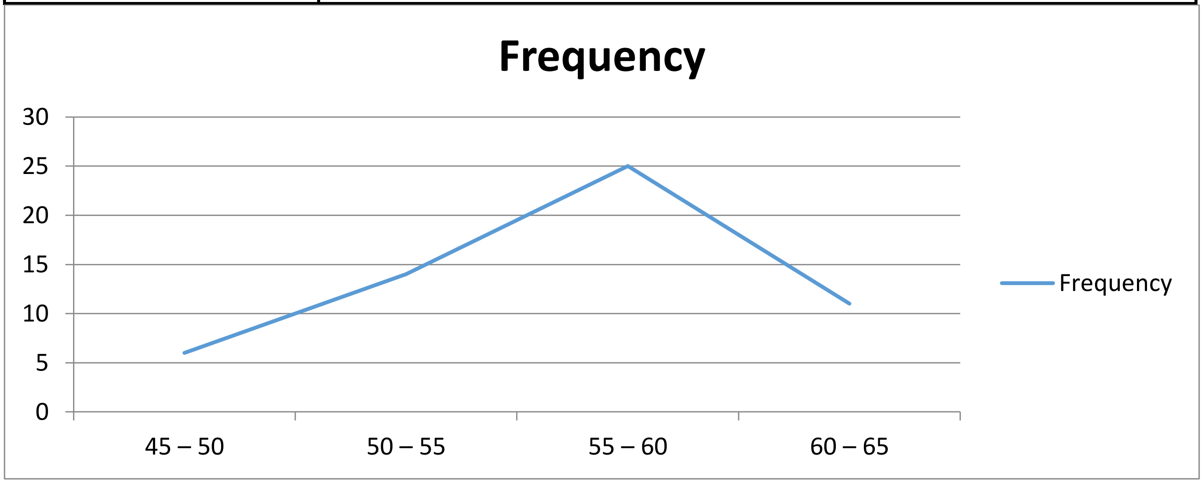
1. Provide an example of quantitative data.
2. How do you identify qualitative data?

2.5 Data Classification

When only certain values occur, such as the number of students in an interval, numerical data is said to be discrete. On the other hand, numerical data can also be categorized as continuous (when you can have intermediate or fractional values – like height or distance). Sometimes continuous data are summarized in tables, with each interval's summary containing the number of data items in that interval.

Example 1

Mass (kg)	Frequency
45 – 50	6
50 – 55	14
55 – 60	25
60 – 65	11



Self-Assessment Exercise 3

1. Each of the following examples of data, determine whether the data type is quantitative or qualitative:

a. The weekly level of the prime interest rate during the past year.

b. The make of car driven by each of a sample of executives.

c. The number of contacts made by each of a company's salespeople during a week.

d. The rating (excellent, good, fair or poor) given to a particular television programme by each of a sample of viewers.

e. The number of shares traded on the New York Stock Exchange each week throughout 2012.

2.6 Summary

This unit defined data and concept of raw data. Data" refers to discrete pieces of factual information that are recorded and utilized for the purpose of analysis. The end product of data analysis is statistical information, which includes both the interpretation and presentation of the data. The unit also defined raw data as a data that has not been processed for use. A distinction is sometimes made between data and information to the effect that information is the end product of data processing

2.7 References/Further Readings/Web Resources

Management College of Southern Africa (MANCOSA), (2016). Quantitative Method in Management. Retrieved from file:///C:/Users/HP/Downloads /Master%20of%n 20Business% 20Administration % 20 Quantitative% 20Methods% 20(%20PDFDrive%20).pdf

1.8 Possible Answers to SAEs

Answers to SAEs 1

1. *Data*" refers to discrete pieces of factual information that are recorded and utilized for the purpose of analysis? The statistics that are compiled come from the raw information that is provided. The end product of data analysis is statistical information, which includes both the interpretation and presentation of the data.
2. *What is raw data*; raw data (sometimes called source data, atomic data or primary data) is data that has not been processed for use.

Answers to SAEs 2

1. The amount of time you have to wait for the next bus is an example of quantitative data. Your height and weight are also examples of quantitative data. Qualitative data are non-numeric. The political party that you support is one example of qualitative data. Other examples include: Your gender.

2. Sometimes qualitative data are given arbitrary numerical values, such as a 1 for male and a 2 for female. In this case, the values for gender are as follows: male = 1, female = 2.

Answers to SAEs 3

1. Quantitative, if the interest rate level is expressed as a percentage. If the level is simply observed as being high, moderate or low, then the data type is qualitative.
2. a. Qualitative. b. Quantitative. c. Qualitative. d. Quantitative. e. Quantitative.

Unit 3 Graphical Technique of Quantitative Methods

Unit Structure

3.1 Introduction

Graphical techniques for quantitative data
This section introduces basic descriptive statistics methods used for organising a set of numerical data in tabular form and presenting it graphically. The presentation of the grouped data enables s the user to quickly grasp the general shape of the distribution of the data.

3.2 Learning Outcomes

By the end of this unit, you will be able to:

- draw and analyze frequency table and graph
- draw and analyze Histogram table and graph
- explain the concept of Frequency Polygon and the diagram

3.3 Graphical techniques for quantitative Method

Quantitative data is information about quantities; that is, information that can be measured and written down with numbers. Some other aspects to consider about quantitative data:

Focuses on numbers
Can be displayed through graphs, charts, tables, and maps
Data can be displayed over time (such as a line chart)

Qualitative data is information about qualities; information that can't actually be measured. Some other aspects to consider about qualitative data:

- Represented through pictures that explore the data in a visual way
- Visual representations focus on the themes found in the data
- Can tell a story
- Can also be displayed graphically as a pie chart or bar graph, the same as quantitative data, however, this can be tricky and can be done incorrectly easily

a. Histogram

A diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

b. **Graphic Timelines**

A timeline is a type of chart which visually shows a series of events in chronological order over a linear timescale. The power of a timeline is that it is graphical, which makes it easy to understand critical milestones, such as the progress of a project schedule.

c. **Pie Chart**

A type of graph in which a circle is divided into sectors that each represent a proportion of the whole.

d. **Scatter Plot**

A scatter plot is a graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.

Self-Assessment Exercises 1

1. Explain the form of data used in graphic
2. Quantitative data is information about quantities. Discuss
3. Define the following terms:
   a. Histogram
   b. Graphic Timelines
   c. Pie Chart
   d. Scatter Plot

3.3.1 Frequency Distribution

A frequency distribution is a table with data summarised into groups known as intervals.
The steps in creating a frequency distribution are given on page 35 of the prescribed textbook.

Example 1
The weights in pounds of a group of workers are:

173	165	171	175	188
183	177	160	151	169
162	179	145	171	175
168	158	186	182	162
154	180	164	166	157

$Range = maximum\ data\ value - minimum\ data\ value = 188 - 145 = 43$

Step 1. Determine the data range.
Step 2. Choose the number of intervals.
Choose five intervals for a small sample size.

$$Interval\ width = \frac{data\ range}{number\ of\ intervals} = \frac{43}{5} = 8{,}6$$

Step 3. Determine the interval width.
Using this calculation as a guide, grouping the data into intervals of width 10 pounds makes practical sense.
Step 4. Set up the interval limits.

Lower limit (weight in lbs)	Upper limit
140	< 150
150	< 160
160	< 170
170	< 180
180	< 190

Note: the upper limit is defined as ‘up to but not including’. Alternatively for discrete data, ie where no fractional values are encountered, the upper limits can be defined as 149, 159, 169 etc.

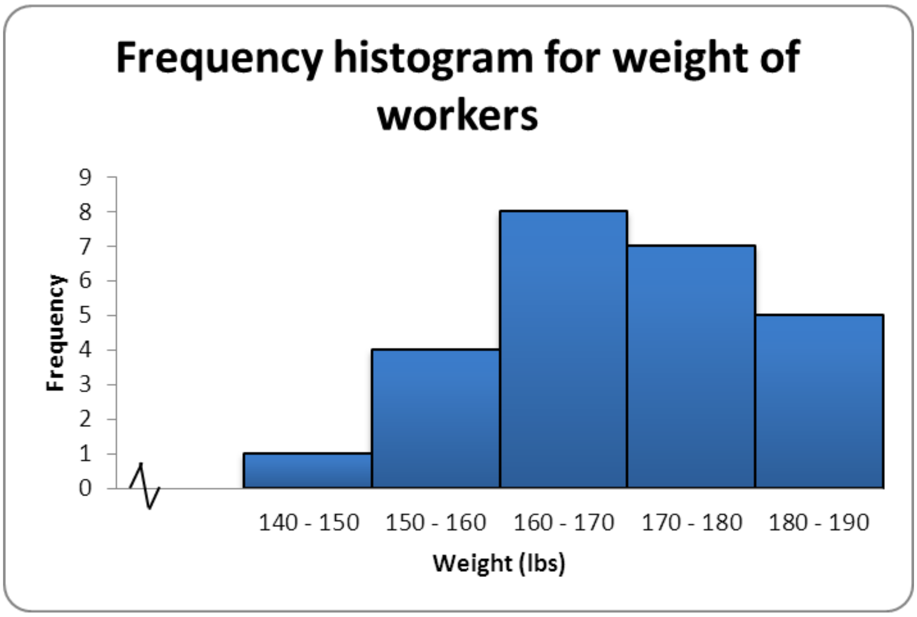
Decide on the upper limit approach you want to use (< 150 or 149) and then be consistent with your approach whenever you need to design a frequency table. This study guide and the textbook have standardised on the ‘less than but not including’, i.e. < 150, approach.

Self-Assessment Exercises 2

1. Use the above table to calculate the following:
   a. Range
   b. Interval width

3.3.2 Histogram

A frequency distribution can be graphically depicted as a histogram. Using the data from the example in section 3.1, the histogram can be depicted as:
Remember that histograms are similar to bar charts, but the bars touch each other. If you graph data and part of an axis is not to scale (in example the x-axis from 0 to 140 is not to scale), show a 'broken' axis.



3.4 Frequency Polygon and the diagram

A frequency polygon is constructed by plotting the frequency of each interval above the midpoint of that interval and then joining the points with straight lines. The polygon is closed by considering one additional interval (with zero frequency) at each end of the distribution and extending a straight line to the midpoint of each of these intervals.
Before constructing a frequency polygon, calculate the midpoints for each interval.

Using the data from the example in section 1.3.1, the midpoints are calculated as:

Interval (weight in lbs)	Frequency	Midpoint
140 – 150	1	145

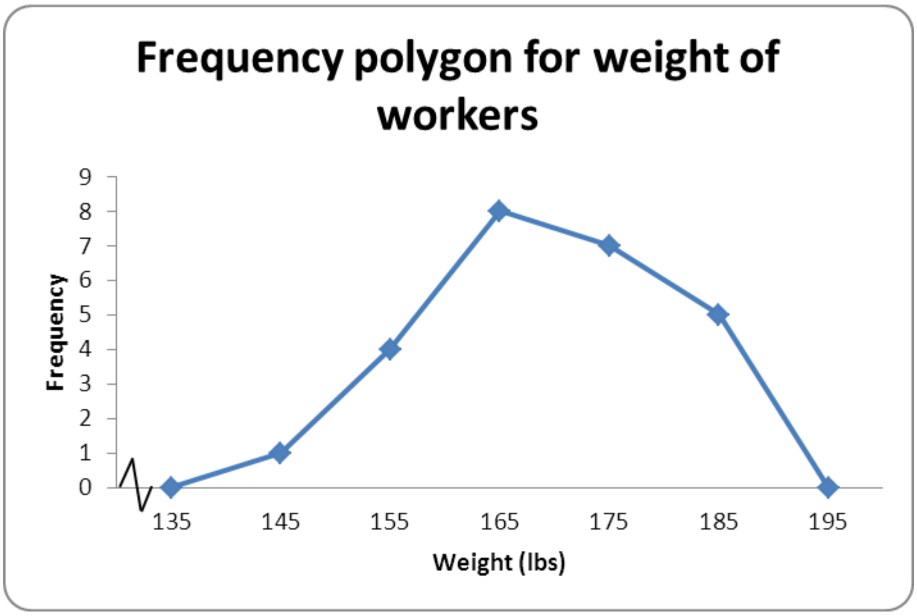
150 – 160	4	155
160 – 170	8	165
170 – 180	7	175
180 – 190	5	185

An easy way to calculate midpoint of an interval is to halve sum of the lower and upper limits. In the case of example, for the first interval:

$$\frac{\text{Lower limit + Upper Limit}}{2} \quad \frac{140+150}{2} = 145$$

The frequency polygon can be depicted as:

$$\frac{\textit{lower limit} + \textit{upper limit}}{2} = \frac{140+150}{2} = 145$$



Self-Assessment Exercise 3

1	Define a frequency polygon

3.5 Summary

This unit defined as Quantitative data as information about quantities; that is, information that can be measured and written down with numbers. Quantitative data can be depicted in: Histogram

Graphic Timelines

A timeline is a type of chart which visually shows a series of events in chronological order over a linear timescale

Pie Chart
A type of graph in which a circle is divided into sectors that each represent a proportion of the whole.

Scatter Plot
A scatter plot is a graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.

3.6 References/Further Readings/Web Resources

Gill, J. and Meier, K.J. (2000). "Public Administration Research and Practice: A Methodological Manifesto." Journal of Public Administration Research and Theory 10(1): 157-199.

Perry, J.L. (2012). "How Can We Improve Our Science to Generate More Usable Knowledge for Public Professionals?" Public Administration Review 72: 479–482.

Raadschelders, J. (2011). Public administration: the interdisciplinary study of government. Oxford: Oxford University Press.

1.7 Possible Answers to SAEs

Answers to SAEs 1

1. Quantitative data is information about quantities; that is, information that can be measured and written down with numbers. Some other aspects to consider about quantitative data: Focuses on numbers.

- Can be displayed through graphs, charts, tables, and maps
- Data can be displayed over time (such as a line chart) Qualitative data is information about qualities; information that can't actually be measured. Some other aspects to consider about qualitative data:
- Represented through pictures that explore the data in a visual way
- Visual representations focus on the themes found in the data
- Can tell a story
- Can also be displayed graphically as a pie chart or bar graph, the same as quantitative data, however, this can be tricky and can be done incorrectly easily
- Can tell a story

- Can also be displayed graphically as a pie chart or bar graph, the same as quantitative data, however, this can be tricky and can be done incorrectly easily

Histogram

A diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

Graphic Timelines

A timeline is a type of chart which visually shows a series of events in chronological order over a linear timescale. The power of a timeline is that it is graphical, which makes it easy to understand critical milestones, such as the progress of a project schedule.

Pie Chart

A type of graph in which a circle is divided into sectors that each represent a proportion of the whole.

Scatter Plot

Answers to SAEs 2

1. The weights in pounds of a group of workers are:

173	165	171	175	188
183	177	160	151	169
162	179	145	171	175
168	158	186	182	162
154	180	164	166	157

$Range = maximum\ data\ value - minimum\ data\ value = 188 - 145 = 43$

Step 1. Determine the data range.

Step 2. Choose the number of intervals.

Choose five intervals for a small sample size.

$$Interval\ width = \frac{data\ range}{number\ of\ intervals} = \frac{43}{5} = 8{,}6$$

Answers to SAEs 3

1. A frequency polygon is constructed by plotting the frequency of each interval above the midpoint of that interval and then joining the points with straight lines. The polygon is closed by considering one additional interval (with zero frequency) at each end of the distribution and extending a straight line to the midpoint of each of these intervals.

Unit 4 Descriptive Statistics

Unit Structure

4.1 Introduction

In our last class, we deal extensively with the problem of Quantitative data, Histogram, graphic timelines**,** Pie Chart and Scatter Plot. In today's class, we are going to be discussing Descriptive Statistics and the component of bar chart.

4.2 Learning Outcomes

- define Descriptive Statistics
- define and explain the Cumulative frequency distribution
- define and show an Ogive curve
- explain and show the Relative distributions
- draw and explain the Pie charts, bar charts and line charts
- explain the concept of Bar charts
- draw and explain Simple bar chart
- define and explain a component or stacked bar chart
- explain and show multiple bar chart
- define and show the Scatter diagrams

4.3 Descriptive Statistics

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

4.3.1 Cumulative frequency distribution

A cumulative frequency distribution summarises the cumulative frequency of a dataset. It results in a 'running total' of frequencies.

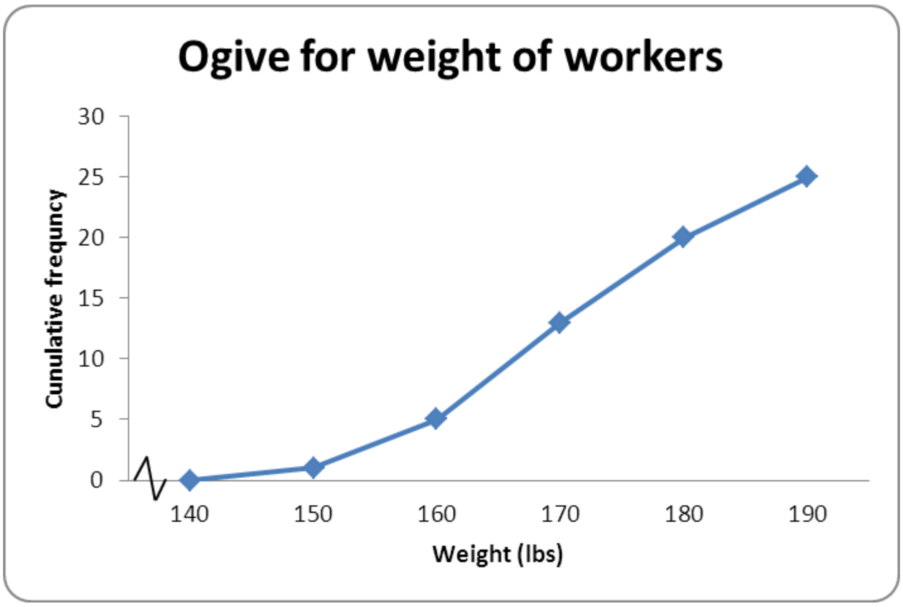
Example:
Using the data from the example in section 1.3.1:
For each interval, calculate the cumulative frequency by adding the frequency count of the interval in question to the cumulative frequency of the interval before.

Interval (weight in lbs)	Frequency	Cumulative frequency
140 – 150	1	1
150 – 160	4	5
160 – 170	8	13
170 – 180	7	20
180 – 190	5	25

4.3.2 An Ogive

An ogive is a graph of the cumulative frequency distribution.
To construct the ogive, the cumulative relative frequency of each interval is plotted above the upper limit of that interval and the points representing the cumulative frequencies are then joined by straight lines. The ogive is closed at the lower end by extending a straight line to the lower limit of the first interval.

Example:
Using the data from the example in section 1.3.1:



Self-Assessment Exercises

1. Define the cumulative frequency distribution
2. Explain the ogive in the cumulative frequency distribution

4.3.3 Relative distributions

For each of the frequency distribution and the cumulative frequency distribution, relative distributions can be calculated. A relative frequency distribution includes the percentage of sample size or relative frequency (frequency relative to the total sample size) for each interval.

Example:
Using the data from the example in section 1.3.1:

Interval (weight in lbs)	Frequency	Relative frequency (factor)	Relative frequency (percentage)
140 – 150	1	0.04	4%
150 – 160	4	0.16	16%
160 – 170	8	0.32	32%
170 – 180	7	0.28	28%
180 – 190	5	0.20	20%

A relative cumulative frequency distribution includes the cumulative percentage of sample size or relative cumulative frequency (cumulative frequency relative to the total sample size) for each interval.

Example:
Using the data from the example in section 1.3.1:

Interval (weight in lbs)	Frequency	Cumulative frequency	Relative frequency (factor)	Relative frequency (percentage)
140 – 150	1	1	0,04	4%
150 – 160	4	5	0,20	20%
160 – 170	8	13	0,52	52%
170 – 180	7	20	0,80	80%
180 – 190	5	25	1,00	100%

4.3.4 Pie charts, bar charts and line charts

The methods described in the previous section are appropriate for summarising quantitative data. But we should also be able to describe data that are qualitative or categorical. These data consist of attributes or names of the categories into which the observations are sorted.

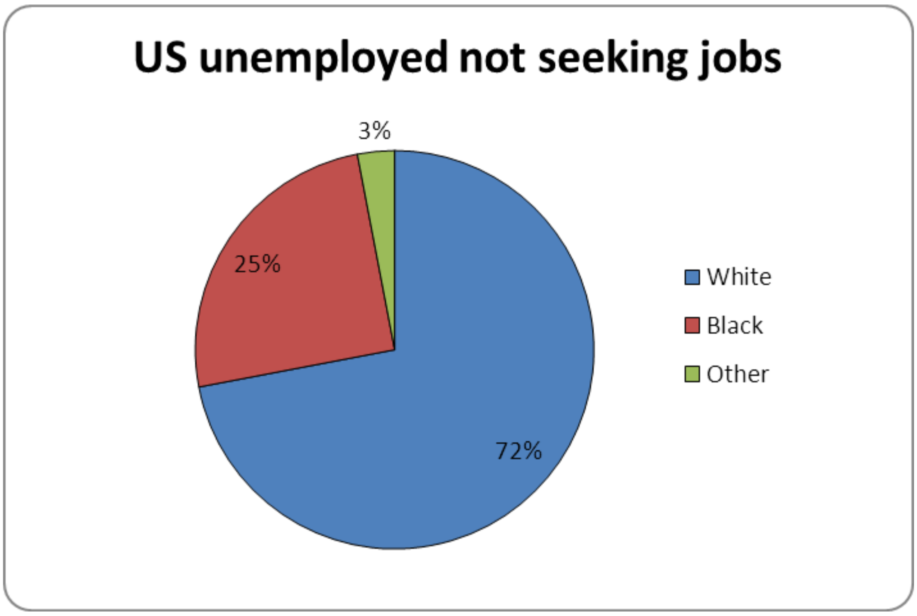
A pie chart is a useful method for displaying the percentage of observations that fall into each cate gory of qualitative data.
A pie chart is an effective method of showing the percentage breakdown of a whole entity.

Example:
A New York Times article reports that "6 million Americans who say they want work are not even seeking jobs". These 6 million Americans are broken down by race:

Race	Frequency
White	4, 320 000
Black	1, 500 000
Other	180, 000
Total	6,000,000

We need to first determine the percentage of the 6 million Americans belonging to each of the three racial categories: 72% white, 25% black and 3% other. Each category is represented by a slice of the pie (a circle) that is proportional in size to the percentage (or relative frequency) corresponding to that category.

Since the entire circle corresponds to 360°, the angle between the lines demarcating the white sector is therefore $(0{,}72)(360) = 259{,}2^\circ$. In a similar manner, we can determine the angles for the black and other sectors as 90° and 10,8°, respectively.



4.4 Bar charts

Bar charts are a quick and easy way of showing variation in or between variables.

Rectangles of equal width are drawn so that the area enclosed by each rectangle is proportional to the size of the variable it represents. This type of graph not only illustrates a general trend, but also allows a quick and accurate comparison of one period with another or the illustration of a situation a particular time.

When drawing up bar charts take care to:

- Make the bars reasonably wide so that they can be clearly seen.
- Draw them neatly and professionally.
- Ensure that the bars all have the same width.
- Ensure that the gaps between the bars have the same width.
- We can produce a variety of bar charts to provide an overview of the data.

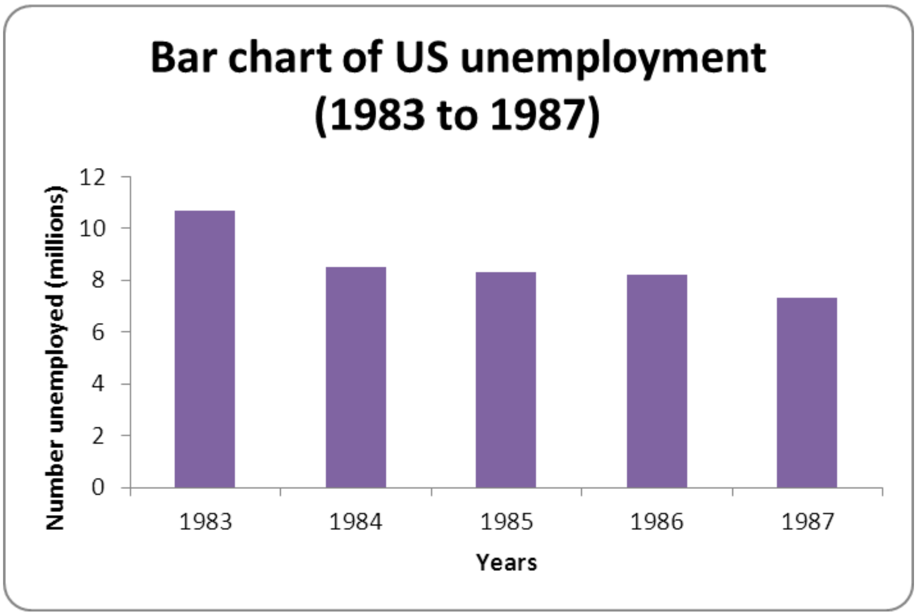
4.4.1 Simple bar chart

A simple bar chart comprises bars representing each variable drawn either vertically or horizontally. While a bar chart can be used to display the frequency of observations that fall into each category, if the categories consist of points in time and the objective is to focus on the trend in frequencies over time, a line chart is useful.

Example:
According to the Nigeria primetimes (27 September 1987), the June levels of unemployment in the Nigeria for five years are:

Year	Unemployed (millions)
1983	10.7
1984	8.5
1985	8.3
1986	8.2
1987	7.3

For the bar chart, the five years or categories are represented by intervals of equal width on the horizontal axis. The height of the vertical bar erected above any year is proportional to the frequency (number of unemployed) corresponding to that year.



A line chart is obtained by plotting the frequency of a category above the point on the horizontal axis representing that category and then joining the points with straight lines.

Self-Assessment Exercises 2

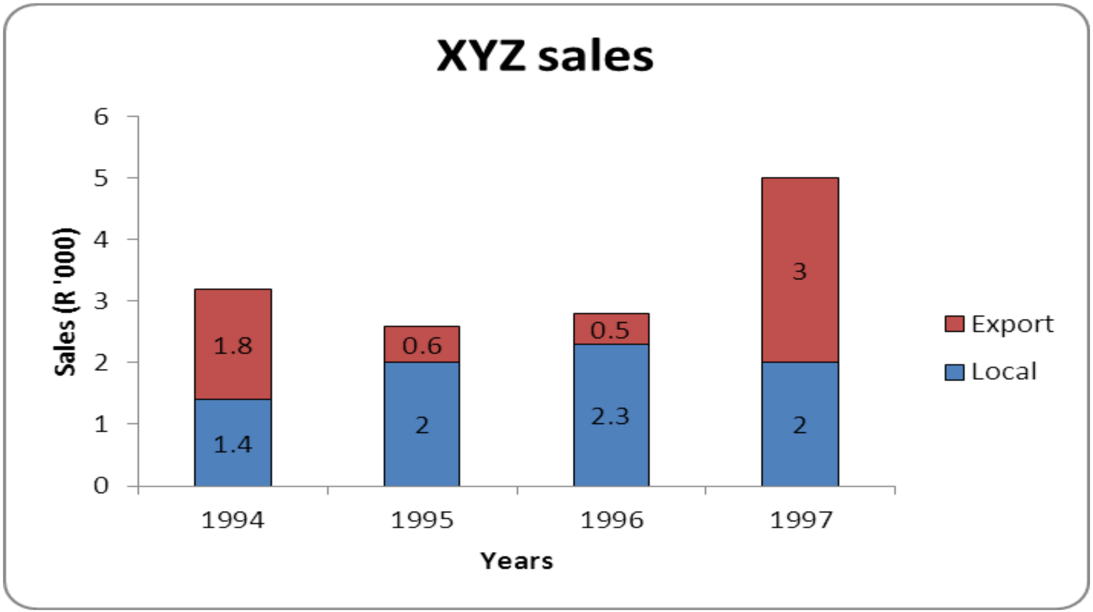
1. Define Bar Chart
2. Explain the concept of Simple Bar Chart

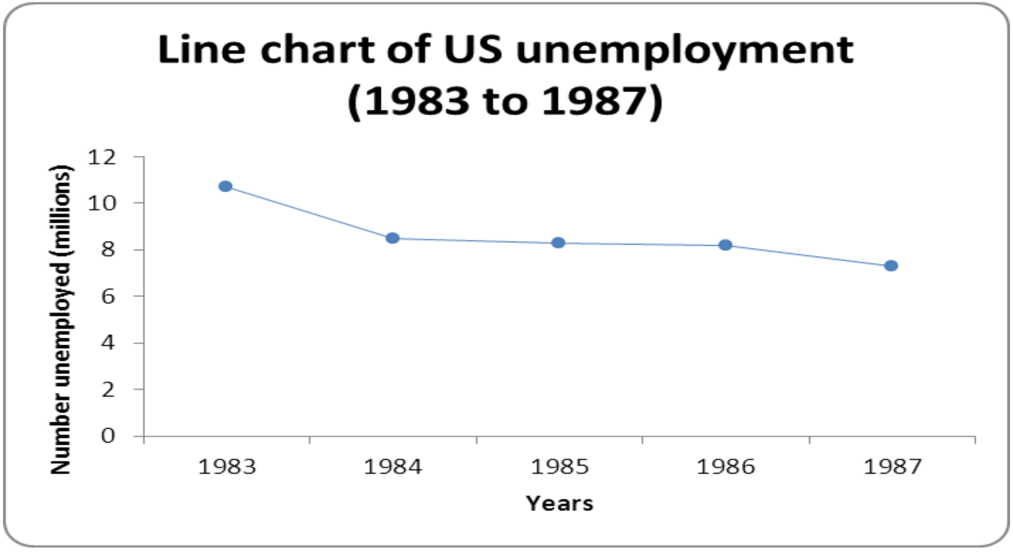
4.4.2 A component or stacked bar chart

In a component or stacked bar chart, a single bar is drawn for each variable, with the heights of the bars representing the totals of the categories. Each bar is then subdivided to show the components that make up the total bar. These components may be identified by colouring or shading, accompanied by an explanatory key to show what each component represents.

Years	XYZ Sales/Categories	
	Local	Export
1994	1.4	1.8
1995	2.0	0.6
1996	2.3	0.5
1997	2.0	3.0

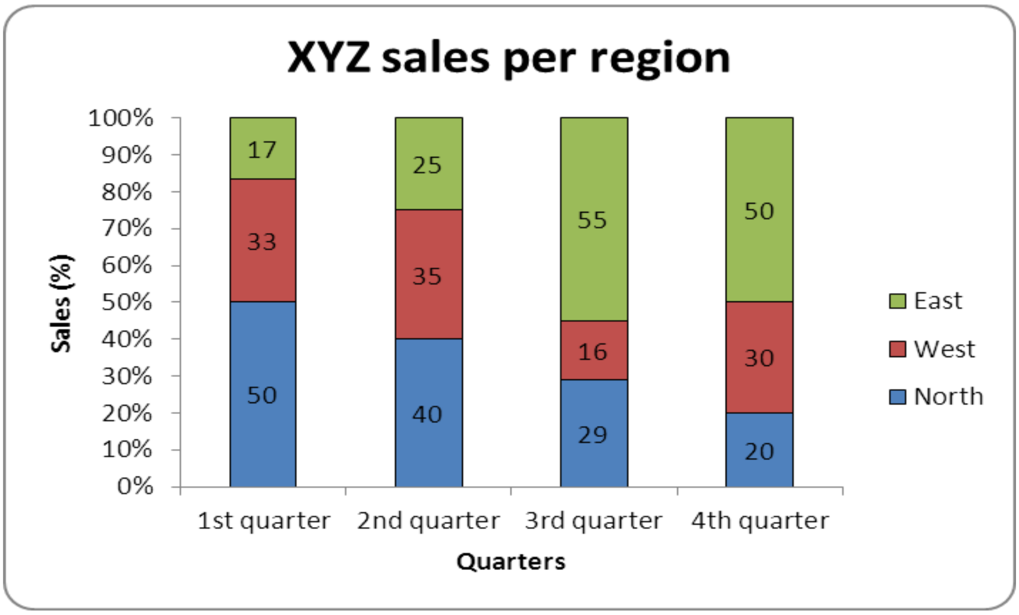
Percentage component bar chart





A percentage component comprises components converted to percentages of the total with the bars divided in proportion to these percentages. The scale is a percentage scale and the height of each bar is therefore 100%.

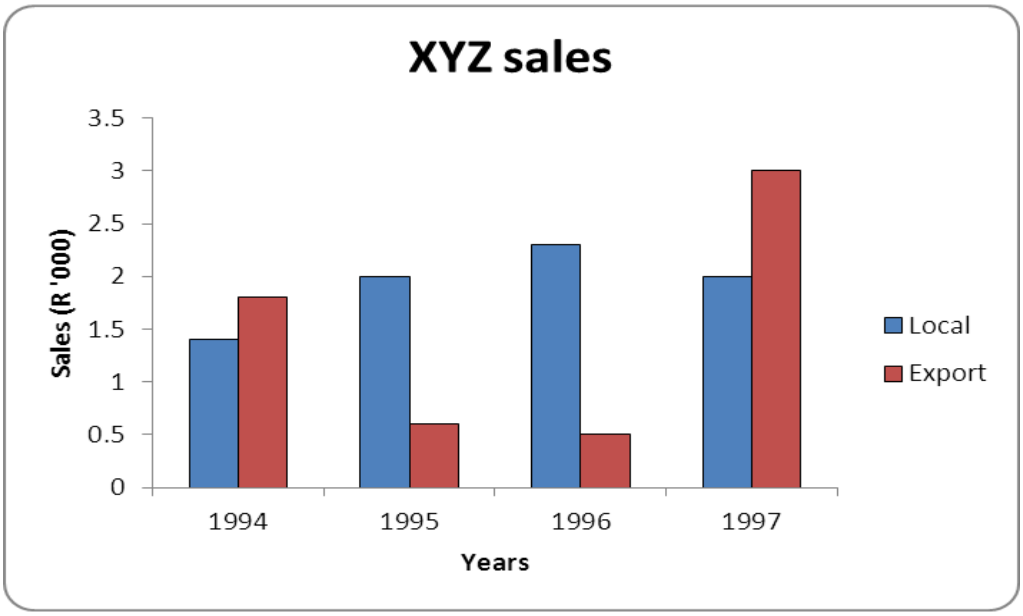
Quarters Sales/Region	North	West	East	Total
1st Quarter	50	33	17	100
2nd Quarter	40	35	25	100
3rd Quarter	29	16	55	100
4th Quarter	20	30	50	100



4.4.3 Multiple Bar Chart

For multiple or cluster bar charts, two or more bars are grouped together in each category. The use of a key helps to distinguish between the categories.

Years	XYZ Sales/Categories	
	Local	Export
1994	1.4	1.8
1995	2.0	0.5
1996	2.3	0.4
1997	1.8	3.0



4.5 Scatter diagrams

The relationship between two quantitative variables can be depicted in a scatter diagram. Economists, for example, are interested in the relationship between inflation rates and unemployment rates. Business owners are interested in many variables, including the relationship between their advertising expenditures and sales levels.
A scatter diagram is a plot of all pairs of values (x, y) for the variables x and y.

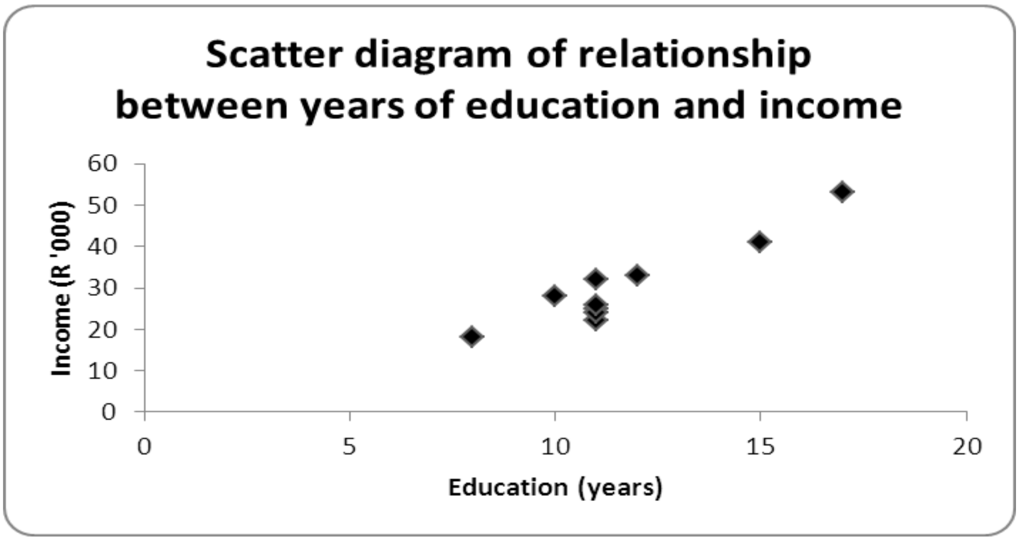
Example:
An educational economist wants to establish the relationship between an individual's income and education. She takes a random sample of 10 individuals and asks for their income (in N´000s) and education (in years).

x (years of education)	y (income in N´000)
11	25
12	33
11	22
15	41
8	18
10	28
11	32
11	24
17	53
11	26

If we feel the value of one variable (such as income) depends to some degree on the value of the other variable (such as years of education), the first variable (income) is called the dependent variable and is plotted on the vertical or y-axis. The second variable is the independent variable and is plotted on the x-axis. Think of the independent variable (x-axis) as the 'cause' and the dependent variable (y-axis) as the 'effect'.

The scatter diagram allows us to observe two characteristics about the relationship between education and income (y): Because these two variables move together, ie their values tend to increase together and decrease together; there is a positive relationship between the two variables. The relationship between income and years of education appears to be linear, since we can imagine drawing a straight line (as opposed to a curved line) through the scatter diagram that approximates the positive relationship between the two variables. The pattern of a scatter diagram provides us with information about the relationship between two variables. Figure 1 depicts a positive linear relationship.

If two variables move in opposite directions and the scatter diagram consists of points that appear to cluster around a straight line, then the variables have a negative linear relationship.



Self-Assessment Exercises 3

1. What is a component or stacked bar chart
2. Explain the concept of Multiple bar chart

4.6 Summary

The unit discussed the descriptive statistics to includes; A cumulative frequency distribution defined as summarises the cumulative frequency of a dataset. It results in a 'running total' of frequencies.
An ogive is a graph of the cumulative frequency distribution and to construct the ogive, the cumulative relative frequency of each interval is p lotted above the upper limit of that interval and the points representing the cumulative frequencies are then joined by straight lines.

For each of the frequency distribution and the cumulative frequency distribution, relative distributions can be calculated. A relative frequency distribution includes the percentage of sample size or relative frequency (frequency relative to the total sample size) for each interval
Bar charts are a quick and easy way of showing variation in or between variables.

A simple bar chart comprises bars representing each variable drawn either vertically or horizontally.

In a component or stacked bar chart, a single bar is drawn for each variable, with the heights of the bars representing the totals of the categories. Each bar is then subdivided to show the components that make up the total bar.

For multiple or cluster bar charts, two or more bars are grouped together in each category. The use of a key helps to distinguish between the categories.

4.7 References/Further Readings/Web Resources

Haessuler, E. F. and Paul, R. S. (1976). *Introductory Mathematical Analysis for Students of Business and Economics,* (2nd edition.) Reston Virginia: Reston Publishing Company.

4.8 Possible Answers to SAEs

Answers to SAEs 1

1. A cumulative frequency distribution defined as summarises the cumulative frequency of a dataset. It results in a ‘running total’ of frequencies
2. An ogive is a graph of the cumulative frequency distribution and to construct the ogive, the cumulative relative frequency of each interval is plotted above the upper limit of that interval and the points representing the cumulative frequencies are then joined by straight lines.

Answers to SAEs 2

1. Bar charts are a quick and easy way of showing variation in or between variables
2. A simple bar chart comprises bars representing each variable drawn either vertically or horizontally

Answers to SAEs 3

1. In a component or stacked bar chart, a single bar is drawn for each variable, with the heights of the bars representing the totals of the categories. Each bar is then subdivided to show the components that make up the total bar
2. For multiple or cluster bar charts, two or more bars are grouped together in each category. The use of a key helps to distinguish between the categories.

Unit 5 Measure of Central Tendency

Unit Structure

1.1 Introduction

One common example of question to be asked in Measure of Central Tendency is, "How many calories should I be eating each day?" how much time do I spend chatting on a daily basis?" Answering this question is difficult because the answer changes daily. In other cases, the question "how many calories do I eat on an average day?" is more appropriate. or "how much time do I spend chatting on a daily basis?" There are three techniques to measure central tendency in data, the mean, the median, and the mode, which we shall cover in this section. Is there a single value for each measurement? What you've described may be regarded the norm.

5.2 Learning Outcomes

By the end of this unit, you will be able to:

- discuss the Measure of Central Tendency
- state the types of Measure of Central Tendency
- explain Grouped data
- analyse the Mean for grouped data
- calculate the mode for a Grouped Data

5.3 The Measure of Central Tendency

For any variable that is being looked at, each member of the population has a certain value for that variable. These numbers are called data, and we can use our measures of central tendency on the whole population to get a single value (or more than one for the mode) that shows the central tendency for the whole population, or we can use our measures on a subset or sample of the population to get an estimate of the central tendency for the whole population.

Usually when two or more different data sets are to be compared it is necessary to condense the data, but for comparison the condensation of data set into a frequency distribution and visual presentation are not enough. It is then necessary to summarize the data set in a single value. Such a value usually somewhere in the center and represent the entire data set and hence it is called measure of central tendency or averages. Since a measure of central tendency (i.e. an average) indicates the location or the general position of the distribution on the X-axis therefore it is also known as a measure of location or position.

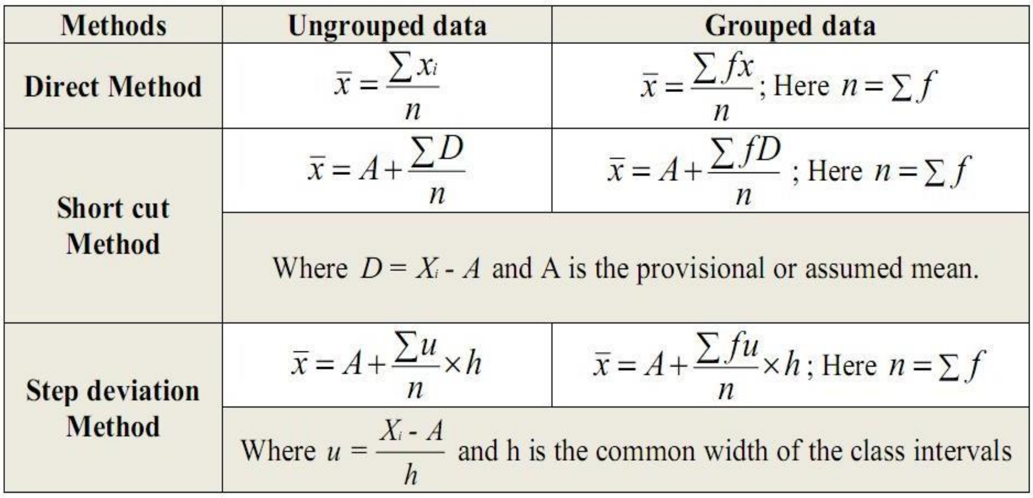
5.3.1 Types of Measure of Central Tendency

1. Arithmetic Mean
2. Median
3. Mode

1. **Arithmetic Mean or Simply Mean:** "A value obtained by dividing the sum of all the observations by the number of observations is called arithmetic Mean".

   The arithmetic mean of a dataset is obtained by adding each value in the dataset and dividing the total by the number of variables in the dataset. It is referred to simply as the mean

Mean □ Sum of All observationNumber of observation



Methods	Ungrouped data	Grouped data
Direct Method	$\bar{x} = \frac{\sum x_i}{n}$	$\bar{x} = \frac{\sum fx}{n}$; Here $n = \sum f$
Short cut Method	$\bar{x} = A + \frac{\sum D}{n}$	$\bar{x} = A + \frac{\sum fD}{n}$; Here $n = \sum f$
	Where $D = X_i - A$ and A is the provisional or assumed mean.	
Step deviation Method	$\bar{x} = A + \frac{\sum u}{n} \times h$	$\bar{x} = A + \frac{\sum fu}{n} \times h$; Here $n = \sum f$
	Where $u = \frac{X_i - A}{h}$ and h is the common width of the class intervals	

Where,
∑ denotes summation of a set of values
x is the variable used to represent raw scores
n represents the number of scores being considered

The result can be denoted by x for the mean of a sample from a larger population

The computed mean of all values of a population is denoted by the Greek letter µ (pronounced mu)

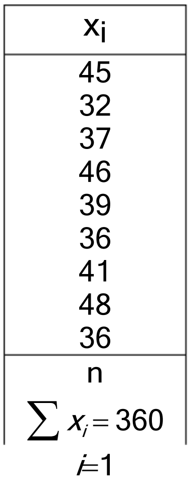
Example
Find the mean of the dataset:

2	3	6	7	12

The mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{2 + 3 + 6 + 7 + 12}{5} = 6$$

Calculate the arithmetic mean for the following the marks obtained by 9 students are given below:
Using formula of arithmetic mean for ungrouped data: Calculate the arithmetic mean for the following data given below:
Numerical Example: X= ∑FX/n 360/9
X = 40 Marks



x_i
45 32 37 46 39 36 41 48 36
$\sum_{i=1}^{n} x_i = 360$

Self-Assessment Exercises 1

Use the information below to calculate the arithmetic mean score of the given data

X	F	FX	CF
1	3	3	3
2	2	4	7
3	5	15	22
4	6	24	46
5	4	20	66
15	20	66	

X= ∑FX/n 66/5
X = 13.5

Use the data below to Calculate the Weighted Average Mean Score of the given student examination

Exam	Weight (wi)	Score (xi)	Wi(xi)
Quiz	5	60	300
Midterm	4	50	200
Semester Exam	3	45	150

∑wi/xi = 650/155
X = 4.19

2. Median

The median is the middle value of an ordered set of numbers.
Note: It is important that the values are in sequential order before you choose the middle value.
Definition: Median:

The median of a dataset is the middle value when the values are arranged in order of increasing (or decreasing) magnitude.
After first arranging the original values in increasing (or decreasing) order, the median will be either of the following:
If the number of values is odd, the median is the number that is exactly in the middle of the list.
If the number of values is even, the median is found by computing the mean of the two middle numbers.

Example
Over a seven-day period, the number of customers (per day) purchasing at Hides Leather Shop is:

4	80	50	10	60	12	5

Array – arranged in order of increasing magnitude:

4	5	10	12	50	60	80

$$median = 12$$

The number of values is odd; therefore, the median is the middle number of the list:

Example
Over an eight-day period, the number of customers observed at the shop per day is:

21	5	11	7	12	15	20	5

Array – arranged in order of increasing magnitude:

5	5	7	11	12	15	20	21

The number of values is even, therefore the median is the mean of the middle two numbers in the list.

$$median = \frac{11 + 12}{2} = 11{,}5$$

Self-Assessment Exercises 2

Consider the following raw data to determine the media Score: 90 80, 85, 88, 89
Consider the following raw data to determine the media Score: Raw data (in =N=) X1 = 950; X2 = 300; X3 = 1000; X4 = 950; X5 = 850 and X6 = 750

3. Mode

The mode of a dataset is the value that occurs most frequently. Where no score is repeated there is no mode. Where two scores occur with the same highest frequency, the dataset is bimodal. If more than two scores occur with the same highest frequency, each is a mode and the dataset is multimodal.

The mode is the most common value. If we look at the set of numbers:

3	4	5	6	6	6	7

The mode is 6 because it is the number that appears most often.

Example
The commission earnings of five salespeople are:

R5 000	R5 200	R5 200	R5 700	R8 600

The modal commission is R5 200.
The lengths of stay (in days) for a sample of nine patients in a hospital are:

17	19	19	4	19	26	4	21	4

The dataset is bimodal with two modes, 19 and 4 days.

Example
There are 40 buck, 25 elephant and 20 smaller animals at a water hole. The modal category is buck since it has the highest frequency.

Example
The hourly income rates (in $) of five students are:

4	9	7	16	10

There is no mode.

5.3.4 Grouped data

Once data is grouped into intervals, the original or raw data is no longer of relevance or may not be known and the frequency distribution data needs to be used for measuring central location.

Formulae can be presented in different ways. In this text we have wherever possible, used the formulae from the textbook.

Remember if a lecturer uses a formula that looks slightly different, it is up to you as a masters' level student to check that it is still the same formula.

5.4 Mean for grouped data

Because the original or raw data is no longer available or of relevance, each dataset observation is assumed to take on the value of the midpoint of its interval. In order to calculate the mean, the total of all values (i.e. midpoint values) is used.

Formula: Mean for grouped data:
Arithmetic Mean for a Grouped Data

When a set of observations is presented in a grouped form or in class limits or boundaries, the computation of the arithmetic mean is done according to the following definitions:
Where Σfx = n
Xg = Grouped mean of the set of data on variable x.
f = frequency of observations
x = mid-value of each class limit defined by
X = Lower Class Limit + Upper Class Limit for each class limit/2.

Example
The following is a data on the daily wages paid to workers in a given factory:

Daily Wages (in N)	F
200 – 300	15
301 – 401	30
402 – 502	45
503 – 603	60
Total	150

Compute the average daily wages.
The following is a data on the daily wages paid to workers in a given factory:

Daily Wages (in N)	F	X	FX
200 – 300	15	250	3750
301 – 401	30	351	10530
402 – 502	45	452	20340
503 – 603	60	553	33180
Total	150	150	= 67800

Solution

Where,

f is the frequency

x is the midpoint of the interval

n is the number of observations in the dataset

Steps in calculating the mean of grouped data:

Step 1. Extend the frequency distribution to add the further columns needed.

Step 2. Calculate the midpoint of each interval in the frequency distribution and include in a new column.

$$midpoint = \frac{lower\ limit + upper\ limit}{2}$$

Each observation is then 'allocated' the midpoint as its value.

Σfx/n = 67800/4. = 16,950

The Median

The average of a given set of data can also be estimated using the median. The median is a measure of central tendency which appears in the "middle" of an ordered sequence of values or observations. That is, half of the observations in a set of data are lower than the median value and half are greater than it. To compute the median from a set of raw data, we must first array the data. If we have odd number of observations, the median is represented by the numerical value of the (n+1)th /2 arrayed observation. On the other hand, if the number of observations in the sample is an even number, the median is represented by the mean or average of the two middle values in the ordered array**.**

Raw data (in =N=) X1 = 950; X2 = 300; X3 = 1000; X4 = 950; X5 = 850 and X6 = 750

Solution

First, re-arrange in order Array

Xi = 300, 750, 850, 950, 950, 1000

Since this involves an "even" number of observations, the median will be the average of the two middle values: $\underline{850+950}$ = 900

Use the data below to calculate the media

S/N	Price (N)	No. of Days (f)
1	110-114	2
2	115-119	6
3	120-124	8
4	125-129	12
5	130-134	14

$$\text{Median Price} = L + \frac{(N/2 - f_b)C}{f_m}$$

Solution
L= Limit media class = 124.5
N= Total Frequency = 46
F_b = Frequency before the media class 6
F_m = Media class 8
C= class interval 5
Where L = 124.5; N = 42; fb = 6+2 = 8; C = 5; fm = 8.
124+(42/2-6)5/8
124+(21-6)5/8 = 124.5+(15)5/8_ = 124.5+ 75/8 = 124.5 + 9.4
X= 133.9
The Mode
The mode is a quick measure of central tendency or average. The mode is the most typical or most frequently observed value in a given set of data. It is the observation with the highest frequency in a given set.
For the set of data, X = 1,2,2,4, the value with the highest frequency is 2. Hence, the mode for the given data is 2.

5.5 Mode for a Grouped Data

The computational process for the mode of a grouped data is similar to that of the median. The process is as follows:

$$\text{Mode} = L + \frac{(fm - f_L)C}{(f_m - f_L) + (f_m - fh)}$$

L = Lower actual class limit of the model group
fm = frequency of the model group
fL = frequency of the group before the model group
fh = frequency of the group after the model group
C = class width for the actual class limit.

Consider the following data on the sales by 90 sales representatives

Sales (N'000s)	No. of Sales Reps.(f)
10 – 15	10
16 – 21	36
22 – 27	28
28 – 33	10
34 – 39	6
Total	

Calculate the modal value of sales.

Solution

Re-tabulating the data, we get:

Consider the following data on the sales by 90 sales representatives

Sales (N'000s)	No. of Sales Reps.(f)	Cumulative Frequency
10 – 15	10	10
16 – 21	36	46
22 – 27	28	74
28 – 33	10	84
34 – 39	6	90
Total	90	

Calculate the modal value of sales.

$$\text{Mode} = L + \frac{(fm - f_L)C}{(f_m - f_L) + (f_m - fh)}$$

$$= 15.5 + \frac{(36-10)(6)}{(36-10)+(36-28)}$$

$$= 15.5 + \frac{156}{34}$$

15.5+4.6 = 20.09

Thus, the modal value of sales is N (20.09x1000) = N20,090 (since values are in thousands).

Self-Assessment Exercises 3

Giving the following table on the share prices of a quoted company over a period of 60 days:

Use the data below to calculate the mode

S/N	Price (N)	No. of Days (f)
1	110-114	2
2	115-119	6
3	120-124	8
4	125-129	12
5	130-134	14

L= Limit model class = 119.5

N= Total Frequency = 52

F_L = Frequency before the media class

F_b = Frequency after the media class

F_m = model class

C= class interval

Where L = 119.5; N = 52; fl = 1+6 = 7; F_h = 4+1 =5, C = 5; fm = 8.

119.5+(8-7)5 = 119.5+(1)5 = 119.5+ 5 119.5+ 5/4 = 119.5+1.25

(8-7)+(8-5)   1+3   4
= 119.5 + 1.25 = 20.75

5.6 Summary

In a nutshell, this unit has exposed you to all aspects of measure of central tendency (i.e. an average) indicates the location or the general position of the distribution on the X-axis therefore it is also known as a measure of location or position. This course discuss the various Measure of Central Tendency such as Arithmetic Mean, Geometric Mean, Harmonic Mean, media and Mode.

5.7 References/Further Readings/Web Resources

Haessuler, E. F. and Paul, R. S. (1976). *Introductory Mathematical Analysis for Students of Business and Economics*, (2nd edition.) Reston Virginia: Reston Publishing Company

Self-Assessment Exercises 1

Use the information below to calculate the arithmetic mean score of the given data

X	F	FX	CF
1	3	3	3
2	2	4	7
3	5	15	22
4	6	24	46
5	4	20	66
15	20	66	

X= ∑FX/n 66/5
X = 13.5

5.8 Possible Answers to SAEs

Answers to SAEs 1

1. Consider the following raw data to determine the media Score: 90 80, 85, 88, 89

The media value for above data is 88. This is because, the 3rd ordered observation, which is 88, since this involves “odd” number of observations.

2 Consider the following raw data to determine the media Score:
Raw data (in =N=) X1 = 950; X2 = 300; X3 = 1000; X4 = 950; X5 = 850 and X6 = 750
Solution
First, re-arrange in order Array
Xi = 300, 750, 850, 950, 950, 1000
Since this involves an “even” number of observations, the median will be the average of the two middle values: $\frac{850+950}{2} = 900$

Answers to SAEs 2

Giving the following table on the share prices of a quoted company over a period of 60
L= Limit model class = 119.5
N= Total Frequency = 52
F_L = Frequency before the media class
F_b = Frequency after the media class
F_m = model class
C= class interval

Where L = 119.5; N = 52; fl = 1+6 = 7; F_h = 4+1 =5, C = 5; fm = 8.

$$119.5+\frac{(8-7)5}{(8-7)+(8-5)} = 119.5+\frac{(1)5}{1+3} = 119.5+\frac{5}{4} \quad 119.5+ 5/4 = 119.5+1.25$$

= 119.5 + 1.25 = 20.75

MODULE 2

Unit 1 Statistical Tools I

Unit Structure

1.1 Introduction

As noted above, for the purpose of this course, we discuss five major measures of dispersion in the behaviour of business and/or economic variables: the range, the mean deviation, the variance, the standard deviation, and the coefficient of variation. Another important related measure that is primarily concerned with distribution of some business data is the Pearson Coefficient of Skewness. You will be exposed to its computation, interpretations, and managerial implications. Dispersion (or spread) refers to the extent to which the data values of a numeric random variable are scattered about their central location value" Wegner (2012). In this unit, the concept of central location was introduced. The variability among data is one characteristic to which averages are not sensitive. It is possible to have two datasets with identical measures of central location but with wider spreads of data

1.2 Learning Outcomes

By the end of the unit, students should be able to:

- calculate the Range
- compute the Mean Deviation (MD)
- calculate the Variance
- compute the Standard Deviation
- calculate the Range Variance and Standard Deviation for a Grouped Data
- calculate the Range Coefficient of Variation
- compute the Measures of Skewness
- calculate the Pearson's No. 1 Coefficient of Skewness
- calculate the Pearson's No. 2 Coefficient of Skewness

1.3 Range, Mean Deviation (MD), Variance and Standard Deviance

1.3.1 The Range

The range measures the difference between the highest and lowest values in a dataset. It is considered a rough measure of spread as it depends on only two values. It is affected by outliers and gives no indication of the clustering of the data.

Formula: Range for ungrouped data: Range **= Highest Value – Lowest Value**

Example 1
Table 1.3: Consider two groups of data:

Dataset A	Dataset B
65	42
66	54
67	58
68	62
71	67
73	77
74	77
77	85
77	93

77	100
Computed measures of central location	
Mean = 71,5	Mean = 71,5
Median = 72	Median = 72
Mode = 77	Mode = 77

Although there is no difference in the computed central measures between the two groups, the scores of dataset B are much more widely scattered than those of dataset A.
The measures that are used to measure dispersion are:

Example
Calculate the range for the set of data

Dataset A 45, 47, 50, 55, 67, 69, 70, 77	Dataset B 42, 63, 65, 74, 82, 91, 100
$range = 77 - 45 = 32$	$range = 100 - 42 = 58$

The ranges indicate that the data in dataset B are more widely spread than that in dataset A.
The time taken to complete an assembling task has been measured for 250 employees:

Time taken (minutes)	Number of people	Cumulative frequency
0 – 5	2	2
5 – 10	2	4
10 – 15	3	7
15 – 20	5	12
20 – 25	5	17
25 – 30	18	35
30 – 35	85	120
35 – 40	92	212
40 – 45	37	249
45 – 50	1	250
Total	250	

Calculate the mid-60% range
Calculate the upper and lower percentiles required for the upper and lower limits of the range:

$$lower\ percentile\ of\ range = \frac{100\% - required\ range\ percentile}{2} = \frac{100\% - 60\%}{2} = 20\%$$

$$upper\ percentile\ of\ range = lower\ percentile\ of\ range + required\ range\ percentile$$
$$= 20\% + 60\% = 80\%$$

1.3.2 The Mean Deviation (MD)

The Mean Deviation can be defined simply by the following relationship:

$$\text{MD} = \frac{\Sigma/\text{X} - \overline{\text{X}}/}{\text{n}}$$

where $\frac{\Sigma/\text{X} - \overline{\text{X}}/}{\text{n}}$ = sum of the absolute values of deviation from Arithmetic mean

n = number of observation

As an example, consider again the arrayed data, X = 2,5,8,9,12,13,18. The mean deviation, MD, can be computed as follows:

$$\overline{\text{X}} = \frac{\Sigma\text{X}}{\text{n}} = \frac{67}{7} = 9.57$$

X	$(\text{X} - \overline{\text{X}})$	$/\text{X} - \overline{\text{X}}/$
2	-7.57	7.57
5	-4.57	2.57
8	-1.57	1.57
9	-0.57	0.57
12	2.43	2.43
13	3.43	3.43
18	8.43	8.43
		Σ /X-X/= 26.57

$$\text{MD}= \frac{\Sigma/\text{X} - \overline{\text{X}}/}{\text{n}} = \frac{26.57}{7} = 3.7957$$

2-9.57 = -7.57
5-9.57 = -4.57
8-9.57 = -1.57
9-9.57 = -0.57
12-9.57 = 2.43
13-9.57= 3.43
18-9.57 = 8.43

Self-Assessment Exercise 1

i. Define Range
ii. State a formula for ungrouped data
iii. Given the set of the following ungroup data: 12, 9, 8, 5, 18, 13, 2, 5
iv. Calculate the Range

1.4 The Variance

The variance measures the average squared deviation from the mean for a dataset

Example

Calculate the variance of the sample scores: 2, 3, 5, 6, 9, 17.

Both variance formulae are used in this example, with all the necessary table columns included for both formulae.

The Variance for a given set of an ungrouped data can be defined by:

$$\text{Variance} = S^2 = \frac{\Sigma X^2 - (\Sigma X/n)^2}{n-1}$$

First it is necessary to calculate the mean:

$$mean, \bar{x} = \frac{\sum x}{n} = \frac{42}{6} = 7$$

where,

x is each value of the dataset is the mean of the dataset

n is the sample size

Where X represents the numerical values of the given set of an ungrouped data.

Continuing with our earlier example, where

X = 2,5,8,9,12,13,18, and by tabulation:

Table 1.4.1: the numerical values of the given set of an ungrouped data

X	**X^2**
2	4
5	25
8	16
9	81
10	48
12	144
13	169
18	324
Total 67	811

Solution:

ΣX = 67; ΣX2 = 811

$$\frac{811-67^2/8}{8-1}$$

$$\frac{811-4489/8}{7}$$

$$\frac{811-561.12}{7}$$

$$\frac{811-561.12}{7}$$

249.88

7
35.70

Thus, the variance of the given set of ungrouped data is 35.70. This is often regarded as

a crude and overestimated measure of dispersion. By interpretation of this result, you can infer that, on the average, each individual observation in the given data differs from the established average of 9.57 by about 28 units. Observe that this type of data will be difficult to make forecasting with and plan with.

$$S = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}}, \text{ for ungrouped data}$$

1.5 The Standard Deviation

Simply stated, the standard deviation is the most useful measure of variation. It can be defined as the square root of the variance for a given set of data.
Standard deviation = S = √S2 or
Using the Table 1.4.1, the standard deviation for the last example is:
S^2=35.70 S= 5.974
Again, by interpretation, this implies that each individual observation in the given set of data deviates from the established average of 9.57 by about 5 units.

Self-Assessment Exercise 2

Calculate Variance and standard deviation (SD) from a given set of grouped data:

S/N	X
1	2
2	5
3	8
4	10
5	11
6	15
7	17
Total	68

1.6 Summary

The unit discussed four major measures of dispersion in the behaviour of business and/or economic variables: the range, the mean deviation, the variance, the standard deviation.

1.7 References/Further Readings/Web Resources

Haessuler, E. F. and Paul, R. S. (1976). *Introductory Mathematical Analysis for Students of Business and Economics,* (2nd edition.) Reston Virginia: Reston Publishing Company

1.8 Possible Answers to SAEs

Answer to SAEs 1

The range measures the difference between the highest and lowest values in a dataset.

R = Xh-XL

Given the arrayed data: X = 2,5,8,9,12,13,18, the range will be: R = 18 – 2 = 16.

Answer to SAEs 2

Calculate Variance and standard deviation (SD) from a given set of grouped data:

S/N	X	$\bar{X}$-X
1	2	-7.7
2	5	-4.7
3	8	-1.7
4	10	0.3
5	11	1.3
6	15	5.3
7	17	7.3
Total	68	0.1

$\bar{X} = 9.7$

$$\frac{\sum(X_1-X)}{n-1}$$

$\underline{\quad(0.1)^2\,}$ $\underline{0.01}$

7-1 6

S^2 = 0.001

Calculate the Standard Deviation (SD)
$\sqrt{S^2} = \sqrt{0.001}$
SD= 0.03

Unit 2 Statistical Tools II

Unit Structure

2.1 Introduction

As noted above, for the purpose of this course, we will discuss another aspect of measures of dispersion in the behaviour of business and/or economic variables: Variance and Standard Deviation for a Grouped Data; the Coefficient of Variation, measures of Skewness such as Pearson's No. 1 Coefficient of Skewness and the Pearson's No. 2 Coefficient of Skewness.

2.2 Learning Outcomes

By the end of this unit, you will be able to:

- calculate the Range Variance and Standard Deviation for a Grouped Data
- calculate the Range Coefficient of Variation
- compute the Measures of Skewness
- calculate the Pearson's No. 1 Coefficient of Skewness
- calculate the Pearson's No. 2 Coefficient of Skewness

2.3 Variance and Standard Deviation

Variance is the average squared deviations from the mean, while standard deviation is the square root of this number. Both measures reflect variability in a distribution, but their units differ: Standard deviation is expressed in the same units as the original values (e.g., minutes or meters).

2.3.1 Variance and Standard Deviation for a Grouped Data

We are

$$\text{Variance} = S^2 = \frac{\Sigma fx^2 - (\Sigma fx)^2/n}{n-1}$$

$$\text{Standard Deviation} = \sqrt{S^2} = \sqrt{\frac{\Sigma fx^2 - (\Sigma fx)^2/n}{n-1}}$$

Example

The following data presents the profit ranges of 100 firms in a given industry.

The computation of variance and standard deviation for a grouped data is illustrated by the following example.

The Variance and Standard Deviation for a grouped data are defined by the following formulations:

The following data presents the profit ranges of 100 firms in a given industry. Profits (N millions)

	No. of Firms (f)	Mid-Value(x)	FX	X^2	FX^2
10 – 15	8	12.5	100	156.25	1250
16 – 21	18	18.5	333	342.25	6160.5
22 – 27	20	24.5	490	600.25	12005
28 – 33	12	30.5	366	930.25	11163
34 – 39	15	36.5	547.5	1332.25	19983.75
40 – 45	17	42.5	722.5	1806 25	30706.25
46 – 51	10	48.5	485	2352.25	23522.50

$\Sigma f = n = 100$

$\Sigma fx=3044$

$\Sigma fx2 = 104791$

It follows that:

$$\frac{(\Sigma fx)^2}{n} = \frac{(3044)^2}{100} = 92659.36$$

$$\text{Variance} = S^2 = \frac{\Sigma fx^2 - (\Sigma fx)^2/n}{n-1} = \frac{104791 - 92659.36}{100-1}$$

$$= \frac{12131.64}{99} = 122.54$$

$$\text{Standard Deviation} = \sqrt{S^2} = \sqrt{122.54} = 11.07$$

Thus, the required variance and standard deviation are 122.54 and 11.07 respectively.

2.3.2 The Coefficient of Variation

Unlike other measures of variability, the coefficient of variation is a relative measure. It is particularly useful when comparing the variability of two or more sets of data that are expressed in different units of measurements.
The coefficient of variation measures the standard deviation relative to the mean and is computed by:

$$\text{Coefficient of Variation} = \text{CV} = \frac{[\text{S}]}{\text{X}}\ 100\%$$

The coefficient of variation is also useful in the comparison of two or more sets of data which are measured in the same units but differ to such an extent that a direct comparison of the respective standard deviations is not very helpful.

As an example, suppose a potential investor is considering the purchase of shares in one or two companies, A or B, which are listed on the Nigerian Stock Exchange (NSE). If neither company offered dividends to its shareholders and if both companies were rated equally high in terms of potential growth, the potential investor might want to consider the volatility of the two stocks to aid in the investment decision.

Now, suppose each share of stock in Company A has an average of N50 over the past months with a standard deviation of N10. In addition, suppose that in this same time period, the price per share for Company B's stock averaged N12 with a standard deviation of N4. Observe that in terms of actual standard deviations, the price of Company A's shares seems to be more volatile than that of Company B. However, since the average prices per share for the two stocks are so different, it would be more appropriate for the potential investor to consider the variability in price relative to the average price in order to examine the volatility/stability of two stocks.

That of Company B is:

$$\text{CV}_\text{B} = \frac{[\text{S}_\text{B}]}{\text{X}_\text{B}}\ 100\% = \frac{[\text{N4}]}{\text{N12}}\ 100\% = 33.3\%$$

It follows that relative to the average, the share price of company B's stock is much more variable/unstable than that of Company A.

2.4 The Measures of Skewness

The measures of skewness are generally called Pearson's first coefficient of skewness and Pearson's second coefficient of skewness. Measures of skewness are used in determining the degree of asymmetry of a distribution; a distribution which is not symmetrical is said to be skewed.

Self-Assessment Exercise 1

1. Explain the measurement of skewness
2. Consider a set of data on monthly sales of a company's product, the mean of which was found to be N240, 000; the mode found to be N135, 000; and the standard deviation found to be N85, 000. The Calculate the Pearson's No. 1 Coefficient of skewness.

2.5 The Pearson's No. 1 Coefficient of Skewness

The formula used in calculating Pearson's No. 1 coefficient is:

$$\text{Sk} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Notice that the mean, the mode, and the standard deviation are all expressed in the units of the original data. When the difference between the mean and the mode is computed as a fraction of the standard deviation (or average spread of the data around the mean), the original units cancel out in the fraction.

The result will be a coefficient of skewness, a number which tells you the extent of the skewness in the distribution.

Example
Consider a set of data on monthly sales of a company's product, the mean of which was found to be N240, 000; the mode found to be N135, 000; and the standard deviation found to be N85, 000. The Pearson's No. 1 Coefficient of skewness would be calculated as follows: = 1.24
Generally, a complete absence of skewness would have a coefficient of skewness equal to zero. In our example, since the mean was larger.

$$\text{Sk} = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{240{,}000 - 135{,}000}{85{,}000}$$

Then the mode, we obtained a positive coefficient of skewness to the extent of 124% of the standard deviation.

2.6 The Pearson's No. 2 Coefficient of Skewness

This type of the Pearson's coefficient of skewness came as a result of the fact that a precise calculation of mode is difficult in many distributions. Hence, Pearson's No. 2 coefficient of skewness uses the difference between the mean and the median of the distribution instead of the difference between the mean and the mode. In this calculation, you have the formula:

$$\text{Sk} = \frac{3(\text{mean} - \text{median})}{\sigma}$$

Consider a set of data on monthly sales of a company's product, the mean of which was found to be N901, 000; the model found to be N282, 000; and the standard deviation found to be N102, 000.
Calculate the Pearson's No. 2 Coefficient of skewness?

$$SK = \frac{3(\text{Mean} - \text{Mode})}{\sigma}$$

$$\frac{3(901{,}000 - 282000)}{102{,}000}$$

$$\frac{2703000 - 846000}{102{,}000}$$

$$\frac{1857000}{102{,}000} \qquad SK = 18.2$$

This formula should give you a more accurate measure of skewness than that of the Pearson's No. 1 formula.

Self-Assessment Exercise 2

1. Explain the Pearson's No. 2 Coefficient of Skewness
2. Consider a set of data on monthly sales of a company's product, the mean of which was found to be N901, 000; the model found to be N282, 000; and the standard deviation found to be N102, 000. Calculate the Pearson's No. 2 Coefficient of skewness?

2.7 Summary

This unit complemented unit 6 in its presentation of the basic principles of dispersion or measures of uniformity or measures of variability in a statistical data set. Five measures of dispersion in the behaviour of business and/or economic variables were presented as: the range, the mean deviation, the variance, the standard deviation, and the coefficient of variation. Another important related measure that is primarily concerned with distribution of some business data is the Pearson Coefficient of Skewness. The computation, interpretations, and managerial implications of these measures were discussed.

We summarise the basic measures of variations as discussed in this cunit in the following form.

1. The Range (R) of a given set of ungrouped data can be determined from an ordered array as the difference between the highest observation and the lowest observation.

Let
Xh = Highest observation
XL = Lowest observation
The n, R = Xh-XL
The Mean Deviation can be defined simply by the following relationship:
n = number of
observation

2. Arithmetic mean
The Variance for a given set of an ungrouped data can be defined by where X represents the numerical values of the given set of an ungrouped data.
The standard deviation is the most useful measure of variation. It can be defined as the square root of the variance for a given set of data.

Thus, Standard deviation = S = √S2

$$S = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}}, \text{ for ungrouped data}$$

The coefficient of variation is a relative measure. It is particularly useful when comparing the variability of two or more sets of data that are expressed in different units of measurements. The coefficient of variation measures the standard deviation relative to the mean and is computed by:

$$\text{Coefficient of Variation} = CV = \frac{[S]}{X}\ 100\%$$

The measures of skewness are generally called Pearsons first coefficient of skewness and Pearson’s second coefficient of skewness. Measures of skewness are used in determining the degree of asymmetry of a distribution; a distribution which is not symmetrical is said to be skewed.

2.8 References/Further Readings/Web Resources

Haessuler, E. F. and Paul, R. S. (1976). *Introductory Mathematical Analysis for Students of Business and Economics, 2nd edition. Reston Virginia:* Reston Publishing Company.

2.9 Possible Answers to SAEs

Answers to SAEs 1

1. Explain the measurement of skewness
   The measures of skewness are generally called Pearson's first coefficient of skewness and Pearson's second coefficient of skewness. Measures of skewness are used in determining the degree of asymmetry of a distribution; a distribution which is not symmetrical is said to be skewed.
2. Consider a set of data on monthly sales of a company's product, the mean of which was found to be N240, 000; the mode found to be N135, 000; and the standard deviation found to be N85, 000. The Calculate the Pearson's No. 1 Coefficient of skewness.

$$SK = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$\frac{240{,}000 - 135{,}000}{85{,}000}$$

$$\frac{105{,}000}{85{,}000} \qquad SK = 1.23$$

Answers to SAEs 2

1. The Pearson's No. 2 Coefficient of Skewness
   This type of the Pearson's coefficient of skewness came as a result of the fact that a precise calculation of mode is difficult in many distributions. Hence, Pearson's No. 2 coefficient of skewness uses the difference between the mean and the median of the distribution instead of the difference between the mean and the mode. In this calculation, you have the formula:

$$\text{Sk} = \frac{3(\text{mean} - \text{median})}{\sigma}$$

2. Consider a set of data on monthly sales of a company's product, the mean of which was found to be N901, 000; the model found to be N282, 000; and the standard deviation found to be N102, 000.
   Calculate the Pearson's No. 2 Coefficient of skewness?

$$SK = 3\frac{(\text{Mean} – \text{Mode})}{\sigma}$$

$$\frac{3(901,000 – 282000)}{102,000}$$

$$\frac{2703000 - 846000}{102,000}$$

$$\frac{1857000}{102,000} \qquad SK = 18.2$$

Unit 3 Statistical Tools III

Unit Structure

3.1 Introduction

In this unit, we develop the basic principles of probabilistic analysis, with special emphasis on the applications of set operations and events.

3.2 Learning Outcomes

By the end of this unit, you will be able to:

- define concept of sets as they are used in mathematical operations
- explain the theory of sets and how it is used in probability analysis.
- explain set enumerations and how they can be applied in solving business-related problems

3.3 Sets Theory

The theory of sets serves as a preliminary concept necessary for the understanding of the theory of probabilities. A mathematical set is a collection of distinct objects, often referred to as elements or members.

Examples
(a) The employees of a company working in the Public Relations Department could be represented as:
PR = {Joseph, Adamu, Adebola, Nkom, Margerate}

(b) The location of shops for a big automobile parts dealer could be represented as:
S = {Abuja, Enugu, Lagos, Aba, Onitsha, Kano, Ikot Ekpene}

3.3.1 Subsets

A subset of a set, say A, is a set which contains some of the elements of set A. For instance:
If set A = {a, e, i, o, u}, then:
X = {a, e, i} is a subset of A
Y = {e, i} is a subset of A
Z = {i, o, u} is a subset of A

3.3.2 The Number of a Set

The *number* of a set A, written as n[A], is defined as the number of elements in set A. For example:
If A = {a, e, i, o, u}, then n[A] = 5 (that is, 5 elements in set A).

3.3.3 Set Equality

Two sets are said to be equal only when they have identical elements.
For example:
If A = {1, 2, 3} and B = {1, 2, 3}, then Set A = Set B.

3.4 Universal Set

A universal set, denoted by U is a set containing different subsets of its elements. For example, a combination of different behaviours in a given population can be considered as universal, while a selected sample of such behaviours are referred to as the subsets. A set of all English alphabets make up the universal set, while a set containing the vowels would be referred to as the subset.
Symbol of Universal Set
The universal set is usually represented by the symbol E or U. It consists of all the elements of its subsets, including its own elements https://www.cuemath.com/algebra/universal-set/.

Example of Universal Set
Let's consider an example with three sets, A, B, and C. Here, A = {2, 4, 6}, B = {1, 3, 7, 9, 11}, and C = {4, 8, 11}. We need to find the

universal set for all three sets A, B, and C. All the elements of the given sets are contained in the universal set. Thus, the universal set U of A, B, and C is given by U = {1, 2, 3, 4, 6, 7, 8, 9, 11}

We can see that all the elements of the three sets are present in the universal set without any repetition. Thus, we can say that all the elements in the universal set are unique. The sets A, B, and C are contained in the universal set, then these sets are also called subsets of the Universal set.

A ⊂ U (A is the subset of U)
B ⊂ U (B is the subset of U)
C ⊂ U (C is the subset of U)

3.5 Complement of Universal Set

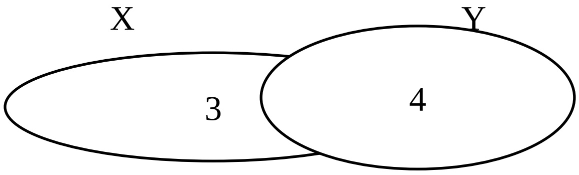
For a subset A of the universal set (U), its complement is represented as A' which includes the elements of the universal set but not the elements of set A. The Universal set consists of a set of all elements of all its related subsets, whereas the empty set contains no elements of the subsets. Thus, the complement of the universal set is an empty set, denoted by ‘{}’ or the symbol 'Φ'.

3.5.1 Complement of a Set

The complement of a set A (denoted by A’) contained in a given universal set, U, is the set of elements in the universal set that are not contained in set A. For example: If set A represents the set of all skilled workers in a given universal set, U, then the complement of set A, A’, is the set of unskilled workers who are members of the universal set.

3.6 Venn Diagrams

Venn diagrams are simple pictorial representations of a set. They are useful for demonstrating relationships between sets.
This can be represented by a Venn diagram as follows:



The elements {3, 4} are contained in the circle common to sets X and Y.

Self-Assessment Exercise 1

1. A mathematical set is a collection of distinct objects, often referred to as-----------
2. State the two basic operations on sets

3.7 Set Enumeration

Set enumeration considers sets in terms of number of elements contained within the various areas defined by their union or intersection. Identifying the number of elements in these areas is known as set enumeration. As an illustration, consider the following enumeration problem.

Suppose an accounting firm currently employs 16 staff members. Given that three staff members have no formal qualifications, and of the seven staff members, who are graduates, 5 are also qualified as chartered accountants, it is possible to evaluate: (a) the number of staff who arenon-graduates, chartered accountants and (b) the number of graduates who are not qualified chartered accountants.

The two values can be calculated as follows:
Since three staff members have no formal qualifications, there should be 16– 3 = 13 staff members with at least one of the two qualifications that is, a graduate or a chartered accountant.

But there are 7 staff members who are graduates, which implies that 16 –7–3= 6 must be non-graduate, qualified chartered accountants, which gives the answer to possibility (a) above.
In addition, since 5 staff members are qualified chartered accountants and graduates, there would be 7 – 5 = 2 staff who are graduates only. This gives answer to possibility (b) above.

In an extended problem, the above approach is not structured enough for the solution. In the following discussions, we present a more structured procedure which solves the above problems and forms a basis for more logical approach for solving enumerations problems in general.

3.7.1 Circle Venn Diagram

Circle Venn Diagram Examples:
For the purposes of an administrative research, a survey of 1000 women is conducted in a town. The results show that 52 % liked watching comedies, 45% liked watching fantasy movies and 60% liked watching romantic movies. In addition, 25% liked watching comedy and fantasy

both, 28% liked watching romantic and fantasy both and 30% liked watching comedy and romantic movies both. 6% liked watching none of these movie genres [(https://www.intellspot.com/venn-diagram-examples/)](https://www.intellspot.com/venn-diagram-examples/).

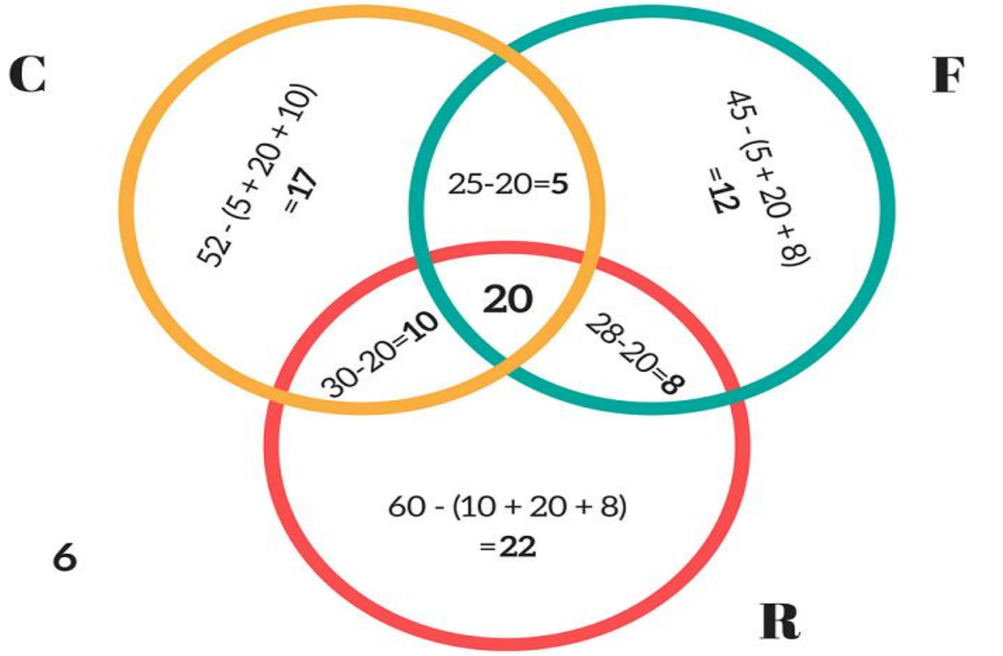
Here are our questions we should find the answer:
How many women like watching all the three movie genres?
Find the number of women who like watching only one of the three genres.

Find the number of women who like watching at least two of the given genres.

Let's represent the data above in a more digestible way using the Venn diagram formula elements:

n(C) = percentage of women who like watching comedy = 52%
n(F) = percentage of women who like watching fantasy = 45%
n(R)= percentage of women who like watching romantic movies= 60%
n(C∩F) = 25%; n(F∩R) = 28%; n(C∩R) = 30%

Since 6% like watching none of the given genres so, n (C ∪ F ∪ R) = 94%.
Now, we are going to apply the Venn diagram formula for 3 circles.
94% = 52% + 45% + 60% – 25% – 28% – 30% + n (C ∩ F ∩ R)
Solving this simple math equation, lead us to:
n (C ∩ F ∩ R) = 20%
It is a great time to make our Venn diagram related to the above situation (problem):



(https://www.intellspot.com/venn-diagram-examples/).
See, the Venn diagram makes our situation much clearer!
From the Venn diagram example, we can answer our questions with ease.
The number of women who like watching all the three genres = 20% of 1000 = 200.

Number of women who like watching only one of the three genres = (17% + 12% + 22%) of 1000 = 510
The number of women who like watching **at least** two of the given genres = (number of women who like watching **only two** of the genres) +(number of women who like watching **all the three** genres) = (10 + 5 + 8 + 20)% i.e. 43% of 1000 = 430. As we mentioned above 2 and 3 circle diagrams are much more common for problem-solving in many areas such as business, statistics, data science and etc. However, 4 circles Venn diagram also have its place.

3.7.2 Summary of the General Enumeration Problems

Note in particular that there are:
(a) 4 distinct areas for two attribute sets, and
(b) 8 distinct areas for three attribute sets.
The general procedure for solving enumeration problems follows the following steps:

Step 1 Identify the attribute sets
Step 2 Draw an outline Venn diagram
Step3 Use the information given to fill in as much of the diagramas possible
Step 4 Evaluate the number of elements in unknown areas.

Consider the following example:

A survey was carried out by a researcher, one of the aims being to discover the extent to which computers are being used by firms in a given area. 32 firms had both stock control and payroll computerized, 65 firms had just one of these two functions computerised, and 90 firms had a computerized payroll. If 22 firms had neither of these functions computerized, how many firms were included in the survey?

Step 1: The two attributes involved are computerised payroll, with set (say P), and computerised stock control, with set (say S).

Step 2: Using standard notations, Let p = number of firms with a computerised payroll only; s = number of firms with a computerised

stock control only; ps = number of firms with both payroll and stock control computerised; and
x = number of firms with neither functions computerised, construct a Venn diagram describing the situation.
Step 3: The following equations can be set up from the given information.

ps = 32
p + s = 65
ps + p = 90
x = 22

Substituting for ps = 32 in equation 3, we get p = 58
Substituting for p = 58 in equation, we get s = 7.

It follows that the number of firms included in the survey equals:
p + s + ps + x = 58 + 7 + 32 + 22 = 119

Self-Assessment Exercise 2

1. Set enumeration considers sets in terms of number of elements contained within the various areas defined by their union or intersection. Identifying the number of elements in these areas is known as -----------
2. State the general procedure for solving enumeration problems

3.8 Summary

The theory of set is currently used in several business operations. It is used in much software programming, in the study of consumer behaviour, and in solving complex business problems. The unit has put together some preliminary concepts in set notations. You must have been exposed to the simple operations on sets.

A mathematical set is a collection of distinct objects, often referred to as Elements or members

There are two basic operations on sets, including:
1. Set union
2. Set intersection

Set enumeration considers sets in terms of number of elements contained within the various areas defined by their union or intersection.

Identifying the number of elements in these areas is known as set enumeration.

State the general procedure for solving enumeration problems:
Step 1 Identify the attribute sets
Step 2 Draw an outline Venn diagram
Step 3 Use the information given to fill in as much of the diagram as possible
Step 4 Evaluate the number of elements in unknown areas.

3.9 References/Further Readings/Web Resources

A. Francis (1998). *Business Mathematics and Statistics, 5th edition.* Great Britain: Ashford Colour Press.

Intellspot.com (2022) Set theory. Retrieved from *(https://www.intellspot.com/venn-diagram-examples/)*. Accessed on 14/08/2022.

3.10 Possible Answer to SAEs

Answer to SAEs 1

1. A mathematical set is a collection of distinct objects, often referred to as Elements or members
2. There are two basic operations on sets, including:
   i. Set union
   ii. Set intersection

Answer to SAEs 2

1. Set enumeration considers sets in terms of number of elements contained within the various areas defined by their union or intersection. Identifying the number of elements in these areas is known as set enumeration.
2. State the general procedure for solving enumeration problems:
   Step 1: Identify the attribute sets
   Step 2: Draw an outline Venn diagram
   Step 3: Use the information given to fill in as much of the diagram as possible
   Step 4: Evaluate the number of elements in unknown areas.

Unit 4 Statistical Tools IV

Unit Structure

4.1 Introduction

In this unit, we pay a special attention to the concept of probability, probability laws, computation of probabilities, and their applications to business decisions. At the end of this lecture, students will be expected to be able to make effective decisions under uncertainties.

The basic elements of probability theory are the outcomes of the process or phenomenon under study.

4.2 Learning Outcomes

By the end of this unit, you will be able to:

- define Probability
- explain the theory of probability
- define probability
- state the laws of probability
- calculate probabilities
- apply probabilities in making decisions involving uncertainties

4.3 Probability

4.3.1 Definitions of Probability

Probability is a concept that used as “chance,” “likelihood” “possibility” and “proportion as part of everyday speech and business situation. For example, most of the following, which might be heard in any business situation, are in fact statements of probability.

Each possible type of occurrence is referred to as **an event.** The collection of all the possible events is called the **sample space.**

A compound or joint event is an event that has two or more characteristics. For example, the event of a student who is “an economics major and B or above average” is a joint or compound event since the student must be an economics major and have a B or above average.

The event “black ace” is also a compound event since the card must be both black and ace in order to qualify as a black ace.

Probability is a concept that most people understand naturally, since such words as “chance,” “likelihood” “possibility” and “proportion are used as part of everyday speech. For example, most of the following, which might be heard in any business situation, are in fact statements of probability.

a) “There is a 30% chance that this job will not be finished in time”.
b) “There is every likelihood that the business will make a profit next year”.
c) “Nine times out of ten, he arrives late for his appointments”.

In statistical sense, probability simply puts a well-defined structure around the concept of everyday probability, enabling a logical approach to problem solving to be followed.

There are basically two separate ways of calculating probability.

1. Calculation based on theoretical probability. This is the name given to probability that is calculated without an experiment that is, using only information that is known about the physical situation.
2. Calculation based on empirical probability. This is probability calculated using the results of an experiment that has been

performed a number of times. Empirical probability is often referred to as *relative frequency* or *Subjective probability.*

1.3.2 Definition of Theoretical Probability

Let E represent an event of an experiment that has an equally likely outcome set, U, then the theoretical probability event E occurring when the experiment is written as Pr (E) and given by:

Where n (E) = the number of outcomes in event set E n (U) = total possible number of outcomes in outcome set, U.

$$\Pr(E) = \frac{\text{number of different ways that the event can occur}}{\text{number of different possible outcomes}} = \frac{n(E)}{n(U)}$$

If, for example, an ordinary six-sided die is to be rolled, the equally likely outcome set, U, is {1,2,3,4,5,6} and the event "even number" has event set {2,4,6}. It follows that the theoretical probability of obtaining an even number can be calculated as:

$$\Pr(\text{even numbers}) = \frac{n(\text{even numbers})}{n(U)} = \frac{3}{6} = 0.50$$

Self-Assessment Exercise 1

1. Define the term Probability
2. State the two ways of calculating probability
3. Explain the theoretical Probability

4.4 Definition of Empirical (Relative Frequency) Probability

If E is some event of an experiment that has been performed a number of times, yielding a frequency distribution of events or outcomes, then the empirical probability of event E occurring when the experiment is performed one more time is given by: Pr(E)□□ number of times the event occurred number of times the experiment was performed f(E)
<u>f(E)</u>
f Where f(E) f = the frequency of event E.
f = total frequency of the experiment.

Put differently, the empirical probability of an event E occurring is simply the proportion of times that event E actually occurred when the experiment was performed.

For example, if, out of 60 orders received so far this financial year, 12 were not completely satisfied, the proportion, 12/60 = 0.2 is the empirical probability that the next order received will not be completely satisfied.

4.5 Laws of Probability

1) Addition Law for mutually exclusive events
2) Addition Law for events that are not mutually exclusive
3) Multiplication Law for Independent events
4) Multiplication Law for Dependent events.

4.5.1 Addition Law for Mutually Exclusive Events

Two events are said to be mutually exclusive events if they cannot occur at the same time. The addition law states that if events A and B are mutually exclusive events, then: Pr (A or B) = Pr (A) + Pr (B)

4.5.2 Multiplication Law for Independent Events

This law states that if A and B are independent events, then:
Pr (A and B) = Pr (A). Pr (B)
As an example, suppose, in any given week, the probability of an assembly line failing is 0.03 and the probability of a raw material shortage is 0.1.

If these two events are independent of each other, then the probability of an assembly line failing and a raw material shortage is given by:
Pr (Assembly line failing and Material shortage) = (0.03)(0.1) = 0.003

4.5.3 Multiplication Law for Dependent Events

This Law states that if A and B are dependent events, then:
Pr (A and B) = Pr(A).Pr(B/A)
Note that Pr (B/A) in interpreted as probability of B given that event A has occurred.

Example
A display of 15 T-shirts in a Sports shop contains three different sizes: small, medium and large. Of the 15 T-shirts: 3 are small 6 are medium 6 are large.

If two T-shirts are randomly selected from the T-shirts, what is the probability of selecting both a small T-shirt and a large T-shirt, the first not being replaced before the second is selected?

Solution

Since the first selected T-shirt is not replaced before the second T-shirt is selected, the two events are said to be dependent events. It follows that:

Pr (Small T-shirt and Large T-shirt)
= Pr(Small).Pr(Large/Small)
= (3/15)(4/14)
= (0.2)(0.429)
= 0.086

4.6 Conditional Probability

Assuming two events, A and B, the probability of event A, given that event B has occurred is referred to as the conditional probability of event A.

In symbolic term:
Pr(*A B*) Pr(*A*)□□ Pr(*B A*)/Pr(*B*) Pr (A□□ B)□
Pr(B)Pr(A)□□ Pr(B)
Pr(B)□ Pr(*A*)
Where Pr (A/B) = conditional probability of event A
Pr (A□B) = joint probability of events A and B
Pr (B) = marginal probability of event B

In general, Pr(A *B*)□□□ Joint Probability of events A and B Marginal Probability of event B

4.7 The Bayes Theorem

Bayes theorem is a formula which can be thought of as "reversing" conditional probability. That is, it finds a conditional probability, A/B given, among other things, its inverse, B/A. According to the theorem, given events A and B, Pr(*A B*)□□□Pr(*A*)□□ Pr(*B A*) Pr(*B*).

As an example in the use of Bayes theorem, if the probability of meeting a business contract date is 0.8, the probability of good weather is 0.5 and the probability of meeting the date given good weather is 0.9, we can calculate the probability that there was good weather given that the contract date was met.

4.8 Probability and Expected Values

The expected value of a set of values, with associated probabilities, is the arithmetic mean of the set of values. If some variable, X, has its values specified with associated probabilities, P, then:
Expected value of X = E (X) =□ PX

Age	Male (M)	Female (F)	Marginal Probability
Below 30 (B)	0.2857	0.3333	0.62
30 and Above (A)	0.2857	0.0952	0.38
Marginal Probability	0.57	0.43	1.00

Example

An ice-cream salesman divides his days into ‘Sunny’ ‘Medium’ or ‘Cold’. He estimates that the probability of a sunny day is 0.2 and that 30% of his days are cold. He has also calculated that his average revenue on the three types of days is N220, N130, and N40 respectively. If his average total cost per day is N80, calculate his expected profit per day.

Solution

We first calculate the different values of profit that are possible since we are required to calculate expected profit per day, as well as their respective probabilities.

Given that Pr (sunny day) = 0.2; Pr (cold day) = 0.3. Since in theory, Pr (sunny day) + Pr (cold day) +Pr (medium day) = 1. It follows that:
Pr (medium day) = 1 - 0.2 - 0.3 = 0.5
The total costs are the same for any day (N80), so that the profits that the salesman makes on each day of the three types of day are:
Sunny day:

N (220-80) = N140
Medium day: N (130-80) = N50
Cold day: N (40-80) = -N40 (loss).

Self-Assessment Exercise 2

1. State the Laws of Probability
2. State the Addition Law for Mutually Exclusive Events
3. Explain the Multiplication Law for Independent Events
4. Since the first selected T-shirt is not replaced before the second T-shirt is selected, the two events are said to be dependent events. It follows that: Pr (Small T-shirt and Large T-shirt). Calculate.

4.9 Summary

Probability is a concept that most people understand naturally, since such words as “chance,” “likelihood,” “possibility” and “proportion are used as part of everyday speech. It is a term used in making decisions involving uncertainty. Though the concept is often viewed as very abstract and difficult to relate to real world activities, it remains the best tool for solving uncertainties problems.

To remove some of the abstract nature of probabilities, this unit has provided you with the simplest approach to understanding and calculating, as well as applying the probability concept. It defines probability in two basic forms:
the theoretical definition; and
the empirical definition

The issues discussed in this unit can be summarised in the following way: there are basically two separate ways of calculating probability which are as stated below:
theoretical probability: this is calculated without an experiment,
ii that is, using only information that is known about the physical situation.

Calculation based on empirical probability. This is probability calculated using the results of an experiment that has been performed a number of times. Empirical probability is often referred to as Relative frequency or Subjective probability.

There are four basic laws of probability:
1. Addition law for mutually exclusive events
2. Addition law for events that are not mutually exclusive
3. Multiplication law for independent events
4. Multiplication law for dependent events

A joint probability implies the probability of joint events. Joint probabilities can be conveniently analyzed with the aid of joint in which all possible events for a variable are recorded in a row and those of other variables are recorded in a column, with the values listed in corresponding cells.

The **Marginal Probability** of an event is its simple probability of occurrence, given the sample space.

Assuming two events, A and B, the probability of event A, given that event B has occurred is referred to as the **conditional probability** of event A.

The expected value of a set of values, with associated probabilities, is the arithmetic mean of the set of values. If some variable, X, has its values specified with associated probabilities, P, then:
Expected value of $X = E (X) = \square PX$

4.10 References/Further Readings/Web Resources

A. Francis (1998). *Business Mathematics and Statistics, 5th edition. Great Britain:* Ashford Colour Press.

4.11 Possible Answers to SAEs

Answers to SAEs 1

1. Probability is a concept that used as “chance,” “likelihood” “possibility” and “proportion as part of everyday speech and business situation. For example, most of the following, which might be heard in any business situation, are in fact statements of probability.
Each possible type of occurrence is referred to as an event. The collection of all the possible events is called the sample space.

A **compound or joint event** is an event that has two or more characteristics. For example, the event of a student who is “an economics major and B or above average” is a joint or compound event since the student must be an economics major and have a B or above average.

The event “black ace” is also a compound event since the card must be both black and ace in order to qualify as a black ace.

a) “There is a 30% chance that this job will not be finished in time”.
b) “There is every likelihood that the business will make a profit next year”.
c) “Nine times out of ten, he arrives late for his appointments”.
In statistical sense, probability simply puts a well-defined structure around the concept of everyday probability, enabling a logical approach to problem solving to be followed.

2. There are basically two separate ways of calculating probability.

i. Calculation based on theoretical probability. This is the name given to probability that is calculated without an experiment that is, using only information that is known about the physical situation.

ii. Calculation based on empirical probability. This is probability calculated using the results of an experiment that has been performed a number of times. Empirical probability is often referred to as *relative frequency* or *Subjective probability.*

3. Definition of Theoretical Probability

Let E represent an event of an experiment that has an equally likely outcome set, U, then the theoretical probability event E occurring when the experiment is written as Pr (E) and given by:
Where n (E) = the number of outcomes in event set E n (U) = total possible number of outcomes in outcome set, U.

$$\Pr(\mathrm{E}) = \frac{\text{number of different ways that the event can occur}}{\text{number of different possible outcomes}} = \frac{\mathrm{n(E)}}{\mathrm{n(U)}}$$

If, for example, an ordinary six-sided die is to be rolled, the equally likely outcome set, U, is {1,2,3,4,5,6} and the event “even number” has event set {2,4,6}. It follows that the theoretical probability of obtaining an even number can be calculated as:

$$\Pr(\text{even numbers}) = \frac{\mathrm{n}(\text{even numbers})}{\mathrm{n(U)}} = \frac{3}{6} = 0.50$$

Answers to SAEs 2

1. Laws of Probability

i. Addition Law for mutually exclusive events
ii. Addition Law for events that are not mutually exclusive
iii. Multiplication Law for Independent events
iv. Multiplication Law for Dependent events.

2. Addition Law for Mutually Exclusive Events

Two events are said to be mutually exclusive events if they cannot occur at the same time. The addition law states that if events A and B are mutually exclusive events, then: Pr (A or B) = Pr (A) + Pr (B)

3. Multiplication Law for Independent Events

This law states that if A and B are independent events, then:
Pr (A and B) = Pr (A). Pr (B)
As an example, suppose, in any given week, the probability of an assembly line failing is 0.03 and the probability of a raw material shortage is 0.1.

If these two events are independent of each other, then the probability of an assembly line failing and a raw material shortage is given by:
Pr (Assembly line failing and Material shortage) = (0.03)(0.1) = 0.003
4. A display of 15 T-shirts in a Sports shop contains three different sizes: small, medium and large. Of the 15 T-shirts: 3 are small 6 are medium 6 are large.

If two T-shirts are randomly selected from the T-shirts, what is the probability of selecting both a small T-shirt and a large T-shirt, the first not being replaced before the second is selected?

Solution
Since the first selected T-shirt is not replaced before the second T-shirt is selected, the two events are said to be dependent events. It follows that:
Pr (Small T-shirt and Large T-shirt)
= Pr(Small).Pr(Large/Small)
= (3/15)(4/14)
= (0.2)(0.429) = 0.086

Unit 5 Basic Advance Mathematics

Unit Structure

1.1 Introduction

Algebra is the branch of mathematics that helps in the representation of problems or situations in the form of mathematical expressions. It involves variables like x, y, z, and mathematical operations like addition, subtraction, multiplication, and division to form a meaningful mathematical expression. All the branches of mathematics such as trigonometry, calculus, coordinate geometry, involve the use of algebra.

Algebra deals with symbols and these symbols are related to each other with the help of operators. It is not just a mathematical concept, but a skill that all of us use in our daily life without even realizing it.

Understanding algebra as a concept is more important than solving equations and finding the right answer, as it is useful in all the other topics of mathematics that you are going to learn in the future or you have already learned in past.

1.2 Learning Outcomes

By the end of this unit, you will be able to:

- define the Basic Algebra and the basic rules.
- calculate the Linear equations
- calculate the Quadratic formula (Factorization)
- calculate the Quadratic formula (Formula Method)
- apply the Quadratic formula to decision making

5.3 Basic Algebra

5.3.1 Linear equations

Linear equations are classified as first-degree equations.
Basic Rules and Properties of Algebra
The basic rules or properties of algebra for variables, algebraic expressions, or real numbers a, b and c are as given below,

Commutative Property of Addition: $a + b = b + a$
Commutative Property of Multiplication: $a \times b = b \times a$
Associative Property of Addition: $a + (b + c) = (a + b) + c$
Associative Property of Multiplication: $a \times (b \times c) = (a \times b) \times c$
Distributive Property: $a \times (b + c) = (a \times b) + (a \times c)$, or, $a \times (b - c) = (a \times b) - (a \times c)$

Reciprocal: Reciprocal of $a = 1/a$
Additive Identity Property: $a + 0 = 0 + a = a$
Multiplicative Identity Property: $a \times 1 = 1 \times a = a$
Additive Inverse: $a + (-a) = 0$

For video demonstration, follow this link: *https://www.khanacademy.org/math/algebra/x2f8bb11595b61c86:foundation-algebra/x2f8bb11595b61c86:intro-variables/v/why-aren-t-we-using-the-multiplication-sign?modal=1*

Example 1
One simple example of an expression in algebra is $2x + 4 = 8$.
Solve the Unknown variables in a linear equation using simple algebraic operations,

a. $5X – 6 = 3X$
b. $2(p + 4) = 7p+2$

Solution
$5X – 6 = 3X$
$5X – 3X = 6$
$(5-3)X = 6$
$2X = 6$
$2X/^2 = 6/2 = 2$

$2(p + 4) = 7p+2$
$2p + 8 = 7p + 2$
$2p - 7p = 2-8$
$(2 - 7)p = -6$
$-5p/5 = -6/5$ $6/5 =1/2$

Example 2

$$\frac{(7X + 3) - (9X + 8)}{2} = \frac{6}{4}$$

Solution

$$\frac{(7X + 3) - (9X + 8)}{2} = \frac{6}{4}$$

$2(7X + 3) - 1(9X + 8) = 4(6)$

$(14X + 6) - (9X + 8) = 24$

Clear the Bracket

$14X + 6 - 9X + 8 = 24$

$14X - 9X = 24 - 6 - 8$

$\frac{5X}{5} = \frac{10}{5}$ $X = 2$

Self-Assessment Exercise 1

1. Calculate $x - 5 = 2$. Five less than a number equals to two. What is the number?
2. Expand $(2x + 3y)^2$ using the algebraic identities.
3. The present age of a person is double the age of his son. Ten years ago, his age was four times the age of his son. Use the concept of algebra and find the present age of the son.

5.4 Quadratic formula (Factorization)

Factorization involves the determination of the factors that form the given quadratic equation. This will then make the solutions for the unknown easy to come by.

Example

Explain the Factorization methods of quadratic equation

Solve for x in the quadratic equation: $X^2 + X - 12 = 0$

Solution

$X^2 + X - 12 = 0$

$(X - 3)$ and $(X + 4)$ are the factors, so that

$(X - 3)(X + 4) = 0$

5.5 Quadratic formula (Formula Method)

The quadratic formula helps us solve any quadratic equation. First, we bring the equation to the form $ax^2+bx+c=0$, where a, b, and c are coefficients. Then, we plug these coefficients in the formula: $(-b\pm\sqrt{(b^2-4ac)})/(2a)$. See examples of using the formula to solve a variety of equations.

Example
Solve the equation using formula method

$4x^2 -17x+15 = 0$

Solution:

$$X = \frac{-b+\_\ b2-4ac}{2a}$$

Using the formula, a = 4; b = -17; c = 15. By substitution to the formula, we get:

$$X = \frac{-(-17)+\_\ (-17)2-4(4)(15)}{2(4)}$$

$17\pm 289-240/8$
$17\pm 49/8$
X = 17+7 or X = 17-7
X = 3 or X = 1.25

5.6 Application of Quadratic formula

Example
XYZ Company produces product A for which cost (including labour and material) is N6/unit. Fixed cost is N80, 000. Each unit is sold for N10. Determine the number of units which must be sold for the company to earn a profit of N60, 000.

Solution:
By definition, Profit = Revenue-Cost
That is = R – C
Let x represent the level of output, so that:
Total Cost = C = Fixed Cost (FC) + Variable Cost (VC),
That is, TC = FC + VC
Therefore: FC = N80,000, VC = 6X, P=10. Expected Profit = 60,000
Where: The variable cost (VC) per unit produced is N6, so that for x units,

a. Determine the C

b. Determine the R

C= FC + VC; R= Price x Quantity P(X); VC = 6X
C = FC + VC = 80,000 + 6X
Revenue (R) = (unit price) (quantity sold)
Thus, R = PX
From the problem, P = 10, so that,
R = PX = 10X
The expected profit ( ) is:
= N60, 000
To find the value of X, R= PX – (FC + VC *X)
60,000 = 10x – (80,000 + 6X)
Solving for X, we get:
10X - (80,000 + 6X) = 60,000
10X - 80,000 - 6X = 60,000
10X - 6X = 60,000 + 80,000
(10 - 6)X = 140,000
4X = 140,000
4X/4 = 140,000/4
X = 35,000
Therefore, 35,000 units must be sold to earn a profit of N60, 000.

Self-Assessment Exercise 2

1. Explain the Factorization methods of quadratic equation
2. Solve for x in the quadratic equation: $X^2 + X – 12 = 0$
3. Solve the equation using formula method. $4x^2 -17x+15 = 0$

https://www.cuemath.com/algebra/

5.7 Summary

This unit discussed the basic Define the Basic Algebra and the basic rules. The unit displayed Linear equations, Quadratic formula (Factorization), Calculate the Quadratic formula (Formula Method) and Application of Quadratic formula to decision making.

5.8 Possible Answers to SAEs

Answers to SAEs 1

1. Calculate x - 5 = 2. Five less than a number equals to two. What is the number?
   Solution:
   Using the concept of Algebra, we will assume the number to be a variable. Let the number be x. As per the question, we can write x - 5 = 2. On solving this, we get x = 7. Therefore, the required number is 7.

2. Expand $(2x + 3y)^2$ using the algebraic identities.
   Solution:
   Here we shall use an identity in algebra, $(a + b)^2 = a^2 + 2ab + b^2$
   (2x + 3y)2 = (2x)2 + 2(2x)(3y) + (3y)2
   $= 4x^2 + 12xy + 9y^2$
   Therefore, the answer is $(2x + 3y)^2 = 4x^2 + 12xy + 9y^2$

3. The present age of a person is double the age of his son. Ten years ago, his age was four times the age of his son. Use the concept of algebra and find the present age of the son.
   Solution:
   Let us consider the present age of the son as 'x' years. It is given that the age of the person is double the age of his son, so the age of the person is '2x' years. Now considering the situation 10 years ago, the age of the son was (x - 10) years and the age of the person was (2x - 10) years. The question says that 10 years ago the age of the person was 4 times the age of his son. Therefore, this can be expressed as,
   2x - 10 = 4(x - 10)
   2x - 10 = 4x - 40
   2x - 4x = -40 + 10
   -2x = -30
   2x = 30
   x = 30/2
   x = 15
   Therefore, the present age of the son is 15 years.

Answers to SAEs 2

1. Explain the Factorization methods of quadratic equation
   This involves the determination of the factors that form the given quadratic equation. This will then make the solutions for the unknown easy to come by.
   Solve for x in the quadratic equation: $X^2 + X – 12 = 0$
   Solution
   $X^2 + X – 12 = 0$
   (X – 3) and (X + 4) are the factors, so that
   (X - 3)(X+ 4) = 0
2. Solve the equation using formula method
   $4x^2 -17x+15 = 0$
   Solution:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Using the formula, a = 4; b = -17; c = 15. By substitution to the formula, we get:

X = $\underline{\text{-(-17)+\_ (-17)2-4(4)(15)}}$
2(4)
17$\underline{+}$ 289-240/8
17$\underline{+}$ 49/8
X = 17+7 or X = 17-7
X = 3 or X = 1.25

MODULE 3

Unit 1 Population vs. Sample

Unit Structure

1.1 Introduction

A population is the entire group that you want to draw conclusions about A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population (Scribbr.com (2022).

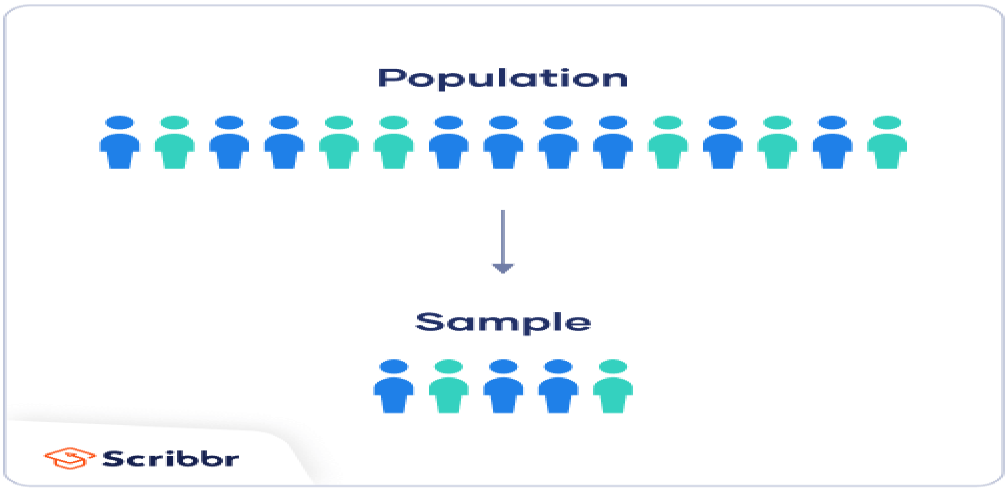
1.2 Learning Outcomes

By the end of this unit, you will be able to:

- state the meaning of Population
- explain Population vs sample
- collecting data from a population
- collecting data from a sample
- seasons for sampling

1.3 Meaning of Population

In research, a population doesn't always refer to people. It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms, etc (Scribbr.com (2022).



Adapted from Scribbr.com (2022).

1.3.1 Population vs sample

Population	**Sample**
Advertisements for IT jobs in the Netherlands	The top 50 search results for advertisements for IT jobs in the Netherlands on May 1, 2020
Songs from the Eurovision Song Contest	Winning songs from the Eurovision Song Contest that were performed in English
Undergraduate students in the Netherlands	300 undergraduate students from three Dutch universities who volunteer for your psychology research study
All countries of the world	Countries with published data available on birth rates and GDP since 2000

1.3.2 Collecting data from a population

Populations are used when your research question requires, or when you have access to, data from every member of the population.

Usually, it is only straightforward to collect data from a whole population when it is small, accessible and cooperative.

Example:
Collecting data from a population
A high school administrator wants to analyze the final exam scores of all graduating seniors to see if there is a trend. Since they are only interested in applying their findings to the graduating seniors in this high school, they use the whole population dataset.

For larger and more dispersed populations, it is often difficult or impossible to collect data from every individual. For example, every 10 years, the federal US government aims to count every person living in the country using the US Census. This data is used to distribute funding across the nation.

However, historically, marginalized and low-income groups have been difficult to contact, locate and encourage participation from. Because of non-responses, the population count is incomplete and biased towards some groups, which results in disproportionate funding across the country.

In cases like this, sampling can be used to make more precise inferences about the population.

Self-Assessment Exercises 1

1. Define Population
2. What is the Method of Collecting data from a population

1.3.3 Collecting data from a sample

When your population is large in size, geographically dispersed, or difficult to contact, it's necessary to use a sample. With statistical analysis, you can use sample data to make estimates or test hypotheses about population data.

Example:
Collecting data from a sample you want to study political attitudes in young people. Your population is the 300,000 undergraduate students in the Netherlands. Because it's not practical to collect data from all of them, you use a sample of 300 undergraduate volunteers from three Dutch universities – this is the group who will complete your online survey.

Ideally, a sample should be randomly selected and representative of the population. Using probability sampling methods (such as simple random sampling or stratified sampling) reduces the risk of sampling bias and enhances both internal and external validity.

For practical reasons, researchers often use non-probability sampling methods. Non-probability samples are chosen for specific criteria; they may be more convenient or cheaper to access. Because of non-random selection methods, any statistical inferences about the broader population will be weaker than with a probability sample.

1.3.4 Reasons for Sampling

i. Necessity: Sometimes it's simply not possible to study the whole population due to its size or inaccessibility.
ii. Practicality: It's easier and more efficient to collect data from a sample.
iii. Cost-effectiveness: There are fewer participant, laboratory, equipment, and researcher costs involved.
iv. Manageability: Storing and running statistical analyses on smaller datasets is easier and reliable.

Self-Assessment Exercises 2

1. Explain the method of Collecting data from a sample
2. Explain the Reasons for sampling

1.4 Summary

In research, a population doesn't always refer to people. It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms collecting data from a population.

Populations are used when your research question requires, or when you have access to, data from every member of the population.

Usually, it is only straightforward to collect data from a whole population when it is small, accessible and cooperative.

Collecting data from a sample: When your population is large in size, geographically dispersed, or difficult to contact, it's necessary to use a sample. With statistical analysis, you can use sample data to make estimates or test hypotheses about population data

Reasons for sampling

Necessity: Sometimes it’s simply not possible to study the whole population due to its size or inaccessibility.

Practicality: It’s easier and more efficient to collect data from a sample.

Cost-effectiveness: There are fewer participant, laboratory, equipment, and researcher costs involved.

Manageability: Storing and running statistical analyses on smaller datasets is easier and reliable

1.8 Possible Answers to SAEs

Answers to SAEs 1

1. In research, a population doesn’t always refer to people. It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms
2. Populations are used when your research question requires, or when you have access to, data from every member of the population.
   Usually, it is only straightforward to collect data from a whole population when it is small, accessible and cooperative.

Answers to SAEs 1

1. When your population is large in size, geographically dispersed, or difficult to contact, it’s necessary to use a sample. With statistical analysis, you can use sample data to make estimates or test hypotheses about population data
2. Reasons for sampling. It is Necessity that sometimes it’s simply not possible to study the whole population due to its size or inaccessibility.
3. Practicality: It’s easier and more efficient to collect data from a sample.
4. Cost-effectiveness: There are fewer participant, laboratory, equipment, and researcher costs involved.
5. Manageability: Storing and running statistical analyses on smaller datasets is easier and reliable

Unit 2 Correlation Analysis

Unit Structure

2.1 Introduction

The simplest methods of measuring relationships existing between economic variables are correlation analysis and regression analysis. Correlation can be defined as the degree of relationship between two or more variables. The degree of relationship between two variables is called simple correlation. The degree of relationship existing among three or more variables is called multiple correlations.

Correlation may be linear for scatter diagram on the values of two variables, (X and Y) are clustered near a straight line, or nonlinear, when all points on the scatter lie near a curve.

Two variables may have a positive correlation or a negative correlation, or they may be unrelated. These correlations are represented are:
Positive linear correlation
Positive non- linear correlation

2.2 Learning Outcomes

By the end of this unit, you will be able to:

- describe the computation of linear correlation coefficients
- explain the computation of rank correlation coefficients
- explain the computation of partial correlation coefficients
- apply the concept of correlations in business decisions.

2.3 Correlation

2.3.1 Linear Correlation

We can determine the kind of correlation between two variables by direct observations. If the points lie close to the line, the correlation is strong. A greater dispersion of points about the line implies weaker correlation.

Simple linear correlation is a measure of the degree to which two variables vary together, or a measure of the intensity of the association between two variables. Correlation often is abused. You need to show that one variable actually is affectinganother variable.

The parameter being measure is □ (rho) and is estimated by the statistic r, the correlation coefficient. r can range from -1 to 1, and is independent of units of measurement.

The strength of the association increases as r approaches the absolute value of 1.0
A value of 0 indicates there is no association between the two variables tested.

A better estimate of r usually can be obtained by calculating r on treatment meansaveraged across replicates.
Correlation does not have to be performed only between independent and dependentvariables.

Correlation can be done on two dependent variables.
We use a parameter to refer to as the correlation coefficient. The sample estimate of this parameter is referred to as r.

Rank Correlation Coefficient is used for qualitative variables, whereby the variables cannot be measured numerically.

Examples of such variables include profession, education, preferences for a particular brand of commodity and the like.

A partial correlation coefficient measures the relationship between any two variables, keeping other variables constant.

The limitations of linear correlations as a technique for the study of economic relations are as follows:

1. The formula for correlation coefficient applies only to linear relationships between variables.

2. That correlation coefficient as a measure of co-variability of variables does not imply any functional relationship between the variables concerned.

The X and Y in the equation to determine r do not necessarily correspond between aindependent and dependent variable, respectively

$$r_{xy} = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{n\Sigma(X_i - \overline{X})^2}\sqrt{\Sigma(Y_i - \overline{Y})^2}}$$

Example 1

X	Y	XY
41	52	2132
73	95	6935
67	72	4824
37	52	1924
58	96	5568
∑X =	∑Y = 367	∑XY =21,383
∑X^2 = 16,232	∑Y^2 = 28,833	n = 5

Step 1: Calculate SSCP
SSCP =21,383 = (276)(367)/5^{-}= 1124.6
Step 2: Calculate SS X
SSCP =16,232= (276)^2/5 = 996.8
Step 3: Calculate SS Y
SSCP =367= (28,833)^2/5= 1895.2
Step 4: Calculate the correlation coefficient r.

$$\frac{SSCP}{\sqrt{(SSX)(SSY)}} \qquad \frac{1124.6}{\sqrt{(996.8)(1895.2)}}$$

r=0.818

Self-Assessment Exercise 1

1. Describe the computation of linear correlation coefficients
2. Explain the computation of rank correlation coefficients

2.3.2 Testing the Hypothesis

That an Association Between X and Y Exists
To determine if an association between two variables exists as determined usingcorrelation, the following hypotheses are tested:
H_o: □ = 0H_A: □ ≠0

Notice that this correlation is testing to see if *r* is significantly different from zero, i.e., there is an association between the two variables evaluated.

You are not testing to determine if there is a "SIGNIFICANT CORRELATION". This cannot be tested.
Critical or tabular values of *r* to test the hypothesis H_o: □ = 0 can be found in tables,in which:
The df are equal to n-2
The number of independent variables will equal one for all simple linear correlation.

The tabular *r*-value, $r_{.05,\ 3\ \mathrm{df}}$ = 0.878
Because the calculated *r* (.818) is less than the table *r* value (.878), we fail to reject H_o: □ = 0 at the 95% level of confidence. We can conclude that there is no association between X and Y.

In this example, it would appear that the association between X and Y is strong because the *r* value is fairly high. Yet, the test of H_o: □ = 0 indicates that there is nota linear relationship.

Points to Consider

The tabular *r* values are highly dependent on n, the number of observations.
As n increases, the tabular r value decreases.
We are more likely to reject Ho: □ = 0 as n increases.

AS an approaches 100, the r value to reject Ho: □ = 0 becomes fairly small. Too many people abuse correlation by not reporting the r value and stating incorrectly that there is a "significant correlation". The failure to accept Ho: □ = 0 says nothing about the strength of the association between the two variables measured.

The correlation coefficient squared equals the coefficient of determination. Yet, you need to be careful if you decide to calculate *r* by taking the square root of the coefficient of determination. You may not have the correct "sign" is there is a negative association between the two variables.

2.3.3 Partial Correlations

A partial correlation coefficient measures the relationship between any two variables, keeping other variables constant.

Assume a multiple relationship between three variables, X1, X2, and X3.

To measure the true correlation between X1 and X2, we find the partial correlation coefficient between X1 and X2, keeping X3 constant.

The partial correlation coefficient is determined in terms of the simple correlation coefficients among the various variables involved in a multiple relationship. For the three variables, X1, X2 and X3, three simple correlation coefficients are involved as follows:

r12 = correlation coefficient between X1 and X2
r13 = correlation coefficient between X1 and X3
r23 = correlation coefficient between X2 and X3.

There are also two partial correlation coefficients involved:
(1) =.3 = partial correlation coefficient between X1 and X2, keeping X3 constant
r12□ (r13)(r23)[1□ (r13) 2][1□ (r23) 2]
r13.2 = partial correlation coefficient between X1 and X3, keeping
X2 constant:
r13.2 = r13□ (r13)(r23)[1□ (r13) 2][1□ (r23) 2

The formula for partial correlation coefficient can be extended to relationships involving any number of explanatory variables.

Assume X is the independent variable and Y is the dependent variable, *n* = 150, and the correlation between the two variables is *r* = 0.30. This value of r is significantly different from zero at the 99% level of confidence.

Calculating r^2 using r, $0.30^2 = 0.09$, we find that 9% of the variation in Y can be explained by having X in the model. This indicates that even though the *r* value is significantly different from zero, the association between X and Y is weak.

Some people feel the coefficient of determination needs to be greater that 0.50 (i.e. *r* = 0.71) before the relationship between X an Y is very meaningful.

Calculating *r* Combined Across Experiments, Locations, Runs, etc.
This is another area where correlation is abused.
https://www.ndsu.edu/faculty/horsley/Corr_revised.pdf

When calculating the “pooled” correlation across experiments, you cannot just put the data into one data set and calculate *r* directly. The value of r that will be calculated is not a reliable estimate of □.

A better method of estimating □ would be to:
Calculate a value of *r* for each environment, and
Average the *r* values across environments.

The proper method of calculating a pooled r value is to test the homogeneity of the correlation coefficients from the different locations. If the r values are homogenous, a pooled r value can be calculated.

Example
The correlation between grain yield and kernel plumpness was 0.43 at Langdon, ND;
0.32 at Prosper, ND; and 0.27 at Carrington, ND. There were 25 cultivars evaluated ateach location.

Step 1: Make and complete the following table

Location	N	r_i	Z'_i	$Z'_i - Z'_w$	$(n_i-3)(Z'_i - Z'_w)^2$
Langdon, ND	25	0.43	0.460	0.104	0.238
Prosper, ND	25	0.32	0.332	-0.024	0.013
Carrington, ND	25	0.27	0.277	-0.079	0.137
	$\sum n_i$=75		Z'_w = 0.356		χ^2 = 0.388

Step 2: Look up tabular χ^2 value at the α = 0.005 level. $\chi^2$0.005, 2 df = 10.6
Step 3: Make conclusions
Because the calculated χ^2 (0.388) is less than the table χ^2 value (10.6), we fail to reject the null hypothesis that the *r*-values from the three locations are equal
Step 4: Calculate pooled *r* (r_p) value
Step 5: Determine if r_p is significantly different from zero using a confidence interval.

Self-Assessment Exercise 2

1. State the step for testing hypotheses
2. Explain the an Association Between X and Y Exists in testing hypothesis

2.4 Summary

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

Three basic types of correlations were discussed in this unit: (i) the simple linear correlation; (ii) rank correlation; and, (iii) partial correlation. You should note that partial correlations involve more than two different variables. Here, the correlation between two identified variables, are obtained holding other variables constant

2.7 References/Further Readings/Web Resources

A. Francis (1998). *Business Mathematics and Statistics, 5th edition. Great Britain:* Ashford Colour Press.

ndsu.edu (2012) https://www.ndsu.edu/faculty/horsley/Corr_revised.pdf

2.8 Possible Answers to SAEs

Answers to SAEs 1

1. Linear Correlation
   We can determine the kind of correlation between two variables by direct observations. If the points lie close to the line, the correlation is strong. A greater dispersion of points about the line implies weaker correlation.

2. Simple linear correlation is a measure of the degree to which two variables vary together, or a measure of the intensity of the association between two variables. Correlation often is abused. You need to show that one variable actually is affecting another variable.

   The parameter being measure is □ (rho) and is estimated by the statistic r, the correlation coefficient. r can range from -1 to 1, and is independent of units of measurement.

The strength of the association increases as r approaches the absolute value of 1.0

A value of 0 indicates there is no association between the two variables tested.

A better estimate of r usually can be obtained by calculating r on treatment meansaveraged across replicates.

Correlation does not have to be performed only between independent and dependentvariables.

Correlation can be done on two dependent variables.

We use a parameter□□ referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.

Rank Correlation Coefficient is used for qualitative variables, whereby the variables cannot be measured numerically.

Examples of such variables include profession, education, preferences for a particular brand of commodity and the like.

A partial correlation coefficient measures the relationship between any two variables, keeping other variables constant.

The limitations of linear correlations as a technique for the study of economic relations are as follows:

1. The formula for correlation coefficient applies only to linear relationships between variables.
2. That correlation coefficient as a measure of co-variability of variables does not imply any functional relationship between the variables concerned.

The X and Y in the equation to determine r do not necessarily correspond between aindependent and dependent variable, respectively

$$r_{xy} = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{n\Sigma(X_i - \overline{X})^2}\,\sqrt{\Sigma(Y_i - \overline{Y})^2}}$$

Answers to SAEs 2

1. State the step for testing hypotheses

Step 1: Make and complete the following table

Step 2: Look up tabular χ^2 value at the $\alpha = 0.005$ level. $\chi^2$0.005, 2 df = 10.6

Step 3: Make conclusions

Because the calculated χ^2 (0.388) is less than the table χ^2 value (10.6), we fail to reject the null hypothesis that the *r*-values from the three locations are equal

Step 4: Calculate pooled *r* (r_p) value

Step 5: Determine if r_p is significantly different from zero using a confidence interval.

2. Explain the an Association Between X and Y Exists in testing hypothesis

That an Association Between X and Y Exists

To determine if an association between two variables exists as determined usingcorrelation, the following hypotheses are tested:

H_o: □ = 0H_A: □ ≠0

Notice that this correlation is testing to see if *r* is significantly different from zero, i.e., there is an association between the two variables evaluated.

You are not testing to determine if there is a “SIGNIFICANT CORRELATION”. This cannot be tested.

Critical or tabular values of *r* to test the hypothesis H_o: □ = 0 can be found in tables,in which:

The df are equal to n-2

The number of independent variables will equal one for all simple linear correlation.

The tabular *r*-value, $r_{.05,\ 3\ \mathrm{df}}$ = 0.878

Because the calculated *r* (.818) is less than the table *r* value (.878), we fail to rejectH$_o$: □ = 0 at the 95% level of confidence. We can conclude that there is no association between X and Y.

In this example, it would appear that the association between X and Y is strong because the *r* value is fairly high. Yet, the test of H_o: □ = 0 indicates that there is nota linear relationship.

Unit 3 Simple Linear Regression

Unit Structure

3.1 Introduction

Simple linear regression is a statistical method you can use to understand the relationship between two variables, x and y. One variable, **x**, is known as the predictor variable. The other variable, **y**, is known as the response variable.

3.2 Learning Outcomes

By the end of this unit, you will be able to:

- define and calculate the Linear Regression
- calculate Simple Linear Regression
- finding the "Line of Best Fit"
- calculate The Coefficient of Determination

3.3 Linear Regression

Simple linear regression is a statistical method you can use to understand the relationship between two variables, x and y.

3.3.1 Simple Linear Regression

For example, suppose we have the following dataset with the weight and height of seven individuals:

$$\Sigma\left(X-\bar{X}_2\right)^2 = 240,000.0$$

Let *weight* be the predictor variable and let *height* be the response variable.

If we graph these two variables using a scatterplot, with weight on the x-axis and height on the y-axis, here's what it would look like:

$$\Sigma\left(X-\bar{X}_3\right)^2=449,\ 750.0$$

Suppose we're interested in understanding the relationship between weight and height. From the scatterplot we can clearly see that as weight increases, height tends to increase as well, but to actually quantify this relationship between weight and height, we need to use linear regression.

Using linear regression, we can find the line that best "fits" our data. This line is known as the least squares regression line and it can be used to help us understand the relationships between weight and height. Usually you would use software like Microsoft Excel, SPSS, or a graphing calculator to actually find the equation for this line.

The formula for the line of best fit is written as:
ŷ = b_0 + b_1x
where ŷ is the predicted value of the response variable, b_0 is the y-intercept, b_1 is the regression coefficient, and x is the value of the predictor variable.

3.4 Finding the "Line of Best Fit"

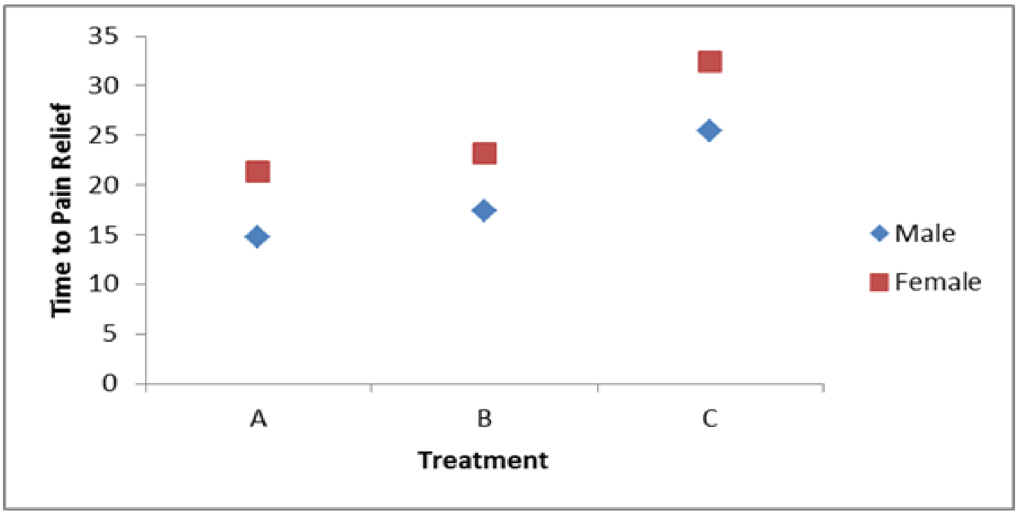
For this example, we can simply plug our data into the Statology Linear Regression Calculator and hit Calculate:

$$\text{SSE}=\Sigma\Sigma\left(X-\bar{X}_j\right)^2=130,\ 083.3+240,\ 000.0+449,\ 750.0=819,\ 833.3$$

The calculator automatically finds the least squares regression line**:**

ŷ = 32.7830 + 0.2001x

If we zoom out on our scatterplot from earlier and added this line to the chart, here's what it would look like:



Notice how our data points are scattered closely around this line. That's because this least squares regression lines is the best fitting line for our data out of all the possible lines we could draw.

How to Interpret a Least Squares Regression Line
Here is how to interpret this least squares regression line: ŷ = 32.7830 + 0.2001x
b0 = 32.7830. This means when the predictor variable weight is zero pounds, the predicted height is 32.7830 inches. Sometimes the value for b0 can be useful to know, but in this specific example it doesn't actually make sense to interpret b0 since a person can't weight zero pounds.

b1 = 0.2001. This means that a one unit increase in x is associated with a 0.2001 unit increase in y. In this case, a one pound increase in weight is associated with a 0.2001 inch increase in height.

How to Use the Least Squares Regression Line
Using this least squares regression line, we can answer questions like:
For a person who weighs 170 pounds, how tall would we expect them to be?
To answer this, we can simply plug in 170 into our regression line for x and solve for y:
ŷ = 32.7830 + 0.2001(170) = **66.8 inches**
For a person who weighs 150 pounds, how tall would we expect them to be?
To answer this, we can plug in 150 into our regression line for x and solve for y:
ŷ = 32.7830 + 0.2001(150) = **62.798 inches**

Caution: When using a regression equation to answer questions like these, make sure you only use values for the predictor variable that are within the range of the predictor variable in the original dataset we used to generate the least squares regression line. For example, the weights in our dataset ranged from 140 lbs to 212 lbs, so it only makes sense to answer questions about predicted height when the weight is between 140 lbs and 212 lbs.

Self-Assessment Exercise 1

1. Suppose we have the following dataset with the weight and height of seven individuals. Calculate.
2. Calculate the Statology Linear Regression

3.5 The Coefficient of Determination

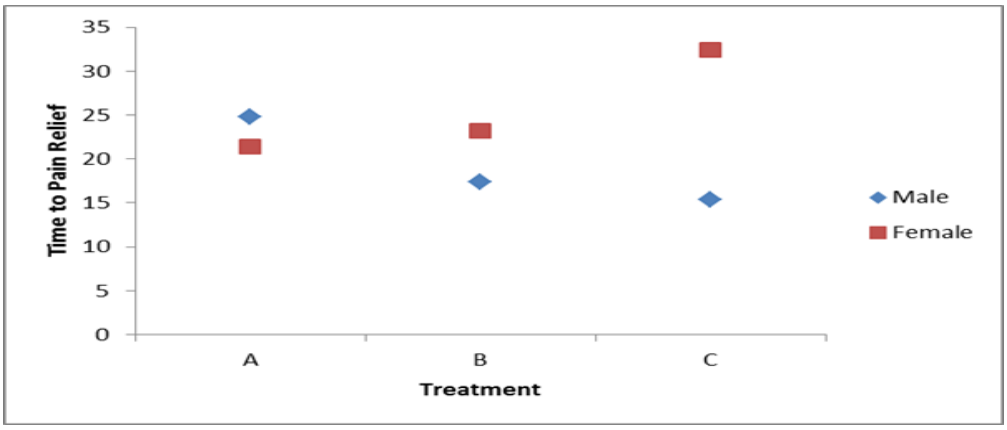
One way to measure how well the least squares regression line "fits" the data is using the coefficient of determination, denoted as R2.

The coefficient of determination is the proportion of the variance in the response variable that can be explained by the predictor variable.

The coefficient of determination can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

An R2 between 0 and 1 indicates just how well the response variable can be explained by the predictor variable. For example, an R2 of 0.2 indicates that 20% of the variance in the response variable can be explained by the predictor variable; an R2 of 0.77 indicates that 77% of the variance in the response variable can be explained by the predictor variable.

Notice in our output from earlier we got an R2 of 0.9311, which indicates that 93.11% of the variability in height can be explained by the predictor variable of weight:



This tells us that weight is a very good predictor of height.

Assumptions of Linear Regression

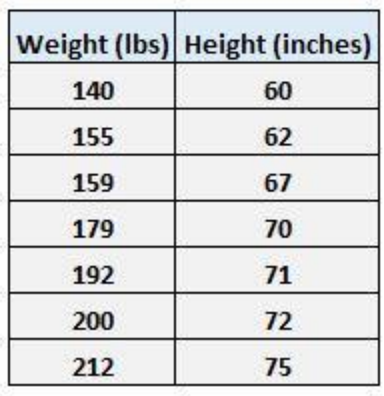
For the results of a linear regression model to be valid and reliable, we need to check that the following four assumptions are met:

i. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
ii. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
iii. Homoscedasticity: The residuals have constant variance at every level of x.
iv. Normality: The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

A linear function

A linear function is represented in a data. Here are two choices.



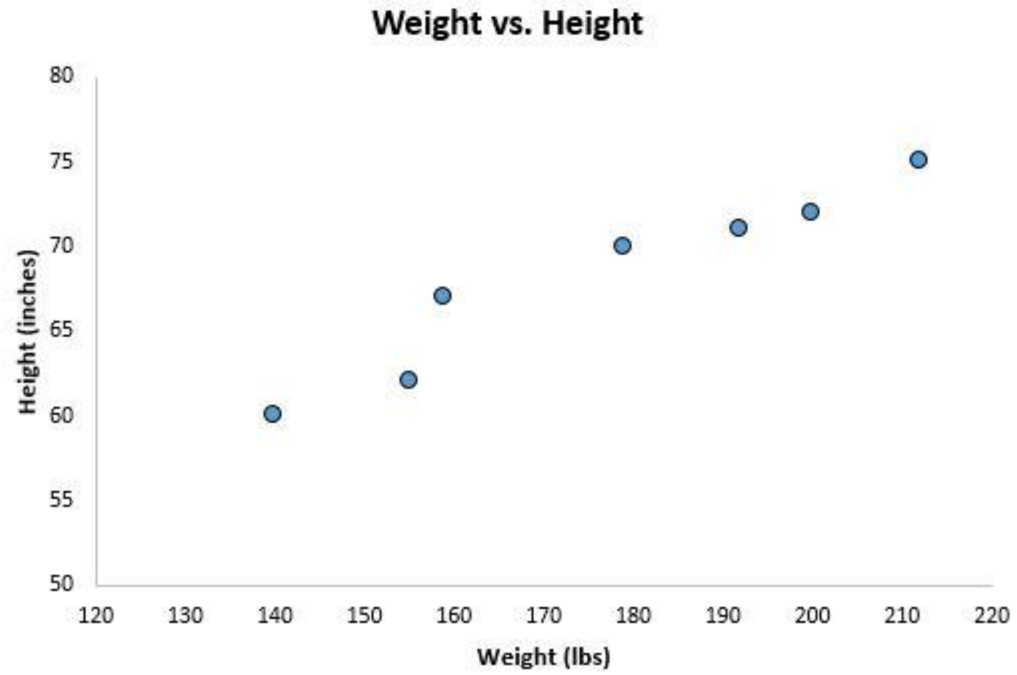
Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

Which is better? The red line or the blue one? How do you decide? Well, you have to make up some criteria for choosing the best line.

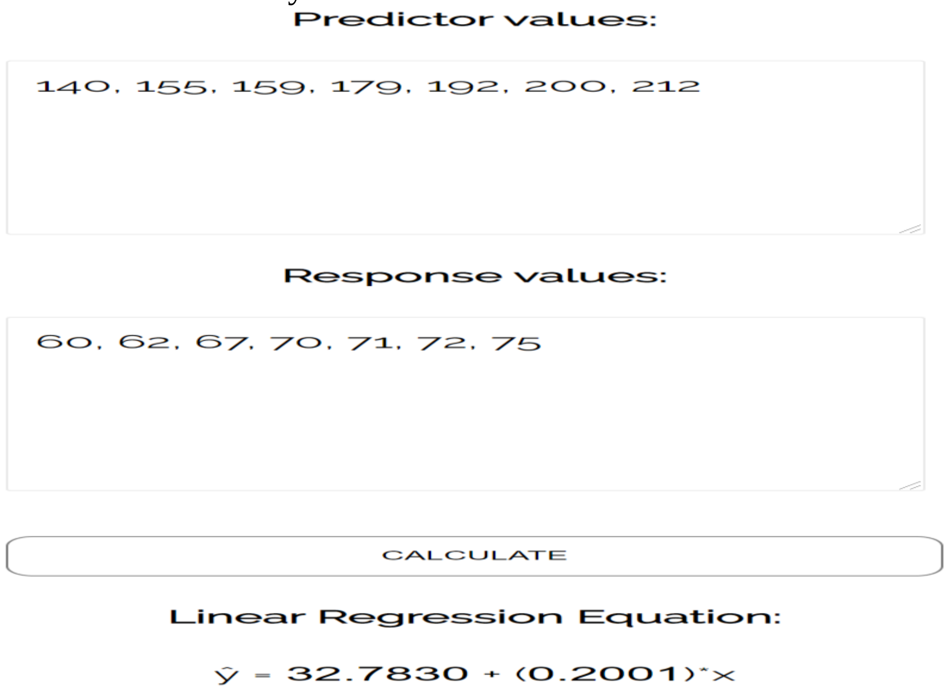
Commonly, it is chosen to pick the line such that the value of the sum of d^2 is minimized. I displayed these *d* values on the graph for you.

Notice that they are the vertical distance from the real data points to the fitting linear function. Why this way? Well, typically, the horizontal variable is your independent variable - so these might be some set values. The vertical data is typically the one with the most error (but not always). You could instead look at the horizontal distance from the data or even the perpendicular.

I don't want to add up these vertical distances because some will be positive and some negative. Instead, I will add up this vertical distance squared such that:



So, let me assume that my best fit linear function has the form:



Predictor values:

140, 155, 159, 179, 192, 200, 212

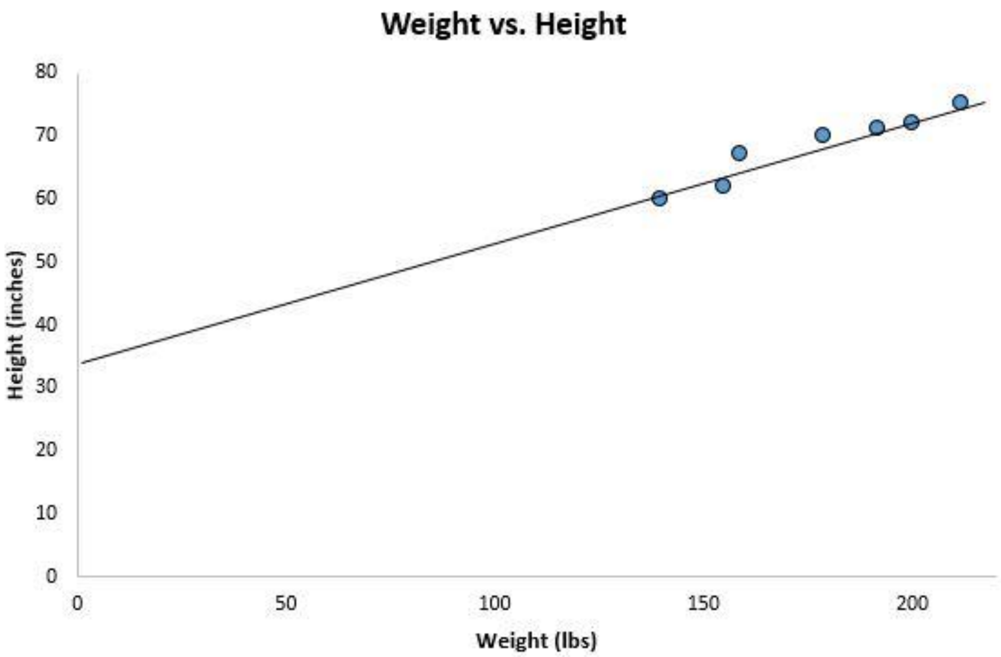
Response values:

60, 62, 67, 70, 71, 72, 75

Linear Regression Equation:

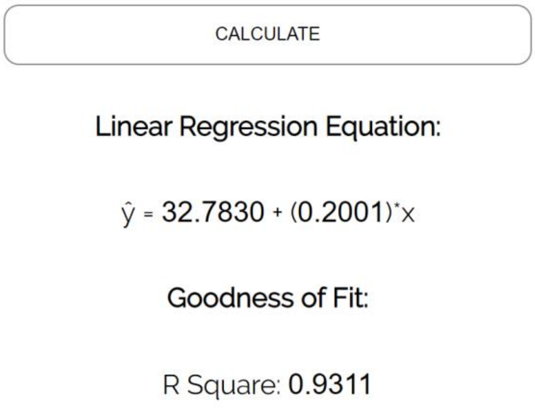
$\hat{y} = 32.7830 + (0.2001)*x$

Let me generically label the data as (x_i, y_i). So, I can write d_i and d_i^2 as:



Well that is just great. Now what? If I let S be the sum of the square of the distances, then I want to pick a line such that S is the smallest. Hint: this is where the term 'least squares fit' comes from. How do you minimize a function? The simple answer is to change the parameters *m* and *b*.

Let me pretend that I changed the parameter *m* and each time calculated the sum of the vertical distances squared (S). Suppose I then made a plot of S for the different values of *m* and it looks like this:



Linear Regression Equation:

$\hat{y}$ = 32.7830 + (0.2001)*x

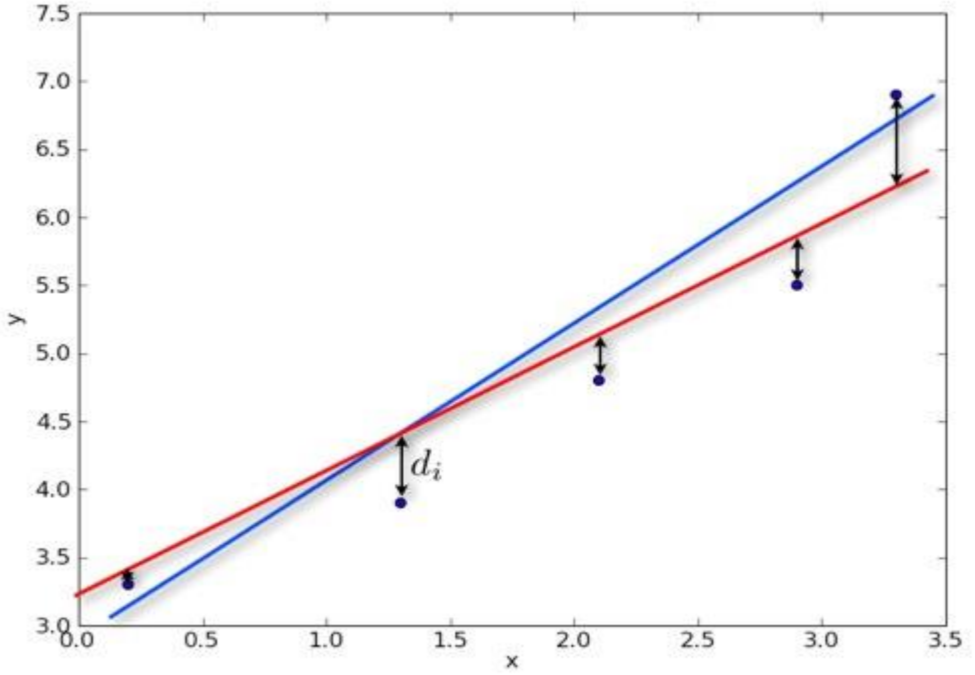
Goodness of Fit:

R Square: 0.9311

On this graph, which labeled point (a - d) is S at a minimum? Go ahead. You can say it. How many of you said 'c'? Well, you would be right.

But, how do you find that lowest point without making a graph? There is one important thing about the lowest point. Right before that lowest point, the function is decreasing. Right after that lowest point, the function is increasing. And so AT the lowest point the function is neither increasing nor decreasing (with respect to changing *m*). Of course, I am talking about the slope of this function. I can find this lowest point by finding where the slope (the derivative with respect to *m*) is zero.

I know, I know. It is possible for a function to have a zero slope and NOT be a minimum. Let me proceed anyway (assuming the only location with a zero slope is a min). There are two things that I can change to get S to be a minimum - *m* and *b*. Let me assume that I can just vary one parameter at time (this means that I can use the partial derivative instead of the full derivative). Here is the partial derivative of S with respect to *m* - note that for sums I will leave off the "i = 1 to n part".



That is the slope. I will set it equal to zero and I get (divide both sides by that pesky -2):

$$S = \sum_{i=0}^{n} d_i^2$$

Now to do a similar thing with how S changes with the parameter *b*.

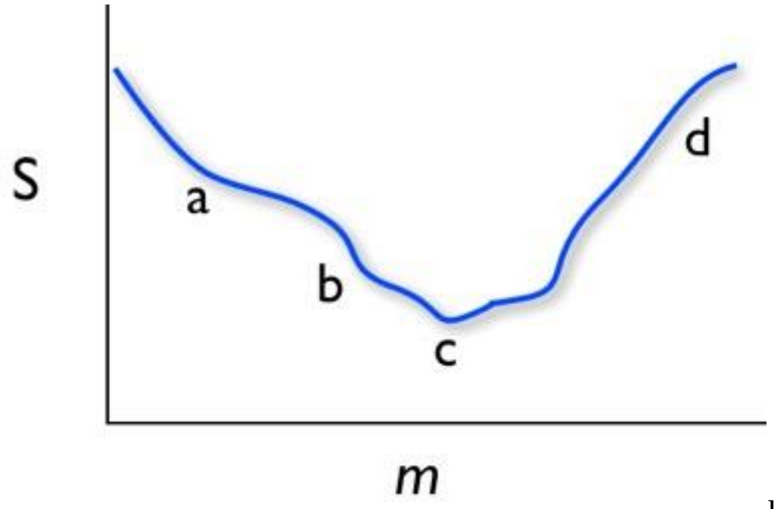
$$y = mx + b$$

And again, setting it equal to zero (and dividing both sides by -2):

$$d_i = y_i - y = y_i - mx_i - b$$
$$d_i^2 = (y_i - mx_i - b)^2$$

Now there are two equations and two unknowns (*m* and *b*). The *n* is the number of data points. All the other stuff (like the sum over x_i) are technically known. What I want to do next it solve for *m* and *b*.

It should be obvious that I skipped some of the algebraic steps. They aren't too difficult. You should be able to go through them yourself.



But, now that I have an expression for *b* and *m*, what to do? Well, if I know all the x and y data points, I can just calculate *m* and then *b* (since I left *b* in terms of *m*). If I don't have too many data points, I could do this by hand. Or I could do it in python - or I could do it in a spread sheet. Randomly, I will choose to do this in a spreadsheet.

Here is that spreadsheet with the same data and with the SLOPE and INTERCEPT function in google docs to show the answer is the same.
There. That is the basic form of linear regression by hand. Note that there are other ways to do this - more complicated ways (assuming different types of distributions for the data). Also, the same basic idea is followed if you want to fit some higher order polynomial. Warning, it gets complicated (algebraically) real quick.

Self-Assessment Exercise 2

1. Explain the Coefficient of Determination
2. State the Assumptions of Linear Regression

3.6 Summary

The unit explains that, using linear regression, we can find the line that best "fits" our data. This line is known as the least squares regression line and it can be used to help us understand the relationships between weight and height. Usually you would use software like Microsoft Excel, SPSS, or a graphing calculator to actually find the equation for this line.

The formula for the line of best fit is written as:
$\hat{y} = b_0 + b_1x$
Regression line "fits" the data is using the coefficient of determination, denoted as R2.

The coefficient of determination is the proportion of the variance in the response variable that can be explained by the predictor variable.

The coefficient of determination can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

An R2 between 0 and 1 indicates just how well the response variable can be explained by the predictor variable. For example, an R2 of 0.2indicates that 20% of the variance in the response variable can be explained by the predictor variable; an R2 of 0.77 indicates that 77% of the variance in the response variable can be explained by the predictor variable.

For the results of a linear regression model to be valid and reliable, we need to check that the following four assumptions are met:

1. **Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. **Independence:** The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
3. **Homoscedasticity:** The residuals have constant variance at every level of x.
4. **Normality:** The residuals of the model are normally distributed.
   If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

3.7 References/Further Readings/Web Resources

Allain, R. (2012). **Linear regression** Southeastern Louisiana University. Retrieved from *https://www.wired.com/2011/01/linear-regression-by-hand*

3.8 Possible Answers to SAEs

Answers to SAEs 1

1. Simple Linear Regression

For example, suppose we have the following dataset with the weight and height of seven individuals:

$$\Sigma\left(X-\bar{X}_2\right)^2=240,\,000.0$$

Let *weight* be the predictor variable and let *height* be the response variable.

If we graph these two variables using a scatterplot, with weight on the x-axis and height on the y-axis, here's what it would look like:

$$\Sigma\left(X-\bar{X}_3\right)^2=449,\,750.0$$

Suppose we're interested in understanding the relationship between weight and height. From the scatterplot we can clearly see that as weight increases, height tends to increase as well, but to actually quantify this relationship between weight and height, we need to use linear regression.

Using linear regression, we can find the line that best "fits" our data. This line is known as the least squares regression line and it can be used to help us understand the relationships between weight and height. Usually you would use software like Microsoft Excel, SPSS, or a graphing calculator to actually find the equation for this line.

The formula for the line of best fit is written as:
$\hat{y} = b_0 + b_1x$
where $\hat{y}$ is the predicted value of the response variable, b_0 is the y-intercept, b_1 is the regression coefficient, and x is the value of the predictor variable.

2. Finding the "Line of Best Fit"

Calculate the Statology Linear Regression Calculator

$$\mathbf{SSE} = \Sigma\Sigma\left(X - \bar{X}_j\right)^2 = \mathbf{130,\ 083.3 + 240,\ 000.0 + 449,\ 750.0 = 819,\ 833.3}$$

The calculator automatically finds the least squares regression line**:**

ŷ = 32.7830 + 0.2001x

Answers to SAEs 2

1. Explain the Coefficient of Determination

One way to measure how well the least squares regression line "fits" the data is using the coefficient of determination, denoted as R2

The coefficient of determination is the proportion of the variance in the response variable that can be explained by the predictor variable.

The coefficient of determination can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

An R2 between 0 and 1 indicates just how well the response variable can be explained by the predictor variable. For example, an R2 of 0.2 indicates that 20% of the variance in the response variable can be explained by the predictor variable; an R2 of 0.77 indicates that 77% of the variance in the response variable can be explained by the predictor variable.

2. State the Assumptions of Linear Regression

For the results of a linear regression model to be valid and reliable, we need to check that the following four assumptions are met:

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
3. Homoscedasticity: The residuals have constant variance at every level of x.
4. Normality: The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

Unit 4 Multiple Linear Regressions

Unit Structure

4.1 Introduction

When we want to understand the relationship between a single predictor variable and a response variable, we often use simple linear regression. However, if we'd like to understand the relationship between multiple predictor variables and a response variable then we can instead use multiple linear regressions. Multiple linear regressions are a method we can use to quantify the relationship between two or more predictor variables and a response variable. This unit explains how to perform multiple linear regressions using manual.

4.2 Learning Outcomes

By the end of this unit, you will be able to:

- state the Multiple linear Regression Model
- interpret of model output
- state the assumptions of Multiple Linear Regressions
- learn how to Use SPSS
- learn the Manual Multiple Linear Regressions

4.3 Multiple linear Regressions

4.3.1 Multiple linear Regression Model

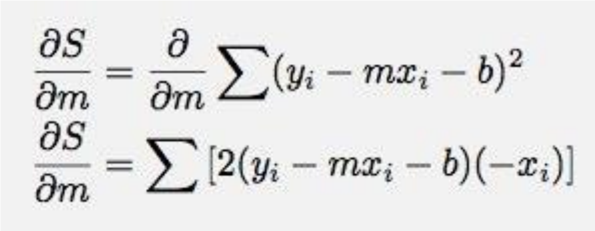
If we have *p* predictor variables, then a multiple linear regression model takes the form:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon}$$
Where:
Y: The response variable
$\mathbf{X_j}$: The jth predictor variable
$\boldsymbol{\beta_j}$: The average effect on Y of a one unit increase in X_j, holding all other predictors fixed
ε: The error term
The values for β_0, β_1, B_2,…, β_p are chosen using the least square method, which minimizes the sum of squared residuals (RSS):
$$\mathbf{RSS = \Sigma(y_i - \hat{y}_i)^2}$$
where:
Σ: A greek symbol that means *sum*
$\mathbf{y_i}$: The actual response value for the i^{th} observation
$\mathbf{\hat{y}_i}$: The predicted response value based on the multiple linear regression model

The method used to find these coefficient estimates relies on matrix algebra and we will not cover the details here. Fortunately, any statistical software can calculate these coefficients for you.

How to Interpret Multiple Linear Regression Output
Suppose we fit a multiple linear regression model using the predictor variables hours studied and prep exams taken and a response variable exam score.

The following screenshot shows what the multiple linear regression output might look like for this model:
Note: The screenshot below shows multiple linear regression output for Excel, but the numbers shown in the output are typical of the regression output you'll see using any statistical software.



$$\frac{\partial S}{\partial m} = \frac{\partial}{\partial m}\sum(y_i - mx_i - b)^2$$
$$\frac{\partial S}{\partial m} = \sum\left[2(y_i - mx_i - b)(-x_i)\right]$$

From the model output, the coefficients allow us to form an estimated multiple linear regression model:
Exam score = 67.67 + 5.56*(hours) – 0.60*(prep exams)
The way to interpret the coefficients is as follows:
Each additional one unit increase in hours studied is associated with an average increase of **5.56** points in exam score, *assuming prep exams is held constant.*

Each additional one unit increase in prep exams taken is associated with an average decrease of **0.60** points in exam score, *assuming hours studied is held constant.*

We can also use this model to find the expected exam score a student will receive based on their total hours studied and prep exams taken. For example, a student who studies for 4 hours and takes 1 prep exam is expected to score a=**89.31** on the exam:
Exam score = 67.67 + 5.56*(4) -0.60*(1) = **89.31.**

Self-Assessment Exercise 1

1. State the Multiple linear Regression Model
2. Interpret of model output

4.3.2 Interpretation of model output

Here is how to interpret the rest of the model output:

R-Square: This is known as the coefficient of determination. It is the proportion of the variance in the response variable that can be explained by the explanatory variables. In this example, 73.4% of the variation in the exam scores can be explained by the number of hours studied and the number of prep exams taken.

Standard error: This is the average distance that the observed values fall from the regression line. In this example, the observed values fall an average of 5.366 units from the regression line.

F: This is the overall F statistic for the regression model, calculated as regression MS / residual MS.

Significance F: This is the p-value associated with the overall F statistic. It tells us whether or not the regression model as a whole is statistically significant. In other words, it tells us if the two explanatory variables combined have a statistically significant association with the response variable. In this case the p-value is less than 0.05, which indicates that the explanatory variables hours studied and prep exams taken combined have a statistically significant association with exam score.

Coefficient P-values. The individual p-values tell us whether or not each explanatory variable is statistically significant. We can see that hours studied is statistically significant (p = 0.00) while prep exams taken (p = 0.52) is not statistically significant at α = 0.05. Since prep exams taken are not statistically significant, we may end up deciding to remove it from the model.

How to Assess the Fit of a Multiple Linear Regression Model
There are two numbers that are commonly used to assess how well a multiple linear regression model "fits" a dataset:
R-Squared: This is the proportion of the variance in the response variable that can be explained by the predictor variables.

The value for R-squared can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

The higher the R-squared of a model, the better the model is able to fit the data.

Standard Error: This is the average distance that the observed values fall from the regression line. The smaller the standard error, the better a model is able to fit the data.
If we're interested in making predictions using a regression model, the standard error of the regression can be a more useful metric to know than R-squared because it gives us an idea of how precise our predictions will be in terms of units.
For a complete explanation of the pros and cons of using R-squared vs. Standard Error for assessing model fit, check out the following articles:

4.3.3 Assumptions of Multiple Linear Regressions

There are four key assumptions that multiple linear regression makes about the data:

1. **Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. **Independence:** The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
3. **Homoscedasticity:** The residuals have constant variance at every level of x.
4. **Normality:** The residuals of the model are normally distributed
   For a complete explanation of how to test these assumptions, check on websites bellows for tutorials on step-by-step examples of how to perform multiple linear regressions using different statistical software:

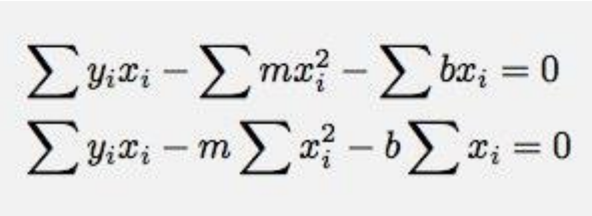
Multiple Linear Regressions Using Software:

i. How to Perform Multiple Linear Regression in SPSS
ii. How to Perform Multiple Linear Regression in R
iii. How to Perform Multiple Linear Regression in Python
iv. How to Perform Multiple Linear Regression in Excel
v. How to Perform Multiple Linear Regression in Stata
vi. How to Perform Linear Regression in Google Sheets

4.3.4 Manual Computation of Multiple Linear Regressions

Multiple Linear Regressions

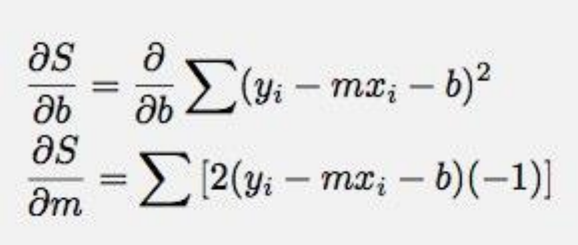
Suppose we have the following dataset with one response variable *y* and two predictor variables X_1 and X_2:



$$\sum y_i x_i - \sum m x_i^2 - \sum b x_i = 0$$
$$\sum y_i x_i - m \sum x_i^2 - b \sum x_i = 0$$

Use the following steps to fit a multiple linear regression model to this dataset.

Step 1: Calculate X_1^2, X_2^2, X_1y, X_2y and X_1X_2.



$$\frac{\partial S}{\partial b} = \frac{\partial}{\partial b} \sum (y_i - m x_i - b)^2$$
$$\frac{\partial S}{\partial m} = \sum [2(y_i - m x_i - b)(-1)]$$

Step 2: Calculate Regression Sums.

Next, make the following regression sum calculations:

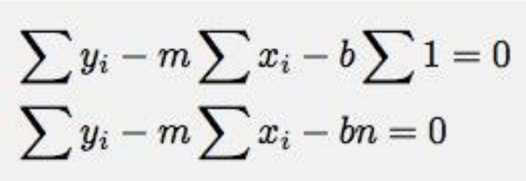
$\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n = 38{,}767 - (555)^2 / 8 =$ **263.875**

$\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n = 2{,}823 - (145)^2 / 8 =$ **194.875**

$\Sigma x_1 y = \Sigma X_1 y - (\Sigma X_1 \Sigma y) / n = 101{,}895 - (555*1{,}452) / 8 =$ **1,162.5**

$\Sigma x_2 y = \Sigma X_2 y - (\Sigma X_2 \Sigma y) / n = 25{,}364 - (145*1{,}452) / 8 =$ **-953.5**

$\Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1 \Sigma X_2) / n = 9{,}859 - (555*145) / 8 =$ **-200.375**



$$\sum y_i - m \sum x_i - b \sum 1 = 0$$
$$\sum y_i - m \sum x_i - bn = 0$$

Step 3: Calculate b_0, b_1, and b_2.

The formula to calculate b_1 is: $[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

Thus, $\mathbf{b_1}$ = $[(194.875)(1162.5) - (-200.375)(-953.5)] / [(263.875)(194.875) - (-200.375)^2]$ = **3.148**

The formula to calculate b_2 is: $[(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

Thus, $\mathbf{b_2}$ = $[(263.875)(-953.5) - (-200.375)(1152.5)] / [(263.875)(194.875) - (-200.375)^2]$ = **-1.656**

The formula to calculate b_0 is: $y - b_1 X_1 - b_2 X_2$

Thus, $\mathbf{b_0}$ = $181.5 - 3.148(69.375) - (-1.656)(18.125)$ = **-6.867**

Step 5: Place b_0, b_1, and b_2 in the estimated linear regression equation.

The estimated linear regression equation is: $\hat{y} = b_0 + b_1{*}x_1 + b_2{*}x_2$

In our example, it is $\boldsymbol{\hat{y} = -6.867 + 3.148x_1 - 1.656x_2}$

How to Interpret a Multiple Linear Regression Equation

Here is how to interpret this estimated linear regression equation: $\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$

$\mathbf{b_0}$ **= -6.867**. When both predictor variables are equal to zero, the mean value for y is -6.867.

$\mathbf{b_1}$ **= 3.148**. A one unit increase in x_1 is associated with a 3.148 unit increase in y, on average, assuming x_2 is held constant.

$\mathbf{b_2}$ **= -1.656**. A one unit increase in x_2 is associated with a 1.656 unit decrease in y, on average, assuming x_1 is held constant.

Self-Assessment Exercise 2

1. State the assumptions of Multiple Linear Regressions
2. Learn how to Use SPSS
3. Learn the Manual Multiple Linear Regressions

4.4 Summary

Multiple linear regressions are a method we can use to quantify the relationship between two or more predictor variables and a response variable. This unit explained how to perform multiple linear regressions using manual

If we have *p* predictor variables, then a multiple linear regression model takes the form:

$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon}$

Where:

Y: The response variable

$\mathbf{X_j}$: The j^th predictor variable

β_j: The average effect on Y of a one unit increase in X_j, holding all other predictors fixed
ε: The error term
The values for β_0, β_1, B_2, … , β_p are chosen using **the least square method**, which minimizes the sum of squared residuals (RSS):
RSS = $\Sigma(y_i - \hat{y}_i)^2$
where:
Σ: A greek symbol that means *sum*
y_i: The actual response value for the ith observation
$\hat{y}_i$: The predicted response value based on the multiple linear regression model.

4.5 References/Further Readings/Web Resources

Zach (2020). Introduction to Multiple Linear Regression. Retrieved from https://www.statology.org/multiple-linear-regression/

4.6 Possible Answers to SAEs

Answers to SAEs 1

1. Multiple linear Regression Model

If we have *p* predictor variables, then a multiple linear regression model takes the form:
$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_pX_p + \varepsilon$
Where:
Y: The response variable
X_j: The jth predictor variable
β_j: The average effect on Y of a one unit increase in X_j, holding all other predictors fixed
ε: The error term
The values for β_0, β_1, B_2, … , β_p are chosen using **the least square method**, which minimizes the sum of squared residuals (RSS):
RSS = $\Sigma(y_i - \hat{y}_i)^2$
where:
Σ: A greek symbol that means *sum*
y_i: The actual response value for the ith observation
$\hat{y}_i$: The predicted response value based on the multiple linear regression model

The method used to find these coefficient estimates relies on matrix algebra and we will not cover the details here. Fortunately, any statistical software can calculate these coefficients for you.

2. How to Interpret Multiple Linear Regression Output

Suppose we fit a multiple linear regression model using the predictor variables hours studied and prep exams taken and a response variable exam score.

The following screenshot shows what the multiple linear regression output might look like for this model:
Note: The screenshot below shows multiple linear regression output for Excel, but the numbers shown in the output are typical of the regression output you'll see using any statistical software.

Interpretation of model output
Here is how to interpret the rest of the model output:

R-Square: This is known as the coefficient of determination. It is the proportion of the variance in the response variable that can be explained by the explanatory variables. In this example, 73.4% of the variation in the exam scores can be explained by the number of hours studied and the number of prep exams taken.

Standard error: This is the average distance that the observed values fall from the regression line. In this example, the observed values fall an average of 5.366 units from the regression line.

F: This is the overall F statistic for the regression model, calculated as regression MS / residual MS.

Significance F: This is the p-value associated with the overall F statistic. It tells us whether or not the regression model as a whole is statistically significant. In other words, it tells us if the two explanatory variables combined have a statistically significant association with the response variable. In this case the p-value is less than 0.05, which indicates that the explanatory variables hours studied and prep exams taken combined have a statistically significant association with exam score.

Coefficient P-values. The individual p-values tell us whether or not each explanatory variable is statistically significant. We can see that hours studied is statistically significant ($p = 0.00$) while prep exams taken ($p = 0.52$) is not statistically significant at $\alpha = 0.05$. Since prep exams taken is not statistically significant, we may end up deciding to remove it from the model.

How to Assess the Fit of a Multiple Linear Regression Model

There are two numbers that are commonly used to assess how well a multiple linear regression model "fits" a dataset:

R-Squared: This is the proportion of the variance in the response variable that can be explained by the predictor variables.

The value for R-squared can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

The higher the R-squared of a model, the better the model is able to fit the data.

Standard Error: This is the average distance that the observed values fall from the regression line.

Answers to SAEs 2

1. Assumptions of Multiple Linear Regressions

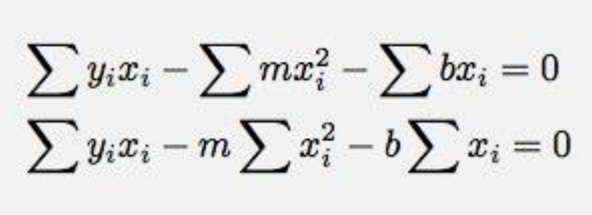
There are four key assumptions that multiple linear regression makes about the data:

i. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
ii. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
iii. Homoscedasticity: The residuals have constant variance at every level of x.
iv. Normality: The residuals of the model are normally distributed

2. Manual Computation of Multiple Linear Regressions

Multiple Linear Regressions

Suppose we have the following dataset with one response variable *y* and two predictor variables X_1 and X_2:



$$\sum y_i x_i - \sum m x_i^2 - \sum b x_i = 0$$
$$\sum y_i x_i - m \sum x_i^2 - b \sum x_i = 0$$

Use the following steps to fit a multiple linear regression model to this dataset.

Step 1: Calculate X_1^2, X_2^2, X_1y, X_2y and X_1X_2.

Unit 5 Spearman’s Rank Correlation

Unit Structure

1.1 Introduction

This unit will discuss the Spearman's Rank correlation coefficient, two variables of linear correlation and linear association between two variables is to use the Pearson Correlation Coefficient as well as the linear association between two variables using the Pearson Correlation Coefficient.

5.2 Learning Outcomes

By the end of this unit, you will be able to:

- state the meaning of Spearman's Rank correlation coefficient
- explain the Spearman’s rank correlation scenarios
- interpret the Statistical Software for Correlation Coefficients

5.3 Spearman's Rank Correlation Coefficient

5.3.1 Meaning of Spearman's Rank correlation coefficient

Spearman's Rank correlation coefficient is a technique which can be used to summarise the strength and direction (negative or positive) of a relationship between two variables.

The most common way to quantify the linear association between two variables is to use the Pearson Correlation Coefficient, which always takes on a value between -1 and 1 where:

-1 indicates a perfectly negative linear correlation
0 indicates no linear correlation
1 indicates a perfectly positive linear correlation
However, this type of correlation coefficient works best when the true underlying relationship between the two variables is *linear*.

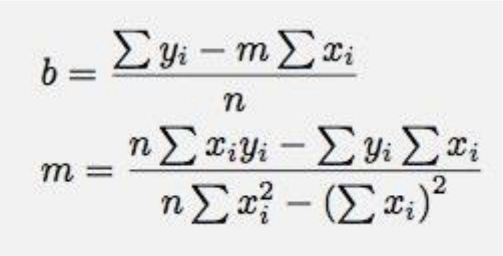
5.3.2 Spearman’s rank correlation scenarios

There is another type of correlation coefficient known as Spearman’s rank correlation that is better to use in two specific scenarios:

Scenario 1: When working with ranked data.
An example could be a dataset that contains the rank of a student’s math exam score along with the rank of their science exam score in a class.

Scenario 2: When one or more extreme outliers are present.
When extreme outliers are present in a dataset, Pearson’s correlation coefficient is highly affected.

The following examples show how to calculate the Spearman Rank Correlation in each of these scenarios.
Scenario 1: Spearman’s Rank Correlation with Ranked Data
Consider the following dataset (and corresponding scatter plot) that shows the relationship between two variables:



$$b = \frac{\sum y_i - m \sum x_i}{n}$$

$$m = \frac{n \sum x_i y_i - \sum y_i \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

Self-Assessment Exercise1

1. Define the Spearman's Rank correlation coefficient
2. Explain the Spearman’s rank correlation scenarios

5.3.3 Interpretation of Statistical Software for Correlation Coefficients

Using statistical software, we can calculate the following correlation coefficients for these two variables:
Pearson’s correlation: **0.79**
Spearman’s rank correlation: **1**

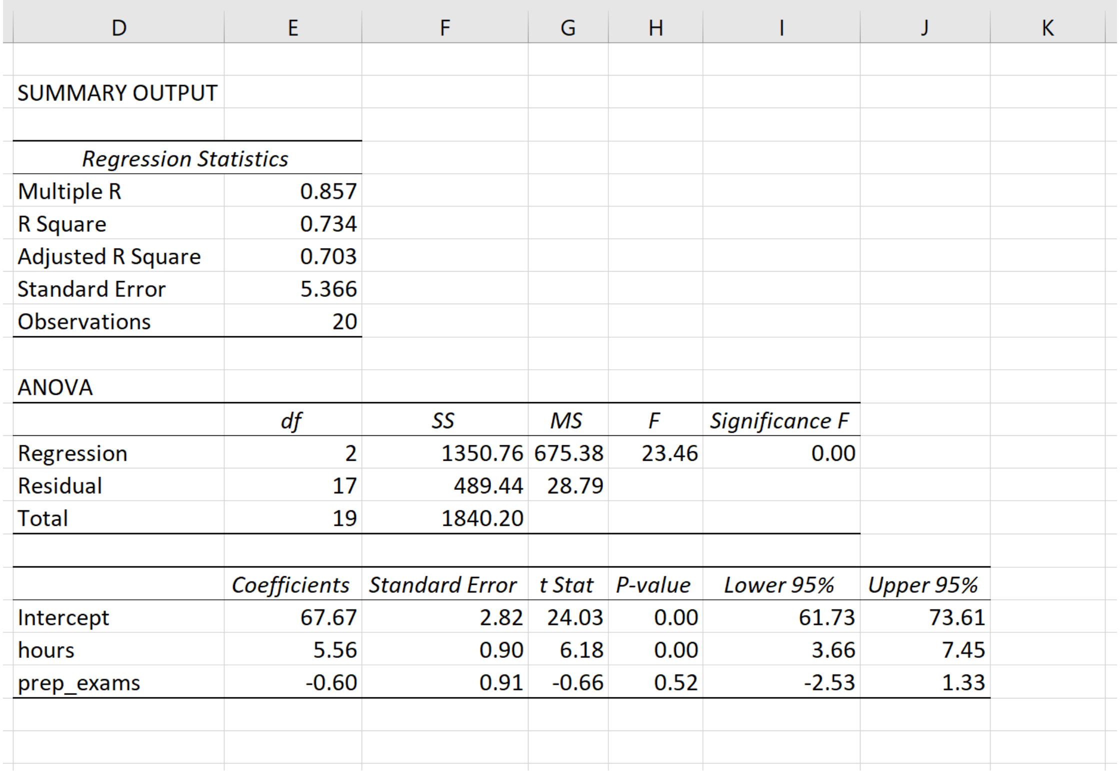
In this scenario, if we only care about the ranks of the data values (when the rank of x increases, does the rank of y also increase?) then Spearman's rank correlation would provide us with a better idea of the correlation between the two variables.

In this particular dataset, as the rank of x increases the rank of y *always* increases.

Spearman's rank correlation captures this behaviour perfectly by telling us that there is a perfect positive relationship (**ρ = 1**) between the ranks of x and the ranks of y.

By contrast, Pearson's correlation tells us the there is a strong linear relationship (**r = 0.79**) between the two variables.
This is true, but it's not useful if we only care about the relationship between the ranks of x and the ranks of y.

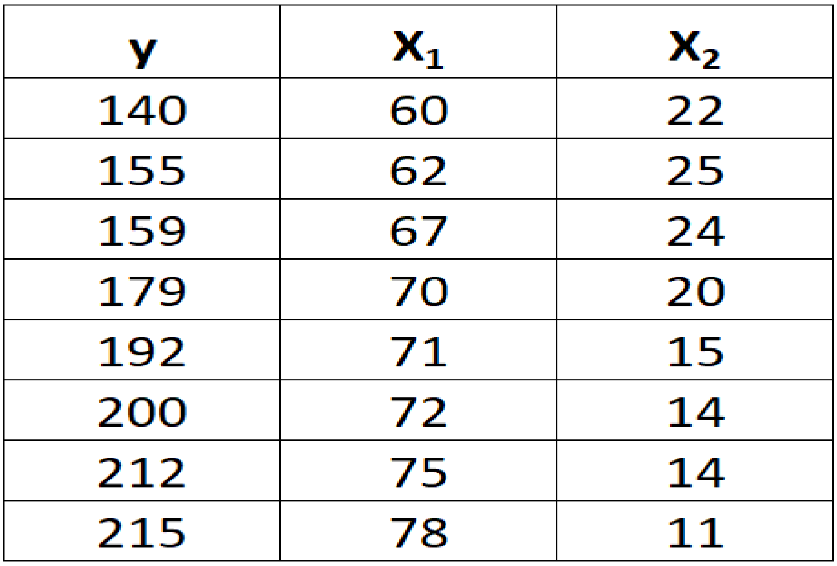
Scenario 2: Spearman's Rank Correlation with Extreme Outliers
Consider the following dataset (and corresponding scatter plot) that shows the relationship between two variables:



D	E	F	G	H	I	J	K
SUMMARY OUTPUT							
Regression Statistics							
Multiple R	0.857						
R Square	0.734						
Adjusted R Square	0.703						
Standard Error	5.366						
Observations	20						
ANOVA							
	df	*SS*	*MS*	*F*	*Significance F*		
Regression	2	1350.76	675.38	23.46	0.00		
Residual	17	489.44	28.79				
Total	19	1840.20					
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	
Intercept	67.67	2.82	24.03	0.00	61.73	73.61	
hours	5.56	0.90	6.18	0.00	3.66	7.45	
prep_exams	-0.60	0.91	-0.66	0.52	-2.53	1.33	

Using statistical software, we can calculate the following correlation coefficients for these two variables:
Pearson's correlation: **0.86**
Spearman's rank correlation: **0.85**

The correlation coefficients are nearly identical because the underlying relationship between the variables is roughly linear and there are no extreme outliers.
Now suppose we change the last y value in the dataset to be an extreme outlier:



y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Using statistical software, we can calculate the correlation coefficients once again:
Pearson's correlation: **0.69**
Spearman's rank correlation: **0.85**
Pearson's correlation coefficient changed dramatically while Spearman's rank correlation coefficient remained the same.

Using statistical jargon, we would say that the relationship between x and y is monotonic (as x increases, y generally increases) but not linear since the outlier influences the data so much.

In this scenario, Spearman's rank correlation does a good job of quantifying this monotonic relationship, while Pearson's correlation does a poor job because it's attempting to calculate the linear relationship between the two variables.

Additional Resources
The following tutorials explain how to calculate the Spearman Rank Correlation using different software:

i. How to Calculate Spearman Rank Correlation in Excel
ii. How to Calculate Spearman Rank Correlation in Google Sheets
iii. How to Calculate Spearman Rank Correlation in R
iv. How to Calculate Spearman Rank Correlation in Python

Self-Assessment Exercise 2

1. Draw a sample of Statistical Software for Correlation Coefficients
2. Explain the interpretation of Statistical Software for Correlation Coefficients

5.6 Summary

This unit discussed the Spearman's Rank correlation coefficient, two variables of linear correlation and linear association between two variables is to use the Pearson Correlation Coefficient as well as the linear association between two variables using the Pearson Correlation Coefficient.

Spearman's Rank correlation coefficient is a technique which can be used to summarise the strength and direction (negative or positive) of a relationship between two variables.

The most common way to quantify the linear association between two variables is to use the Pearson Correlation Coefficient, which always takes on a value between -1 and 1

Spearman's rank correlation scenarios
There is another type of correlation coefficient known as Spearman's rank correlation that is better to use in two specific scenarios:

Scenario 1: When working with ranked data.
An example could be a dataset that contains the rank of a student's math exam score along with the rank of their science exam score in a class.

Scenario 2: When one or more extreme outliers are present.
When extreme outliers are present in a dataset, Pearson's correlation coefficient is highly affected.

5.7 References/Further Readings/Web Resources

Zach (2020). Introduction to Multiple Linear Regression. Retrieved from *https://www.statology.org/multiple-linear-regression*

5.8 Possible Answers to SAEs

Answers to SAEs 1

1. Spearman's Rank correlation coefficient is a technique which can be used to summarise the strength and direction (negative or positive) of a relationship between two variables.
   The most common way to quantify the linear association between two variables is to use the Pearson Correlation Coefficient, which always takes on a value between -1 and 1

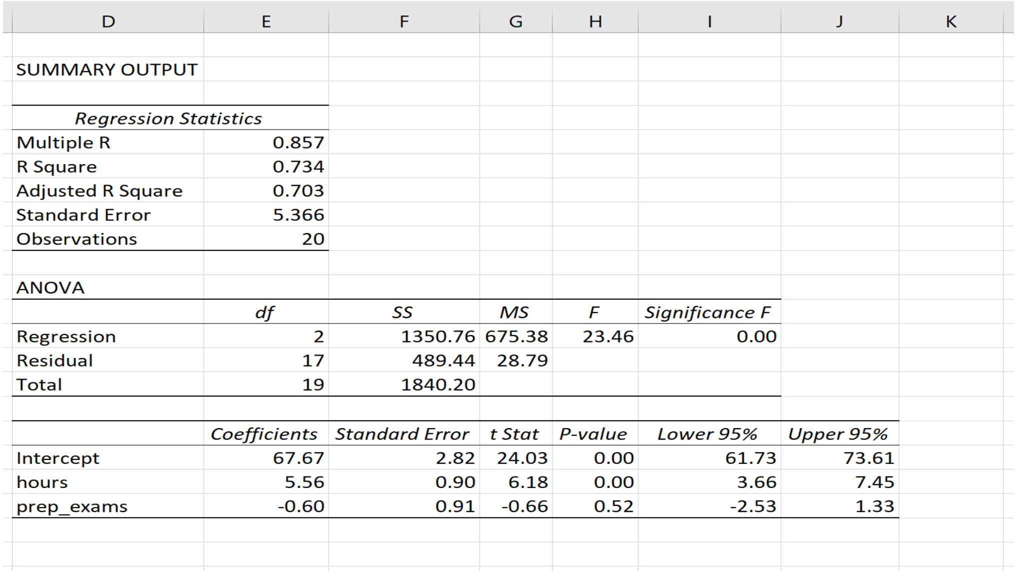
2. **Spearman's rank correlation scenarios**
   There is another type of correlation coefficient known as Spearman's rank correlation that is better to use in two specific scenarios:

Scenario 1: When working with ranked data.
An example could be a dataset that contains the rank of a student's math exam score along with the rank of their science exam score in a class.

Scenario 2: When one or more extreme outliers are present.
When extreme outliers are present in a dataset, Pearson's correlation coefficient is highly affected.

Answers to SAEs 2

1. Consider the following dataset (and corresponding scatter plot) that shows the relationship between two variables:



D	E	F	G	H	I	J	K
SUMMARY OUTPUT							
Regression Statistics							
Multiple R	0.857						
R Square	0.734						
Adjusted R Square	0.703						
Standard Error	5.366						
Observations	20						
ANOVA							
	df	*SS*	*MS*	*F*	*Significance F*		
Regression	2	1350.76	675.38	23.46	0.00		
Residual	17	489.44	28.79				
Total	19	1840.20					
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	
Intercept	67.67	2.82	24.03	0.00	61.73	73.61	
hours	5.56	0.90	6.18	0.00	3.66	7.45	
prep_exams	-0.60	0.91	-0.66	0.52	-2.53	1.33	

Using statistical software, we can calculate the following correlation coefficients for these two variables:
Pearson’s correlation: **0.86**
Spearman’s rank correlation: **0.85**

2. Interpretation of Statistical Software for Correlation Coefficients

Using statistical software, we can calculate the following correlation coefficients for these two variables:
Pearson’s correlation: **0.79**
Spearman’s rank correlation: **1**
In this scenario, if we only care about the ranks of the data values (when the rank of x increases, does the rank of y also increase?) then Spearman’s rank correlation would provide us with a better idea of the correlation between the two variables.

In this particular dataset, as the rank of x increases the rank of y *always* increases.

Spearman’s rank correlation captures this behaviour perfectly by telling us that there is a perfect positive relationship (**ρ = 1**) between the ranks of x and the ranks of y.

By contrast, Pearson’s correlation tells us the there is a strong linear relationship (**r = 0.79**) between the two variables.

This is true, but it’s not useful if we only care about the relationship between the ranks of x and the ranks of y.
Scenario 2: Spearman’s Rank Correlation with Extreme Outliers.

MODULE 4

Unit 1 Pearson Correlation Coefficient

Unit Structure

1.1 Introduction

This unit will discuss and demonstrate the Pearson correlation coefficient, it will explain the Pearson correlation coefficient, and the unit will also State and Demonstrate the Formula for Pearson correlation coefficient. It will explain the Visualizing Correlations as well as Analyse the Testing for Significance of a Pearson Correlation Coefficient

1.2 Learning Outcomes

By the end of this unit, you will be able to:

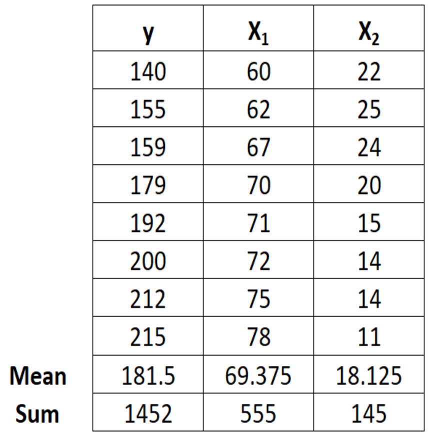
- explain the Pearson correlation coefficient
- state and Demonstrate the Formula for Pearson correlation coefficient
- state and Explain the Visualizing Correlations
- analyse the Testing for Significance of a Pearson Correlation Coefficient

1.3 Pearson Correlation Coefficient

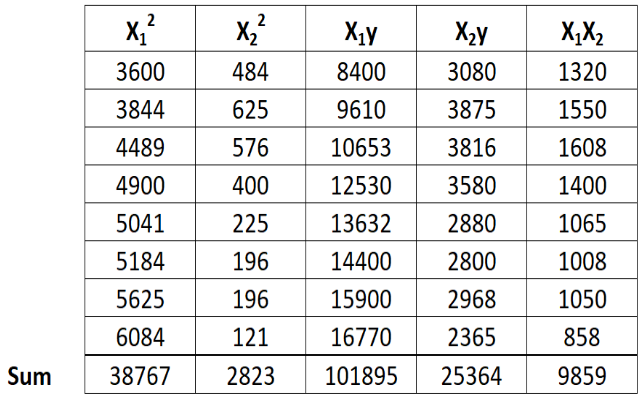
What is Pearson Correlation Coefficient

Pearson correlation coefficient (also known as the “product-moment correlation coefficient”) measure the linear association between two variables *X* and *Y*. It has a value between -1 and 1 where:
-1 indicates a perfectly negative linear correlation between two variables
0 indicates no linear correlation between two variables
1 indicates a perfectly positive linear correlation between two variables

The formula to find the Pearson correlation coefficient, denoted as *r*, for a sample of data is (via Wikipedia):



	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145



	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

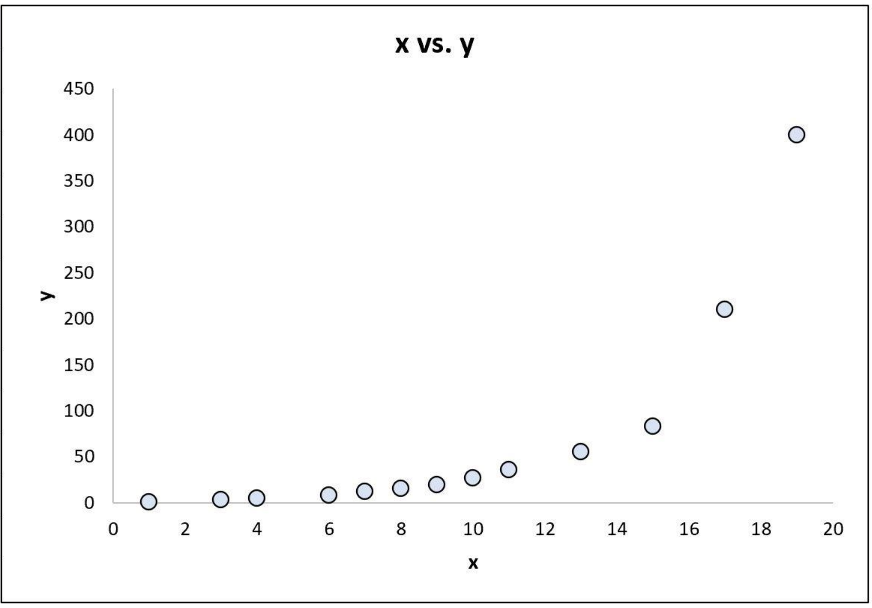
You will likely never have to compute this formula by hand since you can use software to do this for you, but it’s helpful to have an understanding of what exactly this formula is doing by walking through an example.
Suppose we have the following dataset

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859
Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

If we plotted these (X, Y) pairs on a scatterplot, it would look like this:

x	y
1	1
3	3
4	5
6	8
7	12
8	15
9	19
10	27
11	36
13	55
15	83
17	210
19	400

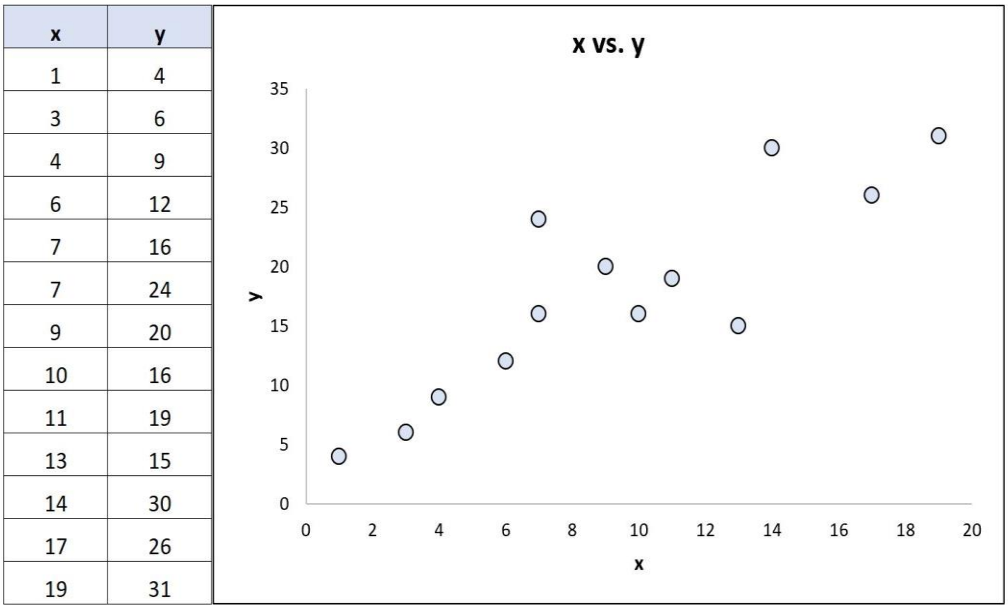


Just from looking at this scatterplot we can tell that there is a positive association between variables X and Y: when X increases, Y tends to increase as well. But to quantify exactly how positively associated these two variables are, we need to find the Pearson correlation coefficient.

Let's focus on just the numerator of the formula

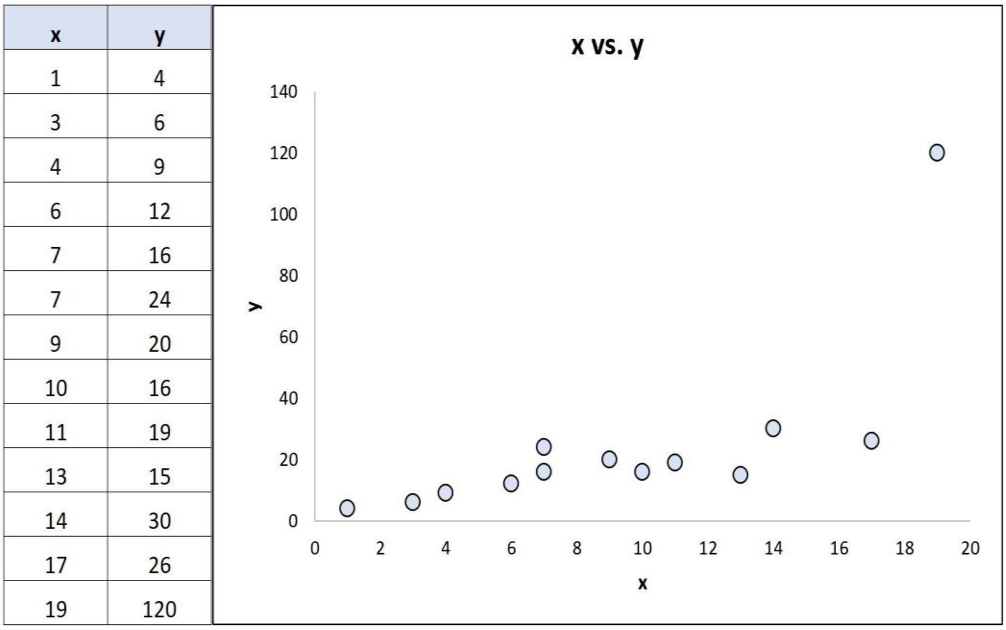
For each (X, Y) pair in our dataset, we need to find the difference between the x value and the mean x value, the difference between the y value and the mean y value, then multiply these two numbers together.

For example, our first (X, Y) pair is (2, 2). The mean x value in this dataset is 5 and the mean y value in this dataset is 7. So, the difference between the x value in this pair and the mean x value is 2 – 5 = -3. The difference between the y value in this pair and the mean y value is 2 – 7 = -5. Then, when we multiply these two numbers together we get -3 * -5 = 15.



x	y
1	4
3	6
4	9
6	12
7	16
7	24
9	20
10	16
11	19
13	15
14	30
17	26
19	31

Here's a visual look at what we just did:

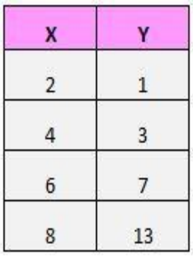


x	y
1	4
3	6
4	9
6	12
7	16
7	24
9	20
10	16
11	19
13	15
14	30
17	26
19	120

1.3.1 Formula for Pearson correlation coefficient

Next, we just need to do this for every single pair:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$



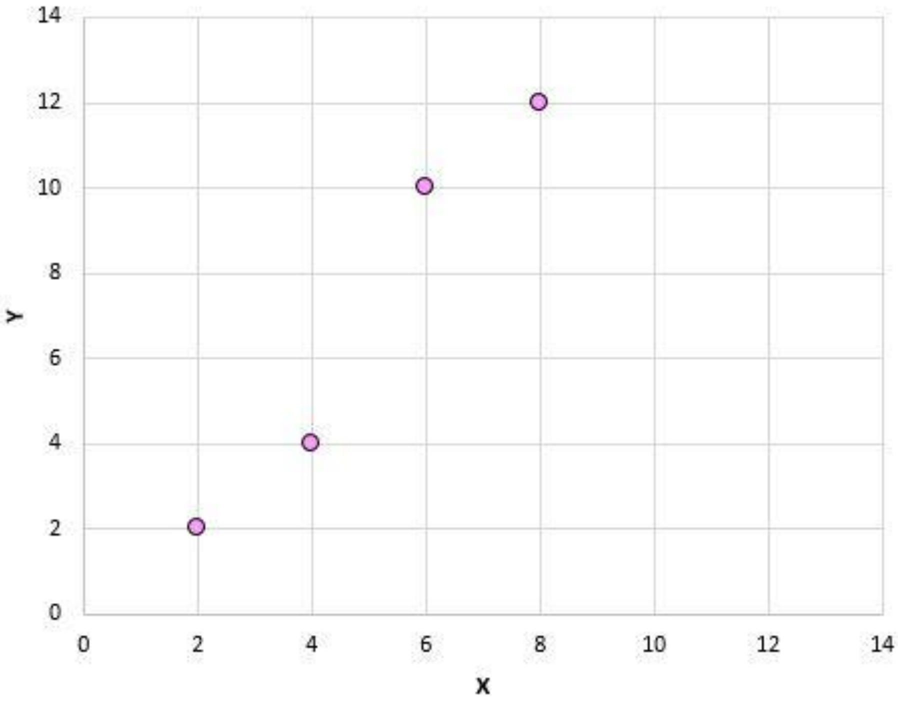
X	Y
2	1
4	3
6	7
8	13

The last step to get the numerator of the formula is to simply add up all of these values:

15 + 3 +3 + 15 = **36**

Next, the denominator of the formula tells us to find the sum of all the squared differences for both x and y, then multiply these two numbers together, then take the square root:

So, first we'll find the sum of the squared differences for both x and y:



Then we'll multiply these two numbers together: 20 * 68 = 1,360.

Lastly, we'll take the square root: √1,360 = **36.88**

So, we found the numerator of the formula to be 36 and the denominator to be 36.88. This means that our Pearson correlation coefficient is r = 36/36.88 = **0.976**

This number is close to 1, which indicates that there is a strong positive linear relationship between our variables *X* and *Y*. This confirms the relationship that we saw in the scatterplot.

Self-Assessment Exercise 1

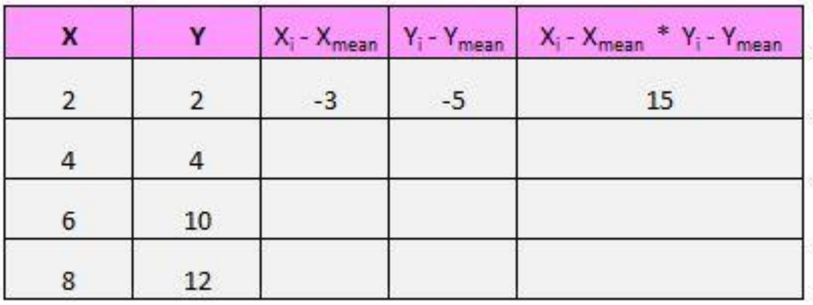
1. Explain the Pearson correlation coefficient
2. State and Demonstrate the Formula for Pearson correlation coefficient

1.3.2 Visualizing Correlations

Recall that a Pearson correlation coefficient tells us the type of linear relationship (positive, negative, none) between two variables as well as the strength of that relationship (weak, moderate, strong).

When we make a scatterplot of two variables, we can *see* the actual relationship between two variables. Here are the many different types of linear relationships we might see:

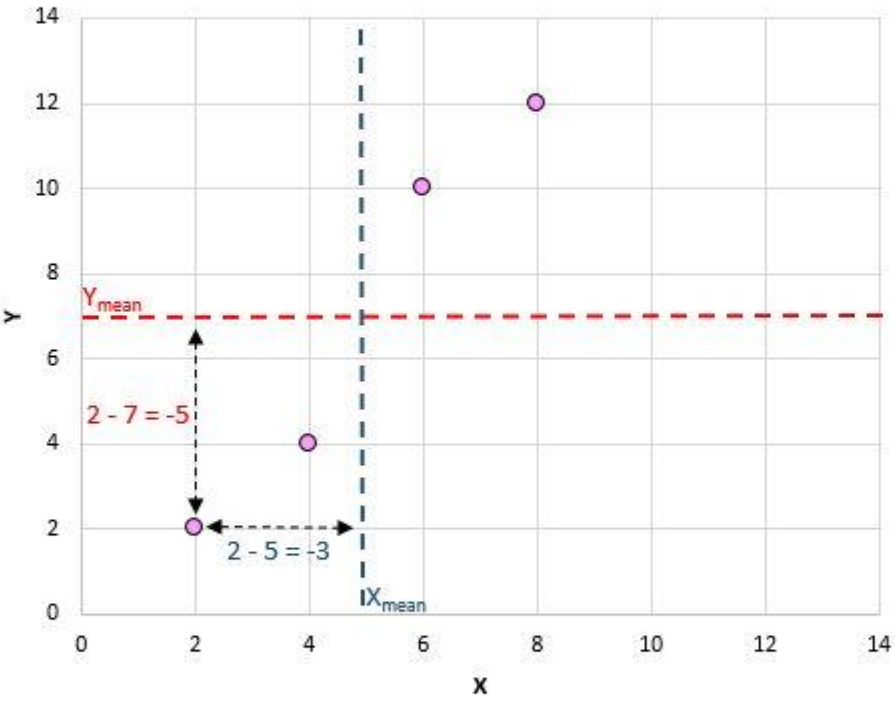
Strong, positive relationship: As the variable on the x-axis increases, the variable on the y-axis increases as well. The dots are packed together tightly, which indicates a strong relationship.



X	Y	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$X_i - X_{mean} * Y_i - Y_{mean}$
2	2	-3	-5	15
4	4			
6	10			
8	12			

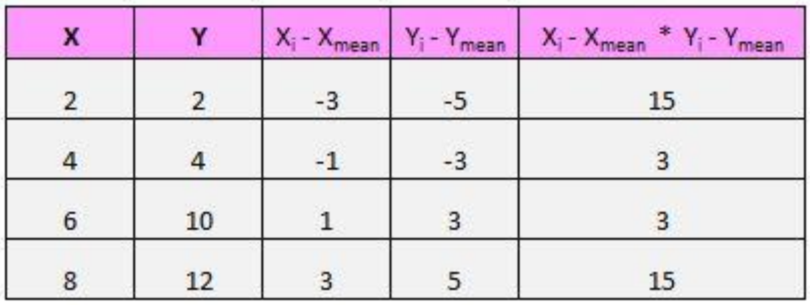
Pearson correlation coefficient: **0.94**

Weak, positive relationship: As the variable on the x-axis increases, the variable on the y-axis increases as well. The dots are fairly spread out, which indicates a weak relationship.



Pearson correlation coefficient: **0.44**

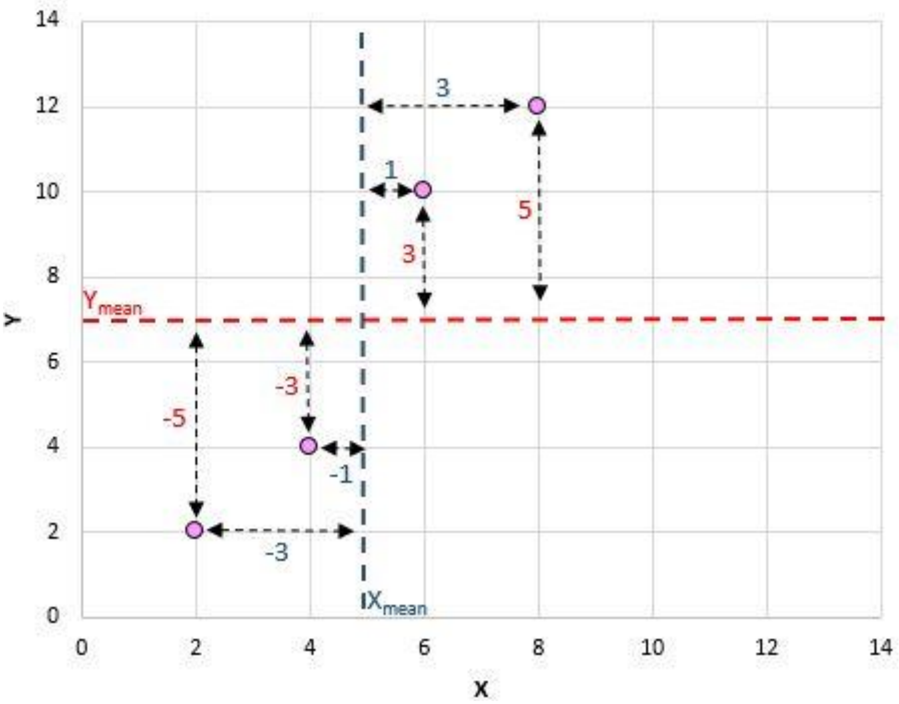
No relationship: There is no clear relationship (positive or negative) between the variables.



X	Y	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$X_i - X_{mean} * Y_i - Y_{mean}$
2	2	-3	-5	15
4	4	-1	-3	3
6	10	1	3	3
8	12	3	5	15

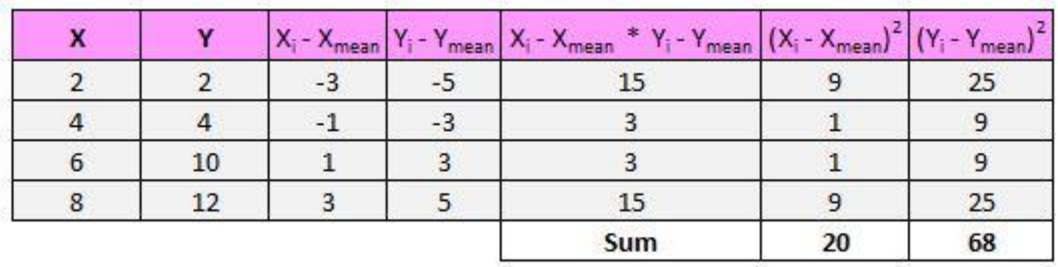
Pearson correlation coefficient: **0.03**

Strong, negative relationship: As the variable on the x-axis increases, the variable on the y-axis decreases. The dots are packed tightly together, which indicates a strong relationship.



Pearson correlation coefficient: **-0.87**

Weak, negative relationship: As the variable on the x-axis increases, the variable on the y-axis decreases. The dots are fairly spread out, which indicates a weak relationship.



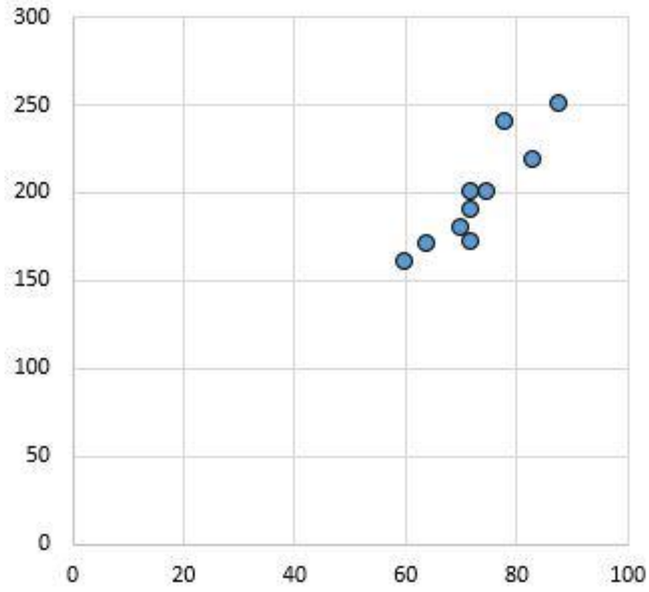
X	Y	X_i - X_{mean}	Y_i - Y_{mean}	X_i - X_{mean} * Y_i - Y_{mean}	$(X_i - X_{mean})^2$	$(Y_i - Y_{mean})^2$
2	2	-3	-5	15	9	25
4	4	-1	-3	3	1	9
6	10	1	3	3	1	9
8	12	3	5	15	9	25
				Sum	**20**	**68**

Pearson correlation coefficient: –**0.46**

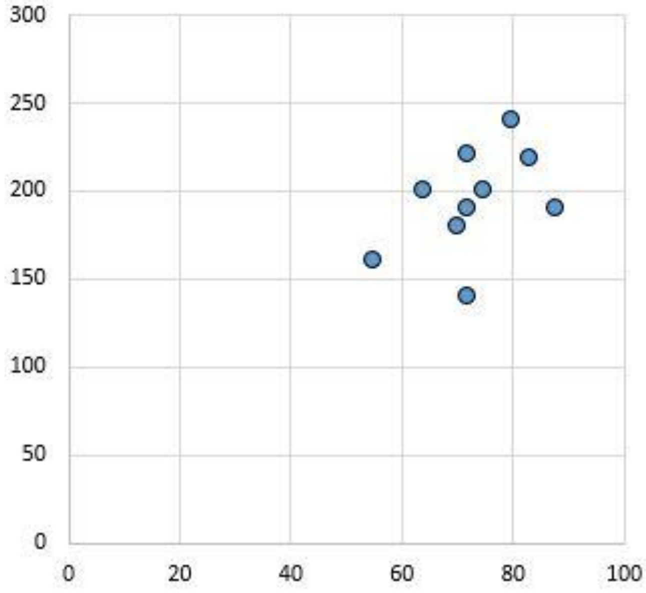
1.3.3 Testing for Significance of a Pearson Correlation Coefficient

When we find the Pearson correlation coefficient for a set of data, we're often working with a *sample* of data that comes from a larger *population*. This means that it's possible to find a non-zero correlation for two variables even if they're actually not correlated in the overall population.

For example, suppose we make a scatterplot for variables *X* and *Y* for every data point in the entire population and it looks like this:



Clearly these two variables are not correlated. However, it's possible that when we take a sample of 10 points from the population, we choose the following points:



We may find that the Pearson correlation coefficient for this sample of points is 0.93, which indicates a strong positive correlation despite the population correlation being zero.

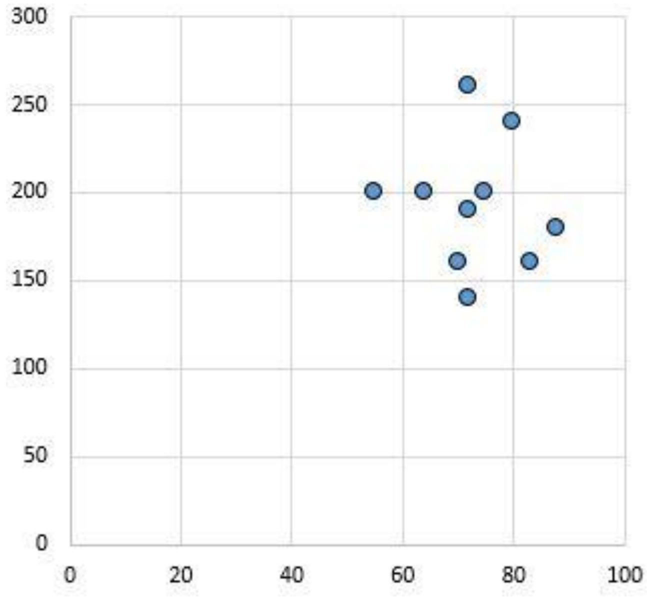
In order to test for whether or not a correlation between two variables is statistically significant, we can find the following test statistic:

Test statistic T = r * $\sqrt{}$(n-2) / (1-r^2)

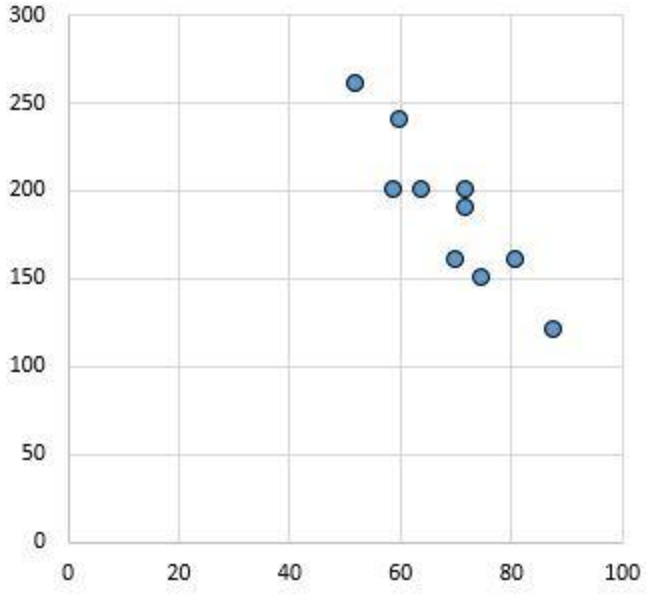
where *n* is the number of pairs in our sample, *r* is the Pearson correlation coefficient, and test statistic T follows a t distribution with n-2 degrees of freedom.

Let’s walk through an example of how to test for the significance of a Pearson correlation coefficient.

Example:
The following dataset shows the height and weight of 12 individuals:



The scatterplot below shows the value of these two variables:



The Pearson correlation coefficient for these two variables is r = 0.836.
The test statistic T = .836 * √(12-2) / (1-.836^2) = 4.804.

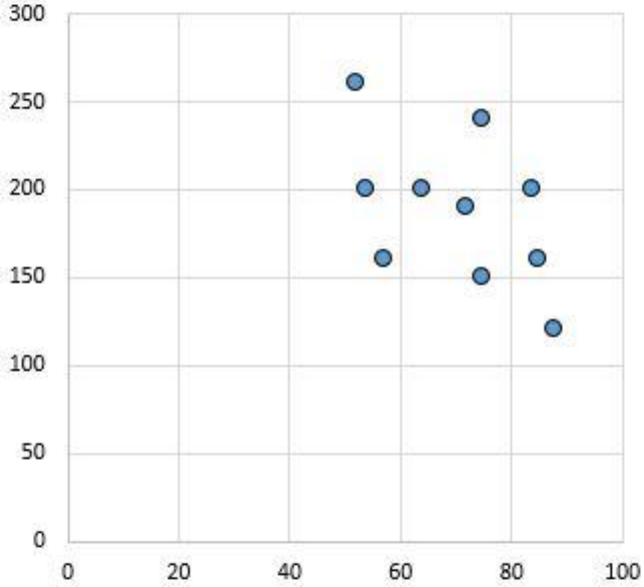
According to our t distribution calculator, a t score of 4.804 with 10 degrees of freedom has a p-value of .0007. Since .0007 < .05, we can conclude that the correlation between weight and height in this example is statistically significant at alpha = .05.

Cautions

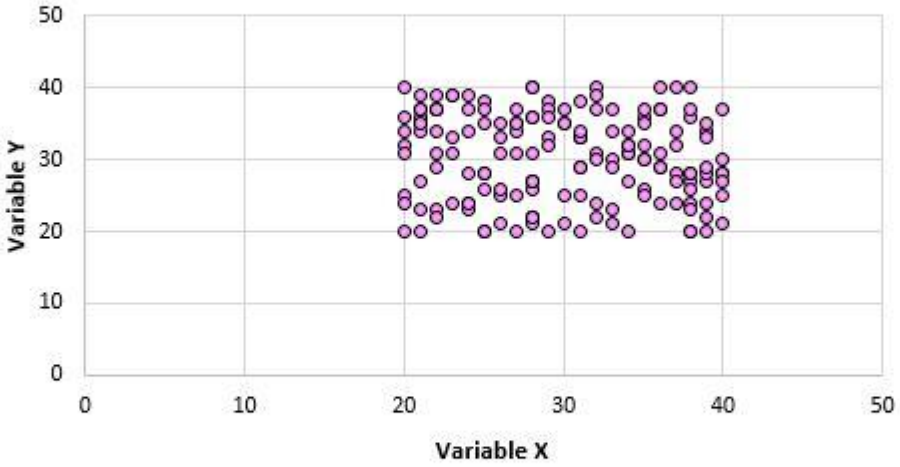
While a Pearson correlation coefficient can be useful in telling us whether or not two variables have a linear association, we must keep three things in mind when interpreting a Pearson correlation coefficient:

1. Correlation does not imply causation**.** Just because two variables are correlated does not mean that one is necessarily *causing* the other to occur more or less often.

   A classic example of this is the positive correlation between ice cream sales and shark attacks. When ice cream sales increase during certain times of the year, shark attacks also tend to increase.
   Does this mean ice cream consumption is *causing* shark attacks? Of course not! It just means that during the summer, both ice cream consumption and shark attacks tend to increase since ice cream is more popular during the summer and more people go in the ocean during the summer.
2. Correlations are sensitive to outliers. One extreme outlier can dramatically change a Pearson correlation coefficient. Consider the example below:

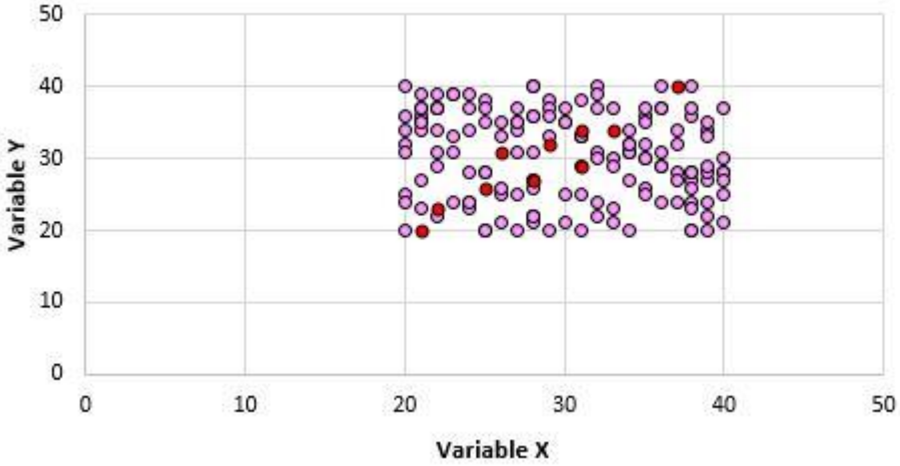


Variables *X* and *Y* have a Pearson correlation coefficient of **0.00**. But imagine that we have one outlier in the dataset:



Now the Pearson correlation coefficient for these two variables is **0.878**. This one outlier changes everything. This is why, when you calculate the correlation for two variables, it's a good idea to visualize the variables using a scatterplot to check for outliers.

A Pearson correlation coefficient does not capture nonlinear relationships between two variables. Imagine that we have two variables with the following relationship:



The Pearson correlation coefficient for these two variables is 0.00 because they have no linear relationship. However, these two variables do have a nonlinear relationship: The y values are simply the x values squared.

When using the Pearson correlation coefficient, keep in mind that you're merely testing to see if two variables are linearly related. Even if a Pearson correlation coefficient tells us that two variables are uncorrelated, they could still have some type of nonlinear relationship. This is another reason that it's helpful to create a scatterplot when analyzing the relationship between two variables – it may help you detect a nonlinear relationship.

Self-Assessment Exercise 2

1. State and Explain the Visualizing Correlations
2. Analyse the Testing for Significance of a Pearson Correlation Coefficient

1.6 Summary

This unit discussed and demonstrated the Pearson correlation coefficient, it explained the Pearson correlation coefficient, and the unit demonstrated the Formula for Pearson correlation coefficient. It explained the Visualizing Correlations as well as Analyse the Testing for Significance of a Pearson Correlation Coefficient.

When using the Pearson correlation coefficient, keep in mind that you're merely testing to see if two variables are linearly related. Even if a Pearson correlation coefficient tells us that two variables are uncorrelated, they could still have some type of nonlinear relationship. This is another reason that it's helpful to create a scatterplot when analysing the relationship between two variables – it may help you detect a nonlinear relationship

1.7 References/Further Readings/Web Resources

Zach (2020). Introduction to Spearman correlations https://www.statology.org/pearson-correlation-coefficient.

1.8 Possible Answers to SAEs

Answers to SAEs 1

1. What is Pearson correlation coefficient?

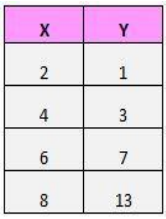
Pearson correlation coefficient (also known as the "product-moment correlation coefficient") measure the linear association between two variables *X* and *Y*. It has a value between -1 and 1 where:
-1 indicates a perfectly negative linear correlation between two variables

0 indicates no linear correlation between two variables
1 indicates a perfectly positive linear correlation between two variables

2. Formula for Pearson correlation coefficient

Next, we just need to do this for every single pair:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$



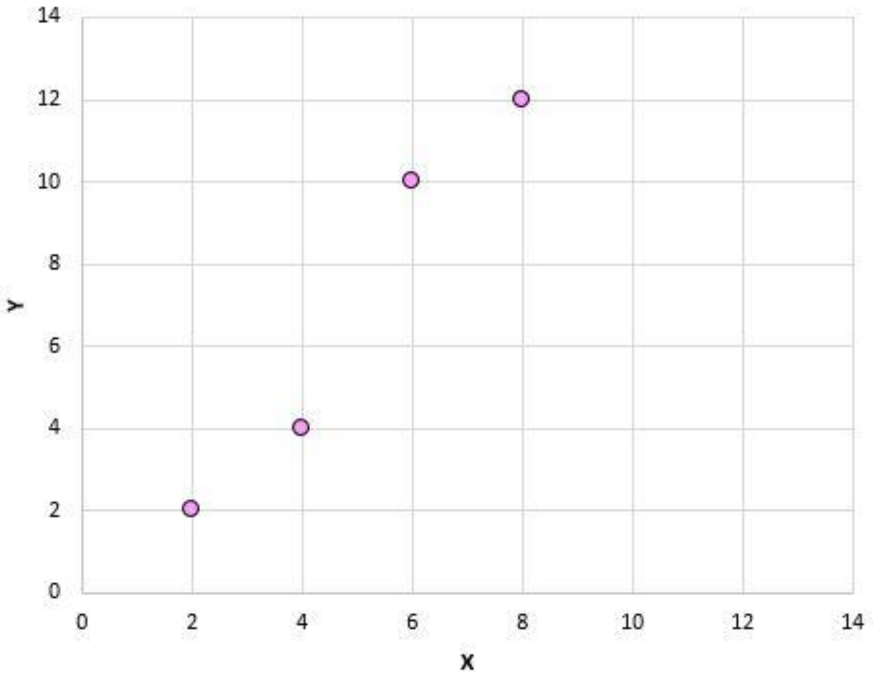
X	Y
2	1
4	3
6	7
8	13

The last step to get the numerator of the formula is to simply add up all of these values:

15 + 3 +3 + 15 = **36**

Next, the denominator of the formula tells us to find the sum of all the squared differences for both x and y, then multiply these two numbers together, then take the square root:

So, first we'll find the sum of the squared differences for both x and y:



Then we'll multiply these two numbers together: 20 * 68 = 1,360.

Lastly, we'll take the square root: √1,360 = **36.88**

So, we found the numerator of the formula to be 36 and the denominator to be 36.88. This means that our Pearson correlation coefficient is r = 36 / 36.88 = **0.976**

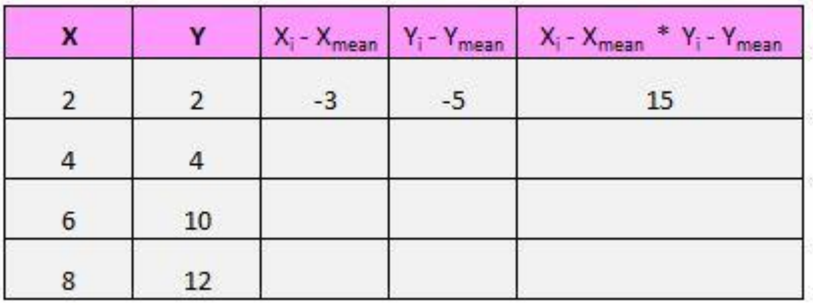
Answers to SAEs 2

1. Visualizing Correlations

Recall that a Pearson correlation coefficient tells us the type of linear relationship (positive, negative, none) between two variables as well as the strength of that relationship (weak, moderate, strong).

When we make a scatterplot of two variables, we can *see* the actual relationship between two variables. Here are the many different types of linear relationships we might see:

Strong, positive relationship: As the variable on the x-axis increases, the variable on the y-axis increases as well. The dots are packed together tightly, which indicates a strong relationship.



X	Y	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$X_i - X_{mean} * Y_i - Y_{mean}$
2	2	-3	-5	15
4	4			
6	10			
8	12			

Pearson correlation coefficient: **0.94**

2. Testing for Significance of a Pearson Correlation Coefficient

When we find the Pearson correlation coefficient for a set of data, we're often working with a *sample* of data that comes from a larger *population*. This means that it's possible to find a non-zero correlation for two variables even if they're actually not correlated in the overall population.

For example, suppose we make a scatterplot for variables *X* and *Y* for every data point in the entire population

Unit 2 Analysis of Variance (ANOVA)

Unit Structure

1.1 Introduction

This unit will discuss analysis of variance (ANOVA) and hypothesis testing, ANOVA in SPSS., One-way ANOVA in SPSS Statistics, Assumptions and Test Procedure in SPSS Statistics The specific test considered here is called analysis of variance (ANOVA) and is a test of hypothesis that is appropriate to compare means of a continuous variable in two or more independent comparison groups.

For example, in some clinical trials there are more than two comparison groups. In a clinical trial to evaluate a new medication for asthma, investigators might compare an experimental medication to a placebo and to a standard treatment (i.e., a medication currently being used). In an observational study such as the Framingham Heart Study, it might be of interest to compare mean blood pressure or mean cholesterol levels in persons who are underweight, normal weight, overweight and obese.

2.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the concept of Analysis of Variance (ANOVA)
- discuss the ANOVA Approach

- demonstrate ANOVA in SPSS.
- explain One-way ANOVA in SPSS Statistics
- state the Assumptions
- highlight the Test Procedure in SPSS Statistics
- explain the SPSS Statistics
- demonstrate the Setup step in SPSS Statistics
- demonstrate descriptive Table
- explain the reporting output of the one-way ANOVA

2.3 Analysis of Variance (ANOVA)

2.3.1 What is Analysis of Variance (ANOVA)?

Analysis of Variance (ANOVA) is the technique use to test the difference between more than two independent means. It is an extension of the two independent samples procedure which applies when there are exactly two independent comparison groups. The ANOVA technique applies when there are two or more than two independent groups. The ANOVA procedure is used to compare the means of the comparison groups and is conducted using the same five step approach used in the scenarios discussed in previous sections. Because there are more than two groups, however, the computation of the test statistic is more involved. The test statistic must take into account the sample sizes, sample means and sample standard deviations in each of the comparison groups.

If one is examining the means observed among, say three groups, it might be tempting to perform three separate group to group comparisons, but this approach is incorrect because each of these comparisons fails to take into account the total data, and it increases the likelihood of incorrectly concluding that there are statistically significate differences, since each comparison adds to the probability of a type I error. Analysis of variance avoids these problemss by asking a more global question, i.e., whether there are significant differences among the groups, without addressing differences between any two groups in particular (although there are additional tests that can do this if the analysis of variance indicates that there are differences among the groups).

The fundamental strategy of ANOVA is to systematically examine variability within groups being compared and also examine variability among the groups being compared

2.3.2 The ANOVA Approach

Consider an example with four independent groups and a continuous outcome measure. The independent groups might be defined by a particular characteristic of the participants such as BMI (e.g., underweight, normal weight, overweight, obese) or by the investigator (e.g., randomizing participants to one of four competing treatments, call them A, B, C and D). Suppose that the outcome is systolic blood pressure, and we wish to test whether there is a statistically significant difference in mean systolic blood pressures among the four groups

Analysis of Variance, i.e. ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. Essentially, ANOVA in SPSS is used as the test of means for two or more populations.

ANOVA in SPSS must have a dependent variable which should be metric (measured using an interval or ratio scale). ANOVA in SPSS must also have one or more independent variables, which should be categorical in nature. In ANOVA in SPSS, categorical independent variables are called factors. A particular combination of factor levels, or categories, is called a treatment.

2.4 ANOVA in SPSS.

In ANOVA in SPSS, there is one way ANOVA which involves only one categorical variable, or a single factor. For example, if a researcher wants to examine whether heavy, medium, light and nonusers of cereals differed in their preference for Total cereal, then the differences can be examined by the one way ANOVA in SPSS. In one way ANOVA in SPSS, a treatment is the same as the factor level.

If two or more factors are involved in ANOVA in SPSS, then it is termed as n way ANOVA. For example, if the researcher also wants to examine the preference for Total cereal by the customers who are loyal to it and those who are not, then we can use N way.

In ANOVA in SPSS, from the menu we choose:
“Analyze” then go to “Compare Means” and click on the “One-Way ANOVA.”

Now, let us discuss in detail how the software operates ANOVA:
The first step is to identify the dependent and independent variables. The dependent variable is generally denoted by Y and the independent

variable is denoted by X. X is a categorical variable having c categories. The sample size in each category of X is generally denoted as n, and the total sample size N=nXc.

The next step in ANOVA in SPSS is to examine the differences among means. This involves decomposition of the total variation observed in the dependent variable. This variation in ANOVA in SPSS is measured by the sums of the squares of the mean.

The total variation in Y in ANOVA in SPSS is denoted by SSy, which can be decomposed into two components:
SSy=SSbetween+SSwithin
where the subscripts between and within refers to the categories of X in ANOVA in SPSS. SSbetween is the portion of the sum of squares in Y related to the independent variable or factor X. Thus it is generally referred to as the sum of squares of X. SSwithin is the variation in Y related to the variation within each category of X. It is generally referred to as the sum of squares for errors in ANOVA in SPSS.

The logic behind decomposing SSy is to examine the differences in group means.

The next task in ANOVA in SPSS is to measure the effects of X on Y, which is generally done by the sum of squares of X, because it is related to the variation in the means of the categories of X. The relative magnitude of the sum of squares of X in ANOVA in SPSS increases as the differences among the means of Y in categories of X increases. The relative magnitude of the sum of squares of X in ANOVA in SPSS increases as the variation in Y within the categories of X decreases.

The strength of the effects of X on Y is measured with the help of $\eta2$ in ANOVA in SPSS .The value of $\eta2$ varies between 0 and 1. It assumes a value 0 in ANOVA in SPSS when all the category means are equal, indicating that X has no effect on Y. The value of $\eta2$ becomes 1, when there is no variability within each category of X but there is still some variability between the categories.

The final step in ANOVA in SPSS is to calculate the mean square which is obtained by dividing the sum of squares by the corresponding degrees of freedom. The null hypothesis of equal means, which is done by an F statistic, is the ratio between the mean square related to the independent variable and the mean square related to the error.

N way ANOVA in ANOVA in SPSS involves simultaneous examination of two or more categorical independent variables, which is also computed in a similar manner.

A major advantage of ANOVA in SPSS is that the interactions between the independent variables can be examined. Follow this link for details.

2.4.1 One-way ANOVA in SPSS Statistics

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups (although you tend to only see it used when there are a minimum of three, rather than two groups). For example, you could use a one-way ANOVA to understand whether exam performance differed based on test anxiety levels amongst students, dividing students into three independent groups (e.g., low, medium and high-stressed students). Also, it is important to realize that the one-way ANOVA is an **omnibus** test statistic and cannot tell you which specific groups were statistically significantly different from each other; it only tells you that at least two groups were different.

Since you may have three, four, five or more groups in your study design, determining which of these groups differ from each other is important. You can do this using a post hoc test (N.B., we discuss post hoc tests later in this guide).

Note: If your study design not only involves one dependent variable and one independent variable, but also a third variable (known as a "covariate") that you want to "statistically control", you may need to perform an ANCOVA (analysis of covariance), which can be thought of as an extension of the one-way ANOVA. To learn more, see our SPSS Statistics guide on ANCOVA. Alternatively, if your dependent variable is the time until an event happens, you might need to run a Kaplan-Meier analysis.

This "quick start" guide shows you how to carry out a one-way ANOVA using SPSS Statistics, as well as interpret and report the results from this test. Since the one-way ANOVA is often followed up with a post hoc test, we also show you how to carry out a post hoc test using SPSS Statistics. However, before we introduce you to this procedure, you need to understand the different assumptions that your data must meet in order for a one-way ANOVA to give you a valid result. We discuss these assumptions next.

Self-Assessment Exercise 1

1. Explain the concept of Analysis of Variance (ANOVA)
2. Discuss the ANOVA Approach
3. Demonstrate ANOVA in SPSS.
4. Explain One-way ANOVA in SPSS Statistics

2.4.2 Assumptions

When you choose to analyse your data using a one-way ANOVA, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using a one-way ANOVA. You need to do this because it is only appropriate to use a one-way ANOVA if your data "passes" six assumptions that are required for a one-way ANOVA to give you a valid result. In practice, checking for these six assumptions just adds a little bit more time to your analysis, requiring you to click a few more buttons in SPSS Statistics when performing your analysis, as well as think a little bit more about your data, but it is not a difficult task.

Before we introduce you to these six assumptions, do not be surprised if, when analysing your own data using SPSS Statistics, one or more of these assumptions is violated (i.e., is not met). This is not uncommon when working with real-world data rather than textbook examples, which often only show you how to carry out a one-way ANOVA when everything goes well! However, don’t worry. Even when your data fails certain assumptions, there is often a solution to overcome this. First, let’s take a look at these six assumptions:

1. Assumption #1: Your dependent variable should be measured at the interval or ratio level (i.e., they are continuous). Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. You can learn more about interval and ratio variables in our article: Types of Variable.

2. Assumption #2: Your independent variable should consist of two or more categorical, independent groups. Typically, a one-way ANOVA is used when you have three or more categorical, independent groups, but it can be used for just two groups (but an independent-samples t-test is more commonly used for two groups). Example independent variables that meet this criterion include ethnicity (e.g., 3 groups: Caucasian, African American and Hispanic), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth.

3. Assumption #3: You should have independence of observations, which means that there is no relationship between the observations in each group or between the groups themselves. For example, there must be different participants in each group with no participant being in more than one group. This is more of

a study design issue than something you can test for, but it is an important assumption of the one-way ANOVA. If your study fails this assumption, you will need to use another statistical test instead of the one-way ANOVA (e.g., a repeated measures design). If you are unsure whether your study meets this assumption, you can use our Statistical Test Selector, which is part of our enhanced guides.

4. Assumption #4: There should be no significant outliers. Outliers are simply single data points within your data that do not follow the usual pattern (e.g., in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally). The problem with outliers is that they can have a negative effect on the one-way ANOVA, reducing the validity of your results. Fortunately, when using SPSS Statistics to run a one-way ANOVA on your data, you can easily detect possible outliers. In our enhanced one-way ANOVA guide, we: (a) show you how to detect outliers using SPSS Statistics; and (b) discuss some of the options you have in order to deal with outliers. You can learn more about our enhanced one-way ANOVA guide on our Features: One-way ANOVA page.

5. Assumption #5: Your dependent variable should be approximately normally distributed for each category of the independent variable. We talk about the one-way ANOVA only requiring approximately normal data because it is quite "robust" to violations of normality, meaning that assumption can be a little violated and still provide valid results. You can test for normality using the Shapiro-Wilk test of normality, which is easily tested for using SPSS Statistics. In addition to showing you how to do this in our enhanced one-way ANOVA guide, we also explain what you can do if your data fails this assumption (i.e., if it fails it more than a little bit). Again, you can learn more on our Features: One-way ANOVA page.

6. Assumption #6: There needs to be homogeneity of variances. You can test this assumption in SPSS Statistics using Levene's test for homogeneity of variances. If your data fails this assumption, you will need to not only carry out a Welch ANOVA instead of a one-way ANOVA, which you can do using SPSS Statistics, but also use a different post hoc test. In our enhanced one-way ANOVA guide, we (a) show you how to perform Levene’s test for homogeneity of variances in SPSS Statistics, (b) explain some of the things you will need to consider when

interpreting your data, and (c) present possible ways to continue with your analysis if your data fails to meet this assumption, including running a Welch ANOVA in SPSS Statistics instead of a one-way ANOVA, and a Games-Howell test instead of a Tukey post hoc test (learn more on our Features: One-way ANOVA page).

You can check assumptions #4, #5 and #6 using SPSS Statistics. Before doing this, you should make sure that your data meets assumptions #1, #2 and #3, although you don't need SPSS Statistics to do this. Remember that if you do not run the statistical tests on these assumptions correctly, the results you get when running a one-way ANOVA might not be valid. This is why we dedicate a number of sections of our enhanced one-way ANOVA guide to help you get this right. You can find out about our enhanced one-way ANOVA guide on our Features: One-way ANOVA page, or more generally, our enhanced content as a whole on our Features: Overview page.

2.4.3 Test Procedure in SPSS Statistics

Follow this link https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php for Test Procedure in SPSS Statistics, we illustrate the SPSS Statistics procedure to perform a one-way ANOVA assuming that no assumptions have been violated. First, we set out the example we use to explain the one-way ANOVA procedure in SPSS Statistics.

2.5 SPSS Statistics

Example:
A manager wants to raise the productivity at his company by increasing the speed at which his employees can use a particular spreadsheet program. As he does not have the skills in-house, he employs an external agency which provides training in this spreadsheet program. They offer 3 courses: a beginner, intermediate and advanced course. He is unsure which course is needed for the type of work they do at his company, so he sends 10 employees on the beginner course, 10 on the intermediate and 10 on the advanced course.

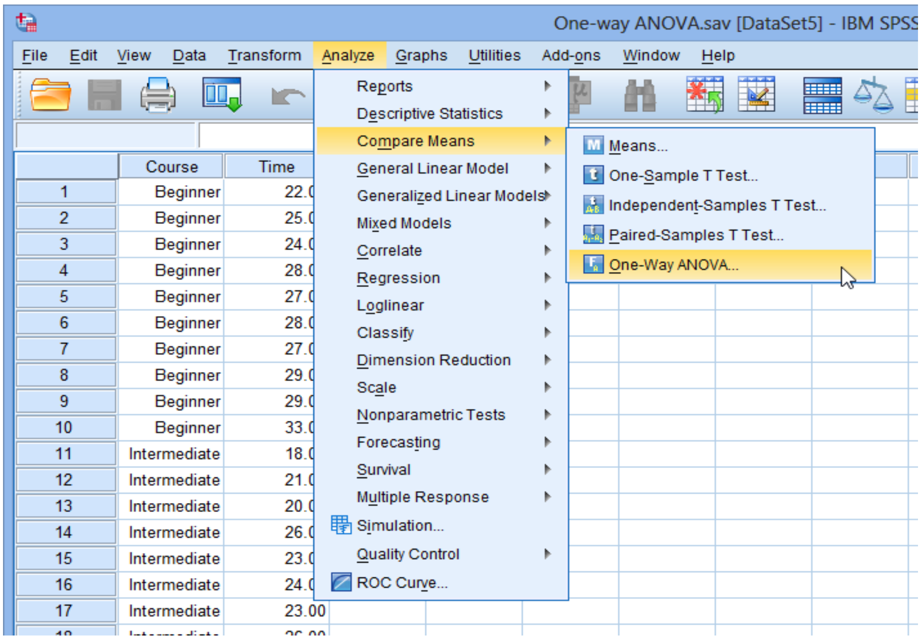
When they all return from the training, he gives them a problem to solve using the spreadsheet program, and times how long it takes them to complete the problem. He then compares the three courses (beginner, intermediate, advanced) to see if there are any differences in the average time, it took to complete the problem SPSS Statistics

1.5.1 Setup in SPSS Statistics

In SPSS Statistics, we separated the groups for analysis by creating a grouping variable called Course (i.e., the independent variable), and gave the beginners course a value of "1", the intermediate course a value of "2" and the advanced course a value of "3". Time to complete the set problem was entered under the variable name Time (i.e., the dependent variable). In our enhanced one-way ANOVA guide, we show you how to correctly enter data in SPSS Statistics to run a one-way ANOVA (see on our Features: One-way ANOVA page). You can learn about our enhanced data setup content in general on our Features: Data Setup. Alternately, see our generic, "quick start" guide: Entering Data in SPSS Statistics.

The eight steps below show you how to analyse your data using a one-way ANOVA in SPSS Statistics when the six assumptions in the previous section, Assumptions, have not been violated. At the end of these eight steps, we show you how to interpret the results from this test.

If you are looking for help to make sure your data meets assumptions #4, #5 and #6, which are required when using a one-way ANOVA, and can be tested using SPSS Statistics, you can learn more on our Features: One-way ANOVA page
Click Analyze > Compare Means > One-Way ANOVA... on the top menu, as shown below.

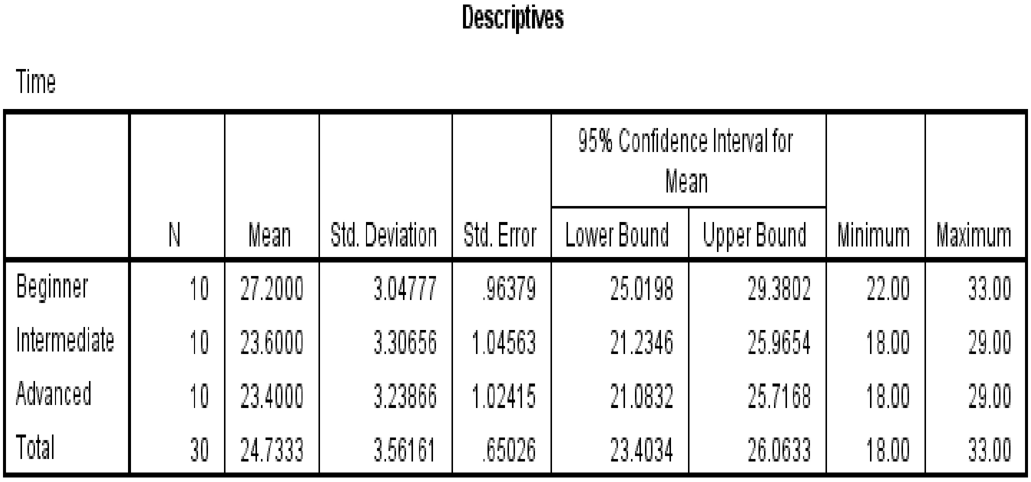


	Course	Time
1	Beginner	22.0
2	Beginner	25.0
3	Beginner	24.0
4	Beginner	28.0
5	Beginner	27.0
6	Beginner	28.0
7	Beginner	27.0
8	Beginner	29.0
9	Beginner	29.0
10	Beginner	33.0
11	Intermediate	18.0
12	Intermediate	21.0
13	Intermediate	20.0
14	Intermediate	26.0
15	Intermediate	23.0
16	Intermediate	24.0
17	Intermediate	23.00

Published with written permission from SPSS Statistics, IBM Corporation.

2.6 Descriptive Table

The descriptive table (see below) provides some very useful descriptive statistics, including the mean, standard deviation and 95% confidence intervals for the dependent variable (Time) for each separate group (Beginners, Intermediate and Advanced), as well as when all groups are combined (Total). These figures are useful when you need to describe your data.



Descriptives

Time

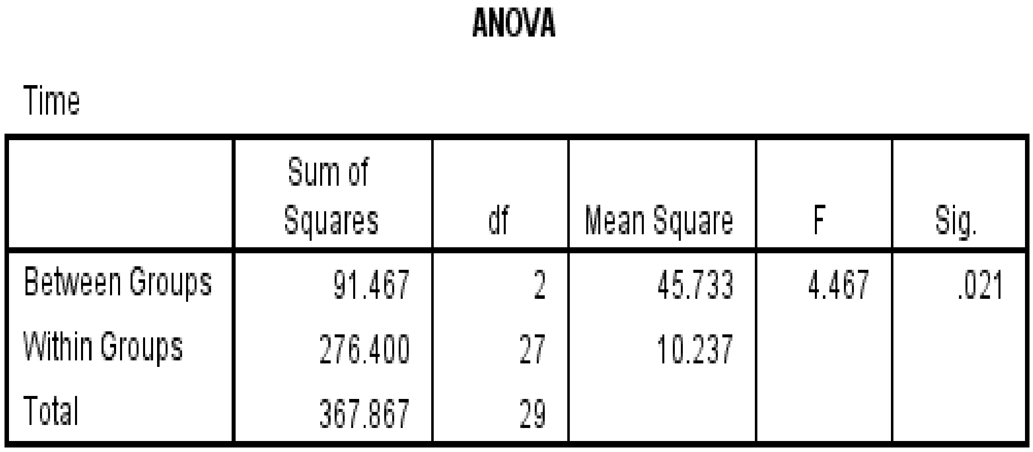
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Beginner	10	27.2000	3.04777	.96379	25.0198	29.3802	22.00	33.00
Intermediate	10	23.6000	3.30656	1.04563	21.2346	25.9654	18.00	29.00
Advanced	10	23.4000	3.23866	1.02415	21.0832	25.7168	18.00	29.00
Total	30	24.7333	3.56161	.65026	23.4034	26.0633	18.00	33.00

Published with written permission from SPSS Statistics, IBM Corporation.

SPSS Statistics

ANOVA Table

This is the table that shows the output of the ANOVA analysis and whether there is a statistically significant difference between our group means. We can see that the significance value is 0.021 (i.e., *p* = .021), which is below 0.05. and, therefore, there is a statistically significant difference in the mean length of time to complete the spreadsheet problem between the different courses taken. This is great to know, but we do not know which of the specific groups differed. Luckily, we can find this out in the **Multiple Comparisons** table which contains the results of the Tukey post hoc test.



ANOVA

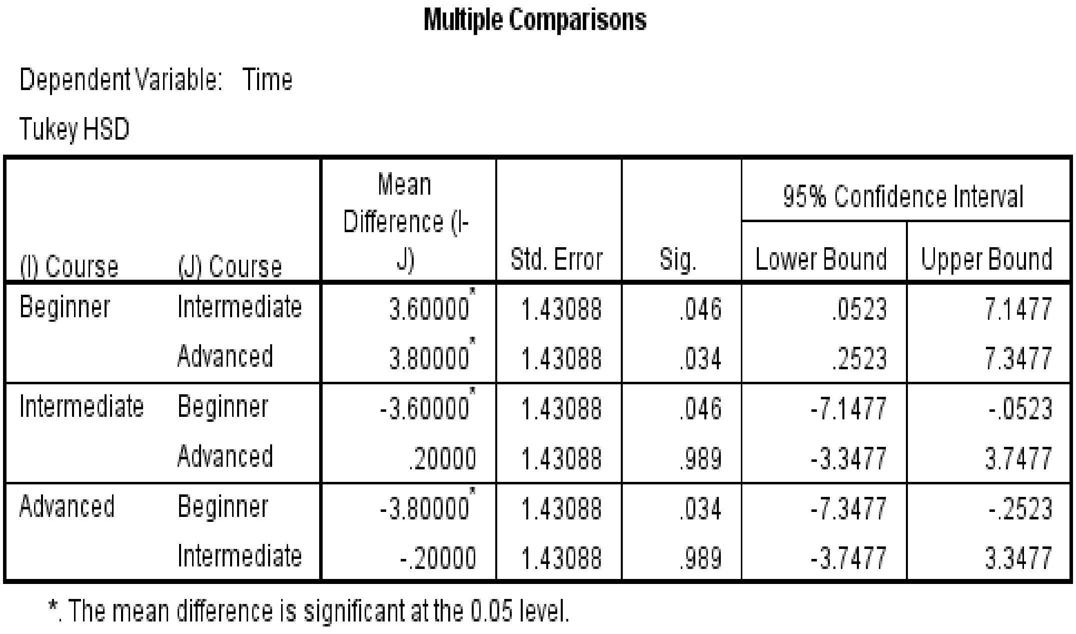
Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	91.467	2	45.733	4.467	.021
Within Groups	276.400	27	10.237		
Total	367.867	29			

Published with written permission from SPSS Statistics, IBM Corporation.

Multiple Comparisons Table

From the results so far, we know that there are statistically significant differences between the groups as a whole. The table below, **Multiple Comparisons**, shows which groups differed from each other. The Tukey post hoc test is generally the preferred test for conducting post hoc tests on a one-way ANOVA, but there are many others. We can see from the table below that there is a statistically significant difference in time to complete the problem between the group that took the beginner course and the intermediate course (*p* = 0.046), as well as between the beginner course and advanced course (*p* = 0.034). However, there were no differences between the groups that took the intermediate and advanced course (*p* = 0.989).



Multiple Comparisons

Dependent Variable: Time

Tukey HSD

(I) Course	(J) Course	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Beginner	Intermediate	3.60000*	1.43088	.046	.0523	7.1477
	Advanced	3.80000*	1.43088	.034	.2523	7.3477
Intermediate	Beginner	-3.60000*	1.43088	.046	-7.1477	-.0523
	Advanced	.20000	1.43088	.989	-3.3477	3.7477
Advanced	Beginner	-3.80000*	1.43088	.034	-7.3477	-.2523
	Intermediate	-.20000	1.43088	.989	-3.7477	3.3477

*. The mean difference is significant at the 0.05 level.

Published with written permission from SPSS Statistics, IBM Corporation.

It is also possible to run comparisons between specific groups that you decided were of interest before you looked at your results. For example, you might have expressed an interest in knowing the difference in the completion time between the beginner and intermediate course groups.

This type of comparison is often called a planned contrast or a simple custom contrast. However, you do not have to confine yourself to the comparison between two time points only. You might have had an interest in understanding the difference in completion time between the beginner course group and the average of the intermediate and advanced course groups. This is called a complex contrast.

All these types of custom contrast are available in SPSS Statistics. In our enhanced guide we show you how to run custom contrasts in SPSS Statistics using syntax (or sometimes a combination of the graphical user interface and syntax) and how to interpret and report the results. In addition, we also show you how to "trick" SPSS Statistics into applying a Bonferroni adjustment for multiple comparisons which it would otherwise not do.

2.6.1 Reporting the output of the one-way ANOVA

Based on the results above, you could report the results of the study as follows (N.B., this does not include the results from your assumptions tests or effect size calculations):

There was a statistically significant difference between groups as determined by one-way ANOVA (*F*(2,27) = 4.467, *p* = .021). A Tukey post hoc test revealed that the time to complete the problem was statistically significantly lower after taking the intermediate (23.6 ± 3.3 min, *p* = .046) and advanced (23.4 ± 3.2 min, *p* = .034) course compared to the beginners course (27.2 ± 3.0 min). There was no statistically significant difference between the intermediate and advanced groups (*p* = .989).

In our enhanced one-way ANOVA guide, we show you how to write up the results from your assumptions tests, one-way ANOVA and Tukey post hoc results if you need to report this in a dissertation, thesis, assignment or research report. We do this using the Harvard and APA styles (see our Features: One-way ANOVA page to learn more).

It is also worth noting that in addition to reporting the results from your assumptions, one-way ANOVA and Tukey post hoc test, you are increasingly expected to report an effect size. Whilst there are many different ways you can do this, we show you how to calculate an effect size from your SPSS Statistics results in our enhanced one-way

ANOVA guide. Effect sizes are important because whilst the one-way ANOVA tells you whether differences between group means are "real" (i.e., different in the population), it does not tell you the "size" of the difference. Providing an effect size in your results helps to overcome this limitation. You can learn more about our enhanced one-way ANOVA guide on our Features: One-way ANOVA page, or our enhanced content in general on our Features: Overview page.

Self-Assessment Exercise 2

1. State the Assumptions
2. Highlight the Test Procedure in SPSS Statistics
3. Explain the SPSS Statistics
4. Demonstrate the Setup step in SPSS Statistics

2.6 Summary

This unit discussed the analysis of variance (ANOVA), ANOVA in SPSS, One-way ANOVA in SPSS Statistics, Assumptions and Test Procedure in SPSS Statistics. The specific test considered here is called analysis of variance (ANOVA) and is a test of hypothesis that is appropriate to compare means of a continuous variable in two or more independent comparison groups. For example, in some clinical trials there are more than two comparison groups. In a clinical trial to evaluate a new medication for asthma, investigators might compare an experimental medication to a placebo and to a standard treatment (i.e., a medication currently being used). In an observational study such as the Framingham Heart Study, it might be of interest to compare mean blood pressure or mean cholesterol levels in persons who are underweight, normal weight, overweight and obese.

2.7 References/Further Reading/Web Resources

Statistics.laerd.com (2022). One-way ANOVA in SPSS Statistics. Retrieved from https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

Statistics.laerd.com (2022). ANOVA in SPSS Statistics. Retrieved from https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

2.8 Possible Answers to SAEs

Answers to SAEs 1

1. Analysis of Variance (ANOVA) is the technique use to test the difference between more than two independent means. It is an extension of the two independent samples procedure which applies when there are exactly two independent comparison groups

2. The ANOVA Approach
   Consider an example with four independent groups and a continuous outcome measure. The independent groups might be defined by a particular characteristic of the participants such as BMI (e.g., underweight, normal weight, overweight, obese) or by the investigator (e.g., randomizing participants to one of four competing treatments, call them A, B, C and D). Suppose that the outcome is systolic blood pressure, and we wish to test whether there is a statistically significant difference in mean systolic blood pressures among the four groups

3. Analysis of Variance, i.e. ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. Essentially, ANOVA in SPSS is used as the test of means for two or more populations

4. ANOVA in SPSS.
   In ANOVA in SPSS, there is one way ANOVA which involves only one categorical variable, or a single factor. For example, if a researcher wants to examine whether heavy, medium, light and nonusers of cereals differed in their preference for Total cereal, then the differences can be examined by the one way ANOVA in SPSS. In one way ANOVA in SPSS, a treatment is the same as the factor level

Answers to SAEs 2

1. Assumptions
   When you choose to analyse your data using a one-way ANOVA, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using a one-way ANOVA. You need to do this because it is only appropriate to

use a one-way ANOVA if your data "passes" six assumptions that are required for a one-way ANOVA to give you a valid result. In practice, checking for these six assumptions just adds a little bit more time to your analysis, requiring you to click a few more buttons in SPSS Statistics when performing your analysis, as well as think a little bit more about your data, but it is not a difficult task

Assumption #1: Your dependent variable should be measured at the interval or ratio level (i.e., they are continuous).

Assumption #2: Your independent variable should consist of two or more categorical, independent groups

Assumption #3: You should have independence of observations, which means that there is no relationship between the observations in each group or between the groups themselves.

Assumption #4: There should be no significant outliers. Outliers are simply single data points within your data that do not follow the usual pattern

Assumption #5: Your dependent variable should be approximately normally distributed for each category of the independent variable

Assumption #6: There needs to be homogeneity of variances. You can test this assumption in SPSS Statistics using Levene's test for homogeneity of variances

2. Test Procedure in SPSS Statistics

Follow this link https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php for Test Procedure in SPSS Statistics, we illustrate the SPSS Statistics procedure to perform a one-way ANOVA assuming that no assumptions have been violated. First, we set out the example we use to explain the one-way ANOVA procedure in SPSS Statistics.

3. SPSS Statistics

A manager wants to raise the productivity at his company by increasing the speed at which his employees can use a particular spreadsheet program. As he does not have the skills in-house, he employs an external agency which provides training in this spreadsheet program. They offer 3 courses: a beginner, intermediate and advanced course. He is unsure which course is needed for the type of work they do at his

company, so he sends 10 employees on the beginner course, 10 on the intermediate and 10 on the advanced course. When they all return from the training, he gives them a problem to solve using the spreadsheet program, and times how long it takes them to complete the problem. He then compares the three courses (beginner, intermediate, advanced) to see if there are any differences in the average time it took to complete the problem.
SPSS Statistics

4. Setup in SPSS Statistics

In SPSS Statistics, we separated the groups for analysis by creating a grouping variable called Course (i.e., the independent variable), and gave the beginners course a value of "1", the intermediate course a value of "2" and the advanced course a value of "3". Time to complete the set problem was entered under the variable name Time (i.e., the dependent variable). In our enhanced one-way ANOVA guide, we show you how to correctly enter data in SPSS Statistics to run a one-way ANOVA (see on our Features: One-way ANOVA page). You can learn about our enhanced data setup content in general on our Features: Data Setup. Alternately, see our generic, "quick start" guide: Entering Data in SPSS Statistics

Unit 3 The use of Statistical Package for Social Science (SPSS) for Analysis

Unit Structure

3.1 Introduction

This unit will examine the procedure for Statistical Package for the Social Sciences in Analysing data.

3.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the meaning of Statistical Package for the Social Sciences (SPSS?)
- state the Common Uses of SPSS
- mention the Data Requirements
- explain the Hypotheses
- demonstrate the Data Set-Up
- explain the problem Statement
- demonstrate the Running the Test
- explain the decision and conclusions

3.3 Statistical Package for the Social Sciences (SPSS)

3.3.1 What is Statistical Package for the Social Sciences (SPSS?)

Statistical Package for the Social Sciences (SPSS) is a software program used by researchers in various disciplines for quantitative analysis of complex data. This introductory level SPSS unit introduces SPSS environment, basic data preparation and management, descriptive statistics, and common statistical analysis (T-test, ANOVA, correlation, regression). Hands-on-activity with example data is provided to learn these basics.

This unit is a good start for people who are new to SPSS and help you harness this useful analysis tool. For a complete SPSS package follow this link: *https://researchcommons.library.ubc.ca/introduction-to-spss-for-statistical-analysis/.* The bivariate Pearson Correlation produces a sample correlation coefficient, *r*, which measures the strength and direction of linear relationships between pairs of continuous variables. By extension, the Pearson Correlation evaluates whether there is statistical evidence for a linear relationship among the same pairs of variables in the population, represented by a population correlation coefficient, ρ (“rho”). The Pearson Correlation is a parametric measure.

This measure is also known as:
Pearson’s correlation
Pearson product-moment correlation (PPMC)

3.3.2 Common Uses of SPSS

The bivariate Pearson Correlation is commonly used to measure the following:
Correlations among pairs of variables
Correlations within and between sets of variables
The bivariate Pearson correlation indicates the following
Whether a statistically significant linear relationship exists between two continuous variables

The strength of a linear relationship (i.e., how close the relationship is to being a perfectly straight line)
The direction of a linear relationship (increasing or decreasing)

Note: The bivariate Pearson Correlation cannot address non-linear relationships or relationships among categorical variables. If you wish to understand relationships that involve categorical variables and/or non-linear relationships, you will need to choose another measure of association.

Note: The bivariate Pearson Correlation only reveals *associations* among continuous variables. The bivariate Pearson Correlation does not provide any inferences about causation, no matter how large the correlation coefficient is.

3.3.3 Data Requirements

To use Pearson correlation, your data must meet the following requirements:
Two or more continuous variables (i.e., interval or ratio level)
Cases must have non-missing values on both variables
Linear relationship between the variables
Independent cases (i.e., independence of observations)
There is no relationship between the values of variables between cases.

This means that:
the values for all variables across cases are unrelated
for any case, the value for any variable cannot influence the value of any variable for other cases
no case can influence another case on any variable

The biviariate Pearson correlation coefficient and corresponding significance test are not robust when independence is violated.
Bivariate normality
Each pair of variables is bivariately normally distributed
Each pair of variables is bivariately normally distributed at all levels of the other variable(s)

This assumption ensures that the variables are linearly related; violations of this assumption may indicate that non-linear relationships among variables exist. Linearity can be assessed visually using a scatterplot of the data.
Random sample of data from the population
No outliers

3.4 Hypotheses

The null hypothesis (H_0) and alternative hypothesis (H_1) of the significance test for correlation can be expressed in the following ways, depending on whether a one-tailed or two-tailed test is requested:
Two-tailed significance test:

H_0: ρ = 0 ("the population correlation coefficient is 0; there is no association")

H_1: $\rho \neq 0$ ("the population correlation coefficient is not 0; a nonzero correlation could exist")

One-tailed significance test:
H_0: ρ = 0 ("the population correlation coefficient is 0; there is no association")

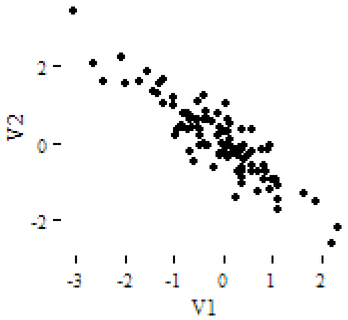
H_1: ρ > 0 ("the population correlation coefficient is greater than 0; a positive correlation could exist")
OR
H_1: ρ < 0 ("the population correlation coefficient is less than 0; a negative correlation could exist") where ρ is the population correlation coefficient.

Test Statistic
The sample correlation coefficient between two variables *x* and *y* is denoted *r* or r_{xy}, and can be computed as:
rxy=cov(x,y)var(x)−−−−−√⋅var(y)−−−−−√rxy=cov(x,y)var(x)⋅var(y)
where cov(*x*, *y*) is the sample covariance of *x* and *y*; var(*x*) is the sample variance of *x*; and var(*y*) is the sample variance of *y*.

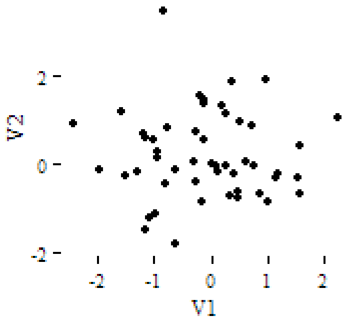
Correlation can take on any value in the range [-1, 1]. The sign of the correlation coefficient indicates the direction of the relationship, while the magnitude of the correlation (how close it is to -1 or +1) indicates the strength of the relationship.

-1 : perfectly negative linear relationship
0 : no relationship
+1 : perfectly positive linear relationship
The strength can be assessed by these general guidelines (which may vary by discipline):
.1 < | *r* | < .3 … small / weak correlation
.3 < | *r* | < .5 … medium / moderate correlation
.5 < | *r* | ……… large / strong correlation

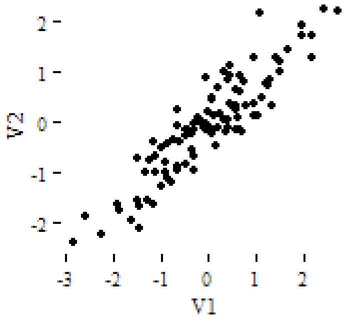
Note: The direction and strength of a correlation are two distinct properties. The scatterplots below show correlations that are *r* = +0.90, *r* = 0.00, and *r* = -0.90, respectively. The strength of the nonzero correlations are the same: 0.90. But the direction of the correlations is different: a negative correlation corresponds to a decreasing relationship, while and a positive correlation corresponds to an increasing relationship.
r = -0.90



r = 0.00



r = 0.90



Note that the *r* = 0.00 correlation has no discernable increasing or decreasing linear pattern in this particular graph. However, keep in mind that Pearson correlation is only capable of detecting *linear* associations, so it is possible to have a pair of variables with a strong nonlinear relationship and a small Pearson correlation coefficient. It is good practice to create scatterplots of your variables to corroborate your correlation coefficients.

Self-Assessment Exercise 1

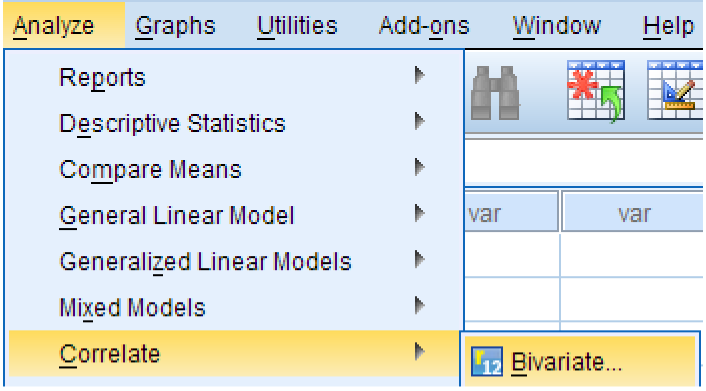
1. Explain the meaning of Statistical Package for the Social Sciences (SPSS?)
2. State the Common Uses of SPSS
3. Mention the Data Requirements
4. Explain the Hypotheses

3.4.1 Data Set-Up

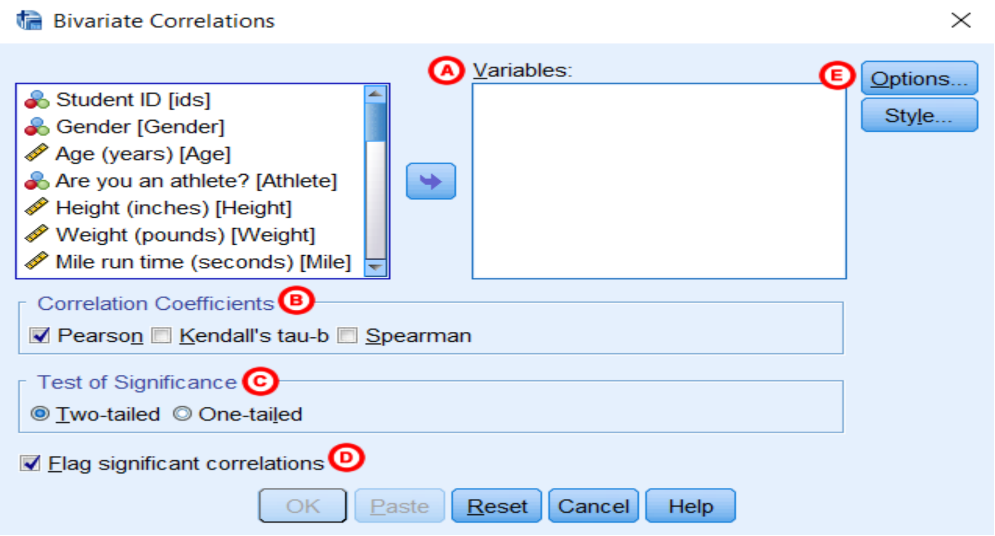
Your dataset should include two or more continuous numeric variables, each defined as scale, which will be used in the analysis.
Each row in the dataset should represent one unique subject, person, or unit. All of the measurements taken on that person or unit should appear in that row. If measurements for one subject appear on multiple rows -- for example, if you have measurements from different time points on separate rows -- you should reshape your data to "wide" format before you compute the correlations.

Run a Bivariate Pearson Correlation

To run a bivariate Pearson Correlation in SPSS, click **Analyze > Correlate > Bivariate**.



The Bivariate Correlations window opens, where you will specify the variables to be used in the analysis. All of the variables in your dataset appear in the list on the left side. To select variables for the analysis, select the variables in the list on the left and click the blue arrow button to move them to the right, in the Variables field.



A **Variables:** The variables to be used in the bivariate Pearson Correlation. You must select at least two continuous variables, but may select more than two. The test will produce correlation coefficients for each pair of variables in this list.

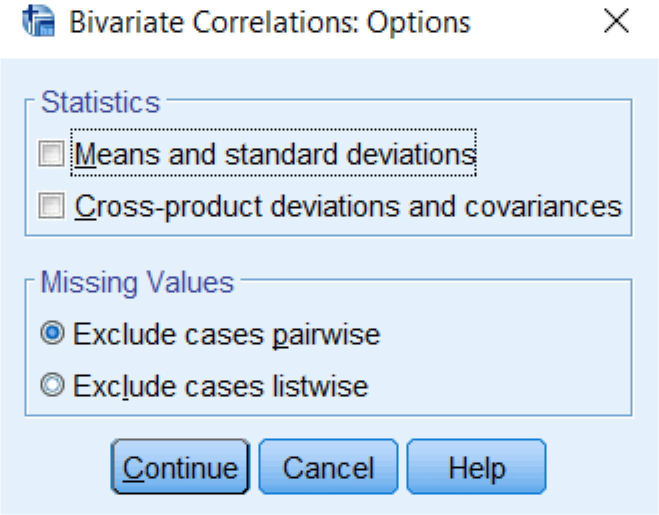
B **Correlation Coefficients:** There are multiple types of correlation coefficients. By default, **Pearson** is selected. Selecting Pearson will produce the test statistics for a bivariate Pearson Correlation.

C **Test of Significance:** Click **Two-tailed** or **One-tailed**, depending on your desired significance test. SPSS uses a two-tailed test by default.

D **Flag significant correlations:** Checking this option will include asterisks (**) next to statistically significant correlations in the output. By default, SPSS marks statistical significance at the alpha = 0.05 and alpha = 0.01 levels, but not at the alpha = 0.001 level (which is treated as alpha = 0.01)

E **Options:** Clicking **Options** will open a window where you can specify which **Statistics** to include (i.e., **Means and standard deviations**, **Cross-product deviations and covariances**) and how to address **Missing Values** (i.e., **Exclude cases pairwise or Exclude cases listwise**).

Note that the pairwise/listwise setting does not affect your computations if you are entering two variable, but can make a very large difference if you are entering three or more variables into the correlation procedure.


Bivariate Correlations: Options

Statistics

□ Means and standard deviations

□ Cross-product deviations and covariances

Missing Values

● Exclude cases pairwise

○ Exclude cases listwise

Example: Understanding the linear association between weight and height

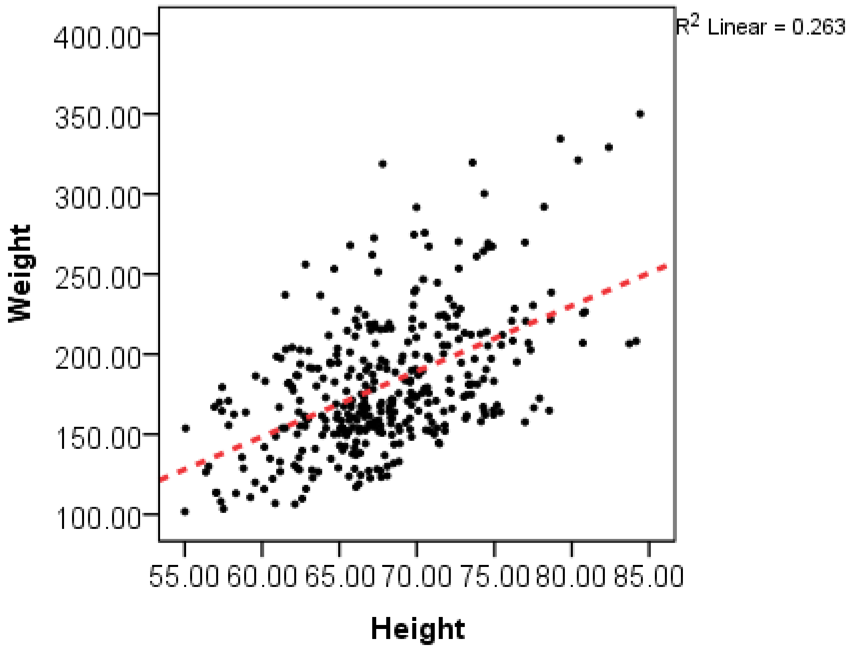
3.4.2 Problem Statement

Perhaps you would like to test whether there is a statistically significant linear relationship between two continuous variables, weight and height (and by extension, infer whether the association is significant in the population). You can use a bivariate Pearson Correlation to test whether there is a statistically significant linear relationship between height and weight, and to determine the strength and direction of the association.

Before the Test

In the sample data, we will use two variables: “Height” and “Weight.” The variable “Height” is a continuous measure of height in inches and exhibits a range of values from 55.00 to 84.41 (**Analyze** > **Descriptive Statistics** > **Descriptives**). The variable “Weight” is a continuous measure of weight in pounds and exhibits a range of values from 101.71 to 350.07.

Before we look at the Pearson correlations, we should look at the scatterplots of our variables to get an idea of what to expect. In particular, we need to determine if it's reasonable to assume that our variables have linear relationships. Click **Graphs > Legacy Dialogs > Scatter/Dot**. In the Scatter/Dot window, click **Simple Scatter**, then click **Define**. Move variable Height to the X Axis box, and move variable Weight to the Y Axis box. When finished, click **OK**.



To add a linear fit like the one depicted, double-click on the plot in the Output Viewer to open the Chart Editor. Click **Elements > Fit Line at Total**. In the Properties window, make sure the Fit Method is set

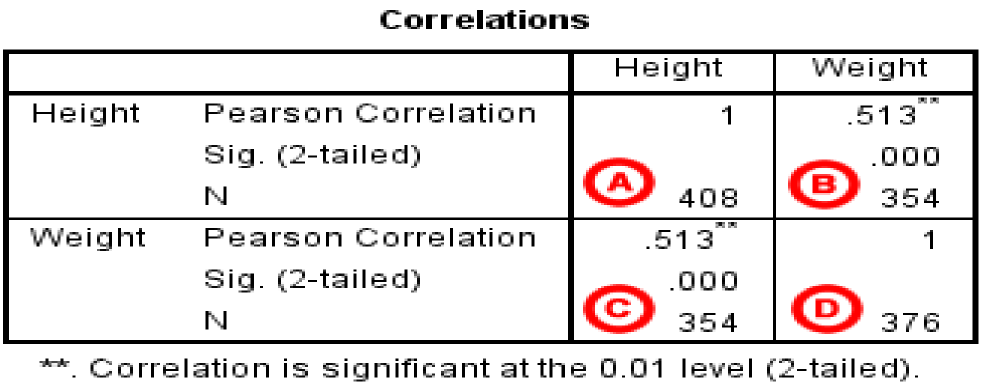
to **Linear**, then click **Apply**. (Notice that adding the linear regression trend line will also add the R-squared value in the margin of the plot. If we take the square root of this number, it should match the value of the Pearson correlation we obtain.)
From the scatterplot, we can see that as height increases, weight also tends to increase. There does appear to be some linear relationship.

3.5 Running the Test

To run the bivariate Pearson Correlation, click **Analyze** > **Correlate** > **B ivariate**. Select the variables Height and Weight and move them to the Variables box. In the **Correlation Coefficients** area, select **Pearson**. In the **Test of Significance** area, select your desired significance test, two-tailed or one-tailed. We will select a two-tailed significance test in this example. Check the box next to **Flag significant correlations**.

Click **OK** to run the bivariate Pearson Correlation. Output for the analysis will display in the Output Viewer.

Syntax
CORRELATIONS
/VARIABLES=Weight Height
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.
OUTPUT
Tables
The results will display the correlations in a table, labeled **Correlations**.



Correlations

		Height	Weight
Height	Pearson Correlation	1	.513**
	Sig. (2-tailed)		.000
	N	408	354
Weight	Pearson Correlation	.513**	1
	Sig. (2-tailed)	.000	
	N	354	376

**. Correlation is significant at the 0.01 level (2-tailed).

A Correlation of Height with itself (r=1), and the number of nonmissing observations for height (n=408).

B Correlation of height and weight (r=0.513), based on n=354 observations with pairwise nonmissing values.

C Correlation of height and weight (r=0.513), based on n=354 observations with pairwise nonmissing values.

D Correlation of weight with itself (r=1), and the number of nonmissing observations for weight (n=376).

The important cells we want to look at are either B or C. (Cells B and C are identical, because they include information about the same pair of variables.) Cells B and C contain the correlation coefficient for the correlation between height and weight, its p-value, and the number of complete pairwise observations that the calculation was based on.

The correlations in the *main diagonal* (cells A and D) are all equal to 1. This is because a variable is always perfectly correlated with itself. Notice, however, that the sample sizes are different in cell A (n=408) versus cell D (n=376). This is because of missing data -- there are more missing observations for variable Weight than there are for variable Height.

If you have opted to flag significant correlations, SPSS will mark a 0.05 significance level with one asterisk (*) and a 0.01 significance level with two asterisks (0.01). In cell B (repeated in cell C), we can see that the Pearson correlation coefficient for height and weight is .513, which is significant ($p <$ .001 for a two-tailed test), based on 354 complete observations (i.e., cases with nonmissing values for both height and weight).

3.5.2 Decision and Conclusions

Based on the results, we can state the following:

- Weight and height have a statistically significant linear relationship (r=.513, p < .001).
- The direction of the relationship is positive (i.e., height and weight are positively correlated), meaning that these variables tend to increase together (i.e., greater height is associated with greater weight).
- The magnitude, or strength, of the association is approximately moderate ($.3 < | r | < .5$).

Self-Assessment Exercise 2

1. Demonstrate the Data Set-Up
2. Explain the problem Statement
3. Demonstrate the Running the Test
4. Explain the decision and conclusions

1.6 Summary

The unit explained Statistical Package for the Social Sciences (SPSS) is a software program used by researchers in various disciplines for quantitative analysis of complex data. This introductory level SPSS unit introduces SPSS environment, basic data preparation and management, descriptive statistics, and common statistical analysis (T-test, ANOVA, correlation, regression).

The bivariate Pearson Correlation is commonly used to measure the following:

a. Correlations among pairs of variables
b. Correlations within and between sets of variables
c. The bivariate Pearson correlation indicates the following
a. Whether a statistically significant linear relationship exists between two continuous variables
b. The strength of a linear relationship (i.e., how close the relationship is to being a perfectly straight line)
c. The direction of a linear relationship (increasing or decreasing)

To use Pearson correlation, your data must meet the following requirements:

i. Two or more continuous variables (i.e., interval or ratio level)
ii. Cases must have non-missing values on both variables
iii. Linear relationship between the variables
iv. Independent cases (i.e., independence of observations)

The null hypothesis (H_0) and alternative hypothesis (H_1) of the significance test for correlation can be expressed in the following ways, depending on whether a one-tailed or two-tailed test is requested:

Two-tailed significance test:

H_0: $\rho = 0$ ("the population correlation coefficient is 0; there is no association")
H_1: $\rho \neq 0$ ("the population correlation coefficient is not 0; a nonzero correlation could exist")

One-tailed significance test:

Your dataset should include two or more continuous numeric variables, each defined as scale, which will be used in the analysis.

Each row in the dataset should represent one unique subject, person, or unit. All of the measurements taken on that person or unit should appear in that row. If measurements for one subject appear on multiple rows

Perhaps you would like to test whether there is a statistically significant linear relationship between two continuous variables, weight and height (and by extension, infer whether the association is significant in the population). You can use a bivariate Pearson Correlation to test whether there is a statistically significant linear relationship between height and weight, and to determine the strength and direction of the association

To run the bivariate Pearson Correlation, click **Analyze** > **Correlate** > **Bivariate**. Select the variables Height and Weight and move them to the Variables box. In the **Correlation Coefficients** area, select **Pearson**. In the **Test of Significance** area, select your desired significance test, two-tailed or one-tailed

Based on the results, we can state the following:

Weight and height have a statistically significant linear relationship (r=.513, p < .001).
The direction of the relationship is positive (i.e., height and weight are positively correlated), meaning that these variables tend to increase together (i.e., greater height is associated with greater weight).
The magnitude, or strength, of the association is approximately moderate (.3 < | r | < .5).

3.7 References/Further Readings/Web Resources

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

UBC Library Research Commons (2022).Introduction to SPSS for statistical analysis. Retrieved from https://researchcommons.library.ubc.ca/ introduction-to-spss-for-statistical-analysis/

Perry H. (2014). SPSS Explained. Retrieved from https://researchcommons.library.ubc.ca/ introduction-to-spss-for-statistical-analysis/

Alan, B. (2012). Social Research Methods. Retrieved from https://researchcommons.library.ubc.ca/ introduction-to-spss-for-statistical-analysis/

Ton J. C. A., Zwinderman, H. (2010). SPSS for Starters. Retrieved from https://researchcommons.library.ubc.ca/ introduction-to-spss-for-statistical-analysis/

Rachad, A. (2012) Interpreting Quantitative Data with IBM SPSS Statistics. Retrieved from https://researchcommons.library.ubc.ca/ introduction-to-spss-for-statistical-analysis/

3.8 Possible Answers to SAEs

Answers to SAEs 1

1. Statistical Package for the Social Sciences (SPSS) is a software program used by researchers in various disciplines for quantitative analysis of complex data. This introductory level SPSS unit introduces SPSS environment, basic data preparation and management, descriptive statistics, and common statistical analysis (T-test, ANOVA, correlation, regression).

2. Common Uses of SPSS

The bivariate Pearson Correlation is commonly used to measure the following:

i. Correlations among pairs of variables
ii. Correlations within and between sets of variables

The bivariate Pearson correlation indicates the following
Whether a statistically significant linear relationship exists between two continuous variables
The strength of a linear relationship (i.e., how close the relationship is to being a perfectly straight line)
The direction of a linear relationship (increasing or decreasing)

3. Data Requirements

To use Pearson correlation, your data must meet the following requirements:

Two or more continuous variables (i.e., interval or ratio level)
Cases must have non-missing values on both variables
Linear relationship between the variables
Independent cases (i.e., independence of observations)

4. Hypotheses

The null hypothesis (H_0) and alternative hypothesis (H_1) of the significance test for correlation can be expressed in the following ways, depending on whether a one-tailed or two-tailed test is requested:
Two-tailed significance test:
H_0: $\rho = 0$ ("the population correlation coefficient is 0; there is no association")

H_1: $\rho \neq 0$ ("the population correlation coefficient is not 0; a nonzero correlation could exist")
One-tailed significance test:

Answers to SAEs 2

1. Data Set-Up

Your dataset should include two or more continuous numeric variables, each defined as scale, which will be used in the analysis.
Each row in the dataset should represent one unique subject, person, or unit. All of the measurements taken on that person or unit should appear in that row. If measurements for one subject appear on multiple rows

2. Problem Statements

Perhaps you would like to test whether there is a statistically significant linear relationship between two continuous variables, weight and height (and by extension, infer whether the association is significant in the population). You can use a bivariate Pearson Correlation to test whether there is a statistically significant linear relationship between height and weight, and to determine the strength and direction of the association

3. Running the Test

To run the bivariate Pearson Correlation, click Analyze > Correlate > Bi variate. Select the variables Height and Weight and move them to the Variables box. In the Correlation Coefficients area, select Pearson. In the Test of Significance area, select your desired significance test, two-tailed or one-tailed

4. Decision and Conclusions

Based on the results, we can state the following:

Weight and height have a statistically significant linear relationship (r=.513, p < .001).

The direction of the relationship is positive (i.e., height and weight are positively correlated), meaning that these variables tend to increase together (i.e., greater height is associated with greater weight).

The magnitude, or strength, of the association is approximately moderate (.3 < | r | < .5).

Unit 4 Time Series Analysis

Unit Structure

4.1 Introduction

This unit will be discussing Time Series Analysis. It will look at why organizations use time series data analysis, the types of Time Series Analysis and Models of time series analysis are discussed. Basic Objectives of the Analysis and Time Series forecasting will be discussed**.**

4.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the concept of Time Series Analysis.
- explain why organizations use time series data analysis,
- state the models of Time Series Analysis
- demonstrate the techniques of time series analysis
- state the Basic Objectives of the Analysis
- explain the Time Series forecasting

4.3 Time Series Analysis

4.3.1 Concept of Time series analysis

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

What sets time series data apart from other data is that the analysis can show how variables change over time. In other words, time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting—predicting future data based on historical data.

4.3.2 Why organizations use time series data analysis

Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time. Using data visualizations, business users can see seasonal trends and dig deeper into why these trends occur. With modern analytics platforms, these visualizations can go far beyond line graphs.

When organizations analyze data over consistent intervals, they can also use time series forecasting to predict the likelihood of future events. Time series forecasting is part of predictive analytics. It can show likely changes in the data, like seasonality or cyclic behaviour, which provides a better understanding of data variables and helps forecast better.

For example, Des Moines Public Schools analyzed five years of student achievement data to identify at-risk students and track progress over time. Today's technology allows us to collect massive amounts of data every day and it's easier than ever to gather enough consistent data for comprehensive analysis

Time series analysis examples

Time series analysis is used for non-stationary data—things that are constantly fluctuating over time or are affected by time. Industries like finance, retail, and economics frequently use time series analysis because currency and sales are always changing. Stock market analysis is an excellent example of time series analysis in action, especially with automated trading algorithms. Likewise, time series analysis is ideal for forecasting weather changes, helping meteorologists predict everything from tomorrow's weather report to future years of climate change. Examples of time series analysis in action include:

Weather data
Rainfall measurements
Temperature readings
Heart rate monitoring (EKG)
Brain monitoring (EEG)
Quarterly sales
Stock prices
Automated stock trading
Industry forecasts
Interest rates

4.4 Time Series Analysis Types

Because time series analysis includes many categories or variations of data, analysts sometimes must make complex models. However, analysts can't account for all variances, and they can't generalize a specific model to every sample. Models that are too complex or that try to do too many things can lead to a lack of fit. Lack of fit or over fitting models lead to those models not distinguishing between random error and true relationships, leaving analysis skewed and forecasts incorrect.

Models of time series analysis include:
Classification: Identifies and assigns categories to the data.

Curve fitting: Plots the data along a curve to study the relationships of variables within the data.

Descriptive analysis: Identifies patterns in time series data, like trends, cycles, or seasonal variation.

Explanative analysis: Attempts to understand the data and the relationships within it, as well as cause and effect.

Exploratory analysis: Highlights the main characteristics of the time series data, usually in a visual format.

Forecasting: Predicts future data. This type is based on historical trends. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points.

Intervention analysis: Studies how an event can change the data.

Segmentation: Splits the data into segments to show the underlying properties of the source information.

Data classification: Further, time series data can be classified into two main categories.

Stock time series data means measuring attributes at a certain point in time, like a static snapshot of the information as it was.

Flow time series data means measuring the activity of the attributes over a certain period, which is generally part of the total whole and makes up a portion of the results.

Data variations: In time series data, variations can occur sporadically throughout the data.

Functional analysis can pick out the patterns and relationships within the data to identify notable events.

Trend analysis means determining consistent movement in a certain direction. There are two types of trends: deterministic, where we can find the underlying cause, and stochastic, which is random and unexplainable.

Seasonal variation describes events that occur at specific and regular intervals during the course of a year. Serial dependence occurs when data points close together in time tend to be related.

Time series analysis and forecasting models must define the types of data relevant to answering the business question. Once analysts have chosen the relevant data they want to analyze, they choose what types of analysis and techniques are the best fit.

4.4.1 Important Considerations for Time Series Analysis

While time series data is data collected over time, there are different types of data that describe how and when that time data was recorded. For example:

Time series data is data that is recorded over consistent intervals of time.
Cross-sectional data consists of several variables recorded at the same time.
Pooled data is a combination of both time series data and cross-sectional data

Self-Assessment Exercise 1

1. Explain the concept of Time Series Analysis.
2. Explain why organizations use time series data analysis,
3. State the models of Time Series Analysis

4.5 Time Series Analysis Techniques

Just as there are many types and models, there are also a variety of methods to study data. Here are the three most common.

4.5.1 Box-Jenkins ARIMA models

These univariate models are used to better understand a single time-dependent variable, such as temperature over time, and to predict future data points of variables. These models work on the assumption that the data is stationary. Analysts have to account for and remove as many differences and seasonalities in past data points as they can. Thankfully, the ARIMA model includes terms to account for moving averages, seasonal difference operators, and autoregressive terms within the model.

4.5.2 Box-Jenkins Multivariate Models

Multivariate models are used to analyze more than one time-dependent variable, such as temperature and humidity, over time.

Holt-Winters Method: The Holt-Winters method is an exponential smoothing technique. It is designed to predict outcomes, provided that the data points include seasonality.

In this lesson, we'll describe some important features that we must consider when describing and modeling a time series. This is meant to be an introductory overview, illustrated by example, and not a complete look at how we model a univariate time series. Here, we'll only consider univariate time series. We'll examine relationships between two or more time series later on.

4.5.3 Univariate Time Series

A univariate time series is a sequence of measurements of the same variable collected over time. Most often, the measurements are made at regular time intervals.

One difference from standard linear regression is that the data are not necessarily independent and not necessarily identically distributed. One defining characteristic of a time series is that it is a list of observations where the ordering matters. Ordering is very important because there is dependency and changing the order could change the meaning of the data.

4.6 Basic Objectives of the Analysis

The basic objective usually is to determine a model that describes the pattern of the time series. Uses for such a model are:

i. To describe the important features of the time series pattern.
ii. To explain how the past affects the future or how two time series can "interact".
iii. To forecast future values of the series.
iv. To possibly serve as a control standard for a variable that measures the quality of product in some manufacturing situations.

4.7 Types of Models

There are two basic types of "time domain" models.

1. Autoregressive Integrated Moving Average

Models that relate the present value of a series to past values and past prediction errors - these are called ARIMA models (for Autoregressive Integrated Moving Average). We'll spend substantial time on these.
Ordinary regression models that use time indices as x-variables. These can be helpful for an initial description of the data and form the basis of several simple forecasting methods.

Important Characteristics to Consider First

Some important questions to first consider when first looking at a time series are:

Is there a **trend**, meaning that, on average, the measurements tend to increase (or decrease) over time?

Is there seasonality, meaning that there is a regularly repeating pattern of highs and lows related to calendar time such as seasons, quarters, months, days of the week, and so on?

Are there outliers? In regression, outliers are far away from your line. With time series data, your outliers are far away from your other data.

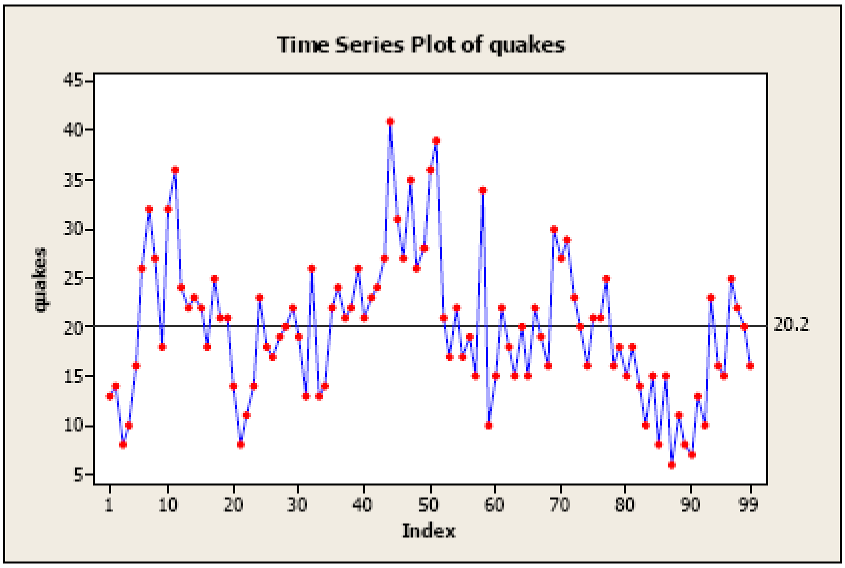
Is there a long-run cycle or period unrelated to seasonality factors?

Is there constant variance over time, or is the variance non-constant?

Are there any abrupt changes to either the level of the series or the variance?

Example 1-1

The following plot is a time series plot of the annual number of earthquakes in the world with seismic magnitude over 7.0, for 99 consecutive years. By a time series plot, we simply mean that the variable is plotted against time.



Some features of the plot:

There is no consistent trend (upward or downward) over the entire time span. The series appears to slowly wander up and down. The horizontal line drawn at quakes = 20.2 indicates the mean of the series. Notice that the series tends to stay on the same side of the mean (above or below) for a while and then wanders to the other side.

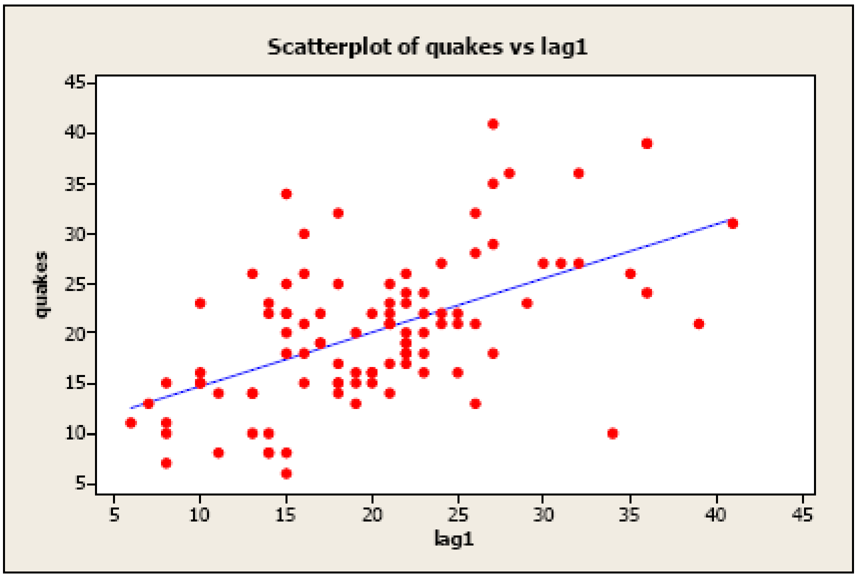
Almost by definition, there is no seasonality as the data are annual data.
There are no obvious outliers.
It's difficult to judge whether the variance is constant or not.

One of the simplest ARIMA type models is a model in which we use a linear model to predict the value at the present time using the value at the previous time. This is called an AR(1) model, standing for autoregressive model of order 1. The order of the model indicates how many previous times we use to predict the present time.

A start in evaluating whether an AR(1) might work is to plot values of the series against **lag 1 values** of the series. Let xt denote the value of the series at any particular time t, so xt−1 denotes the value of the series one time before time t. That is, xt−1 is the lag 1 value of xt. As a short example, here are the first five values in the earthquake series along with their lag 1 values:

T	**xt**	**xt−1 (lag 1 value)**
1	13	*
2	14	13
3	8	14
4	10	8
5	16	10

For the complete earthquake dataset, here's a plot of xt versus xt−1:



Although it's only a moderately strong relationship, there is a positive linear association so an AR(1) model might be a useful model.

The AR(1) model

Theoretically, the AR(1) model is written

$xt=\delta+\phi 1xt-1+wt$

Assumptions:

$wt\sim iidN(0,\sigma w2)$, meaning that the errors are independently distributed with a normal distribution that has mean 0 and constant variance.

Properties of the errors wt are independent of x.

This is essentially the ordinary simple linear regression equation, but there is one difference. Although it's not usually true, in ordinary least squares regression we assume that the x-variable is not random but instead is something we can control. That's not the case here, but in our first encounter with time series we'll overlook that and use ordinary regression methods. We'll do things the "right" way later in the course.

Following is Minitab output for the AR(1) regression in this example:

Quakes = 9.19 + 0.543 lag1

98 cases used, 1 cases contain missing values

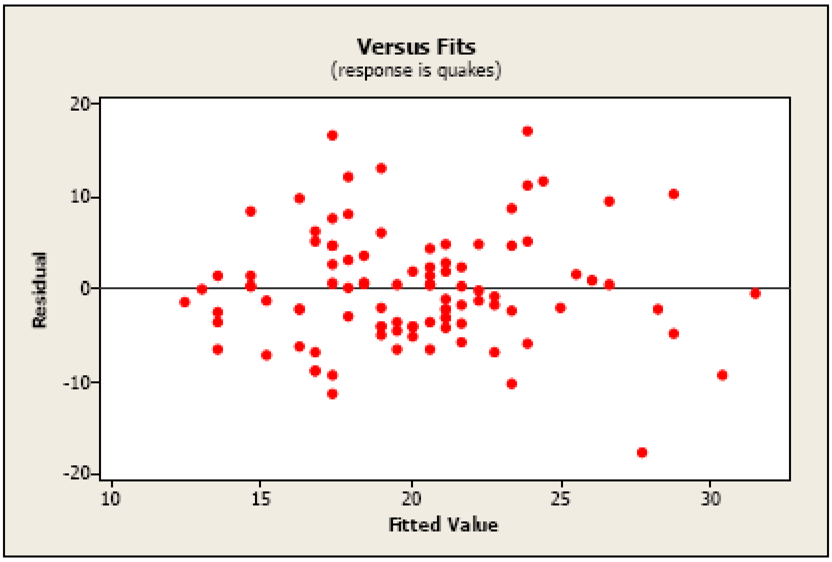
Predictor	**Coef**	**SE Coef**	**T**	**P**
Constant	9.191	1.819	5.05	0.000
lag1	0.54339	0.08528	6.37	0.000

S = 6.12239 R-Sq = 29.7% R-Sq(adj) = 29.0%

We see that the slope coefficient is significantly different from 0, so the lag 1 variable is a helpful predictor. The R2 value is relatively weak at 29.7%, though, so the model won't give us great predictions.

Residual Analysis

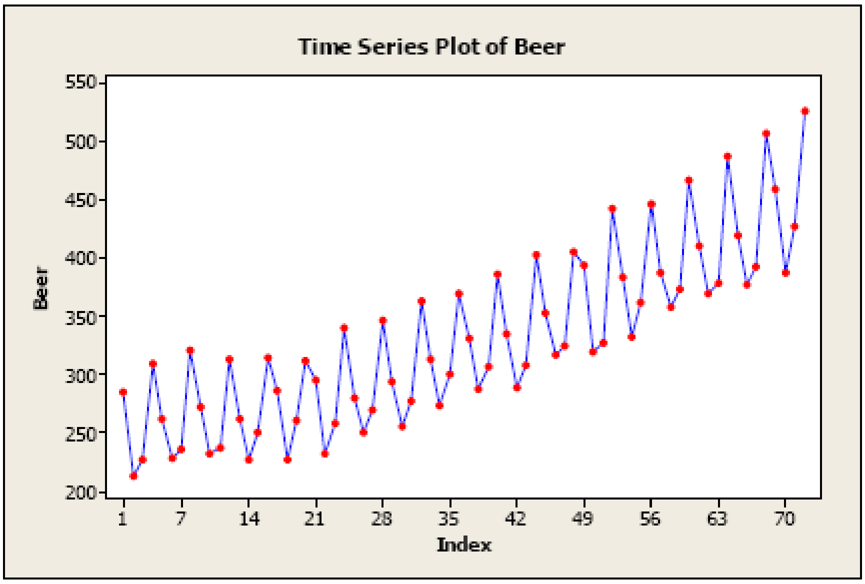
In traditional regression, a plot of residuals versus fits is a useful diagnostic tool. The ideal for this plot is a horizontal band of points. Following is a plot of residuals versus predicted values for our estimated model. It doesn't show any serious problems. There might be one possible outlier at a fitted value of about 28.



Example 1-2
The following plot shows a time series of quarterly production of beer in Australia for 18 years.

Some important features are:

- There is an upward trend, possibly a curved one.
- There is seasonality – a regularly repeating pattern of highs and lows related to quarters of the year.
- There are no obvious outliers.
- There might be increasing variation as we move across time, although that's uncertain.



There are ARIMA methods for dealing with series that exhibit both trend and seasonality, but for this example, we'll use ordinary regression methods.

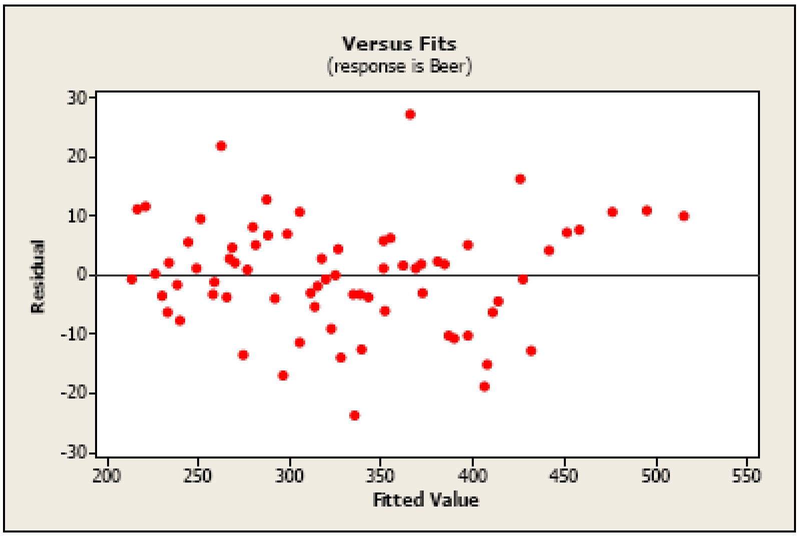
Classical regression methods for trend and seasonal effects
To use traditional regression methods, we might model the pattern in the beer production data as a combination of the trend over time and quarterly effect variables.

Suppose that the observed series is xt, for t=1,2,…,n.
For a linear trend, use t (the time index) as a predictor variable in a regression.

For a quadratic trend, we might consider using both t and t2.
For quarterly data, with possible seasonal (quarterly) effects, we can define indicator variables such as Sj=1 if the observation is in quarter j of a year and 0 otherwise. There are 4 such indicators.

Let $\epsilon t \sim iidN(0,\sigma 2)$. A model with additive components for linear trend and seasonal (quarterly) effects might be written
$xt=\beta 1t+\alpha 1S1+\alpha 2S2+\alpha 3S3+\alpha 4S4+\epsilon t$
To add a quadratic trend, which may be the case in our example, the model is
$xt=\beta 1t+\beta 2t2+\alpha 1S1+\alpha 2S2+\alpha 3S3+\alpha 4S4+\epsilon t$

Note!
We've deleted the "intercept" from the model. This isn't necessary, but if we include it we'll have to drop one of the seasonal effect variables from the model to avoid collinearity issues.



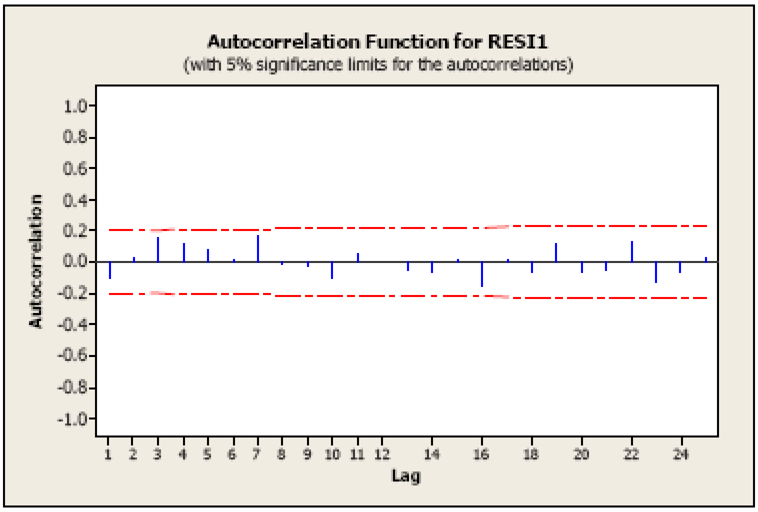
When data are gathered over time, we typically are concerned with whether a value at the present time can be predicted from values at past times. We saw this in the earthquake data of example 1 when we used an AR(1) structure to model the data. For residuals, however, the desirable result is that the correlation is 0 between residuals separated by any given time span. In other words, residuals should be unrelated to each other.

Sample Autocorrelation Function (ACF)

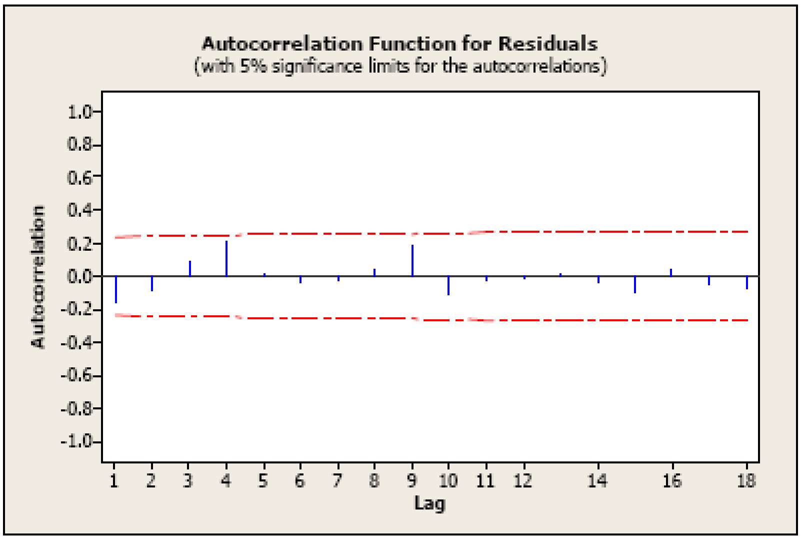
The sample autocorrelation function (ACF) for a series gives correlations between the series xt and lagged values of the series for lags of 1, 2, 3, and so on. The lagged values can be written as xt−1,xt−2,xt−3, and so on. The ACF gives correlations between xt and xt−1, xt and xt−2, and so on.

The ACF can be used to identify the possible structure of time series data. That can be tricky going as there often isn't a single clear-cut interpretation of a sample autocorrelation function. We'll get started on that in Lesson 1.2 this week. The ACF of the residuals for a model is also useful. The ideal for a sample ACF of residuals is that there aren't any significant correlations for any lag.

Following is the ACF of the residuals for Example 1, the earthquake example, where we used an AR(1) model. The "lag" (time span between observations) is shown along the horizontal, and the autocorrelation is on the vertical. The red lines indicated bounds for statistical significance. This is a good ACF for residuals. Nothing is significant; that's what we want for residuals.



The ACF of the residuals for the quadratic trend plus seasonality model we used for Example 2 looks good too. Again, there appears to be no significant autocorrelation in the residuals. The ACF of the residual follows:



4.8 Time Series forecasting

Time series forecasting is a technique for the prediction of events through a sequence of time. The technique is used across many fields of study, from the geology to behavior to economics. The techniques predict future events by analysing the trends of the past, on the assumption that future trends will hold similar to historical trends.

Consider the following periodic data:

Year	Y
1990	50
1991	80
1992	90
1993	49
1994	75
1995	58
1996	82
1997	73
1998	95

From the above data using the equation y = ax + b;
1. Calculate Slope a?
2. Calculate Intercept b
3. Assemble the equation of a line
4. Use $\hat{Y} = \hat{a} + bt$ to forecast the value of output, Y, for year 2003

Solution:
Let t representing period (years) and Y representing output
ty = 30708
y = 625
t = 45

$t^2 = 285$
$(t)^2 = 2025$

a) $b= \frac{N \Sigma(ty) - \Sigma t \Sigma y}{N \Sigma(t^2) - (\Sigma t)^2}$ $\frac{9(3412)-45(625)}{9(285) - 2025}$ $\frac{30708- 29340}{2562 - 2025}$

$b= \frac{2583}{540} = b = 5$

b) $a= \frac{\Sigma y - b(\Sigma t)}{N}$

$\frac{625 - 5(45)}{9}$ $\frac{625- 229.6}{9}$ $= 395.4/9 = 44$

c) Assemble of Regression equation line=
$y = 44x + 5$

d) Forecast the value of output, Y, for year 2003. Following the systematic process, the year 2003 is associated with the numerical value, t = 14, so that for t = 14

To forecast value for output in year 2003, you have to use the
$\hat{Y} = \hat{a} + bt$
Find the value for $\hat{Y}$, $\hat{a}$ and t =
$\hat{Y} = 625/9 = 69.4$
$t = 45/9 = 5$
$\hat{a} = \hat{Y} - bt$
$\hat{a} = 69.4- 5(5) = 44.4$
Fix the least – squares line: $\hat{Y} = \hat{a} + bt$
$\hat{Y}= 44.4+5t$
the forecast value for output in year 2003 is:
$Y = 44.4 + 5(14)$
Therefore, the forecast value for output in year 2003 is114.4

Self-Assessment Exercise 2

1. Demonstrate the techniques of time series analysis
2. State the Basic Objectives of the Analysis
3. Explain the Time Series forecasting

4.9 References/Further Reading/Web Resources

Tableau.com(2022). Time Series Analysis Retrieved from https://www.tableau.com/learn/articles/time-series-analysis#definition.

4.10 Possible Answers to SAEs

Answers to SAEs 1

1. Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.
   What sets time series data apart from other data is that the analysis can show how variables change over time

2. Why organizations use time series data analysis
   Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time. Using data visualizations, business users can see seasonal trends and dig deeper into why these trends occur. With modern analytics platforms, these visualizations can go far beyond line graphs
   Examples of time series analysis in action include:
   Weather data
   Rainfall measurements
   Temperature readings
   Heart rate monitoring (EKG)
   Brain monitoring (EEG)
   Quarterly sales
   Stock prices
   Automated stock trading
   Industry forecasts
   Interest rates

3. Model of Time Series Analysis
   Because time series analysis includes many categories or variations of data, analysts sometimes must make complex models. However, analysts can’t account for all variances, and they can’t generalize a specific model to every sample. Models

that are too complex or that try to do too many things can lead to a lack of fit. Lack of fit or over fitting models lead to those models not distinguishing between random error and true relationships, leaving analysis skewed and forecasts incorrect

Models of time series analysis include:
Classification: Identifies and assigns categories to the data.

Curve fitting: Plots the data along a curve to study the relationships of variables within the data.

Descriptive analysis: Identifies patterns in time series data, like trends, cycles, or seasonal variation.

Explanative analysis: Attempts to understand the data and the relationships within it, as well as cause and effect.

Exploratory analysis: Highlights the main characteristics of the time series data, usually in a visual format.

Forecasting: Predicts future data. This type is based on historical trends. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points.

Intervention analysis: Studies how an event can change the data.

Segmentation: Splits the data into segments to show the underlying properties of the source information

Answers to SAEs 2

1. Time Series Analysis Techniques

Box-Jenkins ARIMA models: These univariate models are used to better understand a single time-dependent variable, such as temperature over time, and to predict future data points of variable

Box-Jenkins Multivariate Models: Multivariate models are used to analyze more than one time-dependent variable, such as temperature and humidity, over time

Holt-Winters Method: The Holt-Winters method is an exponential smoothing technique. It is designed to predict outcomes, provided that the data points include seasonality

2. Basic Objectives of the Analysis

The basic objective usually is to determine a model that describes the pattern of the time series. Uses for such a model are:

a. To describe the important features of the time series pattern.
b. To explain how the past affects the future or how two time series can "interact".
c. To forecast future values of the series.
d. To possibly serve as a control standard for a variable that measures the quality of product in some manufacturing situations

3. Types of Models

There are two basic types of "time domain" models.

i. Autoregressive Integrated Moving Average
Models that relate the present value of a series to past values and past prediction errors - these are called ARIMA models (for Autoregressive Integrated Moving Average). We'll spend substantial time on these

ii. Time Series forecasting
Time series forecasting is a technique for the prediction of events through a sequence of time. The technique is used across many fields of study, from the geology to behavior to economics. The techniques predict future events by analysing the trends of the past, on the assumption that future trends will hold similar to historical trends

Unit 5 Forecasting

Unit Structure

5.1 Introduction

This unit will be discussing forecasting. Forecasts are based on past performances. In other words, future values are predicted from past values. This assumes that the future will be basically the same as the past and present, implying that the relationships underlying the phenomenon of interest are stable overtime. Forecasting can be performed at different levels, depending on the use to which it will be put. Simple guessing, based on previous figures, is occasionally adequate. However, where there is a large investment at stake, structured forecasting is essential.

Any forecasts made, however technical or structured should be treated with caution, since the analysis is based on past data and there could be unknown factors present in the future. However it is often reasonable to assume that patterns that have been identified in the analysis of past data will be broadly continued, at least into the short-term future.

In preparing plans for the future, the management authority has to make some predictions about what is likely to happen in the future. It shows that the managers know something of future happenings even before things actually happen. Forecasting provides them this knowledge.

5.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the concept of Forecasting?
- outline the Steps in Forecasting
- itemize the types of forecasting
- state the Forecasting Techniques
- explain the Role of Forecasting
- explain the Process of Forecasting:
- state and explain the Techniques of Forecasting:

5.3 Forecasting

5.3.1 What is Forecasting?

Forecasting is the process of estimating the relevant events of future, based on the analysis of their past and present behaviour.

Forecasts are based on past performances. In other words, future values are predicted from past values. This assumes that the future will be basically the same as the past and present, implying that the relationships underlying the phenomenon of interest are stable overtime. The future cannot be probed unless one knows how the events have occurred in the past and how they are occurring presently. The past and present analysis of events provides the base helpful for collecting information about their future occurrence.

Thus, forecasting may be defined as the process of assessing the future normally using calculations and projections that take account of the past performance, current trends, and anticipated changes in the foreseeable period ahead.

Whenever the managers plan business operations and organisational set-up for the years ahead, they have to take into account the past, the present and the prevailing economic, political and social conditions. Forecasting provides a logical basis for determining in advance the nature of future business operations and the basis for managerial decisions about the material, personnel and other requirements.

It is, thus, the basis of planning, when a business enterprise makes an attempt to look into the future in a systematic and concentrated way, it may discover certain aspects of its operations requiring special attention.

However, it must be recognised that the process of forecasting involves an element of guesswork and the managers cannot stay satisfied and relaxed after having prepared a forecast.

The forecast will have to be constantly monitored and revised— particularly when it relates to a long- term period. The managers should try to reduce the element of guesswork in preparing forecasts by collecting the relevant data using the scientific techniques of analysis and inference.

5.3.2 Steps in Forecasting

We outline the basic steps in forecasting as follows:

Step 1: Gather past data: daily, weekly, monthly and yearly.

Step 2: Adjust or clean up the raw data against inflationary factors. Index numbers can be used in deflating inflationary factors.

Step 3: Make forecasts from the "refined" data

Step 4: When the future data (which is been forecast) becomes available, compare forecasts with actual values, by so doing, you will be able to establish the error due to forecasting.

5.3.3 Types of forecasting

1. Short-term forecasts: these are forecasts concerning the near future. They, are characterized by few uncertainties and therefore more accurate then distant future forecasts

2. Long – term forecasts: these concern the distant future. They are characterized by more uncertainties than short – term forecasts

3. Extrapolation: these are forecasts based solely on past and present values of the variable to be forecast.

4. Forecasts based on established relationships between the variable to be forecast and other variables.

5.4 Forecasting Techniques

The two generally used methods of forecasting include:

i. Moving averages
ii. Trend lines or least squares.

On the basis of the definition, the following features of forecasting can be identified:

1. Forecasting relates to future events.
2. Forecasting is needed for planning process because it devises the future course of action.
3. It defines the probability of happening of future events. Therefore, the happening of future events can be precise only to a certain extent.
4. Forecasting is made by analysing the past and present factors which are relevant for the functioning of an organisation.
5. The analysis of various factors may require the use of statistical and mathematical tools and techniques.

It helps the managers in planning; Forecasting is the key to planning. It generates the planning process. Planning decides the future course of action which is expected to take place in certain circumstances and conditions. Unless the managers know these conditions, they cannot go for effective planning.

Forecasting provides the knowledge of planning premises within which the managers can analyse their strengths and weaknesses and can take appropriate actions in advance before actually they are put out of market. Forecasting provides the knowledge about the nature of future conditions.

Self-Assessment Exercise 1

1. Explain the concept of Forecasting?
2. Itemize the types of forecasting
3. State the Forecasting Techniques

5.4.1 Role of Forecasting

Since planning involves the future, no usable plan can be made unless the manager is able to take all possible future events into account. This explains why forecasting is a critical element in the planning process. In fact, every decision in the organisation is based on some sort of forecasting.

Though forecasting cannot check the future happenings, it provides clues about those and indicates when the alternative actions should be taken. Managers can save their business and face the unfortunate happenings if they know in advance what is going to happen.

5.4.2 Process of Forecasting

The process of forecasting generally involves the following steps:

1. Developing the Basis:
   The future estimates of various business operations will have to be based on the results obtainable through systematic investigation of the economy, products and industry.

2. Estimation of Future Operations:
   On the basis of the data collected through systematic investigation into the economy and industry situation, the manager has to prepare quantitative estimates of the future scale of business operations. Here the managers will have to take into account the planning premises.

3. Regulation of Forecasts:
   It has already been indicated that the managers cannot take it easy after they have formulated a business forecast. They have to constantly compare the actual operations with the forecasts prepared in order to find out the reasons for any deviations from forecasts. This helps in making more realistic forecasts for future.

4. Review of the Forecasting Process:
   Having determined the deviations of the actual performances from the positions forecast by the managers, it will be necessary to examine the procedures adopted for the purpose so that improvements can be made in the method of forecasting.

5.5 Techniques of Forecasting

There are various methods of forecasting. However, no method can be suggested as universally applicable. In fact, most of the forecasts are done by combining various methods.

A brief discussion of the major forecasting methods is given below:

1. Historical Analogy Method

   Under this method, forecast in regard to a particular situation is based on some analogous conditions elsewhere in the past. The economic situation of a country can be predicted by making comparison with the advanced countries at a particular stage through which the country is presently passing.
   Similarly, it has been observed that if anything is invented in some part of the world, this is adopted in other countries after a gap of a certain time. Thus, based on analogy, a general forecast

can be made about the nature of events in the economic system of the country. It is often suggested that social analogies have helped in indicating the trends of changes in the norms of business behaviour in terms of life.
Likewise, changes in the norms of business behaviour in terms of attitude of the workers against inequality, find similarities in various countries at various stages of the history of industrial growth. Thus, this method gives a broad indication about the future events of general nature.

2. Survey Method

Surveys can be conducted to gather information on the intentions of the concerned people. For example, information may be collected through surveys about the probable expenditure of consumers on various items. Both quantitative and qualitative information may be collected by this method.

On the basis of such surveys, demand for various products can be projected. Survey method is suitable for forecasting demand—both of existing and new products. To limit the cost and time, the survey may be restricted to a sample from the prospective consumers.

3. Opinion Poll

Opinion poll is conducted to assess the opinion of the experienced persons and experts in the particular field whose views carry a lot of weight. For example, opinion polls are very popular to predict the outcome of elections in many countries including India. Similarly, an opinion poll of the sales representatives, wholesalers or marketing experts may be helpful in formulating demand projections (yourarticlelibrary.com, 2022).
If opinion polls give widely divergent views, the experts may be called for discussion and explanation of why they are holding a particular view. They may be asked to comment on the views of the others, to revise their views in the context of the opposite views, and consensus may emerge. Then, it becomes the estimate of future events.

4. Business Barometers

A barometer is used to measure the atmospheric pressure. In the same way, index numbers are used to measure the state of an economy between two or more periods. These index numbers are the device to study the trends, seasonal fluctuations, cyclical movements, and irregular fluctuations.

These index numbers, when used in combination with one another, provide indications as to the direction in which the economy is proceeding. Thus, with the business activity index numbers, it becomes easy to forecast the future course of action.

However, it should be kept in mind that business barometers have their own limitations and they are not sure road to success. All types of business do not follow the general trend but different index numbers have to be prepared for different activities, etc.

5. Time Series Analysis

Time series analysis involves decomposition of historical series into its various components, viz. trend, seasonal variances, cyclical variations, and random variances. When the various components of a time series are separated, the variation of a particular situation, the subject under study, can be known over the period of time and projection can be made about the future.

A trend can be known over the period of time which may be true for the future also. However, time series analysis should be used as a basis for forecasting when data are available for a long period of time and tendencies disclosed by the trend and seasonal factors are fairly clear and stable.

6. Regression Analysis

Regression analysis is meant to disclose the relative movements of two or more inter-related series. It is used to estimate the changes in one variable as a result of specified changes in other variable or variables. In economic and business situations, a number of factors affect a business activity simultaneously.

Regression analysis helps in isolating the effects of such factors to a great extent. For example, if we know that there is a positive relationship between advertising expenditure and volume of sales or between sales and profit, it is possible to have estimate of the sales on the basis of advertising, or of the profit on the basis of projected sales, provided other things remain the same.

7. Input-Output Analysis

According to this method, a forecast of output is based on given input if relationship between input and output is known. Similarly, input requirement can be forecast on the basis of final output with a given input-output relationship. The basis of this

technique is that the various sectors of economy are interrelated and such inter-relationships are well-established.
For example, coal requirement of the country can be predicted on the basis of its usage rate in various sectors like industry, transport, household, etc. and how the various sectors behave in future. This technique yields sector-wise forecasts and is extensively used in forecasting business events as the data required for its application are easily obtained

Example:
A monthly sales of ABC Company is given as:

Months	Jan	Feb	Mar	Apr	May	Jun
Sale(s)	50	55	70	50	45	90

Using a 3 – period moving averages, forecast the sale for the three (3) months.

Solution

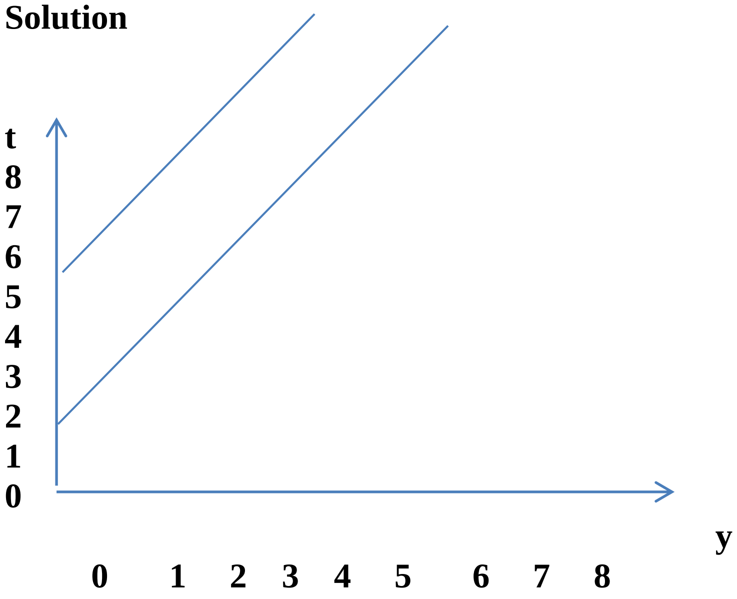
$\frac{50+55+70}{3}$ Jan = 175/3 = 58.3

$\frac{55+70+50}{3}$ Feb = 175/3 = 58.3

$\frac{50+70+45}{3}$ Mar = 165/3 = 55.0

Suppose the line AB in the following straight line reasonably approximates a set of data for 1995 – 2000. Represent the above data in a diagram

Solution



Self-Assessment Exercise 2

A monthly sales of ABC Company is given as:

Months	Jan	Feb	Mar	Apr	May	Jun
Sale(s)	60	65	70	75	80	90

1. Using a 3 – period moving averages, calculate the forecast sale for three months
2. State the Basic step in forecasting

1.6 Summary

Forecasting is the process of estimating the relevant events of future, based on the analysis of their past and present behaviour.

Forecasts are based on past performances. In other words, future values are predicted from past values. This assumes that the future will be basically the same as the past and present, implying that the relationships underlying the phenomenon of interest are stable overtime**.**

The unit itemized the types of forecasting to include; Short-term forecasts: these are forecasts concerning the near future. They, are characterized by few uncertainties and therefore more accurate then distant future forecasts. Long – term forecasts: these concern the distant future. They are characterized by more uncertainties than short – term forecasts. Extrapolation: these are forecasts based solely on past and present values of the variable to be forecast. Forecasts based on established relationships between the variable to be forecast and other variables.

The two generally used methods of forecasting include: Moving averages and Trend lines or least squares.

1.7 References/Further Readings/Web Resources

Yourarticlelibrary.com (2022). forecasting/forecasting. Retrieved from 2022https://www.yourarticlelibrary.com/management/forecasting /forecasting-roles-steps-and-techniques-management-function/70032.

5.8 Possible Answers to SAEs

Answers to SAEs 1

1. What is Forecasting?

Forecasting is the process of estimating the relevant events of future, based on the analysis of their past and present behaviour.

Forecasts are based on past performances. In other words, future values are predicted from past values. This assumes that the future will be basically the same as the past and present, implying that the relationships underlying the phenomenon of interest are stable overtime**.**

2. Types of forecasting

1. Short-term forecasts: these are forecasts concerning the near future. They, are characterized by few uncertainties and therefore more accurate then distant future forecasts

2. Long – term forecasts: these concern the distant future. They are characterized by more uncertainties than short – term forecasts

3. Extrapolation: these are forecasts based solely on past and present values of the variable to be forecast.

4. Forecasts based on established relationships between the variable to be forecast and other variables.

3. Forecasting Techniques

The two generally used methods of forecasting include:
i Moving averages
ii Trend lines or least squares.

Answers to SAEs 2

1.
A monthly sales of ABC Company is given as:

Months	Jan	Feb	Mar	Apr	May	Jun
Sale(s)	60	65	70	75	80	90

Using a 3 – period moving averages, forecast the sale for three months
Solution

$\frac{65+70+75}{3}$ Jan = 210/3 = 70.0

$\frac{70+75+80}{3}$ Feb = 225/3 = 75.0

$\frac{75+80+90}{3}$ Mar = 245/3 = 81.7

2. Basic step in forecasting

Step 1: Gather past data: daily, weekly, monthly and yearly.

Step 2: Adjust or clean up the raw data against inflationary factors.

Step 3: Index numbers can be used in deflating inflationary factors.

Step 4: Make forecasts from the "refined" data. When the future data (which is been forecast) becomes available, compare forecasts with actual values, by so doing, you will be able to establish the error due to forecasting.

MODULE 5

Unit 1 Index Numbers

Unit Structure

1.1 Introduction

This unit will discuss the Index numbers. Index numbers are often used in stabilising the time value of money and in deflating nominal values. An index number measures the percentage change in the value of some economic commodity over a period of time. It will outline the types and Formula for Number Index.

1.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the Index numbers
- outline the types and
- state and Demonstrate the Formula for Number Index

1.3 Index Number

1.3.1 What is Index Number?

An index number is a method of evaluating variations in a variable or group of variables in regards to geographical location, time, and other features. The base value of the index number is usually 100, which indicates price, date, level of production, and more.

There are various kinds of index numbers. However, at present, the most relatable is the price index number that particularly indicates the changes in the overall price level (or in the value of money) for a particular time.

According to Croxton and Cowden, index numbers are devices for measuring differences in the magnitude of a group of related variables.
According to Spiegal, an index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographical locations, or other characteristics.

Here, the value of money is not constant, even if it falls or rises it will affect and change the price level. An increase in the price level determines a decline in the value of money. A decrease in the price level means an increase in the value of money.

Therefore, the differences in the value of money are indicated by the differences in the overall price level for a particular time. Therefore, the changes in the overall prices can be evaluated by a statistical device known as ‘index number (Byjus.com, 2022).’

1.4 Types of Index Numbers

Price index number: It evaluates the relative differences in costs between two particular points in time.

Quantity index number: It measures the differences in the physical quantity of the product’s manufacturing, buying, or selling of one item or a group of items.
Index Numbers: meaning and characteristics

1.4.1 Important characteristics of index numbers

Following are the important characteristics of index numbers.	
(1) Expressed in Percentage	● A change in terms of the absolute values may not be comparable. ● Index numbers are expressed in percentage, so they remove this barrier. Although, we do not use the percentage sign. ● It is possible to compare the agricultural production and industrial production and at the same time being expressed in percentage, we can also compare the change in prices of different commodities.
(2) Relative measures or measures of net changes	● Index numbers measure a net or relative change in a variable or a group of variables. ● For example, if the price of a certain commodity rises from ₹10 in the year 2007 to ₹15 in the year 2017, the price index number will be 150 showing that there is a 50% increase in the prices over this period.
(3) Measure change over a period of time or in two or more places	● Index numbers measure the net change among the related variables over a period of time or at two or more places. ● For example, change in prices, production, and more, over the two periods or at two places.
(4) Specialised average	● Simple averages like, mean, median, mode, and more can be used to compare the variables having similar units. ● Index numbers are specialised average, expressed in percentage, and help in measuring and comparing the change in those variables that are expressed in different units. ● For example, we can compare the change in the production of industrial goods and agricultural goods.
(5) Measuring changes that are not directly measurable	● Cost of living, business activity, and more are complex things that are not directly measurable. ● With the help of index numbers, it is

	possible to study the relative changes in such phenomena.

1.4.2 Advantages of index numbers

What are the advantages of index numbers?	
Index numbers are one of the most widely used statistical tools. Some of the advantages or uses of index numbers are as follows:	
(1) Help in formulating policies	● Most of the economic and business decisions and policies are guided by the index numbers. ● Example: ● To increase DA, the government refers to the cost-of-living index. ● To make any policy related to the industrial or agricultural production, the government refers to their respective index numbers.
(2) Help in study of trends	● Index numbers help in the study of trends in variables like, export-import, industrial and agricultural production, share prices, and more.
(3) Helpful in forecasting	● Index numbers not only help in the study of past and present behaviour, they are also used for forecasting economic and business activities.
(4) Facilitates comparative study	● To make comparisons with respect to time and place especially where units are different, index numbers prove to be very useful. ● For example, change in ‘industrial production’ can be compared with change in ‘agricultural production’ with the help of index numbers.
(5) Measurement of purchasing power of money to maintain standard of living	● Index numbers, such as cost inflation index help in measuring the purchasing power of money at different times between different regions.

	● Such analysis helps the government to frame suitable policies for maintaining or raising the standard of living of the people.
(6) Act as economic barometer	● Index numbers are very useful in knowing the level of economic and business activities of a country. So, these are rightly known as economic barometers.

Problems involved in the construction of index numbers

Following are some of the problems involved in the construction of index numbers:	
(1) Purpose of index numbers	● Many different types of index numbers are constructed with different objectives. ● Example: Price index, quantity index, consumer price index, wholesale price index, and more ● So, the first important issue/problem is to define the objective for which the index number is to be constructed.
(2) Selection of base period	● Base period is the period against which the comparisons are made. ● Selection of a suitable base period is a very crucial step. ● It should be of reasonable length and normal one, i.e., it should not be affected by any abnormalities like, natural calamities, war, extreme business cycle situations. ● It should neither be too close nor too far.
(3) Selection of commodities	● All the items cannot be included in the construction of an index number. ● Nature and number of items to be included in an index number depends upon the type of index to be constructed. ● For example, to construct a ‘consumer price index’ those commodities should be considered that are generally consumed and the number should be neither too small nor too big.
(4) Selection of sources of	● Depending upon the type of index numbers, the correct source should be selected for data.

data	● Like, to construct CPI, we need retail prices and to construct the wholesale price index, we need wholesale prices. Accordingly, the right and reliable source should be selected.
(5) Selection of weights	● The term ‘weight’ refers to the relative importance of different items in the construction of index numbers. ● All the items do not have the same importance. ● So, it is necessary to adopt some suitable measures to assign weight.
(6) Selection of an appropriate formula	● There are various formulas for construction of index numbers like Laspeyres’ method, Paasche’s method, Fisher’s method, and more. ● No single formula is appropriate for all types of index numbers. ● The choice of formula depends upon the purpose of the available data.

(A) Consumer price index number	● Consumer price index (CPI) measures changes in the cost of living due to changes in the retail prices of a basket of goods over a period of time. ● Separate cost of living index is prepared for different classes of people. ● It is also known as the cost of living index numbers or retail price index number.
The uses of consumer price index number:	
(1) Helpful in measuring the purchasing power of money	● Consumer price index has an inverse relation with the purchasing power of money. ● Purchasing power of money = 1/Consumer price index ● As CPI increases, the purchasing power of money decreases.
(2) Helpful in wage negotiations	● CPI helps in determining wages for a particular class. ● It provides the basis for wage negotiations between the workers and employers.
(3) Help government	● These index numbers provide guidelines

in framing policies	for the formulation of wage policy, price policy, taxation policy, and other general economic policies.
(4) Market analysis	● CPI also helps a market analyst to determine the demand for different goods and services.
(5) Help businessmen in forecasting	● On the basis of CPI of different classes of people, a businessman can make predictions about the demand for his products.
The wholesale price index numbers. What is the utility of wholesale price index numbers?	
(A) Meaning of wholesale price index (WPI)	● Wholesale price index measures the changes in the wholesale prices of the commodities. ● It indicates the change in general price level in the economy. ● In India, it is prepared on a weekly basis. ● These days 2004-05 is considered as the base year
Utility of wholesale price index (WPI)	
(1) Indicator of inflation	● Inflation is a persistent rise in the general price level. ● WPI is used to determine the rate of inflation in an economy.
(2) Forecasting demand and supply	● It is often used to forecast demand and supply situations in the economy. ● An increase in WPI indicates a situation of excess demand over supply of goods. ● A decrease in WPI indicates a situation of excess supply of goods.
(3) Helps in determining real changes in aggregates	● WPI helps us to find out the real changes in aggregates like national income, national expenditure, and more. ● Using WPI of the current year and the base year, we can convert national income at current prices into national income at constant prices.
(4) Cost of projects	● It determines the future cost of long-

	run[1] projects. ● If WPI has an increasing trend, it will result in an increase in prices of various goods used in the projects. ● As a result the cost of such a project will go up.
Following are the major limitations of index numbers:	
(1) Difficulty in construction of index numbers	● The decision of objective, selection of base period, selection of commodities, selection of sources of data, selection of 'weights', selection of formula, and more are the several difficulties in the construction of index numbers.
(2) Based on sample items, so only approximate indicators	● Index numbers are generally based on a few sample items. So, the results derived are approximate and not perfect.
(3) Ignores quality of commodities	● These days the quality changes occur very fast and the index numbers ignore this aspect. ● So, the results shown by these may not be appropriate.
(4) Limited use	● There is no 'master index number' or 'all in one index number'. ● Use of each index number is restricted to its specific object.
(5) Useful only for short-term comparison	● Over a period of time, rapid changes occur in habits, tastes, preferences, and more. ● So,the index number constructed in the present may not be comparable with the one constructed a few years back.

Self-Assessment Exercise 1

Multiple Choice Questions

Q.1 According to ______________, the index numbers are devices for measuring differences in the magnitude of a group of related variables.

Q.2 Which of the following is the characteristic of an index number?

Q.3 Index numbers measure a net or relative change in a variable or a group of variables

1.5 Formula for Number Index

Index numbers are expressed in terms of a base of 100.
INDEX NUMBER could be Simple Average or Price Relatives Method or Aggregate Average. In this method, we find out the price relative of individual items and average out the individual values. Price relative refers to the percentage ratio of the value of a variable in the current year to its value in the year chosen as the base.
Price relative Index (R) = (P1÷P2) × 100

Example:
Calculate the Relative Price Index Time Series for Monthly Salaries of workers

Years	Average Monthly Salaries (N)
1985	1,200
1990	2,000
1995	1,800
2000	3,600
2005	5000
2010	6500

Solution

Years	Average Monthly Salaries (N)	Salary Index (1985 =100)
1985	1,200	100%
1990	2,000	200/1200 = 1.66x100 = 167%
1995	1,800	1800/2000 = 0.9x100 = 90%
2000	3,600	3600/1800 = 2%x100 = 200%
2005	5000	5000/3600 = 1.39x100 = 139%
2010	6500	6500/5000 = 1.3x100 = 130%

Calculate the Simple Aggregate for Average Consumer Prices for Selected Staple Food

Items	2000	2006

Sugar	10	40
Wheat Flour	11	20
Butter	71	99
Ground beef	91	186
Frying Chicken	39	88

Solution

For Year 2006

40+20+99+186+88

10+11+71+91+39

=195.06 This implies that the prices of the five items are higher by 95.06 percent in 2006 than in 2000.

Self-Assessment Exercise 2

The weighted average method of forecasting to calculate the Share of the following companies.

Company	No of Share 2004	No of Share 2006	Price Per Share (N) 2004	Price Per Share (N) 2006
A	350	400	0.50	1.25
B	200	180	1.25	3.75
C	140	200	6.25	12.50
D	130	150	12.50	18.75

1.6 Summary

INDEX NUMBER could be Simple Average or Price Relatives Method or Aggregate Average. In this method, we find out the price relative of individual items and average out the individual values. Price relative refers to the percentage ratio of the value of a variable in the current year to its value in the year chosen as the base.

1.7 References/Further Readings/Web Resources

Byjus.com (2022) meaning and characteristics of index numbers https://byjus.com/commerce/meaning-and-characteristics-of-index-numbers/

1.8 Possible Answers to SAEs

Answers to SAEs 1

Q.1 According to _______, the index numbers are devices for measuring differences in the magnitude of a group of related variables.

a. Spiegel
b. Croxton and Cowden
c. Both (a) and (b)
d. None of the above

Q.2 Which of the following is the characteristic of an index number?

a. Measure change over a period of time or in two or more places
b. Specialised average
c. Expressed in percentage
d. All of the above

Q.3 Index numbers measure a net or relative change in a variable or a group of variables.

a. Absolute
b. Relative
c. Both (a) and (b)
d. None of the above

Answers to SAEs 2

1 Solution

1 .25(400)+3.75(180)+12 .50(200)+18.75(150)
0.50(350)+ 1.25(200)+ 6.25(140)+12.5(130)
500+675+2500+281.5
175+250+875+1625X 100
2.218X100
221.79

Unit 2 Inventory Control

Unit Structure

2.1 Introduction

This unit will be discussing inventory control, Why Do We Need Inventory Control, and Types of Inventory Control Systems.

2.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the meaning of inventory control
- find out why Do We Need Inventory Control
- outline the types of Inventory Control Systems

2.3 Inventory control

2.3.1 What do mean by inventory control

Inventory control, also known as stock control, refers to the process of managing a company's warehouse inventory levels. The inventory control process involves managing items from the moment they're ordered; throughout their storage, movement, and usage; and to their final destination or disposal.

An inventory control system is a technology solution that manages and tracks a company's goods through the supply chain. This technology will integrate and manage purchasing, shipping, receiving, warehousing, and returns into a single system. The best inventory control system will automate a lot of manual processes.

The four types of inventory most commonly used are Raw Materials, Work-In-Process (WIP), Finished Goods, and Maintenance, Repair, and Overhaul (MRO). You can practice better inventory control and smarter inventory management when you know the type of inventory you have.

1.3.2 Why do we need Inventory Control?

Inventory control is a vital part of any business's operations and significantly impacts its ability to make a profit.
Here are four major reasons you need inventory control**:**

1. Boost inventory efficiency. Your products may be all over the warehouse, making picking and packing take much longer. Inventory control lets you discover and optimize this and many other issues like an incorrect number to increase efficiency and streamline your processes.
2. Ensure accurate inventory counts. Inaccurate inventory causes headaches and increases costs. Inventory control gives your insight into inventory numbers and helps determine the reorder point and sales trends.
3. Increase sales and mitigate losses. Fill rate is essential for businesses to keep up with customer demand and avoid stockouts. Having a higher fill rate will help increase sales and mitigate losses. Safety stock can ensure that you always have the product available to meet customer demand.
4. Ensure customer satisfaction. Customers hate when products are out of stock or on backorder. Controlling your inventory lets you avoid these issues and keep your customers happy.

2.3.3 Relevance of Inventory

Storing inventory has many associated costs that a manager needs to control. Without inventory control, a business's warehouse can quickly become a problem.

By allowing inventory to move about with no interference, a manager risks running into skyrocketing costs and plummeting profits. This, in turn, will lead to the loss of their job and possibly the closure of the business. That's why it's necessary to monitor cycle inventory to ensure standing inventory is at a sufficient level.

2.4 Inventory Solution

When attempting to track and manage inventory, there are a few ways to track your data. First, you can take the old-school approach and manually track inventory using a pen and paper. While this is quick and easy to do, it's also easy to make mistakes. We don't recommend this approach, regardless of the size of your inventory.

Second, you can use an Excel spreadsheet. We offer a free inventory tracking spreadsheet on our blog, and this is a viable option for a business with a limited inventory or just starting. You'll still need to perform a regular inventory audit or cycle inventory count to ensure accuracy, but it's a step above using a pen and paper.

Finally, the best option is to use an inventory control system and other related inventory tracking software.

Self-Assessment Exercise 1

1. What is Inventory control
2. Outline the Two major reasons you need inventory control

2.5 Types of Inventory Control Systems

Businesses use inventory control systems to measure the number of goods on hand. Large organizations frequently track inventory at several locations, including shops, warehouses, and websites. There are two main types of inventory control systems: periodic and perpetual inventory systems**.**

Now let's take a closer look at these two:

2.5.1 The Periodic Inventory System

The periodic inventory system is one of small businesses' most commonly used inventory systems. Under this system, businesses keep track of their inventory levels at set intervals, typically once a month. This system is simple to use and understand, making it a popular choice for business owners.

At the end of each interval, businesses will count their inventory levels and update their records accordingly. This system can be used with either physical inventories or virtual inventories.

2.5.2 Benefits to using the periodic inventory System

There are four benefits to using the periodic inventory system**:**

i. It is easy to set up and maintain.
ii. It provides accurate information on inventory levels.
iii. It allows businesses to manage their inventories more efficiently.
iv. It helps businesses avoid stock-outs and overstocking.

The Perpetual Inventory System
The perpetual inventory system is crucial to most businesses' inventory management systems. This system provides near-real-time data on inventory levels, allowing businesses to manage their stock more effectively.

There are two benefits to using a perpetual inventory system**:**
It allows businesses to avoid the costly and time-consuming process of physically counting their inventory. This saves businesses significant time and money on the labor cost.

This information can be used to decide production and stocking levels. and stocking levels. It helps businesses to predict future demand more accurately. With accurate data, businesses can avoid overstocking or understocking their products, leading to lost sales or missed opportunities.

2.6 Inventory Control System and Software

One of the best ways to take control of your business's inventory is to invest in inventory management software. This software tracks inventory levels, sales trends, and inventory cycles. There are many options on the market with a variety of capabilities and additional tools.

These programs can be tied to your POS system to provide a perpetual inventory count. This updates your inventory levels each time a sale is made. This lets you make the most informed decisions, calculate optimal reorder points, plan for product lead time, and stock more A-level products.

Automated Inventory Control
Automated inventory control takes a perpetual inventory count to the next level and makes decisions using predetermined rules. This feature is built into some of the best inventory control software platforms and allows you to take a more hands-off role.

Here's an example of how this automatic inventory control can work. You may set a minimum supply threshold for reorder if you have a particular product that you always want to keep in stock. Once you hit this number, the system automatically sends a new purchase order to your manufacturer. You can optimize this process with the optimal economic order quantity.

Self-Assessment Exercise 2

1. Relevance of Inventory
2. Inventory Solution
3. Types of Inventory Control Systems

2.7 Summary

Inventory control, also known as stock control, refers to the process of managing a company's warehouse inventory levels. The inventory control process involves managing items from the moment they're ordered; throughout their storage, movement, and usage; and to their final destination or disposal.

Boost inventory efficiency. Your products may be all over the warehouse, making picking and packing take much longer. Inventory control lets you discover and optimize this and many other issues like an incorrect number to increase efficiency and streamline your processes.

Ensure accurate inventory counts. Inaccurate inventory causes headaches and increases costs. Inventory control gives your insight into inventory numbers and helps determine the reorder point and sales trends.

2.8 References/Further Readings/Web Resources

Byjus.com (2022) meaning and characteristics of index numbers https://byjus.com/commerce/meaning-and-characteristics-of-index-numbers/

2.9 Possible Answers to SAEs

Answers to SAEs 1

1. Inventory control

   Inventory control, also known as stock control, refers to the process of managing a company's warehouse inventory levels. The inventory control process involves managing items from the moment they're ordered; throughout their storage, movement, and usage; and to their final destination or disposal

2. Two major reasons you need inventory control**:**

a. Boost inventory efficiency. Your products may be all over the warehouse, making picking and packing take much longer. Inventory control lets you discover and optimize this and many other issues like an incorrect number to increase efficiency and streamline your processes.

b. Ensure accurate inventory counts. Inaccurate inventory causes headaches and increases costs. Inventory control gives your insight into inventory numbers and helps determine the reorder point and sales trends

Answers to SAEs 2

1. Relevance of Inventory

   Storing inventory has many associated costs that a manager needs to control. Without inventory control, a business's warehouse can quickly become a problem.
   By allowing inventory to move about with no interference, a manager risks running into skyrocketing costs and plummeting profits

2. Inventory Solution

   When attempting to track and manage inventory, there are a few ways to track your data. First, you can take the old-school approach and manually track inventory using a pen and paper. While this is quick and easy to do, it's also easy to make mistakes. We don't recommend this approach, regardless of the size of your inventory

3. Types of Inventory Control Systems

Businesses use inventory control systems to measure the number of goods on hand. Large organizations frequently track inventory at several locations, including shops, warehouses, and websites. There are two main types of inventory control systems: periodic and perpetual inventory systems

Unit 3 Economic Order Quantity (EOQ)

Unit Structure

3.1 Introduction

This unit discussed the Economic Order Quantity and the Importance of Economic Order Quantity. The optimal quantity is the exact amount of inventory you should order and keep on hand to meet demand. Finding your optimal order quantity for a product is the goal of calculating its EOQ. However, this number is very difficult to achieve as any slight variance in demand, cost, or price will throw your numbers off.

3.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the Economic order quantity (EOQ)
- outline the Importance of Economic Order Quantity
- explain the Economic Order Quantity Problems
- state the Advantages of Economic Order Quantity
- demonstrate the EOQ Formula

3.3 Economic order quantity (EOQ)

3.3.1 What is Economic order quantity (EOQ)

Economic order quantity (EOQ) is a production-scheduling method of inventory control that has been used since the early 1900s. This method is built around finding a balance between the amount you sell and the amount you spend on inventory management process.

Economic order quantity is the ideal amount of product a company should purchase to minimize inventory costs. Essentially, it is the amount of product you should order to meet demand without having to store any excess inventory.

3.3.2 Importance of Economic Order Quantity

Managing economic order quantity is an important skill to have when someone is vying for an inventory manager salary. It can help avoid issues like excess stock or dead stock (see what is dead stock) and keep avoidable losses to a minimum. It also helps you establish goals for your inventory KPIs, informs inventory forecasting decisions, and helps increase the company's sales and revenue.

Advantages of Economic Order Quantity
Utilizing EOQ for your business can provide many benefits. Here are just a few.

Minimize costs: All warehouse inventory managers know that storage costs can quickly rise if inventory isn't controlled. By only ordering the amount needed to fulfil customer demand, these costs can be kept very low.

Adapts to your business: Many inventory methods are only viable for certain types of business. EOQ utilizes only your own numbers, so it can benefit any business that uses it.

3.3.3 Economic Order Quantity Problems

Though there are definitely positive aspects of calculating EOQ, there are also a few drawbacks that you need to be aware of.

The math is complicated: You'll see the formula used for EOQ calculations below, and it's safe to say it isn't the easiest to use. Luckily, there are many ways to automate the process and tools to help.

It's based on assumptions: There are a number of assumptions that are required to calculate EOQ. This means any aspect of your business that doesn't match will throw off the numbers and you won't get the optimal quantity. Still, the numbers you find are very helpful for inventory planning

3.4 Advantages of Economic Order Quantity

Utilizing EOQ for your business can provide many benefits. Here are just a few.

Minimize costs: All warehouse inventory managers know that storage costs can quickly rise if inventory isn't controlled. By only ordering the amount needed to fulfill customer demand, these costs can be kept very low. This can be tough if your supplier requires an MOQ (what does MOQ mean?).

Adapts to your business: Many inventory methods are only viable for certain types of business. EOQ utilizes only your own numbers, so it can benefit any business that uses it.

Self-Assessment Exercise 1

1. What is Economic order quantity (EOQ?)
2. Importance of Economic Order Quantity

3.5 EOQ Formula: Economic Order Quantity Formula

Calculating economic order quantity requires you first find a few metrics regarding demand and costs. These are the annual demand for the product in units, the cost per order, and the annual holding cost per unit. Once you've collected this data, it's as easy as plugging them into the formula below.

How to Calculate EOQ
Uncovering the economic order quantity for a product can be done using a slightly complicated formula.

Here's that formula:
EOQ = √ (2 x Demand x Order Cost / Holding Cost)
Economic Order Quantity Formula and Example
If any of that seems confusing to you, let's clear it up a bit with an example. Let's say you are a wholesale supplier for the food industry. You have a particular product you're looking to optimize, in this case cans of creamed corn. The first thing you do is look at your historical data regarding creamed corn (Bluecart.com, 2022).

After poring through your data, you calculate that you normally sell an average of 2,500 cans each year. You also look through your purchase orders and inventory costs to calculate that each shipment of 100 cans of corn costs \$75. And you find that storage of each can costs you \$20 per year.

With these variables in hand, you can now calculate your optimal EOQ for cans of creamed corn. Let's plug them in.
EOQ = √ (2 x Demand x Order Cost / Holding Cost)
EOQ = √ (2 x 2500 x 75 / 20)
EOQ = 136.9 or 137 cans

We discover that the optimal order size is 137 cans of creamed corn. Pair this with calculating the optimal reorder point, and you can maximize the profit you make from cans of corn.

1. Ordering (Replacement) Costs
   These are such costs as transportation costs, clerical and administrative costs associated with the physical movement of the purchased external goods. Where the goods are manufactured internally, there are alternative initial costs to be borne with each production run referred to as set-up costs
2. Holding (Carrying) Costs

These are:
(a) storage costs in terms of staffing, equipment maintenance, and handling;
(b) storage overheads (heat, light, rent, and the like);
(c) cost of capital tied up in inventory;
(d) insurance, security and pilferage;
(e) deterioration or breakages.

3. Stock out Costs

These are costs associated with running out of stock. These include penalty payments, loss of goodwill, idle manpower and machine, and the like.

i. Economic Ordering Quantity (EOQ): This refers to the external order quantity that minimises total inventory costs.
ii. Economic Batch Quantity (EBQ): This refers to the size of the internal production run that minimises total inventory costs.
iii. Safety Stock: This is a term used to describe the stock held to cover possible deviations in demand or supply during the lead time. It is sometimes referred to as buffer or minimum stock.
iv. Maximum Stock: This is a level used as an indicator above which stocks are deemed to be too high.
v. Reorder Level: This is the level of stock, which when reached, signals replenishment order.
vi. Reorder Quantity: This is the level of replenishment order.

Self-Assessment Exercise 2

1. Discuss the Re-order, Lro, Lmin and the Level System in the inventory control system
2. A commodity has a steady rate of demand of 2,000 units per year. Placing an order costs N200 and it costs N50 to hold a unit for a year:

a. Estimate the Economic Order Quantity (EOQ)
b. Find the number of orders placed per year
c. What is the length of the inventory circle?

3.6 Summary

This unit explain that, Economic order quantity (EOQ) is a production-scheduling method of inventory control that has been used since the early 1900s. This method is built around finding a balance between the amount you sell and the amount you spend on inventory management process.

Economic order quantity is the ideal amount of product a company should purchase to minimize inventory costs. Essentially, it is the amount of product you should order to meet demand without having to store any excess inventory.

Managing economic order quantity is an important skill to have when someone is vying for an inventory manager salary. It can help avoid issues like excess stock or dead stock (see what is dead stock) and keep avoidable losses to a minimum. It also helps you establish goals for your inventory KPIs, informs inventory forecasting decisions, and helps increase the company's sales and revenue.

3.7 References/Further Readings/Web Resources

Bluecart.com (2022). Inventory control and economic-order-quantity .Retreived from https://www.bluecart.com/blog/economic-order-quantity#toc-eoq-formula-economic-order-quantity-formula

3.8 Possible Answers to SAEs

Answers to SAEs 1

1. This is the most commonly used control system. It generally results in lower stocks. The system also enables items to be ordered in more economic quantities and is more responsive to fluctuations in demand than the second system discussed below. The system sets the value of three important levels of stock as warning t or action triggers for management:

   Re-order Level: This is an action level of stock which leads to the replenishment order, normally the Economic Order Quantity (EOQ). For a particular time period, the re-order level is computed as follows:
   Lro = maximum usage per period x maximum lead time (in periods)

   Minimum Level: This is a warning level set such that only in extreme cases (above average demand or late replenishment) should it be breached. It is computed as follows:
   Lmin = Re-Order Level – (normal Usage x Average lead time)
2. Given information, D = 2,000; Co = 200; and, Ch = 50
   The Economic Order Quantity is determined by:
   EOQ = {2DCoCh} {2(2000)(200) 50}= 16000= 126.491
   Square root of 16000 is 126.491
   Thus, the economic order quantity is about 127 units.
   (b) Number of orders per year = Yearly Demand
   EOQ= 2000/126.491 = 15.81
   What is average inventory level = 63.5

Answers to SAEs 2

1. What is Economic order quantity (EOQ?)

Economic order quantity (EOQ) is a production-scheduling method of inventory control that has been used since the early 1900s. This method is built around finding a balance between the amount you sell and the amount you spend on inventory management process.
Economic order quantity is the ideal amount of product a company should purchase to minimize inventory costs. Essentially, it is the amount of product you should order to meet demand without having to store any excess inventory

2. Importance of Economic Order Quantity

Managing economic order quantity is an important skill to have when someone is vying for an inventory manager salary. It can help avoid issues like excess stock or dead stock (see what is dead stock) and keep avoidable losses to a minimum. It also helps you establish goals for your inventory KPIs, informs inventory forecasting decisions, and helps increase the company's sales and revenue

Unit 4 Decision Analysis

Unit Structure

4.1 Introduction

This unit will be discussing Decision analysis, Decision trees and backward induction
Typically, more than one decision is involved in decision making, in which case it is best to use a tree instead of a matrix.

4.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the concept of Decision Analysis
- outline the Basic Elements decision theory
- discuss the Decision Criteria
- explain the Decision trees and backward induction
- itemize the Steps in Decision Analysis

4.3 Decision Analysis

4.3.1 What is Decision Analysis?

Decision analysis is a formalized approach to making optimal choices under conditions of uncertainty. It allows the user to enter costs, probabilities, and health-related quality of life values among other inputs of interest, and then calculates probabilistically weighted means of these outcome measures. In public health, these outcome measures usually include costs. Typically, therefore, decision analysis is the heart of cost-effectiveness analyses in public health and medicine (Gold et al., 1998). However, just about any outcome measure can be modeled, including vaccine-preventable illnesses averted, deaths avoided, and so forth.

Therefore, local health departments, pharmaceutical companies, or other agencies can use decision analysis for internal decision-making processes. Decision analysis is often used by non-health businesses interested in deciding whether they should release a product, perform internal restructuring, and so forth.

One great strength of decision analysis modeling is that it allows for the calculation of a range of possible values around a given mean. This approach, called ‘sensitivity analysis,’ allows the user to better understand the chances that he or she will make a bad decision if a given strategy is taken.

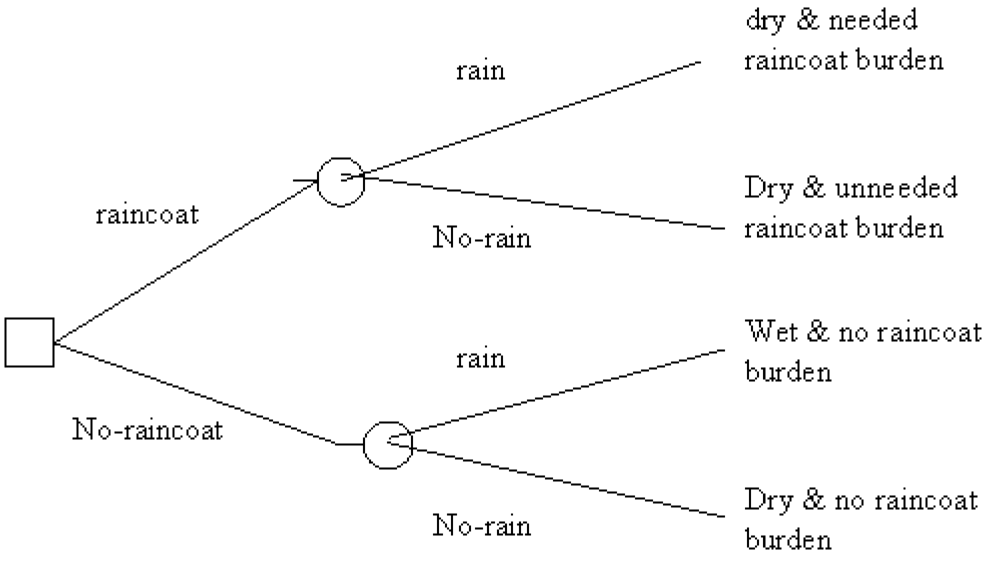
Decision analysis, like cost-effectiveness analysis, is highly dependent on the accuracy and completeness of model inputs, as well as the assumptions that the analysts make. Drugs can have unforeseen side effects, or interventions can have long-term costs that may not be apparent to the analysts. Any of these effects can lead to suboptimal outcomes.

For instance, the optimal treatment strategy for tuberculosis in most instances is a low-cost combination of medications that can be effectively delivered in developing countries. By using the most cost-effective medications, it is possible to maximize the number of lives saved within a given budget. However, as Farmer points out, these medications will be wasted if delivered to populations with a high percentage of drug-resistant tuberculosis (Farmer, 2004). Therefore, decision analysis and cost-effectiveness analysis must be viewed as an adjunct to optimal decision making rather than the final word in health policy.

Decision is sets and recommended rules on how to evaluate the overall environmental, technical, and socioeconomic performance of a set of alternatives and how to choose among them. The steps for decision process are to: define the problem and formulate objectives, find decision alternatives and predict the impacts of each, rank the alternatives and decide. The implementation of the decision is accompanied by regulatory, legal and economic processes, by stakeholders involvement, and by evaluation of the correctness of the decision by following the real responses after the implementation of the project.

4.3.2 Basic Elements decision theory

There are 4 basic elements in decision theory: *acts, events, outcomes,* and *payoffs*. Acts are the actions being considered by the agent -in the example elow, taking the raincoat or not; events are occurrences taking place outside the control of the agent (rain or lack thereof); outcomes are the result of the occurrence (or lack of it) of acts and events (staying dry or not; being burdened by the raincoat or not); payoffs are the values the decision maker is placing on the occurrences (for example, how much being free of the nuisance of carrying an raincoat is worth to one). Payoffs can be positive (staying dry) or negative (the raincoat nuisance). It is often useful to represent a decision problem by a tree.



Here a square indicates a node in the tree where a decision is made and a circle where events take place. The tree does not contain payoffs yet, but they can easily be placed by the outcomes.

In general, we can note two things. First, the nature of the payoffs depends on one's *objectives*. If one is interested only in making money, then payoffs are best accounted for in terms of money. However, if one is interested in, say, safety, then the payoffs are best accounted for in terms of risk of accident, for example. If any numerical approach is possible when disparate objectives are involved, there must be some universal measurable quantity making them comparable. (In fact, utility, of which more later, is such a quantity).

Second, decision making trees can become unmanageable very fast if one tries to account for *too many possibilities*. For example, it would be physically impossible to account for all the possibilities involved in the decision of which 50 out of 200 gadgets should be intensively marketed, as the number of possible combinations, 200!/(50! · 150!) is simply astronomical. Hence one must use good judgment in limiting the options considered; this is potentially problematic, as one may unwittingly fail to consider a possible action which would produce very good outcomes.

4.3.3 Decision Criteria

How one uses a decision tree or a decision matrix depends on the decision criteria one adopts. Consider the following *payoff matrix* where acts are rows, events columns, and the resulting squares contain the payoffs (outcomes are not represented to avoid clutter). So, suppose that we are considering which widget out of 3 to produce and our goal is making money.

	EVENTS	
	Good sales	Bad sales
ACTS		
Produce A	+$5000	-$1000
Produce B	+$10000	-$3000
Produce C	+$3000	-$500

Here producing B is obviously the best option if things go well, while producing C is the best option if things go badly, as losing $500 is the best of the worst payoffs. The decision criterion telling us to choose C is called "Maximin". Obviously maximin is a rather pessimistic strategy, and for this reason it is controversial.

However, if the stakes are very high (for example, suppose that if I lose more than $500 I will be forever ruined), maximin seems a reasonable option. The application of maximin in the original position has played an important role in Rawls' *A Theory of Justice*, the most important work in political philosophy in the last decades. Other decision criteria in cases of uncertainty are *maximax*, *minimax of regret*, and the appeal to subjective probabilities through the *Principle of Indifference*.

Unfortunately, none of these principles is always viable. However, when the probabilities of events are available (that is, in decision under risk) and the agent is indifferent to risk, as when the payoffs involved are *significant but not too significan*t, the criterion usually put forth in decision theory is that of the *expected maximum payoff* (EMP), the counterpart of the principle in gambling enjoining us to choose the bet with the greatest expected value. So, suppose that we could provide the relevant probabilities, as in the following matrix:

	EVENTS			
	Good sales	Bad sales	Payoff	Expected payoff
ACTS				
Produce A Pr(good sales)= 80% Pr(bad sales) =20%	+$5000 x 80% = 4000	-$1000 x 20% = -200	+$4000	+$3800
Produce B Pr(good sales)= 60% Pr(bad sales) =40%	+$10000 x 60% = 6000	-$3000 x 40% = -1200	+$7000	+$4800
Produce C Pr(good sales)= 50% Pr(bad sales) =50%	+$3000 x 50% = $1500	-$500 x 50% = -$250	+$2500	+$1250

Then, EMP would tell us to produce B, as the expected payoff is the greatest. Most business decisions fall into this category. For example, if a company makes dozens of decisions with comparable payoffs every day, then EMP is the best business strategy, as it is for a casino.

Self-Assessment Exercise 1

1. Explain the concept of Decision Analysis
2. Outline the Basic Elements decision theory
3. Discuss the Decision Criteria

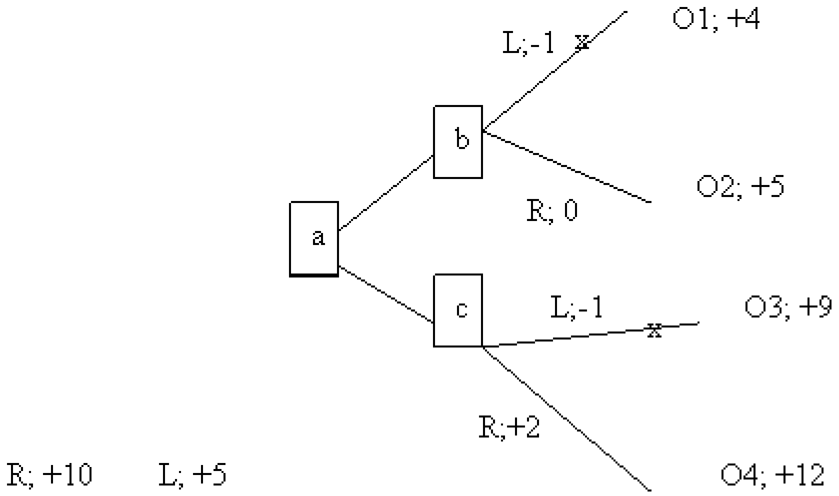
4.4 Decision trees and backward induction

Typically, more than one decision is involved in decision making, in which case it is best to use a tree instead of a matrix. For example, consider the following situation, in which no probabilities are involved. You have arrived at a fork in the road on you way home.

If you go left, you’ll have to go through a very sunny patch in the mid of a very hot day. However, this will also allow you to admire a beautiful bloom of wildflowers growing by the side of the path. You shall then arrive at another fork. If you take another left, you will have to go by a neighbor’s house with a very unpleasant guard dog that will growl at you from the other side of the fence. By contrast, if you go right at the

second fork, you'll go by a very noisy and dusty part of the road. Whichever of the two you take, you shall get home quickly.
If you go right at the first fork, you'll go through the woods, which are very cool this time of the year. However, there will be little to admire until you get to another fork in the road. If you go left at this fork, you will see some beautiful meadows; unfortunately, it will take you longer to reach your home, a bad thing since you are in a bit of a hurry. If you go right, you shall get home in good time.

Suppose that you assign the following utilities:
Getting hot: -10; seeing the wildflowers: +15; being growled at: -3; pleasant coolness: +10; being a bit late getting home: -5; taking a noisy and dusty road: -2; seeing the nice meadows: +4; getting home in good time: +2.
We can construct a decision tree.



Decision trees can be used by applying *backwards induction*. The idea is that in order to determine what to do at a (the decision at the first fork), one needs to decide what one would do at b and c.

In other words, the tree is analyzed form the right (from the outcomes) to the left (to the earlier decisions). So, at b, one would take the right path because it leads to outcome O^2 with utility is +5 while the utility of the left path leading to O1 is +4. We can represent this choice by pruning the left path, that is, by placing an 'x' on it.

By the same token, at c one would choose to go right, and therefore we may place an 'x' over the left option. We are now left with a simplified tree at a: going left will have utility +5, while going right will have utility +12. Hence, we should go right twice.

Adding probabilities

The previous example did not involve probabilities. However, introducing them is not much of a problem, as the following example shows.

You are about to produce a new garment C and must determine whether to merchandise it only nationally (N) or internationally as well (I).
If you choose N and sales are good (G), then you'll make 4, and if they are bad (B) you'll lose 1. (All payoffs are in millions). You believe that the probability of good national sales is .8 and that of bad national sales is .2. You must also decide whether to produce and sell a matching scarf S. If S's sales are good, you'll make an additional 2, and if they are bad you'll lose an additional 1. You think that if C sells well the probability that S's sales are good is .9 and the probability that S sells badly is .1. By contrast, if C sells badly, the probability that S's sales are good is .4 and the probability that S sells badly is .6.

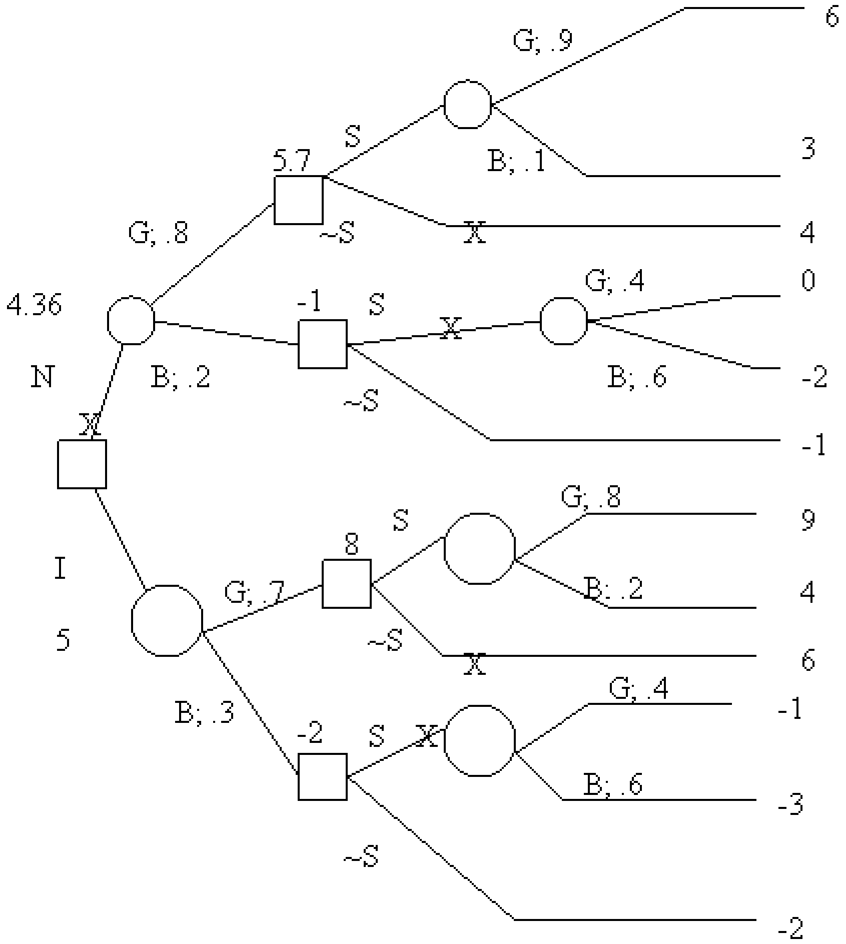
If you choose I and sales are good you'll make 6, and if they are bad you'll lose 2. You believe that the probability of good sales is .7 and that of bad salses is .3. As in the other case, you must decide about S. If S's sales are good, you'll make an additional 3, and if they are bad you'll lose an additional 2. You think that if C sells well the probability that S's sales are good is .8 and the probability that S sells badly is .2. By contrast, if C sells badly, the probability that S's sales are good is .4 and the probability that S sells badly is .6.

Because of the seasonal nature of your business, you need to decide now, a few months in advance, what to do.

The tree below digrams the decision. Obviously, the first decision is whether to choose N or I. Let's follow the subtree tree arising from N, as that stemming from I is analogous. Upon deciding N, there is an act of nature, hence the circle with G (good sales) and B (bad sales). After G, we write .8, which is the probability of good sales if you choose N. Your second decision regards S, which is represented by the square with the two branches S (produce and sell S) and ~S (don't produce S). If you choose S, then the sales can be good or bad (an act of nature) represented by the circle and the relevant probabilities; if you choose ~S, then you are left with the gains or losses from selling C nationally.

The payoff are easy to calculate. Conside the uppermost branch, representing your choice of N, good sales of C, your choice of S and good sales of S. Since C sells well, you make 4 and since S sells well, you make 2, so that the total payoff for this branch is 4 + 2 = 6. The second topmost branch has a payoff of 3 because although C sells well (you gain 4), S does not (you lose 1).

The tree is analyzed by backward induction. Starting at the top, the last decision you make is whether to choose S or ~S. The expected payoff of choosing S is 5.7, which is 6 times .9 plus 3 times .1, exactly as if we were considering a bet in which you get 6 with probability .9 and 3 with probability .1. Since 5.7 > 4, you clip ~S, which simply means that the best decision given N and good sales of C is to choose S. Since 5.7 is the best expected payoff available, we write 5.7 next to the relevant square. By contrast, given N and bad sales of C, the best option is ~S, with payoff of -1, as the option S has an expected payoff of -1.2. Hence, S is clipped. We can now calculate the expected payoff of choosing N, namely 5.7 times .8 plus -1 times .2, which is 4.36. We write this figure next to the choice N. The lower half of the tree is analyzed analogously, resulting in an expectd payoff of 5 for I. Since 5 > 4.36, N is clipped and I is the best choice, that with the highest expected payoff.

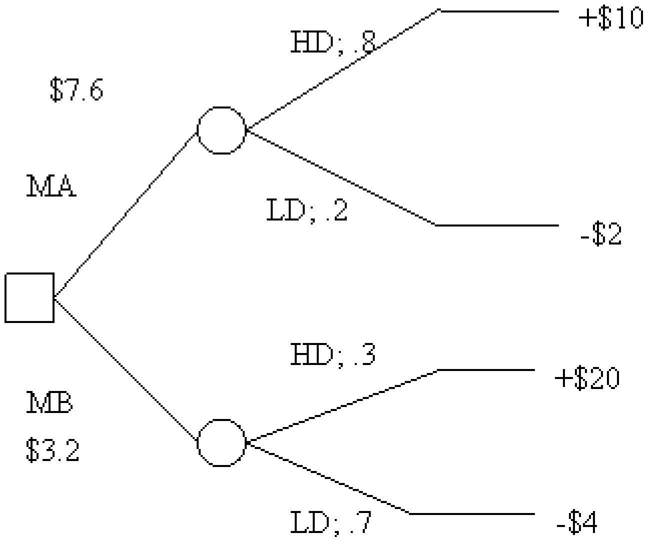


Revising decision trees in the light of new information

Consider the following scenario. You have to choose which of two widgets, A and B, to market. You also believe that the probability of high demand for A is 80% and the probability of high demand for B is 30%. (High and low demands are the only two alternatives). However, while you get only $3 for every sold A, you get $5 for every sold B. Suppose that if an item is in high demand you shall sell 10,000 of them and with low demand only 4000. Item A cost $1 to produce, while item B costs $2. Moreover, time constraints due to the holiday season compel you to produce 10,000 items of whichever of the two widgets you decide to market. What should you do?

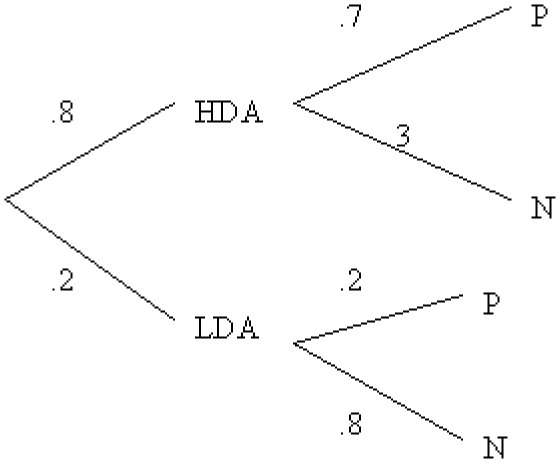
We already know how to construct the decision tree listing the outcomes and the payoffs in thousands. For example, since 10,000 items must be produced no matter what the sales will be, the production cost of A is 10. If A sells well, all 10,000 will be sold, with a gain of 20. Hence, the payoff will be 20-10=10. If demand is low, only 4,000 will be sold, with a gain of 8. Hence, the payoff will be 8-10=-2.

By regressive induction, we see that MA has an expected payoff of 7.6 and MB of 3.2, which means that we should choose MA.



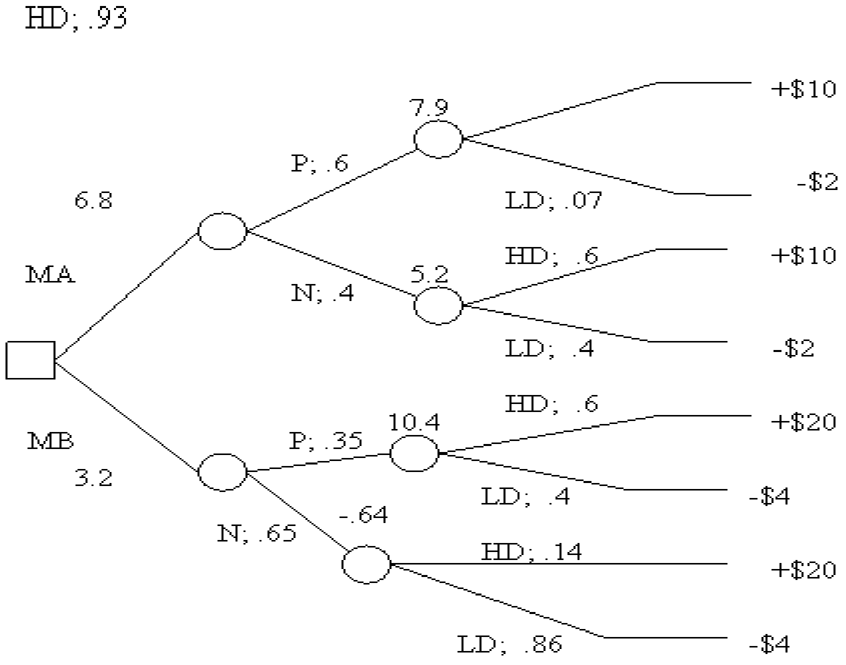
Suppose, however, that you submit A and B to a buyer panel; the reliability of the panel is given by the following figures: in the past, out of 100 high demand items, the panel was positive (P) about 70 and negative (N) about 30; out of 100 low demand items, the panel was positive about 20 and negative about 80. (Note that high or low demand is built into the market, and therefore is the conditioning factor in conditional probability). In short, Pr(P|HD)=70%; Pr(N|HD)=30%; Pr(N|LD)=80%; Pr(P|LD)=20%.

To make use of this new information, we use Bayes' theorem to determine the posterior probabilities of HD and LD, that is, their probabilities given the panel's results. The relevant *probability, not* decision, tree is



Hence, by Bayes' theorem (with figures adjusted to 2 decimals):
Pr(HDA|P) = .93; Pr(LDA|P) = .07
Pr(HDA|N) = .6; Pr(LDA|N) = .4.
By an analogous procedure, one obtains
Pr(HDB|P) = .6; Pr(LDB|P) = .4
Pr(HDB|N) = .14; Pr(LDB|N) = .86.

At this point we can rewrite the decision tree incorporating the new posterior probabilities:



From now on it is just a matter of employing backward induction to determine which widget to market. The expected payoff of marketing A if the panel has a positive reaction is +$7.9; if the panel has a negative reaction the expected payoff drops to +$5.2. With respect to B, if the panel has a positive reaction, the expected payoff is $10.4; if the panel has a negative reaction, it becomes -$.64. The expected value of marketing A is 6.8; that of marketing B is 3.2. Hence, one should market A. Since the expected payoff difference between MA and MB is 3.6, one should pay less than 3.6 for the panel; in other words, given the data of the problem, the extra information is not worth more than 3.6 (siue.edu, 2022).

One can easily see that decision trees could get complex and difficult to construct. For example, getting the information may take time when time is of the essence (perhaps, A and B are holiday season widgets). Moreover, in our examples we came up with the relevant probabilities by fiat. In real situations, of course, coming up with the correct probabilities can be very difficult, especially when appealing to history or statistical surveys does not help. However, such cases are best left to a decision theory course.

The Decision Analysis process transforms a broadly stated decision opportunity into a traceable, defendable and actionable plan. It encompasses one or more discrete analyses at one or more lower (e.g., system element) levels and aggregates them into a higher-level view (e.g., system "scorecard" presentation) relevant to the decision maker and other stakeholders. Decision Analysis can be the central process for formulating, managing and executing an effective and efficient program at any point in the life cycle.

Solution	Evaluation Method	Cost	Performance	Schedule	Total Score
Solution 1	Simulation	4	2	1	29
Solution 2	Discussion	2	3	3	25
Solution 3	Prototype	3	1	4	23
Solution 4	Discussion	1	4	2	23

Sample Evaluation Table

Decision Analysis and associated trade studies should be integrated with, and mutually supportive of, aspects of several SE processes in the early stages of the program, in particular (dau.edu, 2022).
Technical Planning (see Systems SE (SE) Guidebook, Section 4.1.1)
Technical Assessment (see SE Guidebook, Section 4.1.3.)
Stakeholder Requirements Definition (see SE Guidebook, Section 4.2.1)
Requirements Analysis (see SE Guidebook, Section 4.2.2)
Architecture Design (see SE Guidebook, Section 4.2.3)
Results

4.5 Steps in Decision Analysis

Activities and Products
Decision Analysis teams generally include a lead analyst with a suite of reasoning tools, subject matter experts with access to appropriate models and analytical tools and a representative set of end users and other stakeholders. A robust Decision Analysis process acknowledges that the decision maker has full responsibility, authority and accountability for the decision at hand.

Activities
Decision Analysis typically includes the following steps:
Identifying the problem or issue
Reviewing requirements and assumptions to establish the overall decision context
Framing/structuring the decision in terms of supporting program/project objectives

Identifying methods and tools to be used in the analyses (see SE Guidebook, Section 2.2 Tools, Techniques and Lessons Learned)
Developing decision criteria (objectives and measures), criteria weight and associated rationale
Identifying, recording and tracking assumptions
Identifying and defining alternatives to be evaluated (for high-level analyses, these are generally directed, although additional ones may arise during the course of the analysis)
Analyzing and assessing alternatives against criteria
Synthesizing results
Analyzing sensitivities
Developing decision briefing with action/implementation plan(s)
Making appropriate recommendation(s) to decision maker as expected/requested
Products and Tasks

Sound recommendations and action plans are the principal output of a well-framed and well-executed Decision Analysis process. The ability to drill down quickly from overall trade-space visualizations to detailed analyses that support the synthesized views is particularly useful to decision makers in understanding the basis of observations and conclusions.

Self-Assessment Exercise 2

1. Explain the Decision trees and backward induction
2. Itemize the Steps in Decision Analysis

Product	Tasks
12-1-1: Prepare decision analysis	Identify stakeholder and technical requirements, as well as assumptions to establish the overall decision context. Frame the decision in terms of supporting program / project objectives. Identify methods and tools to be used in the decision analysis. For major defense acquisition programs (MDAPS) and major automated information system (MAIS) programs, describe how the tools support the program’s SE approach in the program’s systems engineering plan and incorporate in the documentation of the decision analysis recommendation. Develop decision criteria. Identify and define alternatives to be evaluated. Analyze and assess alternatives against decision criteria. Synthesize results.

	Document analysis and recommend action/implementation to decision maker.

Source: AWQI eWorkbook

4.6 Summary

The unit explained that, A well-executed decision analysis or trade-off analysis helps the Program Manager (PM) and the Systems Engineer understand the impact of various uncertainties, identify one or more course(s) of action that balance competing objectives and objectively communicate the results to decision makers. As such, it provides the basis for selecting a viable and effective alternative from among many under consideration.

Decision Analysis applies to technical decisions at all levels, from evaluating top-level architectural concepts to sizing major system elements to selecting small design details. The breadth and depth of the analysis should be scaled to both the scope of the decision and the needs and expectations of the decision maker(s).

4.7 References/Further Readings/Web Resources

Siue.edu (2022). Concept and Types of Decision Retrieved from https://www.siue.edu/~evailat/decision.htm

dau.edu (2022). Decision Analysis. Retrieved from https://www.dau.edu/tools/se-brainbook/Pages/Management%20Processes/Decision-Analysis.aspx

4.8 Possible Answers to SAEs

Answers to SAEs 1

1. Decision analysis is a formalized approach to making optimal choices under conditions of uncertainty. It allows the user to enter costs, probabilities, and health-related quality of life values

among other inputs of interest, and then calculates probabilistically weighted means of these outcome measures. In public health, these outcome measures usually include costs.

2. Basic Elements decision theory

There are 4 basic elements in decision theory: *acts, events, outcomes,* and *payoffs*. Acts are the actions being considered by the agent -in the example elow, taking the raincoat or not; events are occurrences taking place outside the control of the agent (rain or lack thereof); outcomes are the result of the occurrence (or lack of it) of acts and events (staying dry or not; being burdened by the raincoat or not); payoffs are the values the decision maker is placing on the occurrences (for example, how much being free of the nuisance of carrying an raincoat is worth to one).

3. Decision Criteria

How one uses a decision tree or a decision matrix depends on the decision criteria one adopts. Consider the following *payoff matrix* where acts are rows, events columns, and the resulting squares contain the payoffs (outcomes are not represented to avoid clutter). So, suppose that we are considering which widget out of 3 to produce and our goal is making money

Answers to SAEs 2

1. Decision trees and backward induction

Typically, more than one decision is involved in decision making, in which case it is best to use a tree instead of a matrix. For example, consider the following situation, in which no probabilities are involved. You have arrived at a fork in the road on you way home.

If you go left, you'll have to go through a very sunny patch in the mid of a very hot day. However, this will also allow you to admire a beautiful bloom of wildflowers growing by the side of the path. You shall then arrive at another fork. If you take another left, you will have to go by a neighbor's house with a very unpleasant guard dog that will growl at you from the other side of the fence. By contrast, if you go right at the second fork, you'll go by a very noisy and dusty part of the road. Whichever of the two you take, you shall get home quickly

2. Steps in Decision Analysis

Identifying the problem or issue

Reviewing requirements and assumptions to establish the overall decision context

Framing/structuring the decision in terms of supporting program/project objectives

Identifying methods and tools to be used in the analyses (see SE Guidebook, Section 2.2 Tools, Techniques and Lessons Learned)

Developing decision criteria (objectives and measures), criteria weight and associated rationale

Identifying, recording and tracking assumptions

Identifying and defining alternatives to be evaluated (for high-level analyses, these are generally directed, although additional ones may arise during the course of the analysis)

Analyzing and assessing alternatives against criteria

Synthesizing results

Analyzing sensitivities

Developing decision briefing with action/implementation plan(s)

Making appropriate recommendation(s) to decision maker as expected/requested

Products and Tasks

Sound recommendations and action plans are the principal output of a well-framed and well-executed Decision Analysis process

Unit 5 Network Analysis

Unit Structure

5.1 Introduction

This unit will discuss the Project Network Analysis. Project network is graphic representation of a project's operations. It is the combination of activities and events which are require reaching the end objective of a project.

Various activities are performed at the same time and there are various activities which can be started only at the completion of other activities in a big project. The main work for detailed study of the product is to determine the information about the project and then discover a new, better and quicker way to get the work done. Thus thorough study of project is done through some suitable diagram which shows various activities and their positions in the project.

It is also helpful to know that in what way the delay in any activity can affect the whole project in terms of time and money. A number of nodes (typically shown as small circles or rectangles) and a number of arcs (shown as arrows) that connect two different nodes exist in a project network (figure).

5.2 Learning Outcomes

By the end of this unit, you will be able to:

- explain the concept of Network Analysis
- explain the Objective of Network Analysis
- outline the Advantage of Network Analysis
- outline the Disadvantage of Network Analysis
- state the techniques used in Network Analysis
- outline the Networking Components
- explain the concept of A dummy Activity
- explain the Precedence Relationship
- differentiate between the AOA and AON Approaches
- explain the Drawing Network
- outline the Rules for Drawing Network Diagrams
- state the Common Errors Network Construction

5.3 Network Analysis

5.3.1 Network Analysis

Network analysis involves a group of techniques which are used for presenting information about the time and resources involved in the project so as to assist in the planning, scheduling and controlling of the project. The information usually represented by a network includes the sequences, interdependencies, interrelationships and critical activity of various activities of the project.

5.3.2 Objective of Network Analysis

i. Minimize Production Delay, Interruptions and Conflicts:
This is achieved by identifying all activities involved in the project, their precedence constraints, etc.

ii. Minimization of Total Project Cost
After calculating the total cost of the project the next step is to minimise the total cost. It is done through the calculation of cost of delay in the completion of an activity of the project and calculating the cost of the resources which are required to complete the project in a given time period.

iii. Trade-off between Time and Cost of Project
The duration of same activity can be reduced if additional sources are employed and this is the main idea on which the trade-off

between time and cost of project is based. Due to technical reasons, the duration can be reduced in a specific limit. Similarly, there is also a most cost efficient duration called 'normal point' stretching the activity beyond it may lead to a rise in direct cost.

iv. Minimization of Total Project Duration:
After checking the actual performance against the plan the project duration can be controlled and minimized. If any major difference is found then apply the necessary reschedule process by updating and revising the uncompleted portion of the project.

v. Minimization of Idle Resources:
If there is any variation in the use of scars resources then it can disturb the entire plan and hence it is required that efforts should be made to avoid any increase in cost due to idle resources.

5.3.3 Advantage of Network Analysis

For planning, scheduling and controlling of operations in large and complicated projects network analysis is very important and powerful tool.

For evaluating the performance level of actual performance in comparison to planned target network analysis is a very useful tool.
With the use of network analysis technological interdependence of different activities can be determined for proper integration and co-ordination of various operations.

Network analysis gives the proper co-ordination and communication between various parts of the project.

Network analysis deals with the time-cost trade-off and provides the optimum schedule of the project.

This technique is very simple and suitable for the computer users

5.3.4 Disadvantage of Network Analysis

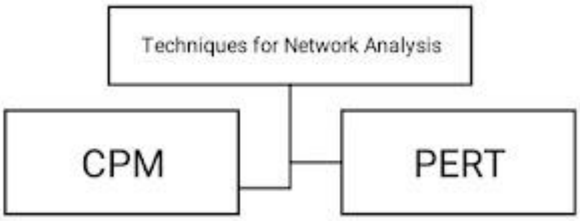
Network construction of complex project is very difficult and time consuming in network analysis.

Actual time estimation of various activities is a difficult exercise.
Analysis of the project is a very difficult work because a number of resource constraints exist in the project.

In many situations time-cost trade off procedure is complicated.

5.4 Techniques Which Are Used In Network Analysis

The two common techniques which are used in network analysis are shown in figure below:



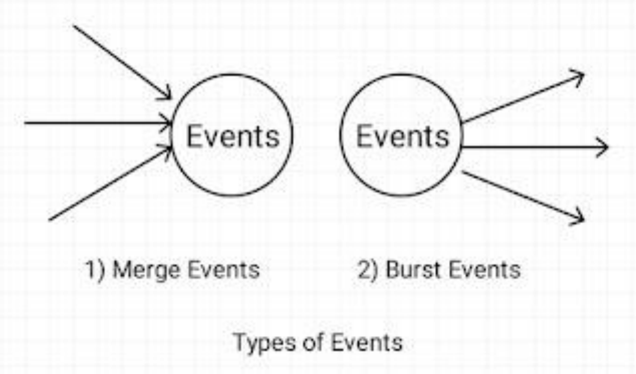
The managers are supported by two well-known network analysis techniques, viz, Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT) in planning and controlling of large scale construction projects, research and development, and so on.

These techniques prove to be very important in supporting the managers in handling such products and performing their project management responsibilities.

5.4.1 Networking Components

1. Events:

In a network diagram events represent the project milestones. For example, start or completion of an activity or activities, and occurrence of the events at a particular instance of time at which some specific portion of project has been or is to be achieved. In the network events are represented by the circles (nodes). The events can be further classified into the following two categories :



i. Merge Event :
The joint completion of more than one activity which shows an activity is called merge event. This is shown in figure.

ii. Burst Event :

An event which shows the beginning of more than one activity is known as burst event. This is shown in figure.

The numbers are used in a network diagram for representing events. For indicating progress of the work, each event is identified by a number which is higher than its immediate preceding event. The numbering of events in the network diagram must start from left (start of the project) to the right (completion of the project) and top to the bottom. It is noted that there should not be any duplication in the numbering of events.

2. Jobs/Activity/Task:

The project operations (or tasks) are represented by activities which are conducted in a network diagram. These activities take a certain amount of time and require resources for completion. An activity is represented by an arrow and its head indicates the direction of progress in the project. The numbering of starting (tail or initial) event and ending (head or terminal) event identifies activities. For example, an arrow (i, j) between two events shows that the tail event i represents starting of the activity and the head event j represents the completion of the activity which is shown in figure. The activities can be further classified into the following three categories:

i. Predecessor Activity

Predecessor activity is an activity which is completed before one or more other activities start.

ii Successor Activity :

Successor activity is an activity which starts immediately after one or more of other activities are completed.

iii Dummy Activity :

The activity which does not use any time or resource for completion is called dummy activity.

Self-Assessment Exercise 1

1. Define the concept of Network Analysis
2 Mention and briefly explain the objectives of Network Analysis
3 Mention and Explain the Networking Components
4. Explain the two common techniques which are used in network analysis

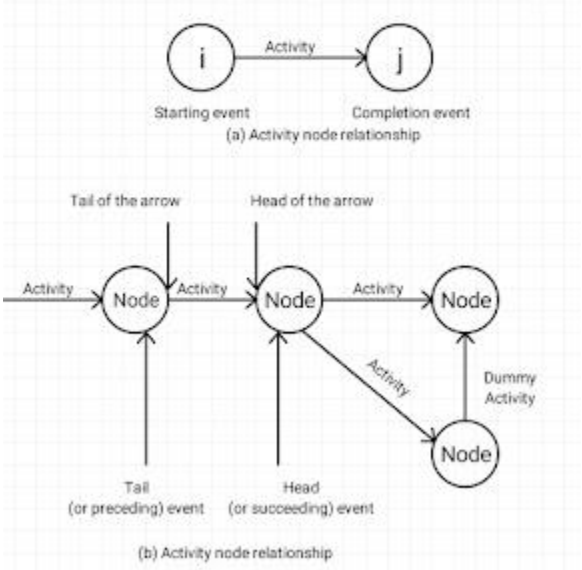
5.5 A dummy Activity

A dummy activity is used in a network to establish the precedence relationship among various activities of the project. It is needed when:

a) Two or more parallel activities in a project have same - the head and tail events.

b) Two or more activities have some (but not all) of their immediate predecessor activities in common.

Dummy activity is represented by a dotted line in the network diagram as shown in figure below.



(a) Activity node relationship

(b) Activity node relationship

Activity-Node Relationship in Network Diagram

5.5.1 Precedence Relationship

Diagramatic representation of project as a network needs the establishment of precedence relationships between activities. For undertaking activities, precedence relationship provides a sequence. It states that any activity cannot start until a preceding activity has been completed.

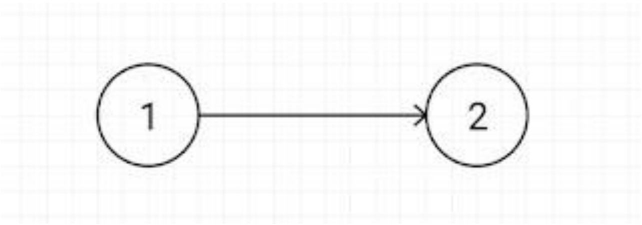
Example:
Brochures announcing a conference for executives must first be designed by the program committee (activity A) before they can be printed (activity B). In other words, activity A must precede activity B.

For large projects, this task is essential because incorrect or omitted precedence relationships will result in costly delays. The precedence relationships are represented by a network diagram.

The following two types of precedence networks are used by network models to show precedence requirements of the activities in the project

1. **Activity-on-Arc (AOA):**

In an AOA network, arrow is used for representing the activity and both the ends of the arrow which are called nodes shows the start and end of the activity.

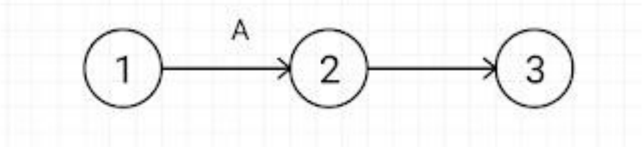


Activities are represented by an arc and events are represented by a node. An activity is separated by a node (an outgoing arc) from each of its immediate predecessors (an incoming arc). One or more activities can be completed at the starting point of any event and one or more events can start from this point. Neither time nor resources are consumed by any event.

AOA approach is an event oriented approach because it focuses on the activity connection points. The precedence relationship explains that an event does not occur until all preceding activities have been completed. AOA approach uses a convention that events are numbered from left to right.

2. **Activity-on-Node (AON)**

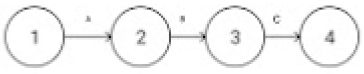
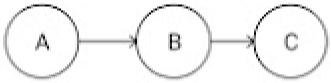
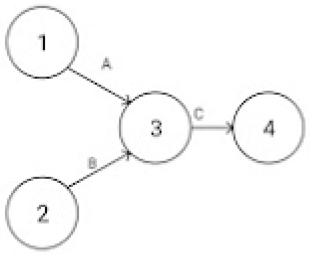
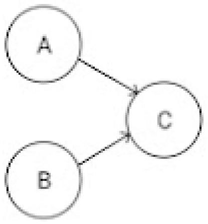
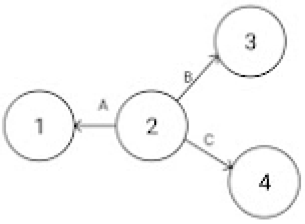
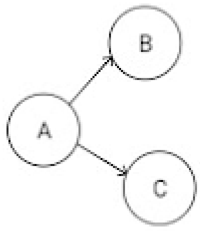
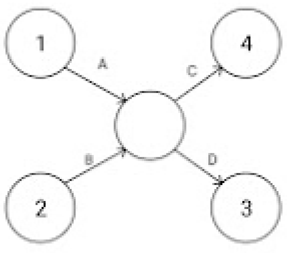
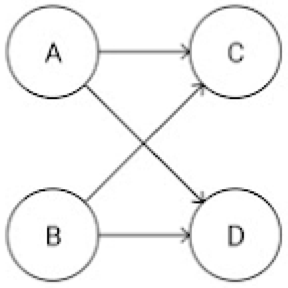
The second approach in the project network is called Activity-on-Node (AON) in which activities are shown on the nodes and precedence relationship between them is represented by arcs. In other words, activities are represented on the nodes and sequencing connection between two different activities is represented by the arrows. Thus, in AOA diagram of following type :

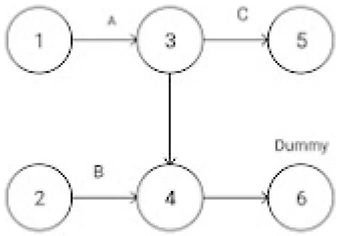
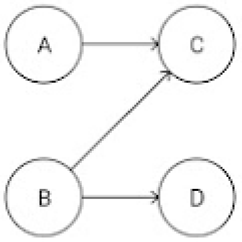
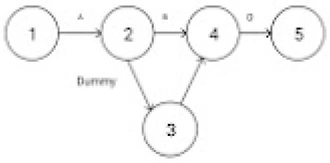
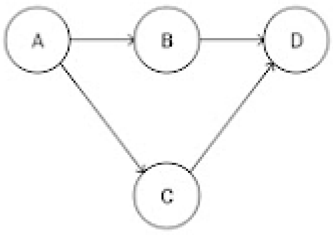


There is no need of dummy activity because this approach is activity based. An AON diagram is better for visual presentation because it is similar to the bar chart. Thus, visual presentation of a project is done better with the use of an AON network diagram.

5.5.2 AOA and AON Approaches

Following Figure Shows AOA and AON Approaches for Several Commonly Encountered Activity Relationship.

Activity – On – Arc (AOA)	Activity – On – Node (AON)	Activity Relationship
		A precedes B, which precedes C.
		A and B must be completed before C can be started.
		B and C cannot begin until A has been completed.
		C and D cannot begin until both A and B have been completed.

		C cannot begin until both A and B have been completed, D cannot begin until B has been completed.
		B and C cannot begin until A has been completed, D cannot begin until both B and C have been completed.

5.6 Drawing Network

The steps of network construction are as follows:

Step 1: Properly define the project and it's all important activities or tasks.

Step 2: Develop the relationships among the activities. Decide which activities must precede the others.

Step 3: Connect all the activities and draw the network.

Step 4: Time and/or cost estimates are assigned to each activity.

Step 5: Calculate the path which has the longest time and this is called critical path.

Step 6: Use the network for planning, scheduling, monitoring and controlling the project.

Rules for Drawing Network Diagrams

For handling events and activities of a project network there are various concepts and rules which should be followed. It provides help in the development of a correct network structure. Some of them are as mentioned below:

One and only one arrow is used for representing each defined activity in the network. Hence, any activity cannot be represented more than once in a network.

All preceding activities must be completed before selecting any new activity.

The arrow which is used for showing the activity is indicative of the logical precedence only.

The direction of the arrow indicates the general progression in time.

When a number of activities terminate at one event, it indicates that no activity emanating from that event may start unless all activities terminating there have been completed.

Numbers are used for representing the events.
The activities are identified by the numbers of their starting and the ending events.

There should be only one initial and one terminal node in a network.

The joint completion of more than one activity which shows an activity is called merge event, while an event which shows the beginning of more than one activity is known as burst event.

Parallel activities between two events, without intervening events are prohibited.

In any network looping is not allowed. Therefore, if A precedes B, and B precedes C, then cannot precede A.

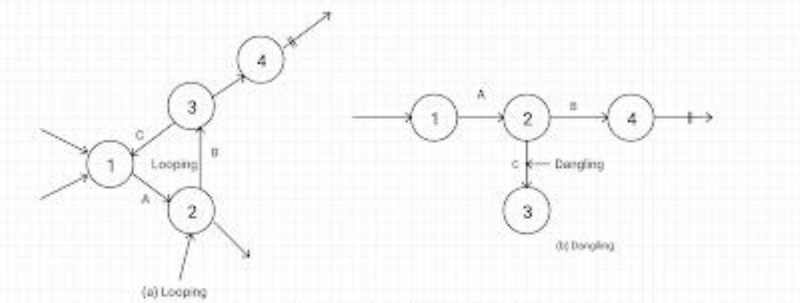
In the development of a network it must be ensured that loops are not present.

5.6.1 Common Errors Network Construction

Following are three common errors in a network construction

1. Looping:

A case of endless loop in a network diagram, which is also known as looping, is shown in figure, where activities A, B and C form a cycle :
Due to precedence relationships, it appears from figure 4.5 that every activity in looping (or cycle) is a predecessor of itself. In this case, it is difficult to number three events associated with activity A, B and C so as to satisfy rule 6 of constructing the network.



(a) Looping

(b) Dangling

Looing and Dangling in the Network Diagram

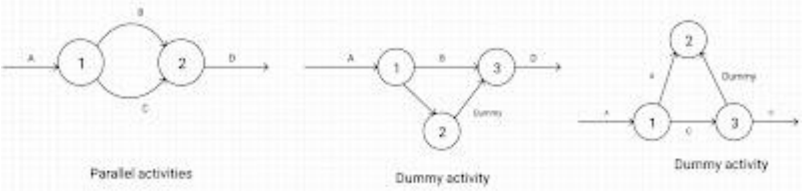
2. Dangling:

A case of disconnected activity before the completion of all activities, which is also known as dangling, is shown in figure. In this case, activity C does not give any result as per the rules of the network. The dangling may be avoided by adopting rule 5 of constructing the network.

3. Redundant Activity

Following are the two cases in which the use of dummy activity may help in drawing the network correctly, as per the various rules:

i. When two or more parallel activities in a project have the same head and tail events, i.e., two events are connected with more than one arrow. In figure, activities B and C have a common predecessor - activity A. At the same time, they have activity D as a common successor. To derive correct network, a dummy activity for the ending event B is required to show that D may not start before B and C, is completed. This is shown in figure:



Parallel activities

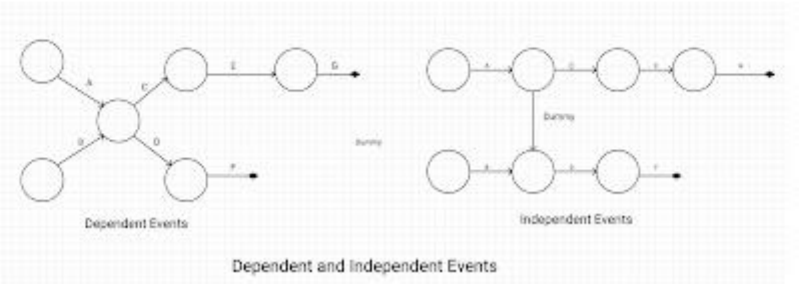
Dummy activity

Dummy activity

Parallel Activities and Dummy Activity

ii. When two chains of activities have a common event, yet are wholly or partly independent of each other, as shown in figure. A dummy which is used in such a case, to establish proper logical relationships, is also known as Logic Dummy Activity.

In figure, if head event of C and E do not depend on the completion of activities A and B, then the network can be re-drawn, as shown in figure. Otherwise, the pattern of figure must be adhered to.



Self-Assessment Exercise 2

1. Explain the concept of a project dummy activity
2. Draw a dummy activity network diagram
3. Outline the step in Drawing Network

5.6 Summary

This unit discussed the Project Network Analysis. Project network is graphic representation of a project's operations. It is the combination of activities and events which are require reaching the end objective of a project.

Various activities are performed at the same time and there are various activities which can be started only at the completion of other activities in a big project. The main work for detailed study of the product is to determine the information about the project and then discover a new, better and quicker way to get the work done.

Thus thorough study of project is done through some suitable diagram which shows various activities and their positions in the project. It is also helpful to know that in what way the delay in any activity can affect the whole project in terms of time and money. A number of nodes (typically shown as small circles or rectangles) and a number of arcs (shown as arrows) that connect two different nodes exist in a project network (figure).

5.7 References/Further Readings/Web Resources

projectmanager.com (2022). Critical Path Method. Retrieved from https://www.projectmanager.com/guides/critical-path-method#:~:text=The%20critical%20path%20method%20(CPM,which%20are%20known%20as%20paths

5.8 Possible Answers to SAEs

Answers to SAEs 1

1. Define the concept of Network Analysis
   Network analysis involves a group of techniques which are used for presenting information about the time and resources involved in the project so as to assist in the planning, scheduling and controlling of the project.

2. Mention and briefly explain the objectives of Network Analysis

i Minimize Production Delay, Interruptions and Conflicts:
This is achieved by identifying all activities involved in the project, their precedence constraints, etc.

ii Minimization of Total Project Cost
After calculating the total cost of the project the next step is to minimise the total cost. It is done through the calculation of cost of delay in the completion of an activity of the project and calculating the cost of the resources which are required to complete the project in a given time period.

iii Trade-off between Time and Cost of Project
The duration of same activity can be reduced if additional sources are employed and this is the main idea on which the trade-off between time and cost of project is based. Due to technical reasons, the duration can be reduced in a specific limit. Similarly, there is also a most cost efficient duration called 'normal point' stretching the activity beyond it may lead to a rise in direct cost.

iv Minimization of Total Project Duration:
After checking the actual performance against the plan the project duration can be controlled and minimized. If any major difference is found then apply the necessary reschedule process by updating and revising the uncompleted portion of the project.

v Minimization of Idle Resources:
If there is any variation in the use of scars resources then it can disturb the entire plan and hence it is required that efforts should be made to avoid any increase in cost due to idle resources.

3. Mention and Explain the Networking Components

a. Events:

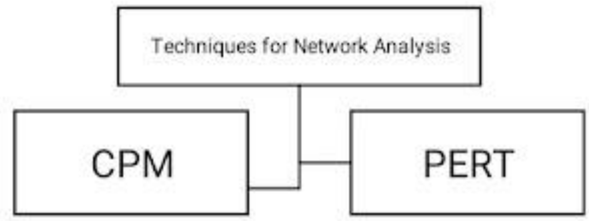
In a network diagram events represent the project milestones. For example, start or completion of an activity or activities, and occurrence of the events at a particular instance of time at which some specific portion of project has been or is to be achieved.

b. Jobs/Activity/Task :

The project operations (or tasks) are represented by activities which are conducted in a network diagram. These activities take a certain amount of time and require resources for completion.

c. Techniques Which Are Used In Network Analysis

The two common techniques which are used in network analysis are shown in figure below :



The managers are supported by two well-known network analysis techniques, viz, Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT) in planning and controlling of large scale construction projects, research and development, and so on.

Answers to SAEs 2

1. Explain the concept of a project dummy activity

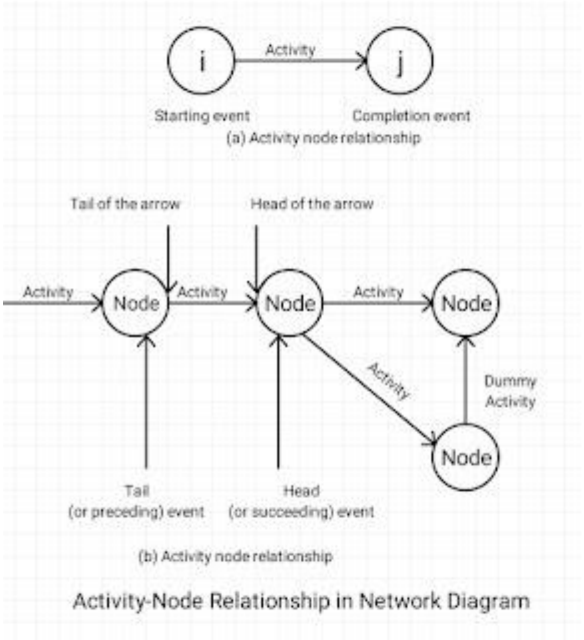
A dummy activity is used in a network to establish the precedence relationship among various activities of the project. It is needed when:

a) Two or more parallel activities in a project have same - the head and tail events.

b) Two or more activities have some (but not all) of their immediate predecessor activities in common.

2. Draw a dummy activity network diagram

Dummy activity is represented by a dotted line in the network diagram as shown in figure below.


i
Activity
j
Starting event
Completion event
(a) Activity node relationship
Tail of the arrow
Head of the arrow
Activity
Node
Activity
Node
Activity
Node
Activity
Dummy Activity
Node
Tail (or preceding) event
Head (or succeeding) event
(b) Activity node relationship
Activity-Node Relationship in Network Diagram

3. Outline the step in Drawing Network

The steps of network construction are as follows:

Step 1: Properly define the project and it's all important activities or tasks.

Step 2: Develop the relationships among the activities. Decide which activities must precede the others.

Step 3: Connect all the activities and draw the network.

Step 4: Time and/or cost estimates are assigned to each activity.

Step 5: Calculate the path which has the longest time and this is called critical path.

Step 6: Use the network for planning, scheduling, monitoring and controlling the project.

MODULE 6

UNIT 1 Critical Path of a Project (CPA) and Program Evaluation Review Technique (PERT)

Unit Structure

1. Introduction

In our last class we discussed the network analysis. This unit will continue with the discussion on the Critical Path of a Project (CPA) and Program Evaluation Review Technique (PERT) and demonstrate the network analysis of a project. CPM History
The critical path method was developed in the late 1950s by Morgan R. Walker and James E. Kelley. The origins of the critical path method are closely related with the Program Evaluation and Review Technique (PERT), a similar method which is commonly used in conjunction with CPM.

1.2 Learning Outcomes

By the end of this unit, will be able to
:

- explain what the Critical Path of a Project is about
- explain what the Critical Path Method (CPM) is about

- outline the reason why CPM is important in Project Management?
- state Key Elements of CPM
- itemize the steps in mapping CPM
- trance the history of Program Evaluation Review Technique (PERT)
- explain what is Program Evaluation Review Technique (PERT) is about
- outline the steps in mapping out a complex project

1.3 Critical Path of a Project (CPA) and Program Evaluation Review Technique (PERT)

1.3.1 What Is the Critical Path of a Project?

In project management, the critical path is the longest sequence of tasks that must be completed to complete a project. The tasks on the critical path are called critical activities because if they're delayed, the whole project completion will be delayed.

1.3.2 What Is the Critical Path Method (CPM)?

The critical path method (CPM) is a technique that's used by project managers to create a project schedule and estimate the total duration of a project.

The CPM method, also known as critical path analysis (CPA), consists in using a network diagram to visually represent the sequences of tasks needed to complete a project. Once these task sequences or paths are defined, their duration is calculated to identify the critical path, which determines the total duration of the project.

1.4 Why Is CPM Important in Project Management?

Important of Critical Path for project managers
Finding the critical path is very important for project managers because it allows them to:

i. accurately estimate the total project duration
ii. identify task dependencies, resource constraints and project risks
iii. prioritize tasks and create realistic project schedules
iv. find the critical path, project managers use the critical path method (CPM) algorithm to define the least amount of time necessary to complete each task with the least amount of slack.

Once done by hand, nowadays the critical path can be calculated automatically with project scheduling software equipped with Gantt charts, which makes the whole CPM method much easier.

Project Manager can calculate the critical path for you on our award-winning Gantt charts—learn more.

Now that we know what's the critical path of a project, we can learn about the critical path method (CPM), an important project management technique that's based on this concept.

Projects are made up of tasks that have to adhere to a schedule in order to meet a timeline. It sounds simple, but without mapping the work, your project scope can quickly get out of hand and you'll find your project off track.

Using the critical path method is important when managing a project because it identifies all the tasks needed to complete the project, then determines the tasks that must be done on time, those that can be delayed if needed and how much float or slack you have.

When done, properly critical path analysis can help you to:

i. Identify task dependencies, resource constraints and project risks
ii. Accurately estimate the duration of each task
iii. Prioritize tasks based on their float or slack time, which helps with project scheduling and resource allocation
iv. Identify critical tasks that have no slack and make sure those are completed on time
v. Monitor your project progress and measure schedule variance
vi. Use schedule compression techniques like crash duration or fast tracking.

Self-Assessment Exercise 1

1. What Is the Critical Path of a Project?
2. What Is the Critical Path Method (CPM)?
3. Why Is CPM Important in Project Management?

1.4.1 CPM Key Elements

Before we learn the steps to calculate the critical path, we'll need to understand some key CPM concepts.

Earliest start time (ES): This is simply the earliest time that a task can be started in your project. You cannot determine this without first knowing if there are any task dependencies

Latest start time (LS): This is the very last minute in which you can start a task before it threatens to delay your project schedule

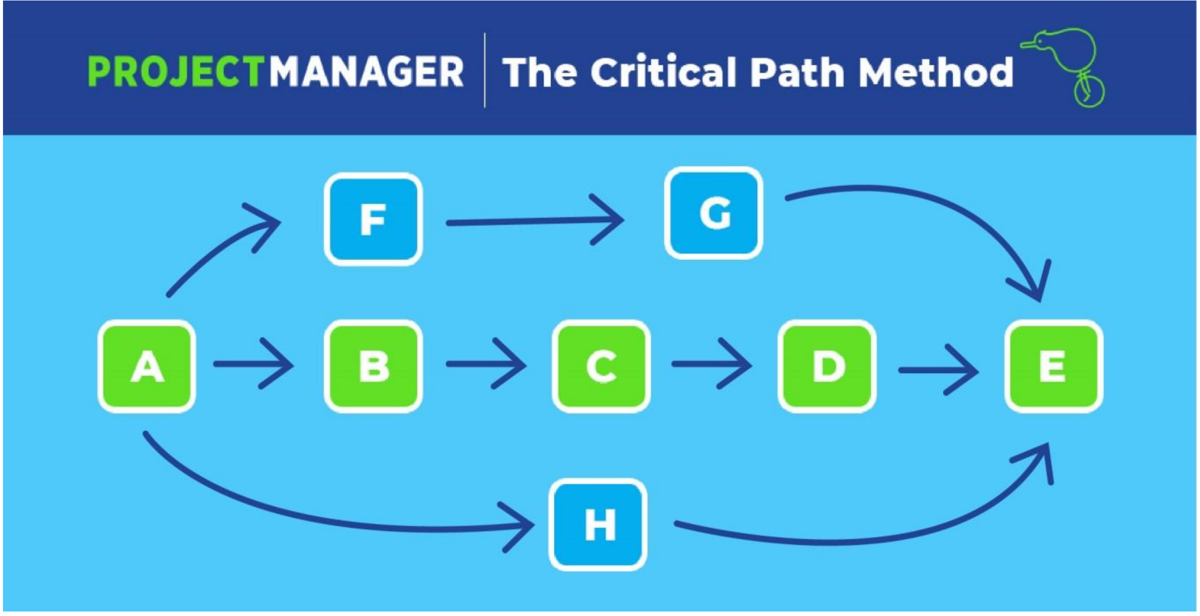
Earliest finish time (EF): The earliest an activity can be completed, based on its duration and its earliest start time

Latest finish time (LF): The latest an activity can be completed, based on its duration and its latest start time

Float: Also known as slack, float is a term that describes how long you can delay a task before it impacts its task sequence and the project schedule. The tasks on the critical path have zero float, because they can't be delayed

Let's take a look at some critical path examples to better understand these critical path analysis elements.

Critical Path Examples
Here's an example of a CPM diagram. Although it's high-level, it can help you visualize the meaning of a critical path for a project schedule. For now, we'll use this critical path diagram to explain the elements that make up the CPM method.

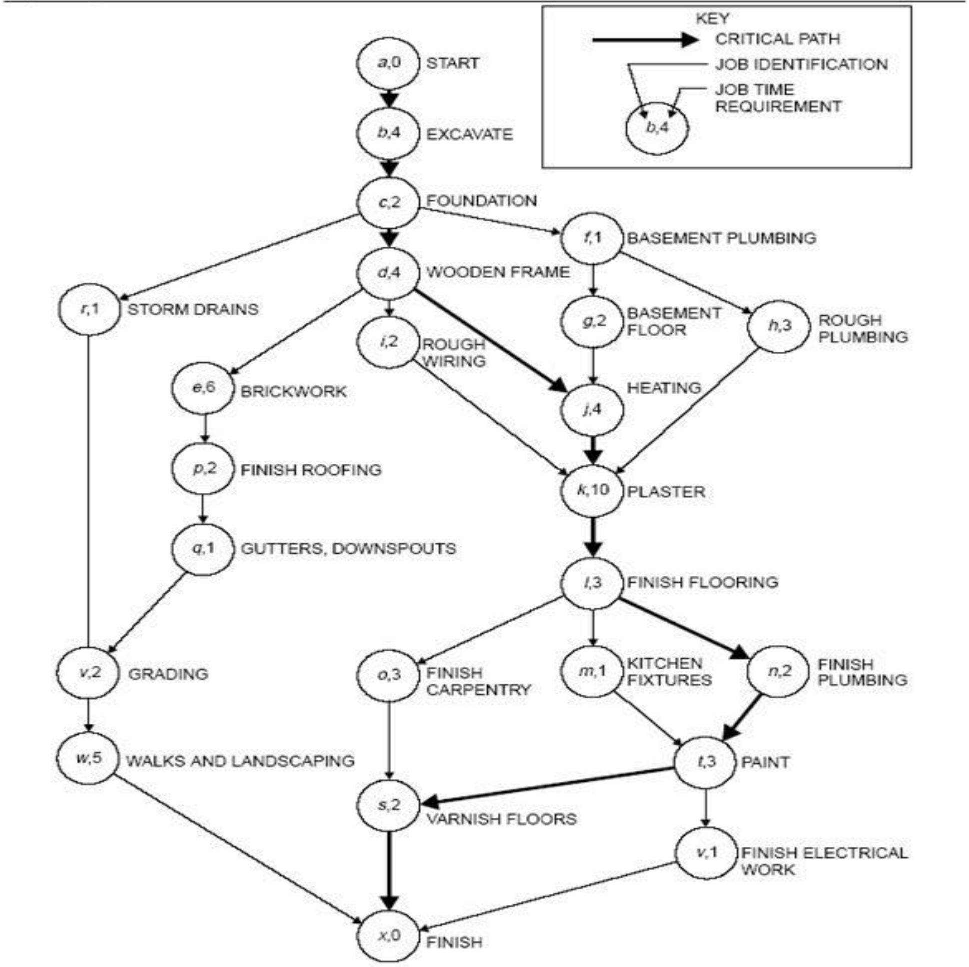


Adopted from https://www.projectmanager.com/guides/critical-path-method.

As you can see in this critical path diagram, project activities are represented by letters and the critical path is highlighted in green. Tasks F, G and H are non-critical activities with float or slack. We can also identify task dependencies between the critical path activities, and also between activities (A, F and G) or (A, H and E), which are parallel tasks.

Here's another critical path example from Harvard Business Review, which shows a critical path schedule for the construction of a house. Each circle in the CPM diagram represents a project activity, as well as it's duration, while the bolded arrows link the critical path activities. As projects become more complex, you'll find more parallel tasks, like in this example.

EXHIBIT II
Project Graph



Source: ***Harvard Business Review***

How to Find the Critical Path of a Project in 8 Steps
Now that you know the key concepts of the critical path method, here's how to calculate

1.4.2 Steps Mapping CPM complex Projects

The 8 steps in critical path are:

1. **Collect Project Activities:** Use a work breakdown structure to collect all the project activities that lead to the final deliverable.
2. **Identify Task Dependencies:** Figure out which tasks are dependent on other tasks before they can begin. Use your judgement and your team members' feedback. Failing to define task dependencies correctly makes the critical path method useless.
3. **Create a Critical Path Diagram**: A critical path analysis chart, or network diagram, depicts the order of activities.
4. **Estimate Timeline**: To use the critical path method, you'll need to estimate the duration of each task. Use data from past projects and other sources of information such as subject matter experts.
5. **Use the Critical Path Algorithm:** The critical path algorithm has two parts; a forward pass and a backwards pass.

Forward Pass
Use the network diagram and the estimated duration of each activity to determine their Earliest Start (ES) and Earliest Finish (EF). The ES of an activity is equal to the EF of its predecessor, and its EF is determined by the formula EF = ES + t (t is the activity duration). The EF of the last activity identifies the expected time required to complete the entire project (projectmanager.com, 2022).

Backward Pass
Begins by assigning the last activity's Earliest Finish as its Latest Finish. Then the formula to find the LS is LS = LF – t (t is the activity duration). For the previous activities, the LF is the smallest of the start times for the activity that immediately follows.

6. **Identify the Float or Slack of Each Activity**: Use this formula to determine the float or slack of each task. Float = LS – ES

7. **Identify the Critical Path:**The activities with 0 float make up the critical path. All of these critical path activities are dependent tasks except for the first task in your CPM schedule. All project tasks with positive slack are parallel tasks to the critical path activities.

8. **Revise during Execution:** Continue to update the critical path network diagram as you go through the execution phase.

These critical path analysis steps determine what tasks are critical and which can float, meaning they can be delayed without negatively impacting the project schedule. Now you have the information you need to plan the critical path schedule more accurately and have more of a guarantee you'll meet your project deadline.

You also need to consider other changes or constraints that might change the project schedule. The more you can account for these unexpected events or risks, the more accurate your critical path schedule will be. If time is added to the project because of these constraints, that is called a critical path drag, which is how much longer a project will take because of the task and constraint (projectmanager.com, 2022).

1.5 History of Program Evaluation Review Technique (PERT)

PERT was developed by the U.S. Navy in the 1950s to help coordinate the thousands of contractors it had working on myriad projects.
While PERT was originally a manual process, today there are computerized PERT systems that enable project charts to be created quickly.

The only real weakness of the PERT process is that the time required for completion of each task is very subjective and sometimes no better than a wild guess. Frequent progress updates help refine the project timeline once it gets underway.

1.5.1 What is Program Evaluation Review Technique (PERT)?

The Program Evaluation Review Technique, or PERT, is a visual tool used in project planning. Using the technique helps project planners identify start and end dates, as well as interim required tasks and timelines. The information is displayed as a network in chart form.

PERT helps project planners identify:

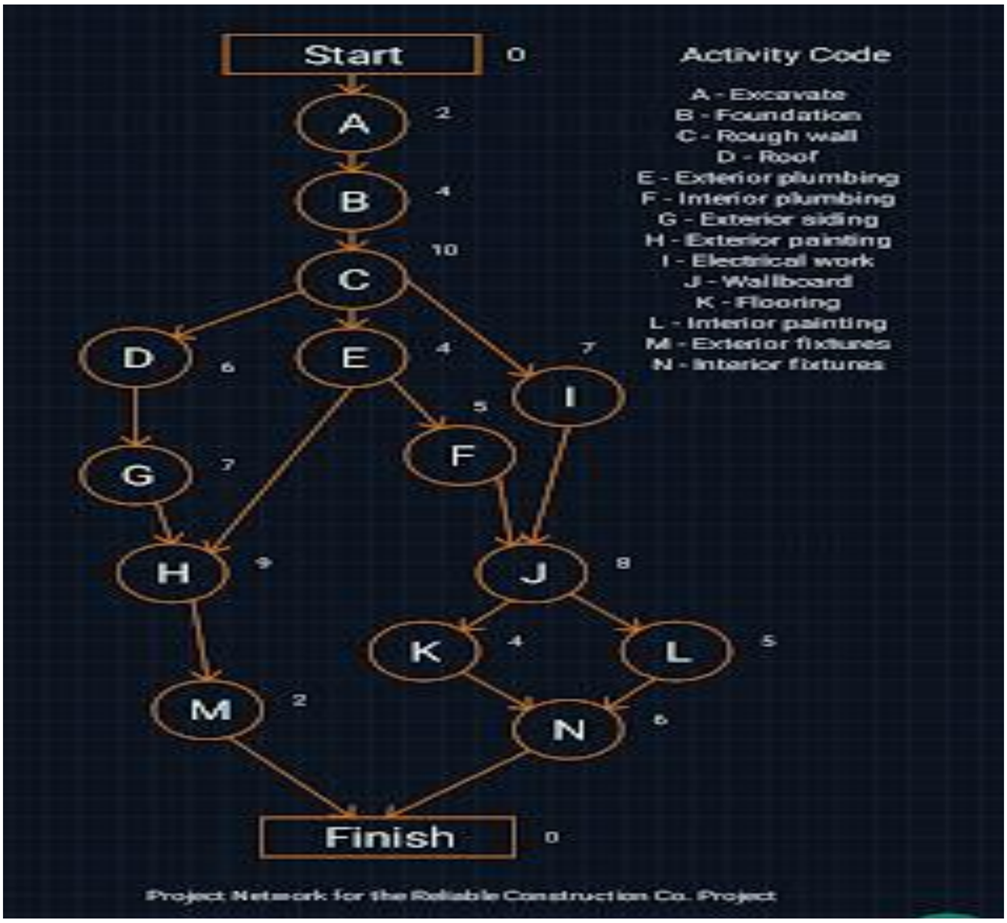
i. Start and end dates
ii. Anticipated total required completion time
iii. All activities, referred to as events on the chart, that impact the completion time
iv. The required sequence of events
v. The probability of completion by a certain date

1.5.2 Steps in mapping out a complex project

PERT has a set series of steps in mapping out a complex project, which include:

1. List all the tasks and milestones (a.k.a. events) required for completion of the project
2. Determine the required sequence of tasks
3. Design a chart to visually display all the steps
4. Estimate the time required for each task
5. Identify the critical path – the longest series of tasks in the project
6. Adjust the chart to reflect progress made once the project starts

A PERT chart uses numbered circles or rectangles to represent milestones and straight lines with arrows at the end to represent tasks to be completed. The direction of the arrows, and the numbers, indicate the required sequence. Typically, the numbers increase by 10 at each milestone, so that new tasks can be added along the way without requiring the whole chart to be redrawn and numbered.



Project Network for the Reliable Construction Co. Project

Self-Assessment Exercise 2

1. Outline the Key Elements of CPM
2. State the Steps Mapping CPM complex Projects
3. What is Program Evaluation Review Technique (PERT)?
4. State the Steps in mapping out a complex project

1.6 Summary

In project management, the critical path is the longest sequence of tasks that must be completed to complete a project. The tasks on the critical path are called critical activities because if they're delayed, the whole project completion will be delayed

The critical path method (CPM) is a technique that's used by project managers to create a project schedule and estimate the total duration of a project.

The CPM method, also known as critical path analysis (CPA), consists in using a network diagram to visually represent the sequences of tasks needed to complete a project

Finding the critical path is very important for project managers because it allows them to:

i. Accurately estimate the total project duration
ii. Identify task dependencies, resource constraints and project risks
iii. Prioritize tasks and create realistic project schedules
iv. To find the critical path, project managers use the critical path method (CPM) algorithm to define the least amount of time necessary to complete each task with the least amount of slack.

Once done by hand, nowadays the critical path can be calculated automatically with project scheduling software equipped with Gantt charts, which makes the whole CPM method much easier.

We learned the steps to calculate the critical path; we also understand some key CPM concepts.

Earliest start time (ES): This is simply the earliest time that a task can be started in your project. You cannot determine this without first knowing if there are any task dependencies
Latest start time (LS): This is the very last minute in which you can start a task before it threatens to delay your project schedule

Earliest finish time (EF): The earliest an activity can be completed, based on its duration and its earliest start time
Latest finish time (LF): The latest an activity can be completed, based on its duration and its latest start time
Float: Also known as slack, float is a term that describes how long you can delay a task before it impacts its task sequence and the project schedule. The tasks on the critical path have zero float, because they can't be delayed

Let's take a look at some critical path examples to better understand these critical path analysis elements
The critical paths in 8 steps are:

1. **Collect Project Activities**: Use a work breakdown structure to collect all the project activities that lead to the final deliverable.
2. **Identify Task Dependencies:** Figure out which tasks are dependent on other tasks before they can begin. Use your judgement and your team members' feedback. Failing to define task dependencies correctly makes the critical path method useless.
3. **Create a Critical Path Diagram**: A critical path analysis chart, or network diagram, depicts the order of activities.
4. **Estimate Timeline:** To use the critical path method, you'll need to estimate the duration of each task. Use data from past projects and other sources of information such as subject matter experts.
5. **Use the Critical Path Algorithm:** The critical path algorithm has two parts; a forward pass and a backwards pas

The Program Evaluation Review Technique, or PERT, is a visual tool used in project planning. Using the technique helps project planners identify start and end dates, as well as interim required tasks and timelines. The information is displayed as a network in chart form.

6. **PERT helps project planners identify:**
   Start and end dates
   Anticipated total required completion time
   All activities, referred to as events on the chart, that impact the completion time
7. **The required sequence of events**
8. **The probability of completion by a certain date**

PERT has a set series of steps in mapping out a complex project, which include:

1. List all the tasks and milestones (a.k.a. events) required for completion of the project
2. Determine the required sequence of tasks

3. Design a chart to visually display all the steps
4. Estimate the time required for each task
5. Identify the critical path – the longest series of tasks in the project
6. Adjust the chart to reflect progress made once the project starts

1.7 References/Further Readings/Web Resources

Projectmanager.com (2022). Critical Path Method. Retrieved from https://www.projectmanager.com/guides/critical-path-method#:~:text=The%20critical%20path%20method%20(CPM,which%20are%20known%20as%20paths.

1.8 Possible Answers to SAEs

Answers to SAEs 1

1. In project management, the critical path is the longest sequence of tasks that must be completed to complete a project. The tasks on the critical path are called critical activities because if they're delayed, the whole project completion will be delayed

2. The critical path method (CPM) is a technique that's used by project managers to create a project schedule and estimate the total duration of a project.

   The CPM method, also known as critical path analysis (CPA), consists in using a network diagram to visually represent the sequences of tasks needed to complete a project

3. Important of Critical Path for project managers

   Finding the critical path is very important for project managers because it allows them to:

i. Accurately estimate the total project duration

ii. Identify task dependencies, resource constraints and project risks

iii. Prioritize tasks and create realistic project schedules

iv. To find the critical path, project managers use the critical path method (CPM) algorithm to define the least amount of time necessary to complete each task with the least amount of slack.

Once done by hand, nowadays the critical path can be calculated automatically with project scheduling software equipped with Gantt charts, which makes the whole CPM method much easier.

Answers to SAEs 1

1. Key Elements of CPM

   Before we learn the steps to calculate the critical path, we'll need to understand some key CPM concepts.

i. **Earliest start time (ES):** This is simply the earliest time that a task can be started in your project. You cannot determine this without first knowing if there are any task dependencies

ii. **Latest start time (LS):** This is the very last minute in which you can start a task before it threatens to delay your project schedule

iii. **Earliest finish time (EF):** The earliest an activity can be completed, based on its duration and its earliest start time

iv. **Latest finish time (LF):** The latest an activity can be completed, based on its duration and its latest start time

v. **Float**: Also known as slack, float is a term that describes how long you can delay a task before it impacts its task sequence and the project schedule. The tasks on the critical path have zero float, because they can't be delayed

Let's take a look at some critical path examples to better understand these critical path analysis elements

2. Steps Mapping CPM complex Projects

The critical path in 8 steps.

i. Collect Project Activities
Use a work breakdown structure to collect all the project activities that lead to the final deliverable

ii. Identify Task Dependencies
Figure out which tasks are dependent on other tasks before they can begin. Use your judgement and your team members' feedback. Failing to define task dependencies correctly makes the critical path method useless.

iii. Create a Critical Path Diagram
A critical path analysis chart, or network diagram, depicts the order of activities.

iv. Estimate Timeline
To use the critical path method, you'll need to estimate the duration of each task. Use data from past projects and other sources of information such as subject matter experts.

v. Use the Critical Path Algorithm
The critical path algorithm has two parts; a forward pass and a backwards pas

3. The Program Evaluation Review Technique, or PERT, is a visual tool used in project planning. Using the technique helps project planners identify start and end dates, as well as interim required tasks and timelines. The information is displayed as a network in chart form.

PERT helps project planners identify:

i. Start and end dates
ii. Anticipated total required completion time
iii. All activities, referred to as events on the chart, that impact the completion time
iv. The required sequence of events
v. The probability of completion by a certain date

4. Steps in mapping out a complex project

PERT has a set series of steps in mapping out a complex project, which include:

1. List all the tasks and milestones (a.k.a. events) required for completion of the project
2. Determine the required sequence of tasks
3. Design a chart to visually display all the steps
4. Estimate the time required for each task
5. Identify the critical path – the longest series of tasks in the project
6. Adjust the chart to reflect progress made once the project starts