BUSINESS ANALYTICS SKILLS FOR THE FUTURE-PROOFS SUPPLY CHAINS -

STATISTICAL METHODS FOR ANALYSING LOGISTICS DATA

Authors:
Sanja Bojić
Kristijan Brglez
Maja Fošner
Roman Gumzej
Rebeka Kovačič Lukman
Benjamin Marcen
Marinko Maslarić
Boško Matović
Dejan Mirčetić

Co-funded by the European Union

Funded by the European Union.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Sanja Bojić, Kristijan Brglez , Maja Fošner, Roman Gumzej, Rebeka Kovačič Lukman, Benjamin Marcen, Marinko Maslarić, Boško Matović, Dejan Mirčetić

STATISTICAL METHODS FOR ANALYSING LOGISTICS DATA

Poznan 2025

Publisher:

Wyższa Szkoła Logistyki
Estkowskiego 6
61-755 Poznan, Poland
www.wsl.com.pl

ISBN 978-83-62285-70-9 (online)

Reviewers:
- prof. Dr. Ajda Fošner, University of Primorska, Faculty of Management
- prof. Dr. Nikša Alfirević, University of Split, Faculty of Economics

Technical Editor: Kristijan Brglez, Faculty of Logistics University of Maribor, Slovenia

Cover design: Michał Adamczak, Poznan School of Logistics, Poznan, Poland

The Monography was written in the framework of the project Business Analytics Skills for the Future-proof Supply Chains (BAS4SC) [2022-1-PL01-KA220-HED-000088856], funded by the Erasmus+ Programme. This project is funded with support from the European Commission. This publication reflects the views only of the author, and the commission cannot be held responsible for the any use which may be made of the information contained therein.

Monography is freely available at: enauka.put.poznan.pl

Foreword

The field of logistics is becoming increasingly reliant on data-driven insights to optimise operations, reduce costs, and ensure efficiency across supply chains. This textbook, *Statistical Methods for Analyzing Logistics Data*, serves as a critical resource for students and professionals alike, equipping them with the necessary skills to navigate the complexities of modern supply chain management. Developed as part of the Business Analytics Skills for Future-proof Supply Chains (BAS4SC) project, this book offers a comprehensive overview of statistical methods, data management, and advanced analytical techniques explicitly tailored to the logistics industry.

Through careful research, this textbook addresses the gaps in current educational offerings by combining theoretical knowledge with practical applications. The ten chapters included a detailed exploration of key topics such as demand forecasting, simulation modelling, regression analysis, and integrating artificial intelligence and machine learning in logistics operations. By using widely recognised tools such as SPSS, R, and SQL, the content of this textbook is designed to bridge the gap between academic learning and industry needs.

We hope that this textbook will provide foundational knowledge in business analytics for logistics and inspire innovation in the field, fostering future leaders equipped to tackle the challenges of a rapidly evolving supply chain landscape.

Prof. Dr. Sanja Bojić
Kristijan Brglez
Prof. Dr. Maja Fošner
Prof. Dr. Roman Gumzej
Prof. Dr. Rebeka Kovačič Lukman
Assist. Prof. Dr. Benjamin Marcen
Prof. Dr. Marinko Maslarić
Assist. Prof. Dr. Boško Matović
Dr. Dejan Mirčetić

TABLE OF CONTENT

INTRODUCTION

This textbook, entitled *Statistical Methods for Analyzing Logistics Data*, is the third in a series developed as part of the Business Analytics Skills for Future-proof Supply Chains (BAS4SC) project. Several preliminary research activities were conducted to determine this textbook's content. First, a comprehensive investigation was conducted to examine the business analytics courses, their content, and the skills they impart to logistics students across the European Union, the United States, and the United Kingdom. This analysis revealed a gap between the logistics knowledge and statistical skills required in the field and those currently offered to students. Based on in-depth interviews with university teaching staff, students, and industry professionals, over 100 business analytics skills were identified as essential. Using the ABC ranking classification method, 33 skills were selected for inclusion in this book, primarily focused on mathematics, computer science, management, applied mathematics, and statistics. Combining these identified needs and skills led to the development of ten content chapters that address the most critical skills required in the field.

The first chapter covers Introductory Statistics and provides a comprehensive overview of statistical concepts and their applications, particularly within supply chain analysis. It begins by emphasising the critical role statistics play in optimising supply chains. It uses descriptive statistics like mean, median, and standard deviation to analyse delivery times, inventory levels, and costs. The chapter introduces predictive techniques like regression and time series analysis for forecasting demand and inventory. It further explores the importance of variables, differentiating between qualitative and quantitative types, and delves into core statistical measures like average, median, mode, variance, and standard deviation.

Additionally, it covers graphical data representation methods, such as histograms and scatter plots, and highlights the difference between descriptive and inferential statistics. Finally, it introduces key concepts in correlation, regression, and probability distributions, offering tools to understand relationships between variables and model random phenomena in data. These statistical techniques help improve supply chain decision-making, efficiency, and risk management.

The second chapter, Statistics for Business Analytics, explores essential statistical concepts and techniques to derive insights from business data. It begins by introducing the importance of data analysis in business decision-making. It explains the foundational role of the normal distribution, which serves as a basis for many statistical methods. The chapter delves into standard deviation, emphasising its importance in measuring data variability. It also covers sampling distributions and the Central Limit Theorem, explaining how they infer population parameters from sample data. Topics such as hypothesis testing, Z-scores, and t-scores are explored to aid decision-making and probability calculations. The chapter concludes by discussing the standard error and confidence intervals, which help quantify the uncertainty surrounding estimates. Ultimately, the chapter equips readers with the statistical tools necessary for business analytics, enabling them to make informed and data-driven decisions.

The chapter on Data Management explores the various facets of managing data in logistics, focusing on data formats, organisation, and technologies. It begins with the role of Electronic Data Interchange (EDI) in exchanging information within supply chains using standardised alphanumeric formats. It explains the concept of information, data, and knowledge, discussing how data is digitised and organised into databases, warehouses, and knowledge bases. The chapter delves into logistic data, particularly the use of barcodes and RFID tags for identification and tracking in logistics. It also introduces data organisation techniques, ranging from spreadsheets to relational databases (RDBs), explaining key concepts such as primary and foreign keys, normalisation, and query languages like SQL. Additionally, the chapter discusses best practices for data filtering and error prevention during data input. Lastly, it discusses the differences between data warehouses and knowledge bases, highlighting their roles in business analysis and decision-making.

The Simulation Modelling and Analysis (SMA) chapter focuses on creating digital models to simulate real-world systems for optimisation and decision-making. It begins by explaining the Conant-Ashby theorem, which suggests that a simulation model must match the complexity of its real-world counterpart to regulate it effectively. SMA optimises supply chains and traffic networks in logistics, allowing managers to simulate and evaluate different scenarios. The chapter outlines critical simulation methodologies, such as Discrete Event Simulation (DES) for process-oriented analysis, System Dynamics (SD) for high-level system performance, Agent-Based Simulation (ABS) for modelling individual entities' behaviour, and Network Simulation (NS) for analysing network flows. Each method provides insights into various aspects of

logistics, from production cycles to traffic optimisation. The chapter concludes with an overview of logistics simulation projects, structured around the Design for Six Sigma and Deming's cycle of improvement, which help in planning, executing, and refining complex logistics systems.

The chapter Linear Regression with Single and Multiple Regressors introduces regression analysis to understand the relationship between dependent and independent variables. It begins with simple linear regression, where a single independent variable, like advertising expenditure, predicts an outcome such as sales. The chapter explains the construction of the regression model and the regression equation used to forecast the dependent variable based on sample data. It also covers the least squares method, an essential technique for estimating the regression line by minimising prediction errors. Next, it introduces the coefficient of determination (R^2) to measure how well the regression model fits the data. The chapter then delves into multiple regression, where two or more independent variables predict a dependent variable, offering a more comprehensive analysis. Examples include predicting journey times based on distance and the number of deliveries.

Introduction to Operations Research chapter focuses on using analytical methods to improve decision-making, particularly in logistics and supply chain management. Operations research employs modelling, statistics, and optimisation techniques to find optimal solutions to complex problems, enabling efficient resource management, inventory control, and process optimisation. The chapter highlights strategic logistics planning, which involves methods like Six Sigma and Just-in-Time production to enhance operational efficiency. Business intelligence (BI) and business analytics (BA) are crucial in data analysis, allowing companies to make informed decisions using forecasting, predictive analytics, and data visualisation. The chapter also introduces multi-criteria decision-making (MCDM), which aids in evaluating and selecting optimal solutions based on various criteria. Finally, decision support systems (DSS) and knowledge-based engineering (KBE) help integrate knowledge and data into decision-making processes, further enhancing operational efficiency and strategic planning.

The chapter on Statistical Data Processing with SPSS introduces IBM's SPSS software as a powerful tool for automating complex statistical analysis, enhancing reliability, and facilitating decision-making. It explains how SPSS allows for data import, manipulation, and preparation through a user-friendly interface. The chapter covers key functionalities like descriptive

statistics, graph creation, and data visualisation. It also introduces fundamental statistical tests—T-tests, correlation, Chi-Square, and ANOVA—guiding readers through the setup and interpretation of each test. Additionally, it explores data management tools such as merging, splitting, and computing variables in datasets, demonstrating how SPSS enhances statistical analysis in logistics and other domains.

Chapter 8 explores Business Analytics (BA) and its application through tools like R and SQL to solve business problems. BA aims to improve decision-making and company performance using data-driven methods. It includes descriptive, predictive, and prescriptive platforms for analysing data and making informed decisions. R is introduced as a robust, open-source statistical analysis and visualisation tool, while SQL is essential for managing and querying large databases. The chapter details the integration of R and SQL for efficient business analytics, emphasising how data stored in SQL can be analysed using R scripts to automate tasks. Practical examples, like querying the Chinook Database, illustrate how R and SQL work together to generate insights, such as identifying top-selling albums. This synergy between BA, R, and SQL enhances the ability to manage and analyse dynamic business data.

Chapter 9 focuses on supply chain demand forecasting, visualisation, and feature engineering. Demand forecasting predicts customer needs, changing the entire supply chain and reducing logistics costs. The chapter outlines critical steps for demand forecasting, including defining the problem, collecting data, analysing trends, selecting models, and evaluating them. Visualisation helps identify patterns such as seasonality and trends, which can inform model selection. S-ARIMA (Seasonal Autoregressive Integrated Moving Average) is highlighted as an effective model for handling complex time series data, especially with seasonal demand patterns. An example in the food industry demonstrates the S-ARIMA model's ability to predict demand and guide decision-making. Finally, the chapter covers how to test and validate forecasting models to ensure their effectiveness in real-world applications, using metrics like RMSE and MAPE for performance evaluation.

The last chapter explores the role of artificial intelligence (AI) and machine learning (ML) in supply chains, beginning with an overview of AI's development from symbolic systems to modern ML approaches. AI refers to the automation of tasks that typically require human intelligence, with ML being a subset of AI that focuses on learning patterns from data. The chapter highlights how AI/ML models, such as supervised and unsupervised learning, are

applied to solve business problems, including demand forecasting, inventory management, and optimisation. An essential case study involves applying AI and ML algorithms in a food factory's central warehouse, optimising forklift usage. By incorporating decision support systems (DSS), the AI/ML models assist managers in selecting the optimal number of forklifts, improving operational efficiency. The chapter emphasises how AI/ML can capture expert knowledge, reduce costs, and improve decision-making processes in supply chain management.

1. Introductory statistics

1.1 The role and importance of statistics in analysing data in supply chains

Statistics play a key role in modern supply chains, where effective management, planning and control are essential. Statistical methods are used to collect, analyse and interpret data, enabling companies to better understand and optimise their supply chains.

Let us outline some of the important roles of statistics in supply chain analysis.

Descriptive statistics are key to describing the basic properties of supply chain data, such as mean, standard deviation, median, quartiles and other measures. These tools help us to understand the distribution and characteristics of data such as average delivery times, quantities in stock and average costs, which contributes to a better understanding and management of the supply chain.

In addition, statistical techniques such as regression, time series analysis and pattern analysis are used to predict future events and trends in supply chains. This includes forecasting demand, inventory and delivery times, allowing better planning and adjustment of supply.

Statistics play a key role in identifying patterns in the data, allowing a better understanding of supply chain behaviour, including seasonal patterns, trends and cycles in demand.

Inventory optimisation is another key area where statistics help to determine the optimal order quantities that minimise storage and ordering costs, using methods such as EOQ (Economic Order Quantity).

In addition, the statistics are also used to assess supply chain risks, such as the likelihood of delays in deliveries, damage during transport and other potential problems.

Through statistical monitoring and process control, we identify deviations from standards, allowing us to improve the quality and efficiency of supply chain processes.

In addition, statistics are used to monitor and improve the quality of products and services in the supply chain, including quality control at suppliers.

Finally, statistics are a key tool for making more informed decisions on procurement, inventory, supplier selection and other aspects of supply management, contributing to the efficient and effective operation of the entire supply chain.

In supply chain analysis, statistics are used to optimise processes, reduce costs, increase efficiency and improve customer satisfaction. It enables a better understanding of supply chain dynamics and better risk management, which is crucial for the successful operation of companies and organisations in today's global environment.

1.2 Basic concepts of statistics

Variables

Variables are basic building blocks in statistics because they represent the properties or characteristics that are measured or observed in a survey, experiment or sample of data. Variables are essential for understanding and analysing data as they allow researchers, analysts and statisticians to describe, analyse and understand phenomena.

It is important to understand the different types of variables and their importance in statistics.

Qualitative (descriptive, categorical) variables are variables that represent qualitative characteristics or categories that cannot be counted or classified according to a mathematical order. Examples include gender (male, female), eye colour (blue, brown, green) or car type (saloon, station wagon, SUV). Qualitative variables are often useful for describing demographic characteristics or traits.

Quantitative (numerical) variables are variables that represent numerical values that can be counted or measured and can be sorted in some mathematical order. Examples include age, height, temperature, income or survey scores. Quantitative variables are often used to analyse and quantitatively investigate phenomena.

Dependent and independent variables. The dependent variable is the one we want to investigate, measure or predict, while the independent variable is the one that is intended to influence the dependent variable. For example, if we want to investigate whether educational attainment affects income, income is the dependent variable and educational attainment is the independent variable.

Discrete and continuous variables. Variables can also be divided into discrete and continuous. Discrete variables have a limited set of possible values and are usually represented by integers. An example is the number of children in a family, where the possible values are 0, 1, 2, etc. Continuous variables, on the other hand, have an infinite number of possible values and are usually measured using decimal numbers. An example is the height of persons, where an infinite number of values are possible within a given range.

Variables are basic tools for research and data analysis. Understanding and correctly defining variables is crucial for carrying out statistical analyses and studying phenomena in research. Variables allow researchers to express and quantify different aspects of reality, enabling better understanding of phenomena, decision-making and prediction of future events. They also allow the use of different statistical techniques to test hypotheses, make predictions and better understand causal links between variables.

1.3 Basic statistical concepts with examples

Average (mean)

The **mean**, also known as the **average**, is one of the basic statistical measures. The mean is the arithmetic average of all the values in a data set. It is calculated by summing all the data and then dividing by the number of data.

Calculating the average:

- Add up all the values in the dataset.
- Divide the sum by the number of values in the set.
- The equation to calculate the average $(\bar{x})$ is: $\bar{x} = (x_1 + x_2 + x_3 + \ldots + x_n) / n$

Where $\bar{x}$ is the average. $x_1, x_2, x_3, \ldots, x_n$ are the values in the dataset. n is the number of values in the dataset.

Example:

Imagine a dataset representing students' grades in a maths exam: 80, 85, 90, 75, 95. To calculate the average, add all these values and divide by the number of grades, which in this case is 5:

Average = (80 + 85 + 90 + 75 + 95) / 5 = 425 / 5 = 85

So the average student score is 85. The average is useful for measuring the central tendency of the data and gives us a rough idea of what to expect as a "typical" value in the data set. However, the mean can change significantly if outliers or outliers are present in the data. It is therefore important to know other statistical measures such as the median and the mode to better understand the distribution of the data.

Median

The median is a statistical concept used to measure the middle value of a set of data. It is the value that divides the ordered data into two equal halves. This means that half of the data has values less than or equal to the median and the other half has values greater than or equal to the median. The median is one of the basic measures of central tendency in statistics and is used to describe the distribution of data, especially when the data are skewed or contain outliers.

How to calculate the median:

- First, you need to sort the dataset from the smallest to the largest value.
- If the number of data is even (n), then the median is the average of the two middle values. This means that the median is equal to the average of the values at position $n/2$ and $(n/2 + 1)$ when the data are sorted in ascending order.
- If the number of data is odd, then the median value is at the middle position.

Example:

Imagine the following data set representing the number of hours of sleep people got in a given period: 7, 6, 5, 8, 6, 9, 7

First, arrange the data in ascending order: 5, 6, 6, 7, 7, 8, 9

Since the number of data is odd (7), the median will be the value at the middle position, which is the 4th value in the ordered data set. So the median in this case is equal to 7 hours. This means that half of the people in this dataset get 7 or less hours of sleep, while the other half get 7 or more hours of sleep.

Modus

Modus is one of the basic statistical metrics used to measure the central tendency of a data set. The modus represents the value that occurs most frequently in the dataset. It is the value that has the highest frequency of occurrence among all the values in the data set.

Modus is useful for identifying the most frequent value in a data set and is particularly useful when analysing qualitative (categorical) variables where the values are non-numerical.

If there are multiple modes in the data set (multiple values occurring with similar maximum frequency), we speak of a multi-modal distribution. If all the data have the same frequency of occurrence, then the data set has no mode.

Example: imagine a dataset representing the colours of the cars in a car park:

Red, Blue, Red, Green, Blue, Blue, Blue, Red

In this case, the modus is "Red", as this value occurs most frequently (three times), while "Blue" and "Green" occur less frequently.

The modus is simple to calculate, as it simply identifies the value with the highest frequency of occurrence in the dataset. Modus is used to describe characteristic values in data and can be useful in understanding which value is most characteristic of a particular situation or group.

Variance range (VR, Range, range)

The difference between the maximum and minimum values in a data set is a statistical concept called the range. This measures how big the difference is between the maximum (maximum) and minimum (minimum) values in the data set. The range is a simple way to estimate the range of values in a data set and to measure the variability between the minimum and maximum values.

Calculating the variation margin is simple:

- First, find the minimum value (min) and the maximum value (max) in the dataset.
- Then calculate the difference between the maximum and minimum value (max - min).

Example: imagine a dataset representing the ages of the participants of an event: 20, 25, 30, 35, 40. To calculate the variation margin, first find the minimum value (20) and the maximum

value (40) in the data set. Then you calculate the difference between the maximum and the minimum value: VR = 40 - 20 = 20

So the variation margin in this case is 20 years. This means that the difference between the oldest and the youngest participant is 20 years.

The variance decomposition is useful for estimating the range of values in a dataset, but it is quite simple and does not take into account all the values in the dataset. For a more detailed analysis of data variability and dispersion, other statistical measures such as the variance or quartiles are commonly used.

Variance and standard deviation

Variance is the average of the squared deviations from the mean. It is the square of the standard deviation. **Standard deviation** is a statistical measure used to measure the dispersion or variability in a set of data. It tells how far the values are from the mean (average) in the set. Standard deviation is one of the most commonly used measures of dispersion in statistics and is calculated by calculating the square root of the variation (variance).

Calculating the standard deviation:

- First, calculate the variation (variance). The variation (variance) is calculated by taking the average of all the values in the set for each value in the set, then squaring and summing these differences.
- Once you have the value of the variation margin (\sigma^2), calculate the standard deviation by calculating the square root of the variation margin. This is done by taking the square root of σ^2:

Standard deviation $\sigma = \sqrt{\sigma^2}$

Standard deviation measures how dispersed the values are around the mean in the data set. A higher value of the standard deviation means that the values are more spread out and differ more from the mean, while a lower value of the standard deviation indicates less spread.

Example: imagine a dataset representing students' grades in a maths exam: 80, 85, 90, 75, 95. The formula that will be presented below is only valid if the five values we started with

form the entire population. First, you calculate the average (mean), which is 85. Then you calculate the variation margin, which is 50.

First, calculate the deviations of each data point from the mean, and square the result of each:

$(80-85)^2 = (-5)^2 = 25, \quad (85-85)^2 = (0)^2 = 0, \quad (90-85)^2 = (5)^2 = 25, \quad (75-85)^2 = (-10)^2 = 100, (95-85)^2 = (10)^2 = 100$

The variance is the mean of these values:

$$\sigma^2 = \frac{25+0+25+100+100}{5} = \frac{250}{5} = 50$$

Finally, you calculate the standard deviation by taking the square root of the variation margin:

Standard deviation = $\sqrt{50} \approx 7.07$

So the standard deviation in this case is about 7.07. This means that on average, students' scores are about 7.07 units away from the mean. The standard deviation is often used in analysing the distribution of data and in assessing the variability of values in a set.

Quantiles

Quantiles are values that divide ordered data into specific parts. For example, quartiles divide data into four equal parts. The first quartile (Q1) divides the bottom 25% of the data, the second quartile (Q2) is equal to the median, and the third quartile (Q3) divides the top 25% of the data.

Example: in the dataset 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, the first quartile (Q1) is equal to 6, the second quartile (Q2) is equal to 11, and the third quartile (Q3) is equal to 16.

1.4 Displaying statistics

The presentation of statistics involves the use of a variety of methods and tools, with the aim of presenting data in a clear, transparent and informative way.

Here are some common ways to display statistics:

Tables

Tables are the basic method for displaying data. Examples include frequency tables, which show the number of occurrences for different values, and data tables, which show more information about the data.

Marks Scored by Students	Tally Marks	Frequency
41 - 49	𝍸	3
50 - 58	𝍸 |	6
59 - 67	𝍸	5
68 - 76	𝍸 |	6
77 - 85	||	2
		Total =22

Figure 1.1 Example of a table.

Graphical representations

Graphical displays are an effective tool for visualising data. They include different types of charts such as bar charts, line charts, pie charts, histograms, box plots, etc.

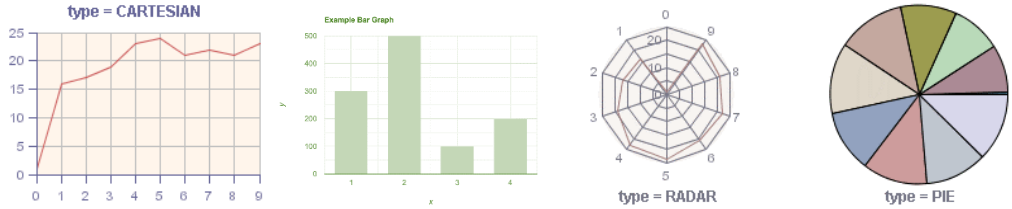

Figure 1.2 Examples of graphical representations of data.

Line charts are used to visualize trends and changes over time, making them ideal for tracking data that evolves continuously. They are particularly effective for showing relationships between variables and highlighting patterns, such as increases, decreases, or fluctuations. Line charts are commonly used in fields like finance, science, and business to analyze time-series data, compare trends across categories, or forecast future developments based on historical data.

Bar charts are used to compare quantities across different categories, making them ideal for presenting discrete data. They are particularly effective for highlighting differences, similarities, and trends between groups. Bar charts are commonly used when you need to show frequencies, percentages, or other numerical measures in a clear and visually

straightforward way. They are widely applied in business, education, and research to analyze and communicate categorical data.

Radar charts, also known as spider charts, are used to display multivariate data across multiple dimensions in a circular format. They are ideal for comparing several variables or entities against the same criteria, highlighting strengths and weaknesses in a clear, visual way. Radar charts are often used in performance analysis, decision-making, and competitive comparisons, such as evaluating product features, team skills, or survey results across different categories.

Pie charts are used to represent proportions or percentages of a whole, making them ideal for visualizing the relative sizes of different categories. They are especially effective when you want to show how parts contribute to a total or to compare proportions at a glance. Pie charts are commonly used in reports, presentations, and surveys to display data like market share, budget allocation, or demographic distribution.

Histograms

Histograms are graphical representations of the distribution of data. They are used to show the frequency of the value of a variable at different intervals.

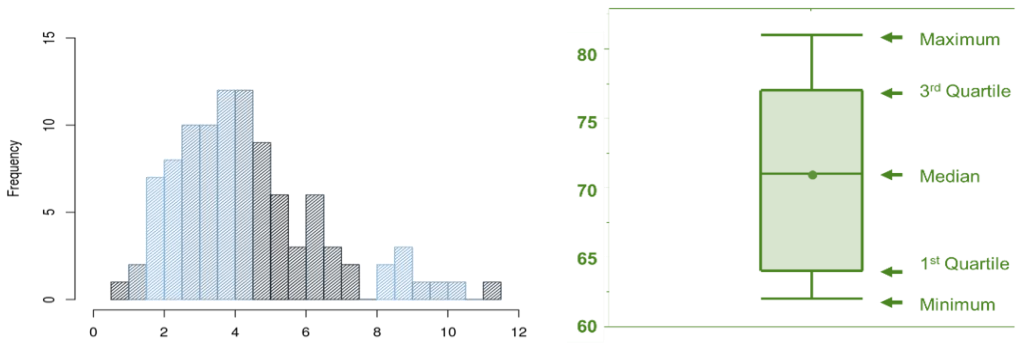

Figure 1.3 Histogram and and Quantile chart (Box plot).

Quantile chart (Box plot)

A quantile plot, or moustache box, is a type of graph used in descriptive statistics as a convenient way of graphically representing groups of numerical data by summarising them with five numbers: minimum, first quartile, median, third quartile and maximum.

The choice of method for displaying statistics depends on the nature of the data, the objectives of the analysis and the target audience. It is important to choose the method that best suits your message and makes the data easier to understand.

1.5 Frequency distribution

A frequency distribution, also known as a frequency table or histogram, is a way of showing the number of occurrences of different values of a variable in a data set. Using a frequency distribution, you can identify patterns, distributions and frequencies of values in the data. It is commonly used for the analysis of qualitative (categorical) variables but can also be used to display discrete values of quantitative (numerical) variables.

The process of creating a frequency distribution involves the following steps:

- Data collection: first, collect the data for which you want to create a frequency distribution.
- Identify different values: identify different values that appear in your data. These are categories or discrete values that you want to analyse.
- Counting occurrences: count how many times each value appears in the dataset.
- Create a frequency table: create a table showing all the different values of the variable and the number of occurrences for each value.
- Drawing a histogram: if you have a large number of different values, you can create a histogram showing the frequency distribution. This is a graphical representation that shows the number of occurrences for each value in the form of bars.

Example of a frequency distribution: Imagine we are analysing the frequency distribution of Marks scored by students. We have collected data of 22 students and we want to see how many student scored a certain number of points.

Marks Scored by Students	Tally Marks	Frequency
41 - 49	𝍸	3
50 - 58	𝍸 |	6
59 - 67	𝍸	5
68 - 76	𝍸 |	6
77 - 85	||	2
		Total =22

Figure 1.4 Frequency distribution table.

A frequency distribution graph (histogram) would show bars for each mark range with the height representing the number of students in every frequency class. This way we can clearly see which frequency class is the most common and how the other marks in the dataset are distributed. Frequency distributions are a useful tool for visualising and analysing qualitative data and for quickly identifying patterns.

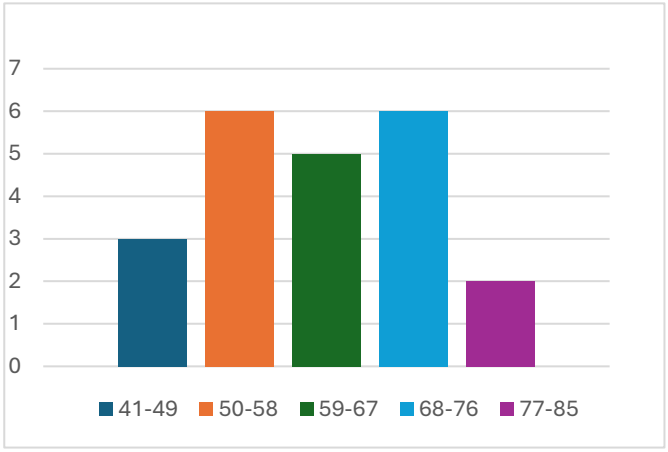

Figure 1.5 Frequency distribution graph.

1.6 Descriptive and inferential statistics

Descriptive statistics: descriptive statistics is concerned with describing and summarising data from the sample or population being studied. It is used to analyse and understand the data, but not to draw conclusions about the population as a whole. The main aim of descriptive statistics is to describe the characteristics of data, for example to calculate the mean, median, range, standard deviation and to create graphical representations such as histograms or graphs. It is used to create summaries and graphs that help to visualise data.

Inferential statistics: inferential statistics deals with making inferences about a population from a sample. This means that inferential statistics allows conclusions to be drawn about the population as a whole from an analysis of a sample. It uses different statistical methods such as hypothesis testing, confidence intervals and regression analysis to understand whether observed sample results can be generalised to a population. For example, if we want to find out whether the mean age in a sample is representative of the population as a whole, we will use inferential statistics.

Inferential statistics

Inferential statistics is the branch of statistics that focuses on the inferences and conclusions we can draw from the data we collect. Its main task is to draw general conclusions about a population or sample from the analysis of a sample of data.

The main objectives of inferential statistics are:

Estimating population parameters: inferential statistics allows us to estimate population parameters such as mean, variance, proportions and other characteristics from a sample.

Hypothesis testing: inferential statistics can be used to test hypotheses about a population based on sampled data. This involves statistical testing, where we compare the sample with assumptions about the population.

Creating confidence intervals: inferential statistics allows us to calculate intervals containing the estimated values of population parameters with a certain level of confidence.

Example of inferential statistics: suppose we want to estimate the average height of all students at a university. Since it is impossible to check all students, we take a sample of 100 students and measure their height.

We then use inferential statistics to calculate a confidence interval for the average height of all students. Our sample has a mean height of 170 cm and a standard deviation of 5 cm.

Assuming that the heights of the students in the population are **approximately normally distributed,** we can use the standard error of the mean to calculate the confidence interval. For example, if we want a 95% confidence interval, we use the standard error and the quantiles of a normal distribution.

An approximate 95% confidence interval for the average height of all students at the university would be:

$$170\ cm\ \pm\ 1.96\ \times\ (\frac{5\ cm}{\sqrt{100}})\ =\ 170\ cm\ \pm\ 0.98\ cm$$

This means that we can say with 95% confidence that the average height of all students is between approximately 169.02 cm and 170.98 cm. This confidence interval allows us to infer the average height of all students at the university from the overall sample.

Together, these statistical methods allow logistics companies to better understand their processes, predict future events and make more informed decisions to improve efficiency and competitiveness.

1.7 Correlation and regression

They are statistical methods used to study relationships between variables and to predict values. Both methods help to understand how one variable affects another and how well one variable can be used to predict another. Here is an explanation of each of these two methods:

Correlation

Correlation is used to measure the degree of association between two quantitative (numerical) variables. It tells whether there is a linear relationship between the two variables and how strong that relationship is. Correlation is measured by the correlation coefficient, which takes the form of **a value between -1** and **1.**

A correlation coefficient of 1 means a perfect positive correlation, which means that the variables are perfectly correlated and moving in the same direction.

A correlation coefficient of -1 means a perfect negative correlation, which means that the two variables are completely inversely correlated and move in opposite directions.

A correlation coefficient of 0 means that there is no linear relationship between the variables.

Example: the correlation between the number of hours of study and the grades students achieve will be positive if an increase in the number of hours of study usually corresponds to higher grades.

Regression

Regression is used to model and predict the value of one quantitative variable (the dependent variable) from the value of another quantitative variable (the independent variable). There are different types of regression, including **simple linear regression, multiple linear regression**, logistic regression, etc.

Simple linear regression: used to model the relationship between one independent variable and one dependent variable. The model is linear and is usually represented by the equation of a straight line ($y = a + bx$), where a is the intercept with the y –axis and b is the slope of the line.

Multiple linear regression: used when you want to model the relationship between several independent variables and one dependent variable.

Example: a simple linear regression can be used to model the relationship between the number of learning tasks completed (independent variable) and the final exam grade (dependent variable).

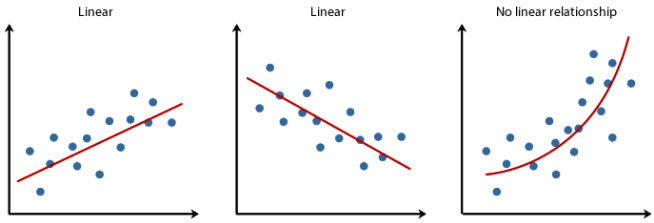

Figure 1.6 Simple linear regression graphs.

1.8 Probability distributions

In statistics, a probability distribution describes the probabilities of different values that a variable can take. It is a mathematical model that helps us to understand and analyse random phenomena and to predict how values will be distributed under certain circumstances. There are several different probability distributions, each with its own characteristics and applications in different situations. Here are some of the most well-known probability distributions in statistics:

Normal (Gaussian) distribution: the normal distribution is one of the most important and widely used distributions. It describes a symmetrical and bell-shaped distribution with known parameters: the mean (μ) and the standard deviation (σ). Many natural phenomena approximate to the normal distribution.

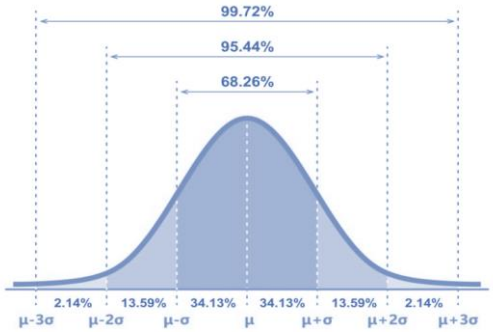

Figure 1.8 Normal distribution graph.

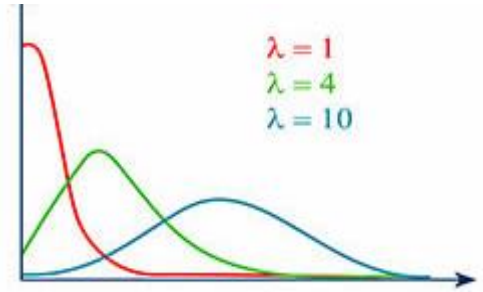

Figure 1.7 Poisson distribution graph.

Binomial distribution: the binomial distribution is used to model the number of successes (e.g. the number of "heads") in a given number of independent Bernoulli trials. It has two parameters: the number of trials (n) and the probability of success (p).

Poisson distribution: the Poisson distribution is used to model the number of events that occur over a period of time or space. It is typically used to model rare events such as accidents, calls to emergency services, etc. The parameter of the distribution is the average rate (λ).

Exponential distribution: the exponential distribution is a special case of the gamma distribution and is used to model the times to the first event in a Poisson process. The parameter of the distribution is the average event rate (λ).

Student's t-distribution: the Student's t-distribution is used to estimate confidence intervals and test hypotheses when you have a small sample size and don't know the population standard deviation. It is important when analysing samples where the assumption of a normal distribution may be fragile.

Chi-square distribution: the Chi-square distribution is used to analyse the frequency distribution in the tables, to test for independence and to test hypotheses. It is often used in statistical tests such as the chi-square test.

F-distribution: the F-distribution is used when comparing the variability between two samples. It is used in analysis of variance (ANOVA) and other statistical tests.

These probability distributions are fundamental building blocks in statistics and are used to model and analyse different types of data in different contexts. Choosing the correct probability distribution is crucial when carrying out statistical analyses and predicting results.

References Chapter 1

- *Introductory Statistics.* Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010l

- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);

- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);

- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011

- Journal of the Royal Statistical Society 2024**,** A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.

- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014

- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®

- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved

Additional links to literature and Youtube videos Chapter 1

- https://open.umn.edu/opentextbooks/textbooks/196
- https://www.scribbr.com/category/statistics/
- https://stats.libretexts.org/Bookshelves/Introductory_Statistics
- https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
- https://saylordotorg.github.io/text_introductory-statistics/

- https://drive.uqu.edu.sa/_/mskhayat/files/MySubjects/20178FS%20Elementary%20Statistics/Introductory%20Statistics%20(7th%20Ed).pdf
- https://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf
- https://www.geeksforgeeks.org/introduction-of-statistics-and-its-types/
- https://onlinestatbook.com/Online_Statistics_Education.pdf
- https://www.researchgate.net/profile/Tareq-Alodat-2/publication/340511098_INTRODUCTION_TO_STATISTICS_MADE_EASY/links/5e8de3dc4585150839c7b58a/INTRODUCTION-TO-STATISTICS-MADE-EASY.pdf
- https://byjus.com/maths/statistics/
- https://www.khanacademy.org/math/statistics-probability
- https://www.youtube.com/watch?v=XZo4xyJXCak
- https://www.youtube.com/watch?v=LMSyiAJm99g
- https://www.youtube.com/watch?v=VPZD_aij8H0
- https://www.youtube.com/watch?v=TLwp5DwcqD4
- https://www.youtube.com/watch?v=fpFj1Re1l84
- https://youtube.com/playlist?list=PLqzoL9-eJTNAB5st3mtP_bmXafGSH1Dtz&si=z-IXQ1iKbw2-ieJW
- https://www.youtube.com/watch?v=44MJyNTxaP8

2. Statistics for Business Analytics

Welcome to the world of business statistics, where data transforms into meaningful insights, guiding decision-making and uncovering hidden truths. In this comprehensive exploration, we embark on a journey to demystify essential statistical concepts and techniques that underpin rigorous business data analysis. From understanding the intricacies of distributions to applying hypothesis testing and constructing confidence intervals, each chapter unfolds a new facet of statistical literacy.

At the heart of statistical analysis lies the normal distribution, a bell-shaped curve that permeates countless phenomena in nature and human behavior. In this part, we delve into the essence of the normal distribution, unraveling its properties and significance in statistical inference. Through visualizations and real-world examples, we illuminate the ubiquity of this fundamental distribution and its role as a cornerstone of statistical theory.

Standard deviation serves as a compass in the statistical landscape, guiding us through the variability inherent in datasets. In this chapter, we dissect the concept of standard deviation, unveiling its importance in quantifying dispersion and assessing the spread of data points. Armed with a deeper understanding of standard deviations, you will navigate data with confidence, discerning patterns and outliers with precision.

Variables form the building blocks of statistical analysis, each possessing distinct characteristics and implications. This chapter elucidates the dichotomy between continuous and discrete variables, shedding light on their respective roles in data modeling and interpretation. By grasping the nuances of variable types, you will harness the full potential of statistical techniques tailored to diverse data structures.

Sampling distribution serves as the bedrock of statistical inference, bridging the gap between sample observations and population parameters. In this chapter, we unravel the concept of sampling distribution, elucidating its relevance in making probabilistic statements about population characteristics. Through concrete examples, you will develop an intuitive understanding of sampling distribution's role in robust statistical analysis.

The Central Limit Theorem is a key concept in statistics that helps us make sense of uncertainty. This chapter explains the Central Limit Theorem in simple terms, showing how it

makes sample means more predictable and aids in hypothesis testing. By understanding this concept, you'll be able to draw meaningful conclusions from data.

Understanding hypothesis testing is essential for making data-driven decisions. It allows us to determine whether observed patterns in data are meaningful or simply due to chance. By applying hypothesis testing, we can evaluate assumptions, compare groups, and assess the statistical significance of results, making it a vital tool in scientific research, business analysis, and many other fields.

Z-scores and Z-tables serve as navigational aids in the sea of standard normal distribution, facilitating standardized comparisons and probability calculations. This chapter elucidates the intricacies of Z-scores, empowering you to interpret standardized scores and harness Z-tables for statistical analysis. With proficiency in Z-scores, you will navigate the vast expanse of normal distribution with confidence and precision.

In situations where sample sizes are small or population standard deviations are unknown, t-scores and t-tables emerge as indispensable tools for statistical analysis. This chapter unravels the mysteries of t-scores, guiding you through their calculation and interpretation using t-tables. Armed with this knowledge, you will navigate the nuances of t-distributions with finesse, ensuring robust inference in diverse statistical scenarios.

Normal and t-distributions stand as pillars of probability theory, each possessing unique characteristics and applications. In this chapter, we elucidate the distinctions between these distributions, empowering you to discern when to employ each in statistical analysis. Through practical examples and comparative analyses, you will develop a nuanced understanding of normal and t-distributions, enriching your statistical toolkit.

Confidence intervals provide a window into the uncertainty surrounding population parameters, empowering us to quantify the precision of our estimates. In this chapter, we explore the construction of confidence intervals for means and proportions, unraveling the methodology and interpretation behind these essential statistical tools. By mastering confidence intervals, you will convey the uncertainty inherent in your findings with transparency and rigor.

While p-values offer a gateway to statistical inference, their misinterpretation can lead to erroneous conclusions and misinformed decisions. This chapter examines the potential pitfalls

of overreliance on p-values, highlighting the importance of context and effect size in statistical analysis. Through critical examination and practical insights, you will navigate the complexities of p-values with vigilance, ensuring the integrity of your statistical conclusions.

Within these pages lie the keys to unlocking the mysteries of statistical analysis, empowering you to navigate the complexities of data with confidence and precision. As we embark on this journey together, let curiosity be our compass and inquiry our guiding light, illuminating the path towards deeper understanding and actionable insights.

2.1 Normal distribution

At the heart of statistical analysis lies the normal distribution, a ubiquitous probability distribution that serves as a benchmark for many statistical techniques. We will delve into its characteristics, its symmetrical bell-shaped curve, and its significance in understanding the distribution of data.

The normal distribution finds applications across various fields, including finance, psychology, engineering, and biology. From modelling stock prices to understanding human height distributions, the normal distribution serves as a versatile tool for analyzing and interpreting data.

Throughout this chapter, we will delve into the mathematical properties of the normal distribution, exploring how to calculate probabilities, percentiles, and z-scores. Moreover, we will discuss practical techniques for visualizing and interpreting normal distributions using histograms, density plots, and cumulative distribution functions.

By the end of this chapter, you will have a deep appreciation for the normal distribution and its significance in statistical analysis. Armed with this knowledge, you will be well-equipped to tackle more advanced statistical concepts and apply them to real-world datasets. Let's embark on this journey to unravel the mysteries of the normal distribution together.

A normal distribution, also known as a Gaussian distribution or bell curve, exhibits symmetrical data distribution without skewness. When graphed, the data forms a bell-shaped curve, with the majority of values congregating around the centre and decreasing as they move away from it.

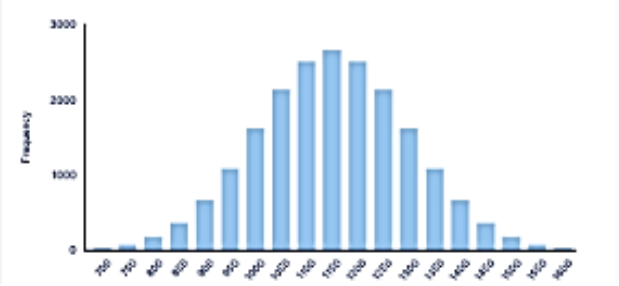

Figure 2.1 Example of a Gaussian distribution or bell curve.

Various variables in both natural and social sciences typically exhibit a normal distribution or an approximation thereof. Examples include height, birth weight, reading ability, job satisfaction, and SAT scores. Due to the prevalence of normally distributed variables, numerous statistical tests are tailored for such populations. Proficiency in comprehending the characteristics of normal distributions empowers individuals to employ inferential statistics for comparing groups and generating population estimates from samples.

Normal distributions have key characteristics that are easy to spot in graphs:

- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

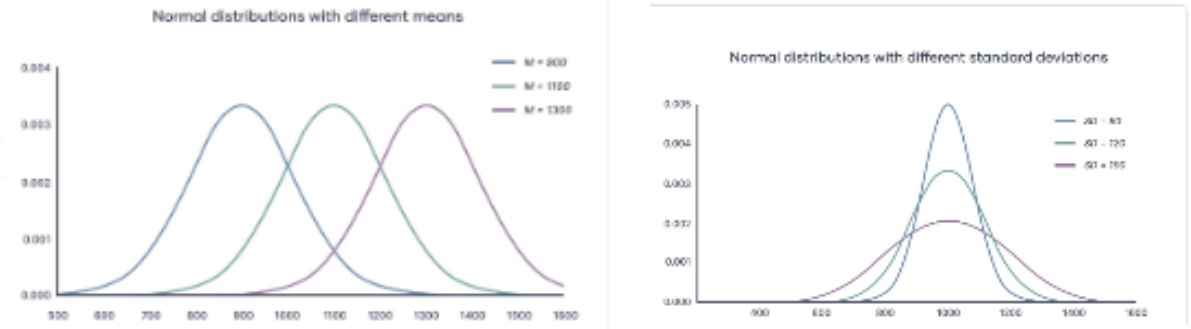

Figure 2.2 Normal distribution with different mean and with different stand deviations.

The mean serves as the location parameter, dictating the centre of the curve's peak. Adjusting the mean shifts the curve accordingly: increasing it shifts the curve to the right, while decreasing it shifts the curve to the left. Meanwhile, the standard deviation functions as the scale parameter, influencing the spread or width of the curve.

The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.

2.2 Empirical rule

The empirical rule, also known as the 68-95-99.7 rule, provides insight into the distribution of values within a normal distribution:

- Approximately 68% of values fall within 1 standard deviation from the mean.
- Roughly 95% of values lie within 2 standard deviations from the mean.
- About 99.7% of values are encompassed within 3 standard deviations from the mean.

For instance, consider a scenario where SAT scores from students in a new test preparation course are gathered, and the data conforms to a normal distribution with a mean score (M) of 1150 and a standard deviation (SD) of 150.

Applying the empirical rule yields the following insights:

- Around 68% of scores fall within the range of 1000 to 1300, corresponding to 1 standard deviation above and below the mean.
- Approximately 95% of scores are within the range of 850 to 1450, representing 2 standard deviations above and below the mean.
- Nearly all scores, around 99.7%, lie within the range of 700 to 1600, encompassing 3 standard deviations above and below the mean.

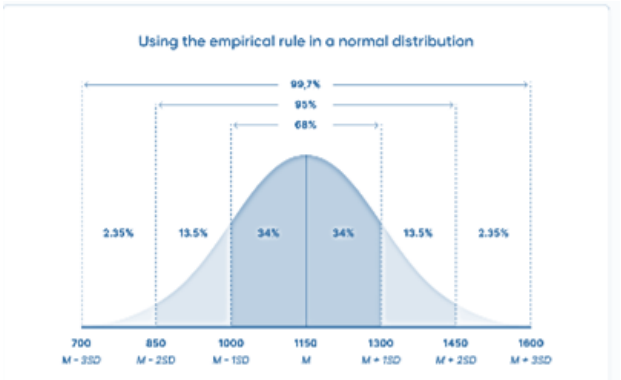

Figure 2.3 Empirical rule in normal distribution.

The empirical rule offers a rapid method to assess data, enabling detection of outliers or exceptional values that deviate from its expected pattern. In cases where data from small samples diverge significantly from this pattern, alternative distributions such as the t-distribution might be more suitable. Identifying the distribution of the variable allows for the application of relevant statistical tests.

2.3 Formula of the normal curve

To construct a normal curve based on a given mean and standard deviation, one can employ a probability density function, thereby accurately representing the distribution of the data.

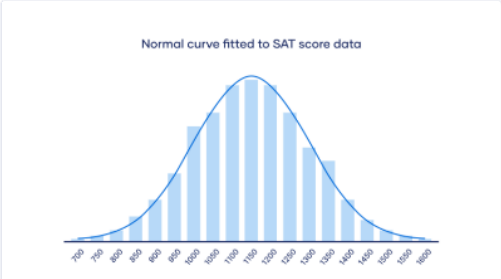

Figure 2.4 Normal curve fitted to SAT score data.

Within a probability density function, the area beneath the curve represents probability. Given that the normal distribution serves as a probability distribution, the cumulative area under the curve invariably sums up to 1 or 100%. Although the formula for the normal probability density function may appear intricate, utilizing it merely necessitates knowledge of the population mean and standard deviation. By substituting these parameters into the formula, one can determine the probability density associated with any given value of x.

- $f(x)$ = probability
- x = value of the variable
- μ = mean
- σ = standard deviation
- σ^2 = variance

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

Example:

Using the probability density function, you want to know the probability that SAT scores in your sample exceed 1380.

On your graph of the probability density function, the probability is the shaded area under the curve that lies to the right of where your SAT scores equal 1380.

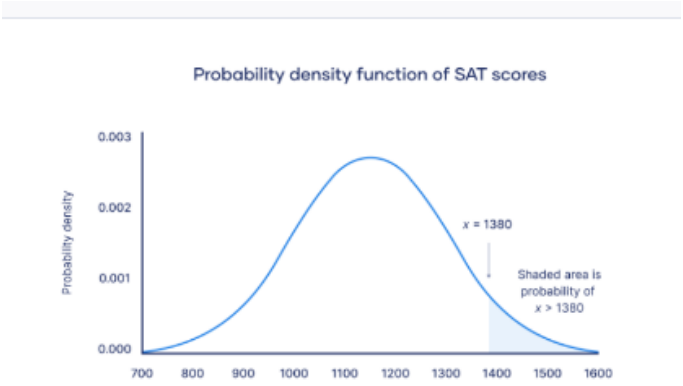

Figure 2.5 Probability density function of SAT scores graph.

You can find the probability value of this score using the standard normal distribution.

2.4 Standard normal distribution

The standard normal distribution, known as the **z-distribution**, is distinct for having a mean of 0 and a standard deviation of 1. Any normal distribution can be seen as a transformation of the standard normal distribution, undergoing adjustments in scale, position, or both.

In the context of the z-distribution, individual observations, which are typically denoted as x in normal distributions, are referred to as z-scores. These z-scores represent the number of standard deviations that each value deviates from the mean. Consequently, converting values from any normal distribution into z-scores facilitates comparison and analysis within the framework of the standard normal distribution.

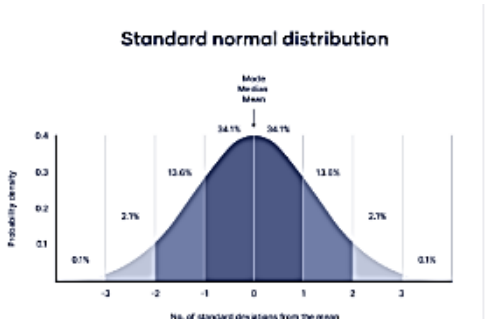

Figure 2.6 Standard normal distribution graph.

You only need to know the mean and standard deviation of your distribution to find the *z*-score of a value.

Z-score Formula Explanation

$$z = \frac{x - \mu}{\sigma}$$

- x = individual value
- μ = mean
- σ = standard deviation

We convert normal distributions into the standard normal distribution for several reasons:

- To find the probability of observations in a distribution falling above or below a given value.
- To find the probability that a sample mean significantly differs from a known population mean.
- To compare scores on different distributions with different means and standard deviations.

2.5 Finding probability using the *z*-distribution

Each z-score corresponds to a probability, often referred to as a p-value, indicating the likelihood of observing values below that specific z-score. By transforming an individual value into a z-score, one can determine the probability of all values up to that point occurring within a normal distribution.

For instance, consider a scenario where you wish to ascertain the probability of SAT scores in your sample surpassing 1380. Initially, you calculate the z-score using the mean and standard deviation of the distribution. With a mean of 1150 and a standard deviation of 150, the z-score reveals the number of standard deviations by which 1380 deviates from the mean.

Formula **Calculation**

$$z = \frac{x - \mu}{\sigma} = \frac{1380 - 1150}{150} = 1.53$$

For a *z*-score of 1.53, the *p*-value is 0.937. This is the probability of SAT scores being 1380 or less (93.7%), and it's the area under the curve left of the shaded area.

To find the shaded area, you take away 0.937 from 1, which is the total area under the curve.

Probability of x > 1380 = 1 – 0.937 = **0.063**

That means it is likely that only 6.3% of SAT scores in your sample exceed 1380.

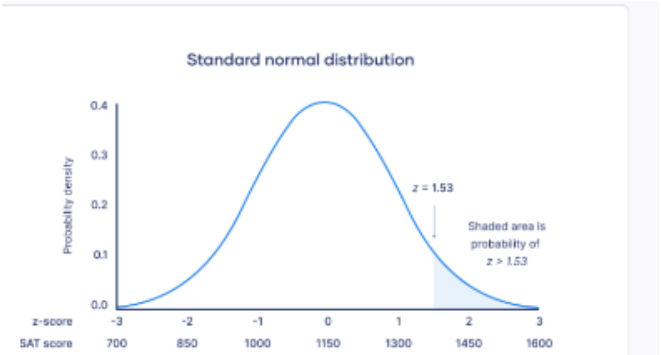

Figure 2.7 Standard normal distribution with SAT score indicated.

2.6 Sampling Distribution

Sampling distributions form the backbone of statistical inference, enabling us to draw conclusions about populations based on sample data. We will delve into the intricacies of sampling distributions, understanding how they reflect the variability of sample statistics and their pivotal role in hypothesis testing.

Sampling distribution refers to the distribution of a statistic, such as the sample mean or sample proportion, obtained from multiple samples of the same size drawn from a population. It provides insights into the behavior of sample statistics and their variability across different samples.

2.7 Central Limit Theorem and Sampling Distribution

The Central Limit Theorem (CLT) is a fundamental concept in statistics that underpins the behavior of sampling distributions. It states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution. This theorem enables us to make robust inferences about population parameters from sample data.

The central limit theorem serves as the cornerstone of understanding normal distributions in statistics. In research settings, obtaining an accurate estimation of a population mean often involves gathering data from numerous random samples within the population. These individual sample means collectively form what is known as a sampling distribution of the mean.

The central limit theorem delineates two key principles:

1. **Law of Large Numbers:** As the sample size or the number of samples increases, the sample mean tends to converge towards the population mean.
2. **Normality of Sampling Distribution:** Despite the original variable's distribution, when working with multiple large samples, the sampling distribution of the mean tends to approximate a normal distribution.

Parametric statistical tests conventionally assume that samples are derived from normally distributed populations. However, the central limit theorem obviates the necessity of this assumption for sufficiently large sample sizes. With large samples, parametric tests can be applied irrespective of the population's distribution, provided other pertinent assumptions are satisfied. A sample size of 30 or more is commonly deemed as sufficiently large.

Conversely, for small samples, ensuring the assumption of normality is crucial due to the uncertainty surrounding the sampling distribution of the mean. Accurate results necessitate confirmation that the population adheres to a normal distribution before employing parametric tests with small sample sizes.

Illustratively, the central limit theorem posits that by obtaining sufficiently large samples from a population, the means of these samples will exhibit a normal distribution, even if the underlying population distribution diverges from normality.

Example: Consider a population following a Poisson distribution (depicted in the left image). Upon drawing 10,000 samples from this population, each consisting of 50 observations, the distribution of sample means aligns closely with a normal distribution, in accordance with the central limit theorem (as illustrated in the right image).

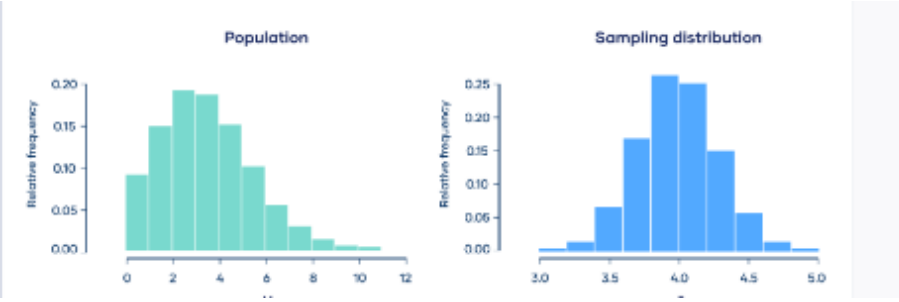

Figure 2.8 Example of population in Poisson distribution and normal distribution.

The central limit theorem hinges upon the notion of a sampling distribution, which represents the probability distribution of a statistic computed from numerous samples drawn from a population.

Conceptualizing an experiment can aid in grasping sampling distributions:

- Let's envision drawing a random sample from a population and computing a statistic, such as the mean.
- Subsequently, another random sample of identical size is drawn, and the mean is recalculated.
- This process is iterated numerous times, resulting in a plethora of means, each corresponding to a sample.

The aggregation of these sample means exemplifies a sampling distribution. According to the central limit theorem, the sampling distribution of the mean tends towards a normal distribution when the sample size is sufficiently large. Remarkably, irrespective of the population's distribution—be it normal, Poisson, binomial, or otherwise—the sampling distribution of the mean exhibit's normality.

Fortunately, one doesn't need to repeatedly sample a population to discern the shape of the sampling distribution. Instead, the parameters of the sampling distribution of the mean are contingent upon the parameters of the population itself.

- The mean of the sampling distribution is the mean of the population.

  $\mu_{\bar{x}} = \mu$

- The standard deviation of the sampling distribution is the standard deviation of the population divided by the square root of the sample size.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

We can describe the sampling distribution of the mean using this notation:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Where:

- $\bar{X}$ is the sampling distribution of the sample means
- $\sim$ means "follows the distribution"
- N is the normal distribution
- μ is the mean of the population
- σ is the standard deviation of the population
- n is the sample size

Sample size, denoted as n, represents the number of observations drawn from the population for each sample, maintaining uniformity across all samples. The sample size significantly influences the sampling distribution of the mean in two key aspects.

1. Sample Size and Normality:

- Larger sample sizes tend to yield sampling distributions that closely adhere to a normal distribution.
- Conversely, with small sample sizes, the sampling distribution of the mean may deviate from normality. This divergence arises because the central limit theorem's validity hinges on having a "sufficiently large" sample size.
- Conventionally, a sample size of 30 or more is considered "sufficiently large."
- When n<30, the central limit theorem doesn't apply, and the sampling distribution mirrors the population distribution. Hence, the sampling distribution is only normal if the population distribution is normal.
- Conversely, when n≥30, the central limit theorem holds true, and the sampling distribution approximates a normal distribution.

2. Sample Size and Standard Deviations:

- The sample size directly impacts the standard deviation of the sampling distribution, reflecting the variability or spread of the distribution.
- With smaller sample sizes, the standard deviation is typically higher, indicating greater variability among sample means due to their imprecise estimation of the population mean.
- Conversely, larger sample sizes correspond to lower standard deviations, indicating less variability among sample means owing to their more accurate estimation of the population mean.

Importance of the Central Limit Theorem:

Parametric tests such as t-tests, ANOVAs, and linear regression possess greater statistical power compared to most non-parametric tests. This enhanced statistical power stems from assumptions regarding the distribution of populations, which are grounded in the central limit theorem.

Continuous distribution

Let's consider the retirement age of individuals in the United States. The population consists of all retired Americans, and the distribution of this population could be represented as follows:

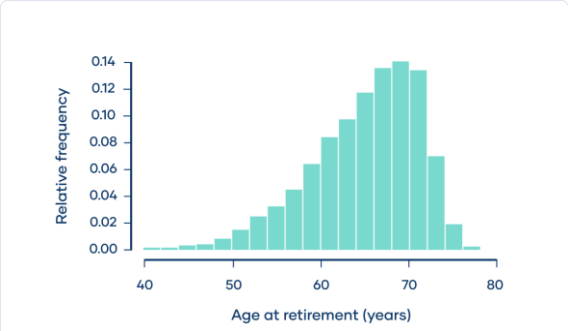

Figure 2.9 Continuous distribution graph.

The distribution of retirement ages skews leftward, with a majority retiring within approximately five years of the mean retirement age of 65 years. However, there exists an extended tail of individuals retiring much earlier, such as at 50 or even 40 years old. The population displays a standard deviation of 6 years.

Imagine conducting a small-scale sampling from this population. Five retirees are randomly selected, and their retirement ages are recorded. For instance: 68, 73, 70, 62, 63

The mean of this sample serves as an approximation of the population mean, albeit with limited precision due to the small sample size of 5. For example: Mean = (68 + 73 + 70 + 62 + 63) / 5 Mean = 67.2 years

Now, suppose this sampling process is repeated 10 times, with each sample comprising five retirees. The mean of each sample is computed, resulting in a distribution known as the sampling distribution of the mean. For instance: 60.8, 57.8, 62.2, 68.6, 67.4, 67.8, 68.3, 65.6, 66.5, 62.1

As this process is repeated numerous times, a histogram depicting the means of these samples will approximate a normal distribution.

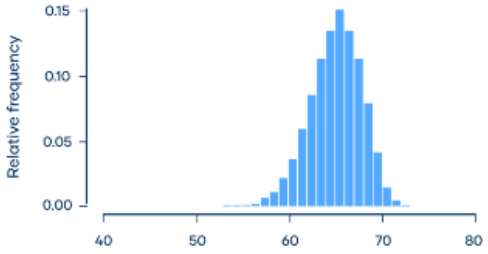

Figure 2.10 Normal distribution of means.

Despite the sampling distribution exhibiting a somewhat more normal shape compared to the population, it still retains a slight leftward skew. Additionally, it's evident that the variability in the sampling distribution is narrower than that of the population.

According to the central limit theorem, the sampling distribution of the mean tends to approximate a normal distribution as the sample size increases. However, the current sampling distribution of the mean deviates from normality due to its relatively small sample size.

2.8 Test statistics

A test statistic represents a numerical value derived from a statistical hypothesis test, indicating the degree of alignment between your observed data and the distribution expected under the null hypothesis of that test.

This statistic plays a crucial role in computing the p-value of your findings, facilitating the determination of whether to accept or reject your null hypothesis.

But what exactly constitutes a test statistic?

A test statistic articulates the similarity between the distribution of your data and the distribution anticipated under the null hypothesis of the statistical test employed. The distribution of data elucidates the frequency of each observation, characterized by its central tendency and the variability around it. Since different statistical tests anticipate different distribution types, selecting the appropriate test aligns with your hypothesis.

The test statistic condenses your observed data into a singular figure, leveraging measures such as central tendency, variability, sample size, and the number of predictor variables in your statistical model.

Typically, the test statistic emerges from discernible patterns in your data (e.g., correlations between variables or discrepancies among groups), divided by the data's variance (i.e., the standard deviation).

Consider this illustration:

You investigate the association between temperature and flowering dates in a specific type of apple tree. Analyzing a comprehensive dataset spanning 25 years, tracking temperature and flowering dates by randomly sampling 100 trees annually from an experimental field.

• Null hypothesis (H_0): No correlation exists between temperature and flowering date.

• Alternate hypothesis (H_A or H_1): A correlation exists between temperature and flowering date.

To scrutinize this hypothesis, you undertake a regression test, yielding a t-value as the test statistic. This t-value juxtaposes the observed correlation between the variables against the null hypothesis of zero correlation.

2.9 Types of test statistics

Outlined below is a synopsis of prevalent test statistics, along with their corresponding hypotheses and the categories of statistical tests in which they are employed. While various statistical tests may employ distinct methodologies for computing these statistics, the fundamental hypotheses and interpretations of the test statistic remain consistent.

Test statistic	Null and alternative hypotheses	Statistical tests that use it
***t* value**	**Null:** The means of two groups are equal **Alternative:** The means of two groups are not equal	• *T* test • Regression tests
***z* value**	**Null:** The means of two groups are equal **Alternative:** The means of two groups are not equal	• *Z* test
***F* value**	**Null:** The variation among two or more groups is greater than or equal to the variation between the groups **Alternative:** The variation among two or more groups is smaller than the variation between the groups	• ANOVA • ANCOVA • MANOVA
***X*²-value**	**Null:** Two samples are independent **Alternative:** Two samples are not independent (i.e., they are correlated)	• Chi-squared test • Non-parametric correlation tests

In real-world scenarios, you'll typically compute your test statistic using a statistical software package such as R, SPSS, or Excel, which will also furnish the p-value associated with the test statistic. Nevertheless, formulas for manual computation of these statistics can be sourced online.

For instance, in testing your hypothesis concerning temperature and flowering dates, you conduct a regression analysis. The regression test yields: • a regression coefficient of 0.36 • a t-value comparing this coefficient to the anticipated range of regression coefficients under the null hypothesis of no relationship.

The resultant t-value from the regression test, 2.36, represents your test statistic.

2.10 Standard Error

The standard error of the mean (SE or SEM) serves as an indicator of the probable disparity between the population mean and a sample mean. It offers insight into the degree of variability one would anticipate in the sample mean if the study were replicated using fresh samples drawn from the same population.

While the standard error of the mean is the most frequently cited form of standard error, similar measures exist for other statistical parameters such as medians or proportions. Standard error functions as a prevalent gauge of sampling error, depicting the disparity between a population parameter and a sample statistic.

To mitigate standard error, increasing the sample size is recommended. Employing a large, randomized sample serves as the most effective strategy for minimizing sampling bias and enhancing the reliability of findings.

Standard error and standard deviation are both measures of variability:

- The **standard deviation** describes variability **within a single sample**.
- The **standard error** estimates the variability **across multiple samples** of a population.

The standard deviation serves as a descriptive statistic derived directly from sample data, whereas the standard error represents an inferential statistic, typically estimated unless the exact population parameter is known.

2.11 Standard error formula

The standard error of the mean is determined by employing the standard deviation alongside the sample size. Through the formula, it becomes apparent that the sample size and the standard error share an inverse relationship. In simpler terms, as the sample size increases, the standard error decreases. This phenomenon occurs because a larger sample tends to yield a sample statistic closer to the population parameter.

Various formulas are employed based on whether the population standard deviation is known. These formulas are applicable to samples comprising more than 20 elements ($n > 20$).

When population parameters are known

When the population standard deviation is known, you can use it in the below formula to calculate standard error precisely.

Formula	Explanation
$SE = \frac{\sigma}{\sqrt{n}}$	• SE is standard error • σ is population standard deviation • n is the number of elements in the sample

When population parameters are unknown

When the population standard deviation is unknown, you can use the below formula to only estimate standard error. This formula takes the sample standard deviation as a point estimate for the population standard deviation.

Formula	Explanation
$SE = \frac{s}{\sqrt{n}}$	• SE is standard error • s is sample standard deviation • n is the number of elements in the sample

Example: Using the standard error formula to estimate the standard error for math SAT scores. Follow next two steps.

First, find the square root of your sample size (n).

Formula	Calculation
$n = 200$	$\sqrt{n} = \sqrt{200} = 14.1$

Next, divide the sample standard deviation by the number you found in step one.

Formula	Calculation		
$SE = \frac{s}{\sqrt{n}}$	$s = 180$	$\sqrt{n} = 14.1$	$\frac{s}{\sqrt{n}} = \frac{180}{14.1} = 12.8$

The standard error of math SAT scores is 12.8.

You can present the standard error alongside the mean or incorporate it into a confidence interval to convey the uncertainty surrounding the mean.

For instance: Example: Presenting the mean and standard error The mean math SAT score for a random sample of test takers is 550 ± 12.8 (SE).

Reporting the standard error within a confidence interval is preferable as it eliminates the need for readers to perform additional calculations to derive a meaningful range.

A confidence interval denotes a span of values where an unknown population parameter is anticipated to lie most frequently if the study were to be replicated with new random samples.

At a 95% confidence level, it's expected that 95% of all sample means will fall within a confidence interval encompassing ± 1.96 standard errors of the sample mean. This interval serves as an estimate within which the true population parameter is believed to lie with 95% confidence.

For example: Example: Constructing a 95% confidence interval You construct a 95% confidence interval (CI) to estimate the population mean math SAT score. Given a normally distributed characteristic like SAT scores, roughly 95% of all sample means fall within approximately 4 standard errors of the sample mean.

Confidence interval formula

$CI = \bar{x} \pm (1.96 \times SE)$

$\bar{x}$ =sample mean = 550

SE = standard error = 12.8

Lower limit	**Upper limit**
$\bar{x} - (1.96 \times SE)$	$\bar{x} + (1.96 \times SE)$
$550 - (1.96 \times 12.8) =$ **525**	$550 + (1.96 \times 12.8) =$ **575**

With random sampling, a 95% CI [525 575] tells you that there is a 0.95 probability that the population mean math SAT score is between 525 and 575.

References Chapter 2

- *Introductory Statistics.* Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010l

- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);

- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);

- Journal of the Royal Statistical Society 2024**,** A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.

- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011

- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014

- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®

- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved

Additional links to literature and Youtube videos Chapter 2

- https://open.umn.edu/opentextbooks/textbooks/196
- https://www.scribbr.com/category/statistics/
- https://stats.libretexts.org/Bookshelves/Introductory_Statistics
- https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
- https://saylordotorg.github.io/text_introductory-statistics/
- https://drive.uqu.edu.sa/_/mskhayat/files/MySubjects/20178FS%20Elementary%20Statistics/Introductory%20Statistics%20(7th%20Ed).pdf
- https://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf
- https://www.geeksforgeeks.org/introduction-of-statistics-and-its-types/
- https://onlinestatbook.com/Online_Statistics_Education.pdf

- https://www.researchgate.net/profile/Tareq-Alodat-2/publication/340511098_INTRODUCTION_TO_STATISTICS_MADE_EASY/links/5e8de3dc4585150839c7b58a/INTRODUCTION-TO-STATISTICS-MADE-EASY.pdf
- https://byjus.com/maths/statistics/
- https://www.khanacademy.org/math/statistics-probability

3. Data Management

In B2B electronic data interchange (EDI) messages, comprising product or service codes, transport unit identifications, as well as legal documents are exchanged among partners in a supply chain. They usually have the form of alphanumeric strings.

3.1 Information-Data-Knowledge

In computer-based information systems the representation of information varies depending on its purpose and use. It may be alphanumeric or binary considering the fact whether it represents a text, figure, sound or executable program… To be able to store, retrieve, process and transmit data of various types (e.g., numbers, characters, dates, currencies, etc.) they need to be properly encoded. Alphanumeric formats of data representation have their origins in the ASCII (ask-key) alphabet, having evolved in the sense of national character sets (e.g., standards 8859-1, Latin 1 and 8859-2, Latin 2) and finally progressed into international UTF-8 and UTF-16 formats. Thus, they enable common data interpretation by business partners belonging to different ethnic groups and geographic environments. While alphanumeric strings depend on their encoding, numeric data mainly differ in their size and/or precision.

The process of transforming data from their original analogue into the digital form is popularly termed digitization. Appropriately designed utility and application programs accepting data from various sources (e.g., optical scanners, electrical sensors, EDI, etc.) enable organizations to automate their data acquisition, storage, processing and transmission within and between their computer-based information systems.

When collected, data of various types may be joined and organized into tables of data, data bases, data warehouses (chronological order) or knowledge bases (conceptual order). Higher levels of data organization allow for automated classification, reasoning and representation of the hereby accumulated knowledge for analytic purposes.

3.2 Logistic Data

In logistics EDI is used to transmit transaction data between business partners. Since they may use different languages and applications, their conversion to a common format (e.g., XML, JSON) is necessary to be interpretable by different partner's information systems (W3schools, 2023). For quick identification and manipulation barcodes and RFID tags were devised.

To allow for international cooperation, globally accepted data formats needed to be defined for logistics purposes. Logistics data formats correspond with service and product labels, transport unit identifications and transaction codes, usually having the form of time-stamped alphanumeric strings. For simplicity of manipulation and processing speed these codes were standardized and encoded as optically readable bar-codes or electromagnetically readable radio-frequency identification (RFID) codes.

Bar-code (EAN/UCC) is a multi-sector and international form of numbering items (POS EAN-8 and EAN-13, changeable EAN-128, data bar, packaging ITF-14, QR, data matrix, etc.). They are used to identify products, product batches or shipments (1D codes) as well as services (2D codes). Among 2D bar codes the QR code is the most established one, readable also by smartphones, which increases their usability in different application areas.

RFID codes are primarily used in the same way as bar codes. They uniquely identify items or services. Typically, besides an optional bar code, RFID labels carry even more information on a chip the size of a pin head. Besides identification RFID tags enable the logging of tracking information, often required in logistics applications. In contrast to bar codes, RFID enables their scanning in without a direct line of sight and also multiple labels at once.

By the GS1 standard EPC Gen2 (ISO/IEC 18000-6:2013) a technological standard determining communication between RFID tags and readers was established. Similar to bar codes, EPCglobal standards link RFID technology with EPC tagging of products, logistic transport units, locations, inventory, returnable items, documents, etc. for direct, automated identification and tracking of logistics units within supply chains.

EPCglobal standards also represent the basis of the GDSN (Global Data Synchronization Network). It enables automated acquisition and interchange of specification data on products

and their packaging, hereby enabling enterprises to centrally manage these data to be used by them and their partners interchangeably.

Table 3.1 summarizes the various identification technologies with their applications. It reveals a variety of one and two-dimensional bar codes as well as different classes of RFID codes with their capabilities.

Table 3.1 Tagging technologies.

Technology	Application
Bar 1D	Retail items and product components
Bar 2D	Services (e.g., UPS, air-tickets), wholesale items requiring tracking
RFID class 1 (passive, R-tags)	Items requiring mass identification, access control
RFID class 2 (passive, RW-tags)	Items requiring tracking
RFID class 3 (semi-active, RW-tags)	Access control with added tracking information
RFID class 4 (active, RW-tags)	Closed space tracking and tracing
RFID class 5 (active tags/interrogators)	Open space tracking and tracing, proximity services with enabled devices, location-based services

Future trends in tagging, tracking and tracing follow two main directions: miniaturization and diversity. The data bar (1D) codes shall also enable tagging miniature items (e.g., medical capsules). Novel data matrix (2D) codes shall not only enable error correction while scanning but also data encryption.

RFID continues to spread to other usage areas like the identification by service providers (e.g., railcard, registration at work, etc.), contactless payments (e.g., wireless money transfers, payments at vending machines) and smart solutions (e.g., smart home management, remotely operating smart devices, etc.) as well as e-currencies.

3.3 Data organization

Apart from having a certain format the data may be organized in different ways to facilitate their management, processing and presentation. Although data on their input are mostly unstructured, by their storage, transmission and processing their organization increases. In the sequel the usual forms of data organization are presented by increasing complexity from semi-structured (e.g., CSV) to structured (e.g., spreadsheets, databases, etc.) formats.

Spreadsheets

The first form of data organization is by two-dimensional arrays of fields, also called tables or spreadsheets. Usually, the first line of a spreadsheet denotes the meanings of values stored in the underlying columns, followed by lines of data.

A table field or cell is the smallest unit of data. It has a certain type (string, number, date, currency, etc.). Its contents are addressable by the row and column markings (e.g., A1, representing the first row of column A).

Each table line is a group of related fields, representing a record (e.g.: transaction, student record, product data, etc.). Since all table lines have the same structure, we may define the type of a record as a list of attributes (e.g., Student data (name, family name, birthdate, birthplace, ID…)) of the corresponding data types.

Databases

A database file or table is a collection of records of the same type. A database (DB) is made up of multiple inter-connected tables. Hence, the ANSI definition of a database:

- DB data are interconnected and sorted
- DB data can be simultaneously used by multiple users
- Data in the DB are not repeated
- DB is stored in a computer

From the above definition one may draw some conclusions about the client-server architecture where the server holds the DB, being accessed by its clients. Of course, to be able to access the DB a communication network has to be established between the server and its clients. The

DB server is usually termed as its "back-end", while the clients represent its "front-end". The database management system (DBMS) at the server enables its clients access to the data stored in the DB via its application program interface (API) and DBM functions. The DBM functions are mechanisms that enable DB data input, retrieving, processing and presentation. To call up these functions standard query languages (SQL) have been defined.

Relational database model

There are various forms of DB organization with the relational model (RDB) being the most common. The basic idea behind this model is the fact that a user cannot know all the possible uses data stored in a DB upfront. Since there are usually no fixed paths of searching through the database files, various query languages have been devised for data retrieval and manipulation. The RDB model is based on the concept of entities and relations:

- An entity is a person/thing/concept that can be uniquely identified and has attributes.
- A relation represents a way to associate two or more entities.

RDB tables, representing entities or relations, are interconnected by means of keys. The set of attributes that uniquely identify an entity is termed its primary key. When a primary key appears as a field in another table with the goal to fulfil a relation with its original table, it is termed secondary or foreign key. While a table may contain only one primary key to uniquely identify its records, it may contain multiple secondary keys.

In general, there are two approaches to constructing an RDB: analytic and synthetic.

The analytic approach to RDB construction comprises the following four steps:

1. Real world analysis - global model
2. Determine entities and relations – conceptual model (e.g., E-R diagram)
3. Determine logical model – relational schema
4. Build the database (DBMS) – physical model

To illustrate the approach, let us consider an example of a convenience store chain and their suppliers (Figure 3.1). Each store has multiple suppliers. Every supplier may supply different stores. The replenishment cycle starts by an order from the store. In return a supplier delivers goods to the store. Orders are transactions in which the data on the store, supplier and goods delivered are combined (Figure 3.2).

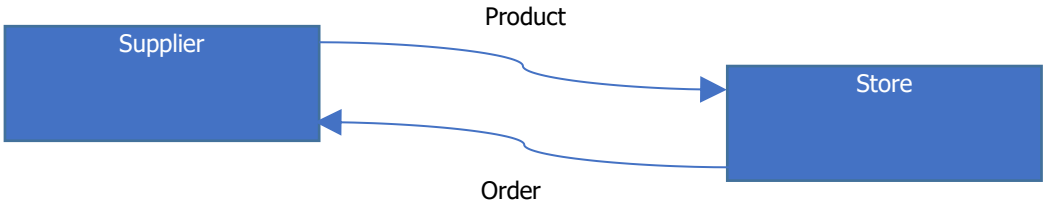

Figure 3.1 Global model.

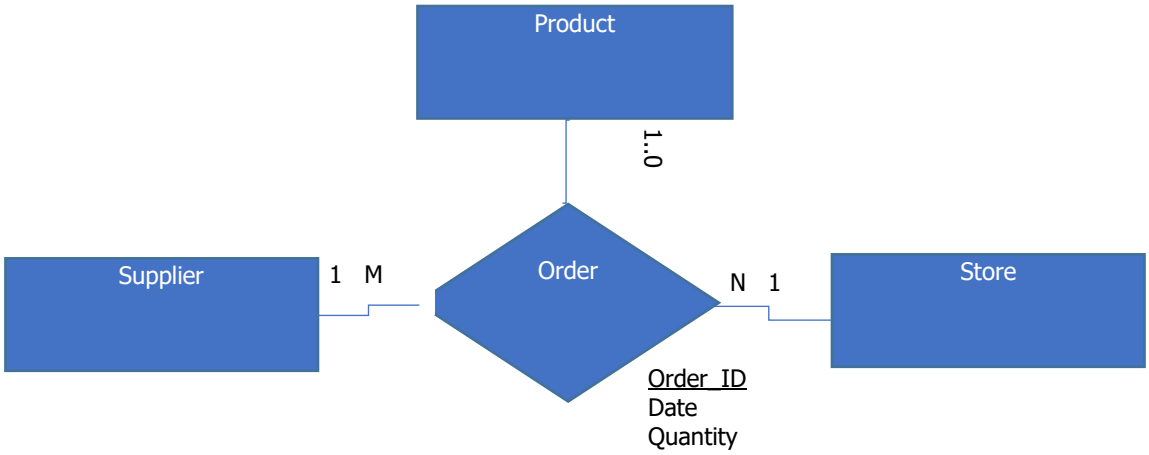

Figure 3.2 Conceptual model.

SUPPLIER (Supplier_ID#, Supplier_name, Supplier_contact)

STORE (Store_ID#, Store_name, Store_address)

ORDER (Supplier_ID#, Store_ID#, Order_ID#, Date, Quantity, EPC)

PRODUCT (EPC#, Product_name, Product_price)

Figure 3.3 Logical model.

The logical model represents the RDB tables by the types of their records. In the logical model (Figure 3.3) certain attributes contain the hashtag (#) symbol. This signifies that the field they represent is or belongs to a (composite) primary key. Some fields are underlined. They are called secondary or foreign keys, since they reference the primary keys of related tables.

The synthetic approach to an RDB comprises the following three steps:

1. Data analysis – list of all relevant attributes
2. Determine logical model by normalization – relational schema

3. Physical model (DBMS)

Normalization or canonical synthesis (Kent, 1983) ensures that by reverse engineering from the relevant attributes a DB is formed fulfilling the conditions of a RDB (cp. Figure 4). While the initial normal form (0NF) of attributes represents a table of unordered attributes, the subsequent normal forms, as defined by (Codd, 1970), represent higher levels of data organization. One may claim that by reaching the 3rd normal form one has achieved a schema fulfilling the requirements for a RDB logical model.

A table is in 1NF, if it represents a relation. By this it is ensured that all repeating groups of data are kept separately and hence do not repeat.

A table in 2NF is in 1NF. Additionally, no key-attributes may be partially functionally dependent on the primary key. Hereby, all keys that uniquely identify certain attributes are kept separately. Mainly this is to ensure that only attributes that are dependent on all (parts of) the primary key are kept in a single table.

A table in 3NF is in 2NF. Additionally, no non-key attributes are transitively dependent on the primary key. This means that all non-key attributes that may represent a key to certain other non-key attributes are kept in a separate table, whereby only their key is maintained in the original table as foreign key.

By following the normalization steps from the 1NF to the 3NF one ends up with a logical model that corresponds to the regulations of a relational database (RDB). Higher forms of normalization are mainly for RDB model optimization.

A table in Boyce-Codd NF (BCNF) is in 3NF; additionally, every determinant is a key. This removes all relations not covered by existing ordering of keys from the original table, hereby forming additional tables for every candidate key. A table in 4NF is in 3NF and BCNF. Additionally, every multi-valued attribute that is partially dependent on the key is in its own table. 4NF is meant to remove all possible remaining multi-valued attributes from the original table. A table is 5NF is in 4NF; additionally, every JOIN operation is foreseen by the keys. A table in 6NF is in 5NF; additionally, all non-trivial JOIN-dependencies are accounted for.

One can observe that by higher forms of organization, the number of tables increases with each step. Hence, it is sensible to observe the fragmentation of data in order to prevent unnecessary infrequently accessed tables from being created.

Table 3.2 RBD Normalization example.

ONF ORDER • Supplier_ID* • Supplier_name • Supplier_contact • Store_ID* • Store_name • Store_address • Order_ID* • Date • EPC • Quantity • Product_name • Product_price	1NF ORDER • Supplier_ID# • Supplier_name • Supplier_contact • Store_ID# • Store_name • Store_address • Order_ID# • Date • EPC • Quantity • Product_name • Product_price
* Candidate keys of the repeating group of attributes, uniquely identifying an order	
2NF SUPPLIER • Supplier_ID# • Supplier_name • Supplier_contact STORE • Store_ID# • Store_name • Store_address	ORDER • Order_ID# • Supplier_ID# • Store_ID# • EPC • Product_name • Product_price • Date • Quantity
3NF ORDER • Supplier_ID# • Store_ID# • Order_ID# • Date • Quantity • Product_ID	PRODUCT • EPC# • Product_name • Product_price

Query languages

They are mainly of two types: Structured Query Language (SQL) and Query by Example (QBE). While SQL is considered a programming language and a DBMS's API, QBE is mainly used with the DBMS directly for DB management and data warehousing.

The standard SQL (ISO/IEC 9075, 1986-2016) is a fourth-generation programming language for database manipulation. It enables searching, adding, modifying as well as deleting data records. Despite its standardization there are slight differences in its implementation with different database management systems (DBMS).

In the sequel the language is briefly presented with the most common options. By convention the SQL keywords are written in capital letters and every sentence ends with a semicolon. The sentences are presented with links to associated reference materials offering further information.

Every database manipulation starts by its creation. The sentence

CREATE DATABASE *database_name*;

creates a new empty database with the specified name.

As elaborated above, the data within databases are organized in tables of data records of a certain type where all rows share a common structure. To create a table the following sentence is used:

CREATE TABLE *table_name* (
column1 type1,
column2 type2,
column3 type3,
….);

Every named column represents an attribute with a certain data type. For example, in:

CREATE TABLE *Store* (Store_*ID* int NOT NULL PRIMARY KEY,…);

CREATE TABLE *Order* (Order_*ID* int NOT NULL PRIMARY KEY, …, Product_*Id* int FOREIGN KEY REFERENCES *Product* (*EPC*));

two tables are created. The first contains the data on the customers, while the second contains the data on their orders, referencing the first table via the customer number as foreign key.

While data in a table is already sorted by the primary key, it can be additionally sorted by other attributes, provided it is indexed. We can index it by creating an index on the provided attribute(s) by the following sentence:

CREATE INDEX *index_name* ON *table_name* (*column_name*);

Every data manipulation on an indexed table takes a little longer, since for its consistency, not only the data provided by keys needs to be checked and the data ordered appropriately, but also other attributes from the specified index.

The most common operation on a database is data query enabled by the SELECT statement:

SELECT *column1, column2*, …
FROM *table_name*;

This data query returns data in column1, column2, etc. from the table. The query sentences are usually formed by providing additional options, filtering out data, fulfilling the specified condition(s):

WHERE specifies a condition, which determines the records selection criteria.

GROUP BY joins records, having a common property to enable aggregate functions.

HAVING specifies aggregate functions on groups defined by GROUP BY statements.

ORDER BY specifies the attributes on which the return records are ordered.

For example:

```
SELECT "Store"."Store_Name", "Product"."Product_name", "Order"."Quantity" FROM "Order",
"Product", "Supplier", "Store" WHERE "Order"."Product_ID" = "Product"."EPC" AND
"Order"."Supplier_ID" = "Supplier"."Supplier_ID" AND "Order"."Store_ID" =
"Store"."Store_ID" ORDER BY "Store"."Store_Name" ASC
```

returns a list of stores with their ordered products and quantities, ordered by store name.

The most important operation in the selection process is the JOIN operation. Often, it replaces the WHERE condition as JOIN ON, followed by the condition. It compares the column values and determines, based on the comparison, whether or not they should be included in the result. In LEFT JOIN the record is returned, if the criteria are fulfilled in the left table and vice versa in the RIGHT JOIN operation. As outlined above, the condition needs to be fulfilled in

both tables in order to comply with the INNER JOIN or FULL JOIN operation. Since the latter is most commonly used, one can use JOIN as synonym. Referring to the conditions of the 5th and 6th normal forms, this is the same JOIN operation, which needs to be fulfilled to comply with the terms of the corresponding NF.

To input new data into the table the INSERT INTO operation is used:

INSERT INTO *table_name* (*column1*, [*column2*, ...])
VALUES (*value1*, [*value2*, ...]);

To be successful, the values in the operation need to fulfil all the conditions of the attributes denoted by column names. One does not need to specify column names in case all values are listed. In case some DEFAULT values are foreseen in the table, one does not need to specify them, unless they are different.

Once the data are input, they can be modified by an UPDATE statement:

UPDATE *table_name*
SET *column1=value1, column2=value2*,...
WHERE *some_column=some_value*;

In the statement new values for the fields in the listed columns are given. The row selection criteria are denoted by the WHERE specifier, which determines all column values to which the UPDATE statement applies. In order to prevent unwanted changes, extra caution needs to be applied with the formulation of the selection criteria.

A record or multiple records can be deleted from a table by the DELETE operation:

DELETE FROM *table_name*
WHERE *some_column=some_value*;

As with the UPDATE statement, the WHERE specifier is used to determine all rows that should be deleted.

Of course, database management doesn't end here. Every element of the database can also be removed, altered and/or replaced by a new one. In case an index, a table or a database is to be removed, the following statements can be applied:

DROP INDEX *index_name* ON *table_name*;

DROP TABLE *table_name*;

DROP DATABASE *database_name*;

If one only wishes to remove data from a table the TRUNCATE statement may be used:

TRUNCATE TABLE *table_name*;

In case one wishes to add or remove an attribute (column) to/from a table, one can do so by the ALTER statement:

ALTER TABLE *table_name* ADD *column_name datatype*;

ALTER TABLE *table_name* DROP COLUMN *column_name*;

This concludes this short overview of the SQL language and its most common usage scenarios. SQL is commonly used in client-server architectures with the DBMS hosted by the server. To access it, SQL statements are issued, either by a client application programme or the server's DBMS Web-interface.

On the other hand, QBE is also often used with relational DBMS with a graphical user interface (GUI), like the MS Access or LibreOffice Base. With QBE, the database and its tables are created much more interactively and their structure is easier to maintain. As its name suggests, it also offers a simpler form of data input and discovery. To perform searches, one needs to assemble all tables that are used in the search and then establish conditions as patterns in column fields to filter out the relevant data (Figure 3.4). The query formulation is supplemented by the existing relations between tables. As usual, the result from such a query is another table with the resulting data, which can be further processed later on. This way also cascading or multi-phase queries can be formed.

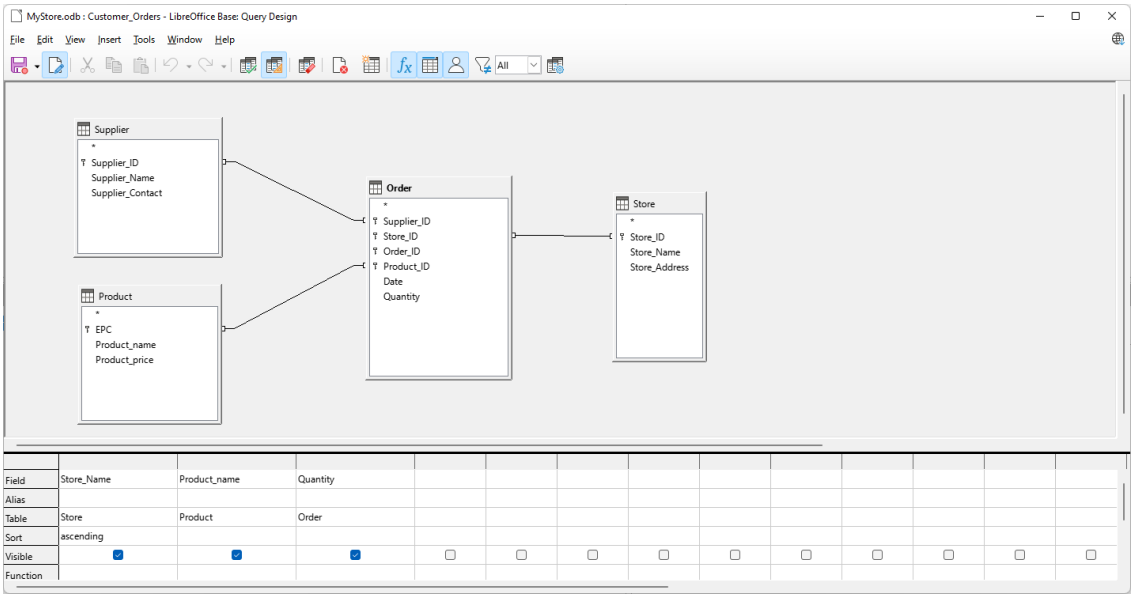

Figure 3.4 Query by Example equivalent to the above SQL query.

Data filters and masks

In order to prevent erroneous or incomplete data, which could interfere with their processing and interpretation, additional precautions should be applied:

1. Filtering out empty rows and columns
2. Applying strong data typing to prevent computational errors
3. Defining input masks to prevent input of wrong data
4. Data biasing to prevent erroneous results

Empty rows and columns are a common source of errors that originate mainly from poor interfaces in data collection applications. Cancelled or incomplete transactions usually result in empty lines or missing data in transaction logs. They can only partly be dealt with spreadsheet applications where empty lines and missing data can be detected by checking the total number vs. the number of non-zero data in rows/columns. They can be filtered out by removing the empty rows/columns, however, this may not always be the desired action, since we might lose some valuable data as well. The best way to prevent this from happening is by using a DBMS which would only allow complete transaction data to be entered on one hand, whereas on the other, it would also prevent empty rows/columns, since in databases there are none.

Weak data typing is another common source of errors. If alphanumeric data, like date or the amount of currency, is entered in place of numeric data, this would result in errors when processing these data. In spreadsheets as well as in databases the individual data cells, representing attribute values in data records, can be assigned data types, which would alarm us, when entering data in the wrong format. Thus, by this measure one can prevent wrong data to interfere with their processing.

When constructing databases, restrictions can be applied on the data fields representing entity or relation attributes. Apart from assigning them an appropriate type, input masks can be defined allowing data, like dates, currencies, EAN codes, etc., to be input only in a certain format. This usually resolves many misconceptions, which could otherwise occur while processing the data.

Another common source of errors are unbiased data, representing data that are orders of magnitude greater or smaller than expected. Again, they could interfere with our processing, yielding erroneous results. They are harder to detect and can only be filtered out by looking at the data. In spreadsheet applications a good common practice would be to determine the minimum, maximum and mean values of data in the corresponding columns to detect possible deviations. If they are detected, they can then be highlighted and dealt with manually, if they are a few, or filtered out and modified by a query on a database table, if they are many. Either way they should be assessed carefully, in order not to make the situation even worse, in which case it would be better to remove these data.

Data warehouses and knowledge bases

Data in data warehouses are collected from RDB and catalogued in chronological order. Usually, there are some business analyses performed on the data and the results stored for later references by the analytical department as well. Once they are stored in the warehouse, these data are usually not changed to preserve their consistency.

Besides chronological order, contextual orders are considered when building knowledge bases. Here, the entities are represented as derivatives of a top-level entity or a subsidiary entity thereof. Their relations are established more freely as they are meant to be updated and upgraded as they are used. They are established in the form of rules, based on entity properties. Hence, the form in which they are stored is somewhat different. Often, they are stored in the form of ontologies containing a deeper knowledge on the collected data. Similar

to storing query results in data warehouses, queries in knowledge bases are also stored for later use to render current results, as entities, relations and data instances change.

As opposed to databases and warehouses, which are application specific, knowledge bases may be application independent and are often used cross-domain by different applications. An example is presented in (Gumzej et. al., 2023).

3.4 Conclusion

In this chapter different aspects of data management in logistics have been dealt with. Apart from data representations and data storage standards, data organization and retrieval mechanisms have been presented. Finally, some common mistakes in automated data processing have been addressed to keep the conscious reader alert. Besides the listed examples, more can be discovered in the associated learning materials.

Reference Chapter 3

- Codd E.F. (1970). A relational model of data for large shared data banks. Communications. ACM 13, 6, pp. 377–387.
- Gumzej, R., Kramberger, T., Dujak, D. (2023). A knowledge base for strategic logistics planning, Proceedings of the 23rd International Scientific Conference Business Logistics in Modern Management: October 5-6, 2023, Osijek, Croatia, Dujak, Davor (ed.) Osijek: Josip Juraj Strossmayer University of Osijek, Faculty of Economics and Business, pp. 317-330. [available at: https://blmm-conference.com/past-issues/, access November 3rd, 2023]
- GS1 (2023). GS1 Standards. [available at: https://www.gs1.org/standards, access October 27th, 2023]
- Kent, W. (1983). A Simple Guide to Five Normal Forms in Relational Database Theory, Communications of the ACM, vol. 26, pp. 120-125.
- W3schools (2023). XML Tutorial [available at: https://www.w3schools.com/xml/, access November 3rd, 2023]
- W3schools (2023). JSON - Introduction [available at: https://www.w3schools.com/js/js_json_intro.asp, access November 3rd, 2023]

- W3schools (2023). Database Normalization [available at: https://www.w3schools.in/DBMS/database-normalization/, access December 7th, 2023]

4. Simulation modelling and analysis

By simulation modelling and analysis (SMA) one strives to fulfil the demands of the Conant–Ashby theorem (Conant and Ashby, 1970), defining a model of the system as a good regulator, having as many handles, parts and states as its original physical counterpart; thus, providing the possibility to build its digital model and establish a digital laboratory that will enable its exploration, adaption and optimization. The resulting simulation models are abstract, dynamic and in most cases stochastic, since their system variables are modelled by probability distributions.

4.1 Simulation in logistics

In logistics SMA can provide valuable inputs to Supply Chain (SC) and traffic network (TN) optimization. Simulation modelling can be used to graphically visualize temporal flows through complex SC and TN processes and resources, allowing the prediction and quantification of possible outcomes from different scenarios. This helps SC and TN entities to gain valuable insights and understand the effects of their potential decisions on SC and TN performance including SC lead-times, TN travel times and costs. Hence, SMA in SC and TN modelling can contribute to the SC and TN analysis and to the improvement of their designs toward achieving higher performance and sustainability.

There are many facets of a SC, representing different SC management perspectives. A production manager's view of the SC differs from the marketing manager's view, which again differs from supply manager's view, etc. Hence, the models used are different, even for the same company, let alone the whole SC.

When solving SMA problems in logistics, managers need to make decisions on strategic, tactical and operational levels, depending on their effect on the SC or TN as a whole. Due to their interdependences, managers are often unable to solve problems on any single level. At the same time, it is also difficult to observe all three levels from the perspective of any individual entity. From SMA perspective, one can observe a SC or TN on two levels:

1. Macro level

- self-organization,
- co-evolution of entities,
- dependency on connections/transport routes.

2. Micro level

- multiple and heterogeneous entities,
- local interactions among entities,
- structured entities,
- adaptive entities.

Although they are performed in real time, the temporal aspect of SC operations is somewhat ambiguous. Depending on the level and perspective the durations of operations may be measured in days, weeks or even months when considering inter-organizational activities, while on the other hand, intra-organizational operations are measured in hours or even seconds. Depending on the nature of the modelled problem, the duration of the shortest operation or the maximum frequency of incoming/outgoing requests determines not only the representation of time in an SMA model, but also its granularity. The shorter the minimum duration of the shortest operation or the higher the highest frequency of requests is, the finer is the granularity of time or in other words the precision of time keeping in the model. This is important for the modeller, since the model's reaction time cannot be shorter than the predefined time granularity. Hence, one needs to estimate the duration of all operations and inter-arrival times of incoming/outgoing signals in advance to be able to determine the time units of a system model correctly.

In a simulation model, time can either progress by critical events from transaction to transaction or continuously. In the latter case, the progression of time in the model is independent from the frequency of operations. With critical event triggered time flow, the operations are invoked according to their occurrence times, namely critical events. The benefit of SMA is that during a simulation, one may speed up the progression of time in the model, so the processes perform faster than in real time. Thus, one can make early predictions of following events.

The times between incoming simulation units and their processing/transit times may result from observations and measurements. If they don't vary, they are deterministic. However, usually they are stochastic in their nature. Hence, the introduction of constructs modelling their probability distribution functions (e.g., triangular, uniform, exponential, etc.).

4.2 Discrete event simulation

DES analysis offers the most detailed insight into a logistic (production) process to the production manager by a consistent and coherent model. Thus, DES is highly regarded tool to determine real-time behaviour and resource utilization in process industry, including logistics.

Constructs:

- Flow units represent simulation units (e.g., orders, materials, etc.) which enter the system on the input(s) and progress through the system model.
- Processors represent mobile (e.g., people, forklifts, etc.) and fixed (e.g., machines, production lines, etc.) resources which process the simulation units.
- Queues store the flow units until their transition to the next available processor.
- Connectors define the promotion of units through the system model.

Properties:

- Process-oriented.
- Focuses on detailed process modelling.
- Heterogeneous entities.
- Micro-entities are passive objects.
- Events introduce dynamics into the system.
- Discrete time-progression; from one (time) event to the next.
- Flexibility is achieved by changing the structure of the model; system structure during simulation is fixed.

Example

The DES example (Figure 4.1, extracted from the JaamSim (JaamSim Development Team, 2023) simulation environment) comprises a model of variant production, where four different products are being produced (Gumzej and Rakovska, 2020). According to the production plan, some 10, 30, 40, 20% of product types 1, 2, 3, and 4, respectively, are being produced. Choosing a product type is induced by the triangular distribution between 1 and 4 with modulo at 3. Each product type has a dedicated production line. The production orders are fulfilled according to the exponential distribution around the 30 s mean time value. The production of every single product takes 100–120s according to the uniform distribution. After they are finalized, the products are checked for quality at a dedicated test site. The quality check takes 10 s. From the company's experience, on average every 1 out of 10 products doesn't pass inspection. Products of insufficient quality are transported back to the original production line. Their reprocessing takes 120–130s according to the uniform distribution. The durations of production and quality inspection and reprocessing don't depend on product type. After they have successfully passed their quality control the finished products are transported from the production site to the finished products warehouse. Re-manufacturing defective products while still in production is an effective way to reduce both environmental impacts and manufacturing costs.

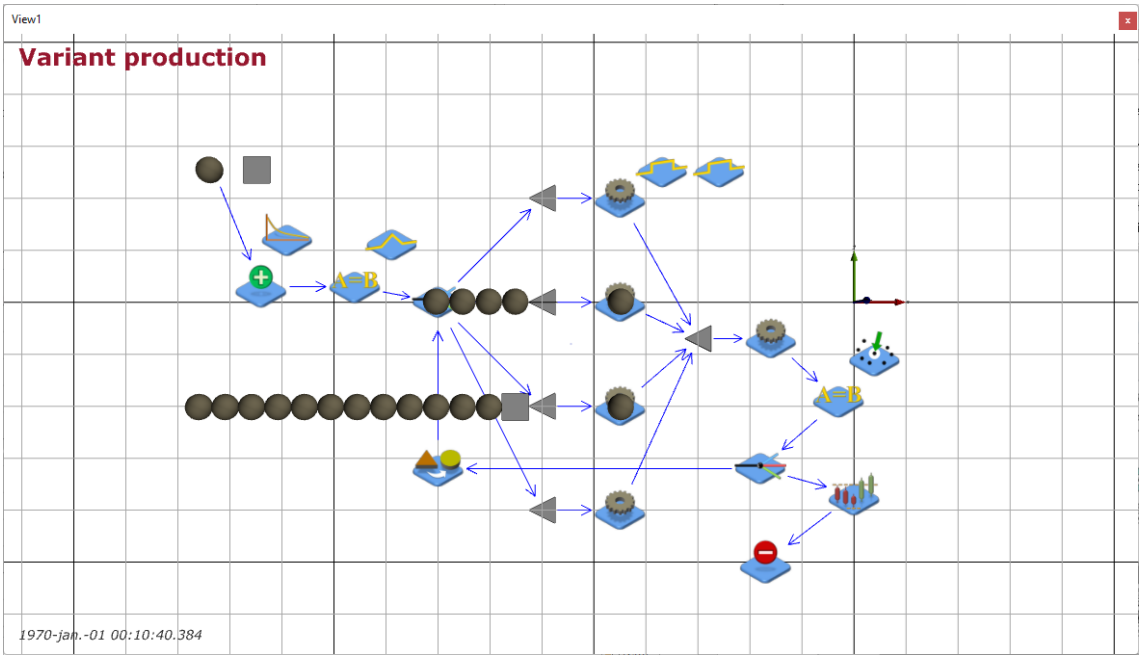

Figure 4.1 Variant production with quality control.

Synopsis

The following process parameters can be analysed and optimized by DES:

- Production cycle-time and performance.
- Utilization of production cells and spaces.
- Capacity of storage spaces as well as storage unit's dwell-times.
- Utilization of mobile resources (e.g., operators, conveyors, forklifts).

4.3 System dynamics

SD analysis represents a SC manager´s view of a production process by a consistent and coherent model. SD is regarded as a tool best suited to determine the structure as well as optimal volumes (when and how much of individuals site´s inputs, stocks, and outputs. Therefore, it allows for the efficient utilization of production and storage facilities.

Constructs:

- Stocks represent buffers which can store delivery items on the supply chain.
- Flows represent supply channels.
- Feedback loops represent fine-tuning parameters for stock replenishment.

Properties:

- System-centred.
- Key performance-indicators' oriented modelling of system variables.
- Homogeneous entities.
- Entities on micro-level are disregarded.
- Dynamics is introduced by feedback loop coupling.
- Continuous time-progression; time progresses synchronously for all components of the system model.
- Flexibility is achieved by changing the structure of the model.

- System structure during simulation is fixed.

Example

The SD example (Figure 4.2, extracted from the NetLogo (Wilensky, 1999) simulation environment) comprises a home appliance company's SC and describes material flows between its subsidiaries (Gumzej and Rakovska, 2020). The company has multiple production sites: main site in Slovenia (SI) as well as affiliate firms in Germany (DE), Poland (PL), Hungary (H), and Bosnia–Herzegovina (BIH). In addition to production sites, its gross-sales sites are situated in Russia (RUS), Ukraine (UKR), and Romania (RU). The production sites supply their own markets with finished products and each other with product components.

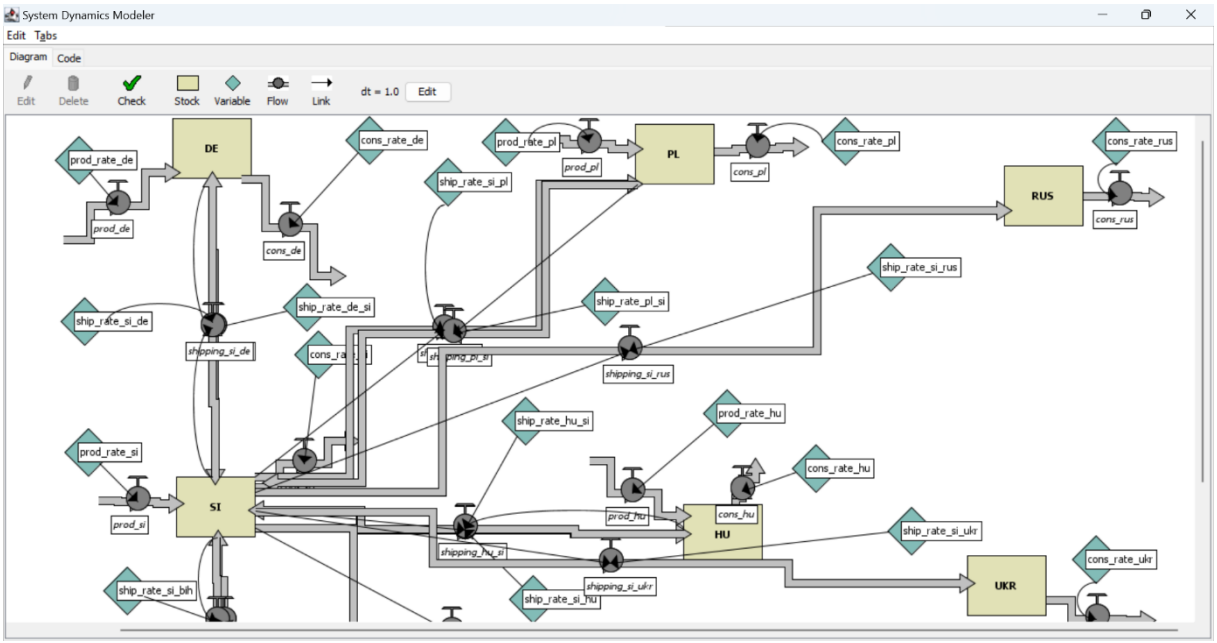

Figure 4.2 SC layout.

The associated NetLogo dashboard (Figure 4.3) serves as a decision support tool (DST), to covenant the production- and stock quantities with the predispositions and their physical distribution. The time flow is continuous throughout every day's transactions, i.e., every day a certain number of components are shipped between production sites and a certain number of finished products are consumed on site or shipped to the distribution sites. Based on an initial stock of 300 units at SI location and 0 stock at other locations and the distribution model, the stock quantities at individual locations represent the average stock according to given production (pcs), consumption (%) and shipping (%) rates.

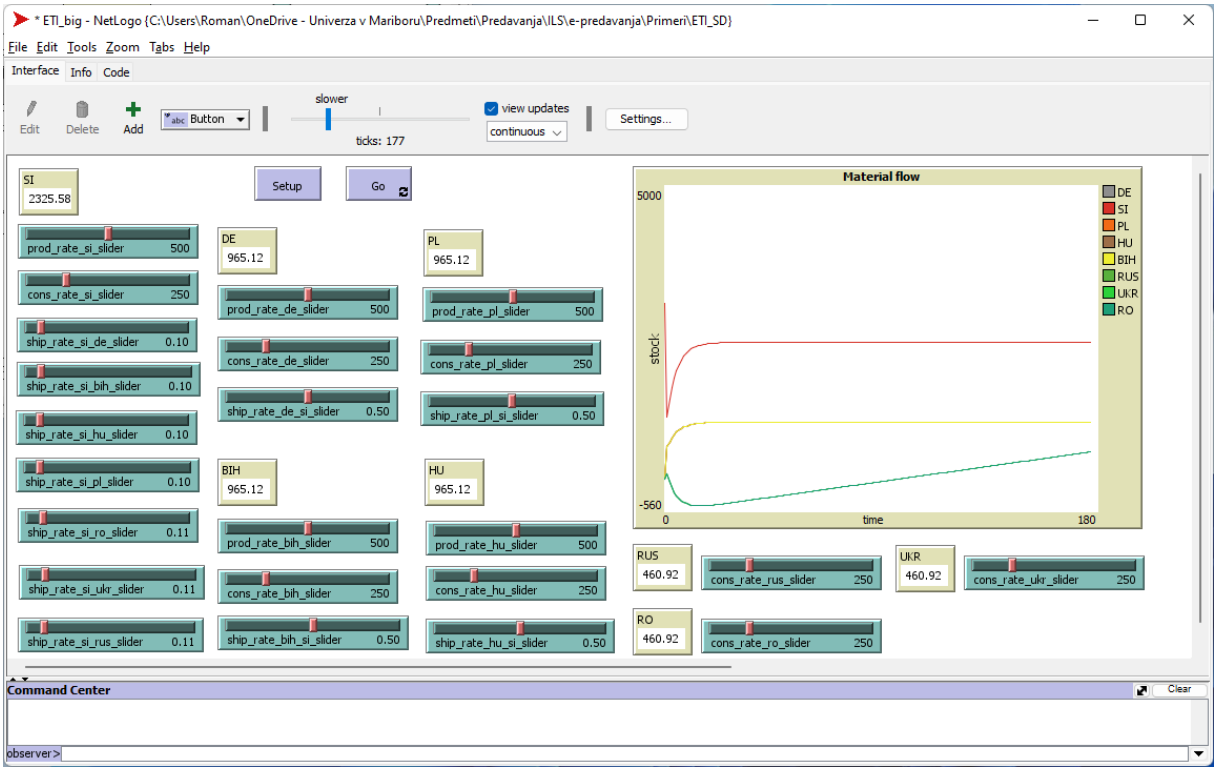

Figure 4.3 SC Dashboard.

Synopsis

System dynamics simulation allows for:

- Planning the layout of a SC.
- Optimization of production and distribution capacities.
- Estimation of distribution channels' loads and associated costs.

4.4 Agent-based Simulation

ABS analysis offers a strategic manager´s or market regulator´s view of the marketplace. Hence, ABS is regarded as a tool best suited to determine the optimal structure and layout/assortment of one´s market and/or SC by considering their global characteristics (e.g., demography, climate, GDP, quality, awareness, etc.).

Constructs:

Agents representing supply chain nodes (e.g., suppliers, retailers and inspectors) with their properties, relations and behaviour.

Properties:

- Entity-centred.
- Problem-oriented modelling of entities and their interactions.
- Heterogeneity of entities.
- Micro-entities are active objects that act in their environments, communicate among each other and autonomously make decisions.
- Decisions and interactions between agents introduce dynamics into systems.
- Agents and their environments constitute formal models.
- Time flow is discrete and universal on model-level; model timing is consistent with the frequency of SC transactions and the life cycles of SC nodes.
- Model flexibility is achieved by the changing system structure and behaviour of agents.
- System structure during simulation is variable.

Example

The ABS example (Figure 4.4, extracted from the NetLogo (Wilensky, 1999) simulation environment) was used to analyse the behaviour SC echelons in an open market (Gumzej and Rakovska, 2020), with respect to their Quality of Service (QoS). In the example, the different policies concerning the company's total quality management were investigated by a model, comprising its suppliers, customers and market regulators.

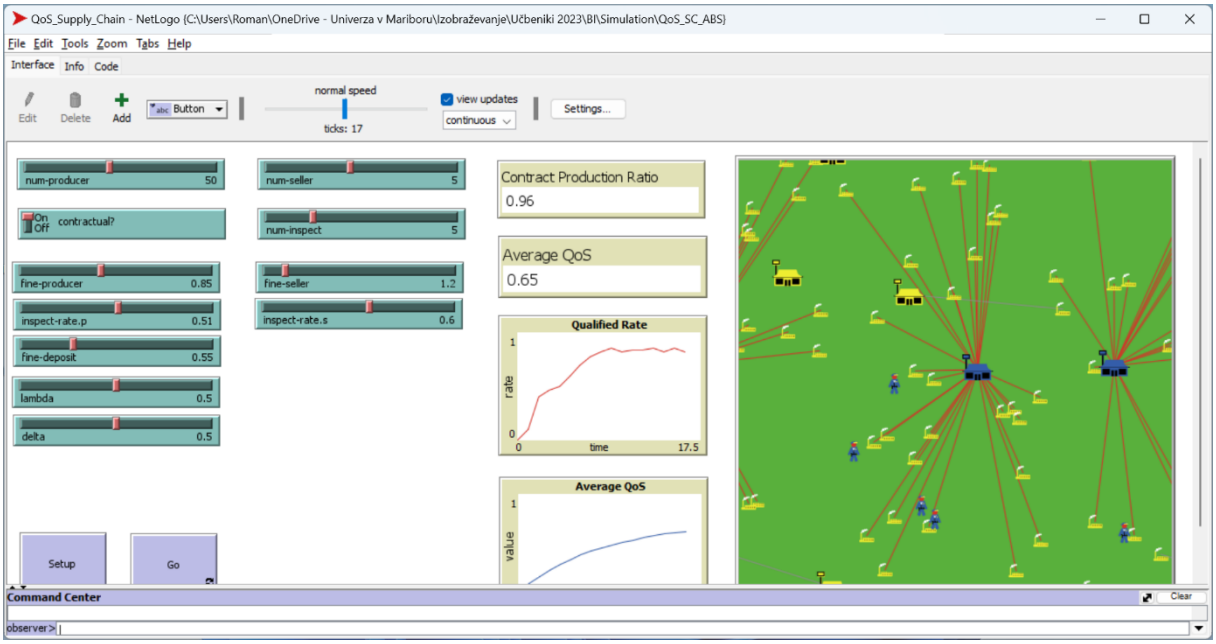

Figure 4.4 Marketplace regulation.

Synopsis

Agent-based simulation allows for:

- Planning the layout of an SC.
- Modelling the dynamic growth of an SC.
- Modelling the behaviour of partners within SCs.
- Optimization of global indicators.

4.5 Network simulation

NS analysis offers a network regulator´s view of the network. Hence, NS is regarded as a tool best suited to determine the optimal structure, layout and assortment of one´s network by considering its global characteristics (e.g., throughput, emissions, QoS indicators, etc.).

Constructs:

Agents representing the flow objects with their properties, relations and behaviour.

Network representing the overlay network (e.g., traffic network) on which the flow objects commute.

Properties:

- System-centred.
- Problem-oriented modelling of entities and their interactions.
- Heterogeneity of entities.
- Micro-entities are active objects that act in their environments, communicate among each other and autonomously make decisions.
- Decisions and interactions between agents introduce dynamics into systems.
- Agents and their environments constitute formal models.
- Time flow is discrete and universal on model-level; timing is consistent with the relative speeds of flow objects.
- Model flexibility is achieved by changing the network structure, which is fixed during simulation, and behaviours of agents which vary according to the (traffic) network state and their goals.

Example

The presented example (Figure 4.5, extracted from the SUMO (Pablo et.al., 2018) simulation environment) was used to determine the traffic flows and throughputs of streets in a city centre affected by a planned road blockage (Šinko and Gumzej, 2021). In addition, traffic related indicators like travel times, fuel consumptions and emissions were measured.

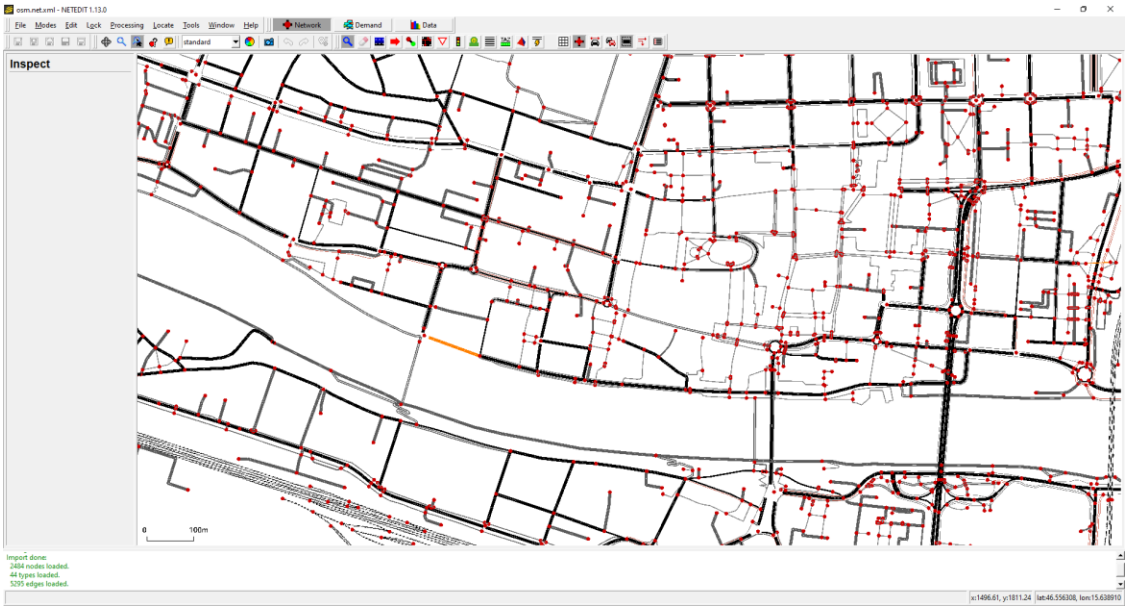

Figure 4.5 Traffic situation & network.

Synopsis

Network simulation allows for:

- Planning the layout of a network.
- Modelling the dynamic behaviour of a network to determine bottlenecks and weak links.
- Modelling the flow of network items.
- Optimization of global network indicators.

4.6 Logistics simulation projects

Logistics simulation projects are designed in concordance with Design for Six Sigma (DFSS) paradigm and are based on the Deming's cycle of improvement:

- Planning: definition of the system and goals.
- Execution: design of the simulation model.
- Analysis: experimenting with the simulation model and evaluating alternatives.

- Action: utilization of simulation results to implement improvements.

Any logistics simulation project (Figure 4.6) is composed of seven phases:

1. Strategic plan: analysis of existing and suggested resources and processes.
2. Conceptual model: abstract system model and predispositions definition, data collection.
3. Logical model: object-flow, stock and flow or network diagram of the system model.
4. Simulation model: development of an adequate simulation model.
5. Verification and validation of the simulation model: checking model consistency and coherence.
6. Analysis based on the simulation model: design and execution of experiments.
7. Utilization of simulation results to devise an action plan: projection of system improvements.

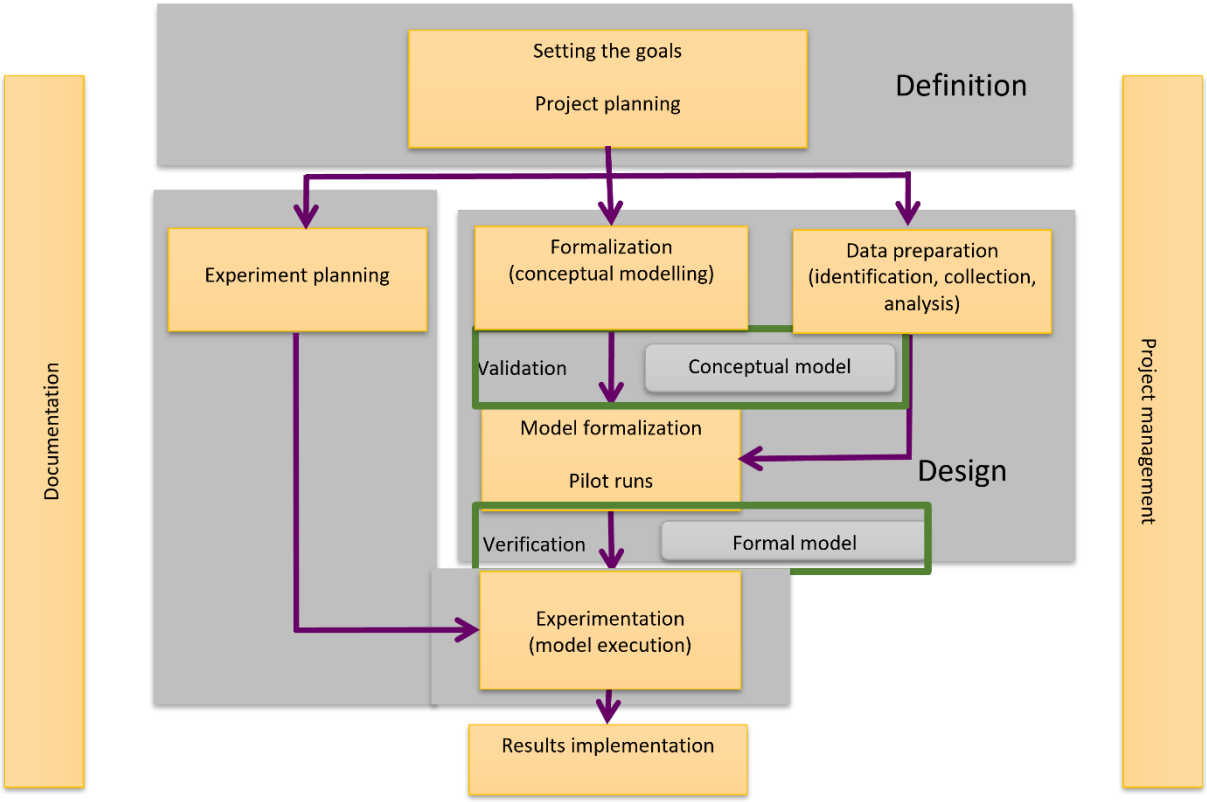

Figure 4.6 Simulation modelling and analysis (SMA) process.

4.7 Conclusion

In logistics SMA is an important component of operations research (OR) to enable process optimization.

System dynamics (SD) is a methodology for analysing complex, dynamic and non-linear interactions in systems, resulting in new structures and policies to improve the system behaviour. Here, physical and information flows are addressed with the aim to reduce their delay and ultimately SC inventory.

Another popular process-oriented methodology is Discrete event simulation (DES). It is one of the most widely used and flexible analytical tools in SMA of manufacturing systems. It successfully handles uncertainty and provides possibilities to compare alternative ways for lead-time reduction as well as optimization of machine and resource utilization.

A helpful methodology to understanding the behaviours of organizations and their interactions (e.g., SCs and their entities) is Agent-Based Simulation (ABS). Network simulation (NS), being a special kind of ABS, allows for modelling and optimization of (traffic) overlay networks.

This leads to the conclusion that a holistic approach of applying SMA in logistics contributes greatly to complex system design decisions, where there are many variables interacting with each other. A useful integrated approach, including SD, DES and ABS methodologies, which can quantify the workflows on different supply chain levels, has been presented in (Gumzej & Rakovska, 2020). In traffic flow analysis and optimization, the NS methodology provides for the necessary framework, regarding the monitoring and fine-tuning of key performance indicators (Šinko and Gumzej, 2021).

References Chapter 4

- Conant R.C. and Ashby W.R. (1970). Every good regulator of a system must be a model of that system, Int. J. Systems Sci., 1(2), pp. 89-97.
- Gumzej, R. and Rakovska, M. (2020). Simulation modeling and analysis for sustainable supply chains. In Ecoproduction – Sustainable logistics and production in industry 4.0 : new opportunities and challenges, Grzybowska, K., Awasthi, A., Sawhney, R. (ed.). Springer Nature, pp. 145-160.

- JaamSim Development Team (2023). JaamSim: Discrete-Event Simulation Software. Version 2023-08. [Available at: https://jaamsim.com, access November 8th, 2023]
- Šinko, S. and Gumzej, R. (2021). Towards smart traffic planning by traffic simulation on microscopic level. International journal of applied logistics, 11(1), pp. 1-17.
- Wilensky, U. (1999). NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. [Available at: http://ccl.northwestern.edu/netlogo/, access November 8th, 2023]
- Pablo A.L., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P. and Wießner, E. (2018). Microscopic Traffic Simulation using SUMO. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), IEEE. The 21st IEEE International Conference on Intelligent Transportation Systems, 4.-7. Nov. 2018, Maui, USA, pp. 2575-2582.

5. Linear Regression with Single and Multiple Regressors

Management decisions are often based on the relationship between two or more variables. For example, a marketing manager may try to forecast sales at a certain level of advertising expenditure after examining the relationship between that expenditure and sales.

In the second case, the public undertaking may use the ratio between the daily maximum value temperature and electricity demand to predict electricity consumption. Sometimes the manager relies on intuition. Intuitively, he judges how the two variables are related. However, if it is possible to obtain the data, it makes sense to use a statistical procedure called regression analysis to show how the two variables are related to each other.

In regression terminology, the predicted variable is called the dependent variable.

The variable or variables used to predict the value of the dependent variable are called independent variables.

In analysing the effect of advertising expenditure on sales, sales would thus be the dependent variable. Advertising expenditure would be the independent variable. In statistical notation y denotes the dependent variable, and x denotes the independent variable.

In this section, we will look at the simplest type of regression analysis, which involves one independent variable and one dependent variable. The relationship between the two variables will be approximated by a straight line. It is called simple linear regression. Regression analysis involving two or more independent variables is called multiple regression analysis.

5.1 Simple linear regression model

Best Burger is a chain of fast-food restaurants located in a multi-state area. Best Burger locations are located near university campuses. Managers believe that the quarterly sales of these restaurants (indicated by y) is positively correlated with the size of the student population (denoted by x). Restaurants near campuses with a large number of students tend to generate more sales than those near

campuses with a small number of students. Using regression analysis, we can develop an equation that shows how the dependent variable y is related to the independent variable x.

5.2 Regression model and regression equation

In the case of Best Burger, the population is all Best Burger restaurants. For each restaurant in the population there is a value x (student population) and a corresponding value y (quarterly sales). The equation describing how the y is related to x is called a regression model.

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0 and $\beta_1$are called the model parameters, ϵ (Greek letter epsilon) is a random variable called the model error. The error represents the variability y which cannot be explained by a linear relationship between $\mathrm{x\ in\ y}$.

The population of all Best Burger restaurants can also be seen as a collection of sub-populations, one for each separate value x. For example, one subpopulation consists of all Best Burger restaurants near university campuses with 8000 students. The second subpopulation consists of all Best Burger restaurants located near university campuses with 9000 students and so on. Each subpopulation has a corresponding distribution of values y. Each value distribution y has its mean or expected value. The equation describing what the expected value is y, denoted by $E(y)$, which is related to x is called the regression equation. The regression equation for a simple linear regression is as follows

$$E(y) = \beta_0 + \beta_1 x$$

The graph of a simple linear regression equation is a straight line. β_0 represents the initial value of the regression line, β_1 is the direction coefficient of the line, and $E(y)$ the mean value or expected value y for a given value of x.

Examples of possible regression lines are shown in the figure 5.1 below. The regression line in case A shows that the value of y is positively correlated with x. As the values increase x, the values also increase $E(y)$. The regression line in Panel B shows the value of y, which is negatively correlated with x. Where smaller values are $E(y)$ are associated with higher values x. The regression line in Panel C shows the case where the value of y is not associated with x. This means that the value y is the same for each value x.

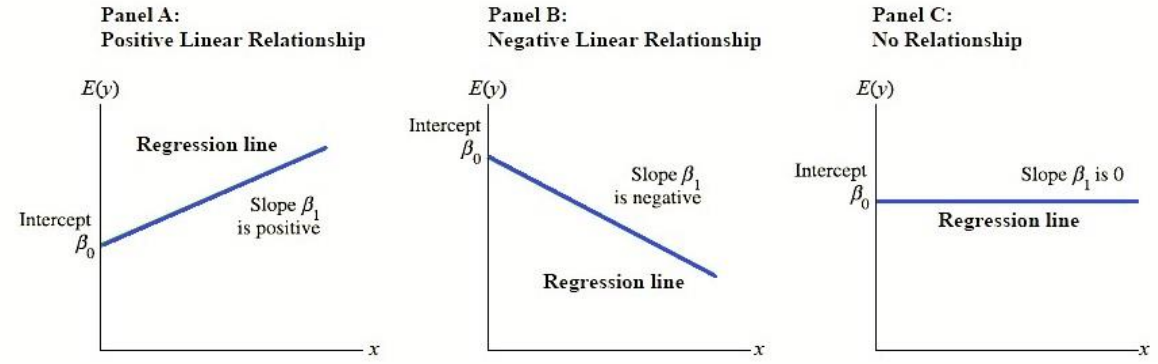

Figure 5.1 Graph examples of Linear Relationship.

5.3 Estimated regression equation

If the values of the population parameters were known β_0 and β_1, we could use the above equation to calculate the values of y for a given value x. In practice, these parameters are difficult to access, so they are simply estimated using sample data. The sample statistics (denoted by b_0 and b_1) are calculated as estimates of the population parameters β_0 and β_1. Replacing the values of the sample statistics b_0 and b_1 instead of $\beta_0$and $\beta_1$in the regression equation gives us a new, estimated regression equation. The estimated regression equation for a simple linear regression is as follows

$$\hat{y} = b_0 + b_1 x$$

The graph of an estimated simple linear regression is called the estimated regression line. b_0 represents the initial value of the regression line, b_1 is the direction coefficient of the line.

Below we show how to use the least squares method to calculate the values of b_0 and b_1 in the estimated regression equation.

In general $\hat{y}$ (score for $E(y)$) average value y for a given value x. If we now wanted to estimate the expected value of quarterly sales for all Best Burger restaurants located close to campuses with 10000 students, the value would be x would be replaced by the value 10000 in the last equation. In some cases, however, we may be more interested in forecasting sales for only one specific restaurant. For example, suppose you wanted to forecast quarterly sales for a restaurant that you plan to build near a college with 10000 students. As it turns out, even in this case, the best predictor of the value of the y for a given value x value $\hat{y}$.

5.4 Least squares method

The least squares method is a procedure where, using sample data, we find the equation of the estimated regression line. To illustrate the least squares method, let us assume that the data were collected from a sample of 10 Best Burger restaurants near university campuses. With x_i will denote the size of the student population (in thousands) and by y_i the size of quarterly sales (in thousands of EUR). $x_i \; in \; y_i$ for the 10 sample restaurants are summarised in the table below. We see that restaurant 1, z x_1 = 2 and y_1 = 58, is close to a campus with 2000 students and has quarterly sales of € 58,000. Restaurant 2, with x_2 = 6 and y_2 = 105, is close to a campus with 6000 students and has quarterly sales of 105.000 €. The restaurant with the highest sales value is restaurant 10, which is close to the campus with 26,000 students and has quarterly sales of € 202,000.

The following is a scatter plot of the data in the figure 5.2 below. Student population is shown on the horizontal axis and quarterly sales on the vertical axis. The scatter diagrams for the regression analysis are constructed with the independent variable x on the horizontal axis and the dependent variable y on the vertical axis. The scatter diagram thus allows us to draw preliminary conclusions about the possible relationship between the variables.

Restaurant i	Student Population (1000s) x_i	Quarterly Sales (€1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Figure 5.2 Scatter plot of data.

What preliminary conclusions can be drawn from the figure below 5.3? Higher quarterly sales occur in campuses with a larger student population. In addition, there is a constant relationship between the size of the student population and quarterly sales, which can be described by a straight line. Between $x \; in \; y$ a positive linear relationship is indeed implied. Therefore, we have chosen a

simple linear regression model to represent the relationship between quarterly sales and the student population. Given this choice, our next task is to use the sample data table to determine the values of b_0 and b_1, which are important parameters in the estimation of a simple linear regression equation. For the i-th restaurant, the estimated regression equation is

$$\hat{y}_i = b_0 + b_1 x_i$$

Where

$\hat{y}_i$ _ estimated value of quarterly sales (€1000) for the i-th restaurant

b_0 _ initial value of the estimated regression line

b_1 _ direction coefficient of the estimated regression line

x_i _size of the student population (1000) for the i-th restaurant

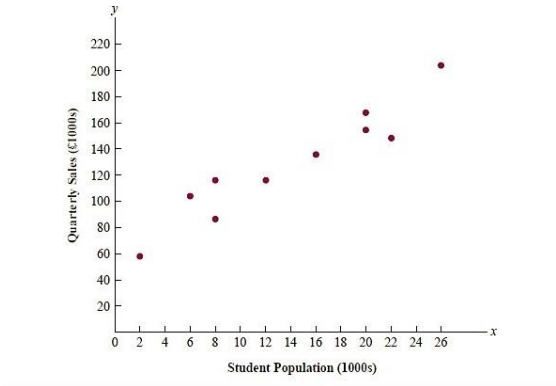

Figure 5.3 Scatter plot graph.

y_i denotes the observed (actual) sales for the restaurant i and $\hat{y}_i$, representing the estimated value of sales for the restaurant i, each restaurant in the sample will have an observed sales value of y_i and the predicted sales value $\hat{y}_i$. For the estimated regression line to ensure a good fit to the data, we want the differences between the observed sales values and the predicted sales values to be as small as possible.

The least squares method uses sample data to provide values b_0 and b_1.

Minimise the sum of the squares of the deviations between the observed values of the dependent variable y_i and the predicted value of the dependent variable $\hat{y}_i$. The starting point for calculating the minimum sum by the least squares method is given by the expression

Minimum Sum Criterion: $\min \sum(y_i - \hat{y}_i)^2$

Where

y_i =observed value of the dependent variable for the i-th observation

$\hat{y}_i$ =predicted value of the dependent variable for the i-th observation

The directional coefficient of the regression line and the initial value:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

x_i _ value of the independent variable for the i-th observation

y_i _ value of the dependent variable for the i-th observation

$\bar{x}$ _ average value for the independent variable

$\bar{y}$ _ average value for the dependent variable

n _total number of observations

Some of the calculations needed to develop the estimated least squares regression line are shown below. With a sample of 10 restaurants, we have n=10 observations. The above equations first require the calculation of the mean value of x and the average value y.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14, \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130$$

Alternative calculation equation b_1:

$$b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

Using the last equations and the information in Figure 5.4, we can calculate the directional coefficient of the regression line for the Best Burger restaurants example. Calculating the slope (b_1) is as follows.

Figure 5.5 shows a plot of this equation on a scatter diagram.

The slope of the estimated regression equation or the directional coefficient of the equation ($b_1 = 5$) is positive.

Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

Figure 5.4 Plot of equation on a scatter diagram.

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

This is followed by the calculation of the initial value (b_0).

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

This is how the regression equation is estimated:

$$\hat{y} = 60 + 5x$$

Figure shows a plot of this equation on a scatter plot.

The slope of the estimated regression equation (b_1 = 5) is positive, which means that as a student population increases, sales increase. In fact, we can infer (based on measured sales in the 1000s and student population in the 1000s), meaning an increase in the student population of 1000 is associated with an increase in expected sales of 5000; i.e. quarterly sales are expected to increase by 5€ per student.

If we believe that the regression equation, estimated by least squares, adequately describes the relationship between $x\ in\ y$, it seems reasonable to use the estimated regression equation predict the value y for a given value x. For example, if you wanted to predict quarterly sales for a restaurant located near a campus of 16,000 students would calculated by

$$\hat{y} = 60 + 5(16) = 140$$

Therefore, we would assume quarterly sales of 140,000 for this restaurant. In the following sections we discuss methods for assessing the appropriateness of using the estimated regression equation for estimation and forecasting.

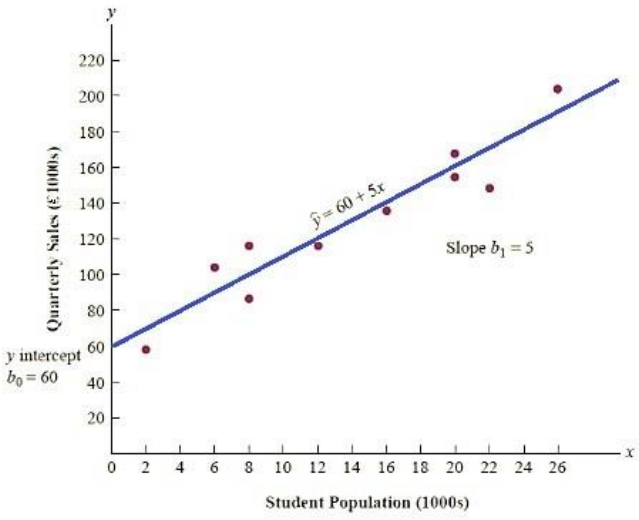

Figure 5.5 Scatter plot of student population and quarterly sales.

5.5 Coefficient of determination

For the Best Burger restaurants example, we developed an estimated regression equation $y = 60 + 5x$ for an approximately linear relationship between the size of the student population x and quarterly sales y. Now the question is: how well does the estimated regression equation fit the data? In this section, we show that the coefficient of determination provides a goodness-of-fit measure for the estimated regression equation. For the i-th observation, the difference between the observed value of the dependent variable $y_i$and the predicted value of the dependent variable is called the i-th residual.

The sum of the squares of these residuals or errors is the quantity that is minimised by the least squares method. This quantity, also known as the sum of squares squared to the error, is denoted by SSE.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

The SSE value is a measure of the error in using the estimated regression equation to predict the values of the dependent variable in the sample. Figure 5.6 shows the calculations needed to calculate the sum of squares due to the error for the Best Burger case.

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales (€1000s)	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1530

Figure 5.6 Squares of errors in Best Burger case.

Suppose we are asked to produce an estimate of quarterly sales without knowing the size of the student population. Without knowing any associated variables, we would use the sample average as an estimate of quarterly sales at any restaurant. Table in Figure 5.6 showed that for sales data y_i =1300. Therefore, the average quarterly sales value for a sample of 10 Best Burger restaurants is y_i/n = 1300/10 = 130. In Table 14.4 we show the sum of squares of the deviations obtained by using the sample mean of 130 to predict the value of quarterly sales for each restaurant in the sample. For the i-th restaurant in the sample, the difference y_i provides a measure of the error that is included in the application for sales forecasting. The corresponding sum of squares, called the total sum of squares, is denoted by SST.

$$SST = \sum (y_i - \bar{y})^2$$

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales (€1000s)	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
				SST = 15,730

Figure 5.7 Sum of squares.

The sum at the bottom of the last column in Figure 5.7 is the total sum of squares for BestBurger's restaurants SST = 15,730. In Figure 5.8 we show the estimated regression line $y = 60 + 5x$ and the line corresponding to y = 130. Note that the points cluster more closely around the estimated regression line than about the line y = 130. For example, for the 10th restaurant in the sample, we see that the error is much larger when 130 is used to predict y = 10 than when 130 is used $y = 60 + 5x$ and is 190. We can think of SST as a measure of how well the observations cluster around the line and SSE as a measure of how well the observations cluster around the line.

To measure how much the values on the estimated regression line deviate from the following, another sum of squares is calculated. This sum of squares, called the sum of squares due to regression, is denoted as SSR.

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

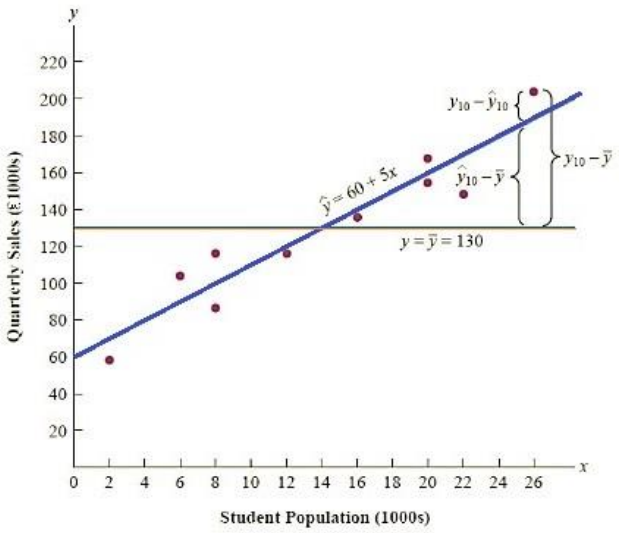

Figure 5.8 Regression line for Best Burger case.

From the previous discussion, we should expect that SST, SSR and SSE are linked. In fact, the relationship between these three sums of squares is one of the most important results in statistics.

5.6 The relationship between SST, SSR and SSE:

$$SST = SSR + SSE$$

Where it is:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Equation ($SST = SSR + SSE$) shows that the total sum of squares can be divided into two components, sum of squares due to regression and sum of squares due to error. So if the values of any two of these sums of squares are known, the third sum of squares can be known easily by calculation. For example, in the case of Best Burger restaurants, we already know that SSE = 1530 and SST = 15,730; therefore, by solving for the SSR in equation above, we find that the sum of squares due to regression is

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

Now let's see how we can use the three sums of squares, SST, SSR and SSE, to provide

a goodness-of-fit criterion for the estimated regression equation. The estimated regression equation would provide a perfect fit if each value of the dependent variable $yi\_$ would lie randomly on the estimated regression line. In this case it would be zero for each observation, resulting in SSE =0. Since SST = SSR + SSE, we see that for a perfect fit SSR must equal SST and the ratio (SSR/SST) must equal one. A worse fit will result in larger values for SSE. Solving for SSE in equation (14.11), we see that SSE = SST - SSR. Therefore, the largest value for SSE (and hence the worst fit) occurs when SSR = 0 and SSE = SST.

The SSR/SST ratio is used for the estimation, which has values between zero and one fit to the estimated regression equation.

This ratio is called the coefficient of determination and is denoted by r^2 .

$$r^2 = \frac{SSR}{SST}$$

For the example of Best Burger restaurants, the value of the coefficient of determination is

$$r^2 = \frac{SSR}{SST} = \frac{14200}{15730} = 0.9027$$

When the coefficient of determination is expressed as a percentage, we can r^2 can be interpreted as the percentage of the total sum of squares that can be explained using the estimated regression equation. For Best Burger Restaurants we can conclude that 90.27% of the total sum of squares can be explained using the estimated regression equation $y = 60 + 5x$ to predict quarterly sales. In other words, 90.27% of the variability in sales can be explained by a linear relationship between the size of the student population and sales. We should be pleased to see that it fits the estimated regression equation so well.

5.7 Correlation coefficient

The correlation coefficient can be thought of as a descriptive measure of the strength of the linear relationship between two variables, x and y. The values of the correlation coefficient are always between -1 and +1. A value of +1 means that the two variables $x\ in\ y$ are perfectly correlated in a positive linear sense. This means that all data points on a are a straight line with a positive slope. A value of -1 means that $x\ in\ y$ perfectly related in a negative linear sense, with all data points on a straight line having a negative slope. Correlation coefficient values close to zero mean that x and y are not linearly related.

If a regression analysis has already been carried out and the coefficient of determination r^2 has been calculated, the sample correlation coefficient can be calculated as follows.

$$r_{xy} = (sign\ of\ b_1)\sqrt{coefficient\ of\ determination}$$

$$r_{xy} = (sign\ of\ b_1)\sqrt{r^2}$$

PEARSON CORRELATION COEFFICIENT: SAMPLE DATA

$$rxy = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

$$rxy = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Where they are:

$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$, $s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$, $s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$

The sign for the sample correlation coefficient is positive if the estimated regression equation has a positive slope ($b_1 > 0$) and negative if the estimated regression equation has

negative slope ($b_1 < 0$).

For the Best Burger case, the value of the coefficient of determination corresponding to the estimated regression equation $y = 60 + 5x$ is 0.9027. Because the slope of the estimated regression equation is positive, equation (14.13) shows that the sample correlation coefficient is By the sample correlation coefficient

Rxy=0.9501, we would conclude that a strong positive linear relationship exists between $x\ in\ y$.

In the case of a linear relationship between two variables, both coefficients of determination and the sample correlation coefficient provide a measure of the strength of the relationship.

The coefficient of determination provides a measure between zero and one, while the sample correlation coefficient provides a measure between -1 and +1. Although the sample correlation coefficient is limited to a linear relationship between two variables, the coefficient of determination can be applied to non-linear relationships and to relationships that have two or more independent variables. Thus, the coefficient of determination provides a wider range of applicability.

5.8 Multiple Regression Model

In the following sections, we continue our study of regression analysis by considering situations involving two or more independent variables. This subject area, called multiple regression analysis, allows us to take more factors into account and thus obtain better predictions than are possible with simple linear regression.

Multiple regression analysis is the study of how the dependent variable y is related to two or more independent variables. In the general case, we will denote by p the number of independent variables.

5.9 Regression model and regression equation

The concepts of regression model and regression equation introduced in the previous section apply in the case of multiple regression. The equation describing how the dependent variable y is related to the independent variables $x_1, x_2, \ldots, x_p$ and the error term is called a multiple regression model. We start by assuming that the multiple regression model has the following form.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

In a multiple regression model $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters and the error term (ϵ) is a random variable. A close examination of this model reveals that y is a linear function of the variables $x_1, x_2, \ldots, x_p$ plus the error term ϵ epsilon. The error term takes into account the variability $y$which cannot be explained by the linear effect of p independent variables.

In section 5.10 we discuss the assumptions for the multiple regression model and epsilon. One of the assumptions is that the mean or expected value (ϵ) is zero. The implication of this assumption is that the mean or expected value of y, is denoted by $E(y)$, is equal to $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. The equation describing how the mean value is y is related to $x_1, x_2, \ldots, x_p$ is called a multiple regression equation.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

5.10 Estimated multiple regression equation

If the values are $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are known, equation (5.9) can be used to calculate the average value of y at given values of $x_1, x_2, \ldots, x_p$. Unfortunately, these parameter values will generally not be known and must be estimated from the sample data. A simple random sample is used to calculate the sample statistic $b_0, b_1, b_2, \ldots, b_p$to be used as point estimators of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. These sample statistics provide the following multiple regression equation estimation:

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

Where they are

$b_0, b_1, b_2, \ldots, b_p$ are estimates $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

$\hat{y}$ = the predicted value of the dependent variable

Example: Frigo Transport Company

As an illustration of multiple regression analysis, we will consider the problem faced by Frigo Trucking Company, an independent trucking company in Southern Italy. The largest part of Frigo's business involves deliveries throughout the local area. For better development work schedules, managers want to plan a common daily travel time for their drivers.

Initially, managers believed that the total daily journey time would be closely linked to the number of km travelled in daily deliveries. A simple random sample of 10 driver assignments provided the data shown in Figure 5.9 and a scatter plot. After reviewing this scatter diagram, the Managers assumed that a simple linear regression model could be used to describe the relationship $y = \beta_0 + \beta_1 x_1 + \epsilon$ between total journey time (y) and the number of km travelled (x_1).

Driving Assignment	x_1 = KM Traveled	y = Travel Time (hours)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

Figure 5.9 Data for Frigo Transport Company example.

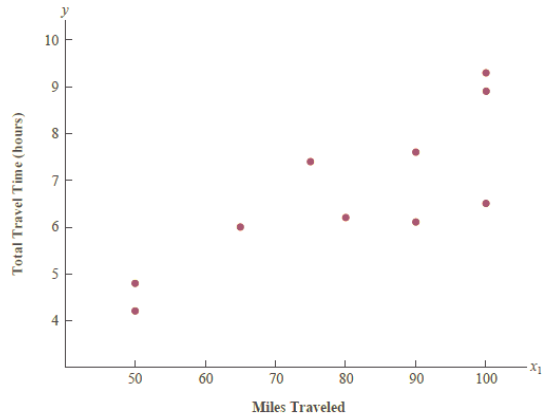

Figure 5.10 Scatter plot for Frigo Transport Company example.

To estimate the parameters β_0 and β_1 the least squares equation method was used to develop the estimated regression.

$$\hat{y} = b_0 + b_1 x_1$$

Figure above shows the output of Minitab using simple linear regression to the data in Table above. The estimated regression equation is

$$\hat{y} = 1.27 + 0.0678 x_1$$

At the 0.05 significance level, an F-value of 15.81 and a corresponding p-value of 0.004 indicate that the relationship is significant. This means that we can reject $H_0: \beta_1 = 0$ because the p-value is less than $\alpha = 0{,}05$. Note that the same conclusion follows from the value of $t = 3{,}98$ and the associated p-value of 0.004. Thus, we can conclude that the relationship between total journey time and number of miles travelled is significant. Longer journey times are associated with more kilometres travelled. With the coefficient of determination (expressed as a percentage) $R - Sq = 66{,}4\ \%$, we see that 66.4% of the variability in travel time can be explained by a linear effect of the number of miles travelled.

This finding is quite good, but managers may want to consider adding a second independent variable to explain some of the remaining variables in the dependent variable.

MINITAB OUTPUT FOR FRIGO TRUCKING WITH ONE INDEPENDENT VARIABLE

```
The regression equation is
Time = 1.27 + 0.0678 kM

Predictor      Coef  SE Coef     T      p
Constant      1.274    1.401  0.91  0.390
kM          0.06783  0.01706  3.98  0.004

S = 1.00179   R-Sq = 66.4%   R-Sq(adj) = 62.2%

Analysis of Variance

SOURCE           DF      SS      MS      F      p
Regression        1  15.871  15.871  15.81  0.004
Residual Error    8   8.029   1.004
Total             9  23.900
```

Figure 5.11 Results with one independent variable.

When trying to identify the second independent variable, managers considered that the number of deliveries may also contribute to the total journey time. The Frigo Trucking data, with the number of deliveries added, is shown in Figure below. (x_1) and the number of deliveries (x_2), as independent variables, is shown in Figure 5.12. The estimated regression equation is

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$$

DATA FOR FRIGO TRUCKING WITH KM TRAVELED (x_1) AND NUMBER OF DELIVERIES (x_2) AS THE INDEPENDENT VARIABLES

Driving Assignment	x_1 = kM Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Figure 5.12 Frigo Trucing data and independent variables.

Let's take a closer look at the values b_1= 0.0611 and b_2= 0.923 in last equation.

Note on the interpretation of the coefficients

At this point we can make one comment on the relationship between the estimated regression equation with only miles travelled as the independent variable and the equation including the number of deliveries as the other independent variable. Value b_1 is not the same in both cases. In a simple linear regression we interpret b_1 as an estimate of the change y for a one-unit change in the independent variable. In multiple regression analysis this interpretation needs to be modified slightly. That is, in multiple regression analysis, each regression coefficient is interpreted as follows: would represent an estimate of the change in the y corresponding to the change in x_i by one unit when all other independent variables are held constant.

In the case of Frigo Trucking, that involves two independent variables, b_1=0.0611 and b_2= 0.923.

MINITAB OUTPUT FOR FRIGO TRUCKING WITH TWO INDEPENDENT VARIABLES

```
The regression equation is
Time = - 0.869 + 0.0611 kM  + 0.923 Deliveries

Predictor       Coef    SE Coef      T      p
Constant     -0.8687     0.9515  -0.91  0.392
kM          0.061135   0.009888   6.18  0.000
Deliveries    0.9234     0.2211   4.18  0.004

S = 0.573142   R-Sq = 90.4%   R-Sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       2  21.601  10.800  32.88  0.000
Residual Error   7   2.299   0.328
Total            9  23.900
```

Figure 5.13 Results for Frigo Trucking with two independent variables.

Thus, 0.0611 hours is an estimate of the expected increase in travel time corresponding to an increase in one mile per distance travelled when the number of deliveries is constant. Similarly, since b_2 =0.923, the estimate of the expected increase in journey time corresponding to an increase of one delivery when the number of miles travelled is constant is 0.923 hours.

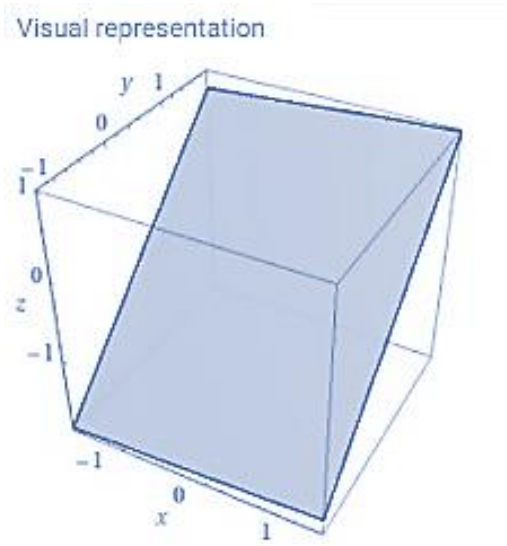

Figure 5.14 Visual representation of results for Frigo Trucking case.

References Chapter 5

- *Introductory Statistics.* Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/97898151231351230101

- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);

- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);

- Journal of the Royal Statistical Society 2024**,** A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.

- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011

- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014

- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®

- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved

6. Introduction to Operations Research

Operations research (British English: operational research, U.S. Air Force Specialty Code: Operations Analysis), often shortened to the initialism OR, is a discipline that deals with the development and application of analytical methods to improve decision-making. The term management science is occasionally used as a synonym.

OR methods are used to analyze resource capacities, bottlenecks, lead and cycle times, demand patterns, inventory, resource distribution, maintenance, operator dispatching, product mixing, product output, reliability, resource utilization, rules and policies, schedule and dispatch efficiency, process throughput, etc. (Ueda, 2010)

Employing techniques from other mathematical sciences, such as modeling, statistics, and optimization, operations research arrives at optimal or near-optimal solutions to decision-making problems. Because of its emphasis on practical applications, operations research has overlapped with many other disciplines, notably industrial engineering and logistics, making it an integral part of their Knowledge Management Systems (KMS).

6.1 Strategic logistics planning

According to (Robinson, 2004) operations research mainly deals with the interaction among planned technical and human systems. They are characterized by variability, interdependence among components and structural as well as behavioral complexity. To manage these characteristics a wide array of methods has been devised to appropriately address them as a whole. They are used in their strategic planning to:

- enable complex what-if analysis;
- manage complexity: interdependence + variability + dynamics;
- involve less costs and interference with the process than by experimentation with a real system;

- focus on details;
- improve the understanding of a system;
- improve communication between management and experts.

Strategic logistics planning (Figure 6.1) comprises all activities which need to be performed on the strategic, tactical and operational levels in order to provide for Total quality management (TQM) (Ciampa, 1992) and Just-in-time production (JIT) (Britannica, 2023).

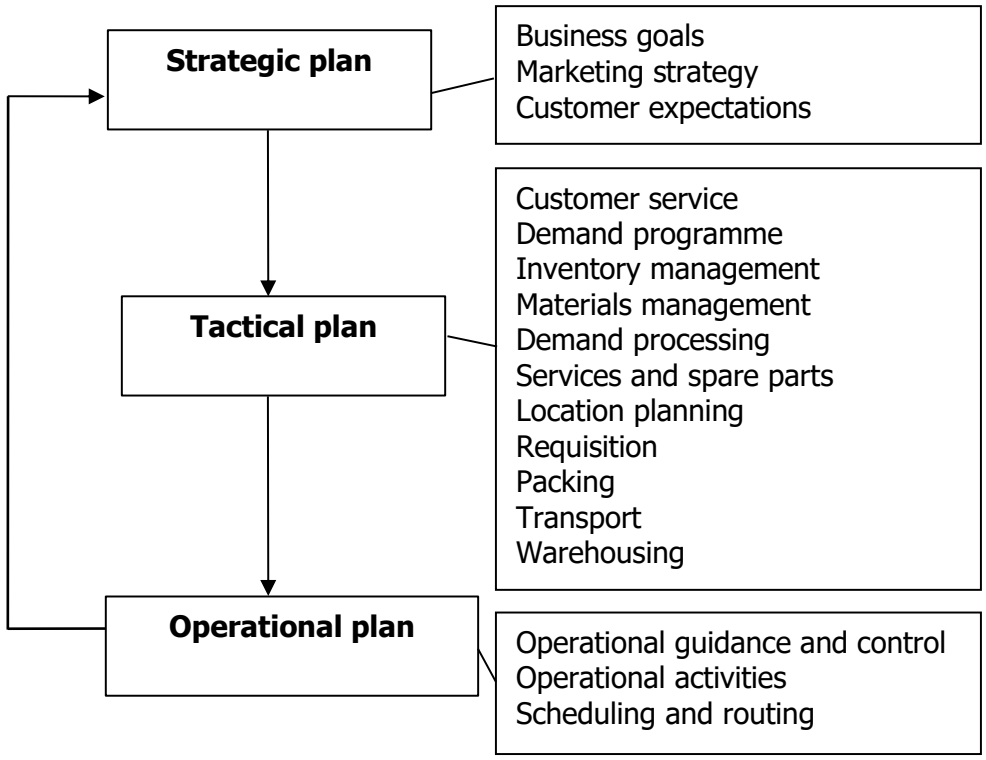

Figure 6.1 Strategic logistics planning.

6.2 Six-Sigma

In strategic logistics planning the SCOR reference model (AIMS, 2021) helps businesses evaluate and perfect supply chain management for reliability, consistency, and efficiency. It recognizes 6 major business processes — Plan, Source, Make, Deliver, Return and Enable.

The SCOR planning process comprises all activities associated with developing plans for supply chain management and improvement. Continuous efforts to achieve stable and predictable

process results by reducing process variation (6-sigma) are of vital importance to business success.

DMAIC and DMADV explained

The manufacturing processes (sourcing, making, delivering, enabling, as well as handling returns) have characteristics that can be defined, measured, analyzed, improved, and controlled. Hence, these phases constitute the production process management methodology, abbreviated as DMAIC.

Some practitioners have combined 6-sigma ideas with lean manufacturing to create a methodology named Lean Six Sigma (Wheat & Mills & Carnell, 2003). The Lean Six Sigma methodology considers lean manufacturing ("just-in-time" production), which addresses process efficiency, and 6-sigma, with its focus on reducing variation and waste, as complementary disciplines that promote business and operational excellence.

The DMADV methodology (define, measure, analyze, design and verify), also known as DFSS ("Design for Six Sigma"), is consistent with KBE ("Knowledge Based Engineering"). DFSS methodology's (Chowdhury, 2002) phases are:

1. Define design goals that are consistent with customer demands and the enterprise strategy.
2. Measure and identify characteristics that are Critical to Quality (CTQ), measure product capabilities, production process capacity, and measure risks.
3. Analyze to develop and design alternatives.
4. Design an improved alternative, best suited per analysis in the previous step.
5. Verify the design, set up pilot runs, implement the production process and hand it over to the process owner(s).

Six Sigma (Tennant, 2001) business improvement projects, inspired by W. Edwards Deming's "Plan–Do–Study–Act" cycle (Tague, 2005), depending on their nature, follow either of the aforementioned methodologies, each with five phases:

1. DMAIC is used for projects aimed at improving an existing business process.
2. DMADV is used for projects aimed at creating new products or process designs.

6.3 Business intelligence

Business intelligence (BI) comprises all strategies and technologies used by enterprises for the data analysis of past and current business information (Tableau, 2023). It is supported by Knowledge Management Systems (KMS) that represent the part of Logistics Information Systems (LIS) that enables experts in different fields to advise and provide support to different levels of management.

Business analytics

Business analytics (BA) is a process based on BI, enabling new insights into the business process and better strategic decision making for the future. It originates from Data Mining (DM) being the process of finding anomalies, patterns, and correlations in larger data sets, to predict the results.

The BA process is composed of:

1. Data Aggregation: prior to analysis, data must first be gathered, organized, and filtered, either through volunteered data or transactional records.
2. Data Mining: data mining sorts through large datasets using databases, statistics, and machine learning to identify trends and establish relationships.
3. Association and Sequence Identification: the identification of predictable actions that are performed in association with other actions or sequentially.
4. Text Mining: explores and organizes large, unstructured text datasets for the purpose of qualitative and quantitative analysis.
5. Forecasting: analyses historical data from a specific period in order to make informed estimates that are predictive in determining future events or behaviors.
6. Predictive Analytics: predictive business analytics uses a variety of statistical techniques to create predictive models, which extract information from datasets, identify patterns and provide predictive score for an array of organizational outcomes.

7. Optimization: once trends have been identified and predictions have been made, businesses can engage simulation techniques to test best-case scenarios.
8. Data Visualization: provides visual representations such as charts and graphs for easy and quick data analysis.

Sales and operations planning

Sales and operations planning (SOP) is a flexible tool to forecast and plan production activities. SOP steps:

1. sales plan,
2. production plan and
3. capacity planning.

SOP operates on data from different information sources throughout the company: Sales, Marketing, Production, Accounting, Human resources and Requisition. They are usually provided by the corresponding departments through the company's enterprise resource planning (ERP) system.

While SOP operates on the strategic level, ERP operates on the tactical level of a company's logistics information system. They are joined by the Demand Management program, linking the strategic sales and operations planning (SOP) and detailed production planning (Master Production Scheduling / Material Requirements Planning) on the operational level. Here the previously mentioned scheduling and simulation come into play to render a feasible and optimal production plan.

Demand management program (Figure 6.2) is composed of two types of forecasts:

1. planned independent requirements (PIR) from projected sales volumes based on marketing and
2. customer independent requirements (CIR) from data based on existing and planned sales orders.

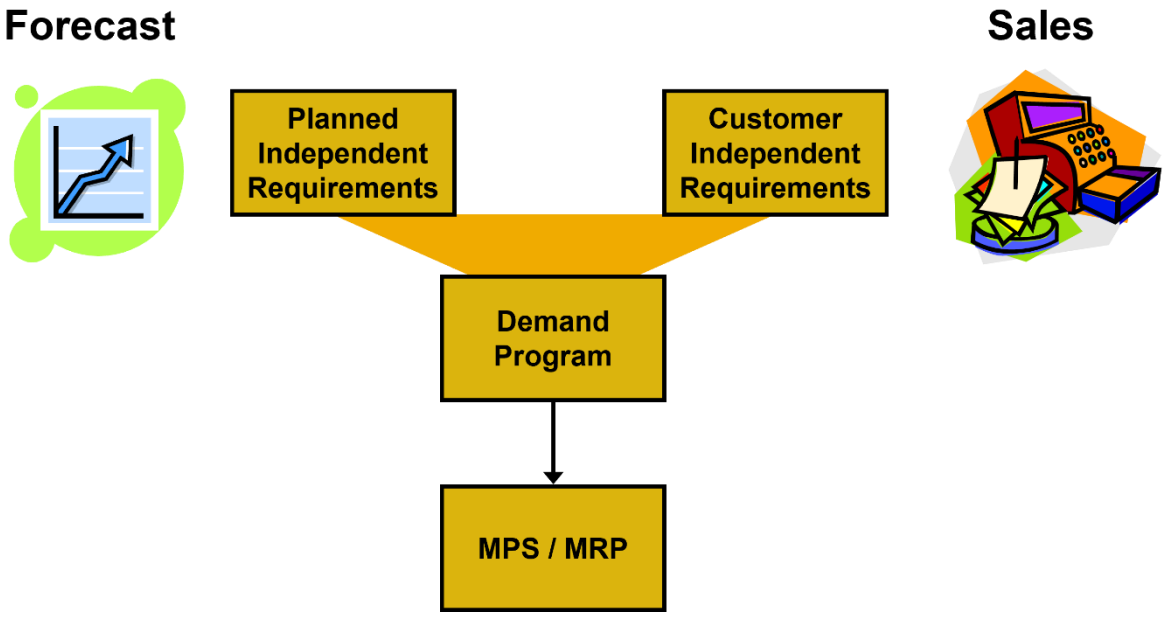

Figure 6.2 Demand management.

SOP Example

A company trades several types of products across its sales network. Sales transactions are stored centrally to be able to monitor stock status as well as perform sales analysis for demand management. They are logged in CSV (Comma Separated Values) format, which is easy to process by the company's ERP system as well as by the analytics department, which uses spreadsheets.

In sales analysis the sales transactions are initially filtered to determine whether they are complete and correctly formatted. Only then are they ready to be statistically assessed, since otherwise any missing or ill-formed data could result in misinterpretation of the results.

Date	Seller ID	Customer ID	Transaction ID	Product ID	Product Price
1.10.24	1	12	1	101	195,00 €
1.10.24	1	12	1	102	45,00 €
1.10.24	1	12	1	103	35,00 €
1.10.24	2	14	2	104	55,00 €
1.10.24	2	14	3	101	195,00 €
2.10.24	3	15	4	105	85,00 €
2.10.24	3	15	4	101	195,00 €
2.10.24	3	15	4	103	35,00 €
2.10.24	3	16	5	104	55,00 €
2.10.24	1	17	6	101	195,00 €
2.10.24	1	17	6	102	45,00 €
2.10.24	1	17	6	105	85,00 €
3.10.24	2	18	7	106	35,00 €
3.10.24	2	18	7	107	65,00 €
3.10.24	2	18	7	108	86,00 €
3.10.24	4	19	8	105	85,00 €
3.10.24	4	19	8	101	195,00 €
3.10.24	4	19	8	103	35,00 €
3.10.24	4	19	9	104	55,00 €
4.10.24	5	20	10	105	110,00 €
4.10.24	5	20	10	106	125,00 €
4.10.24	5	20	10	104	55,00 €
4.10.24	5	20	10	101	195,00 €
4.10.24	1	21	11	102	45,00 €
4.10.24	1	21	11	105	85,00 €
4.10.24	1	21	12	106	35,00 €
4.10.24	3	12	13	103	35,00 €
4.10.24	3	12	13	104	55,00 €
4.10.24	3	12	13	105	110,00 €
4.10.24	3	12	13	101	195,00 €
5.10.24	1	22	14	107	35,00 €
5.10.24	1	22	14	108	25,00 €

Figure 6.3 Weekly sales data.

Usually by analytics, various sensible insights into the collected "raw data" (Figure 6.3) are enabled. One can achieve this with pivot tables that enable to group data by chosen attributes and perform statistical analysis. In principle, any attribute (column) of input data can be considered a pivot. Hence, we are often speaking of a "data cube" of multiple dimensions. Since we cannot graphically represent more than two or three dimensions, the simplest, yet usually the most useful, representations of input data are formed from two or three pivot attributes. Based on the given data sample, some examples of pivot tables are given in the sequel.

Date	Vsota - Product Price
1.10.24	525,00 €
2.10.24	695,00 €
3.10.24	556,00 €
4.10.24	1.045,00 €
5.10.24	875,00 €
6.10.24	605,00 €
7.10.24	556,00 €
Skupaj Rezul	**4.857,00 €**

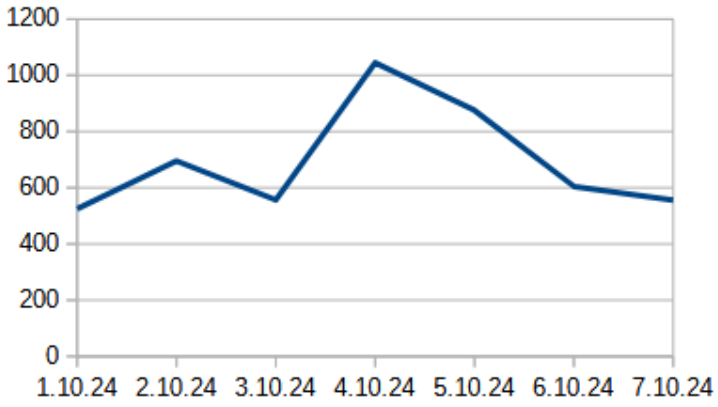

Figure 6.4 Sales statistics by weekday.

Sales statistics by weekday or month (Figure 6.4) enable insights into seasonal trends.

Seller ID	Vsota - Product Price
1	1.110,00 €
2	1.072,00 €
3	1.585,00 €
4	605,00 €
5	485,00 €
Skupaj Rezul	**4.857,00 €**

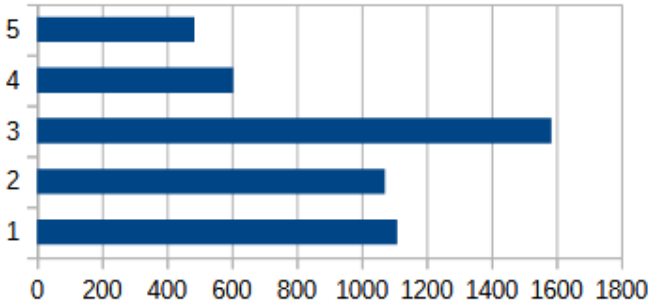

Figure 6.5 Sales statistics by sales-office.

Sales statistics by the sales office (Figure 6.5) determine which offices are the busiest and/or create the most revenue.

Product ID	Vsota - Product Price
101	1.950,00 €
102	180,00 €
103	245,00 €
104	440,00 €
105	925,00 €
106	425,00 €
107	165,00 €
108	197,00 €
109	35,00 €
110	95,00 €
111	75,00 €
112	125,00 €
Skupaj Rezult	**4.857,00 €**

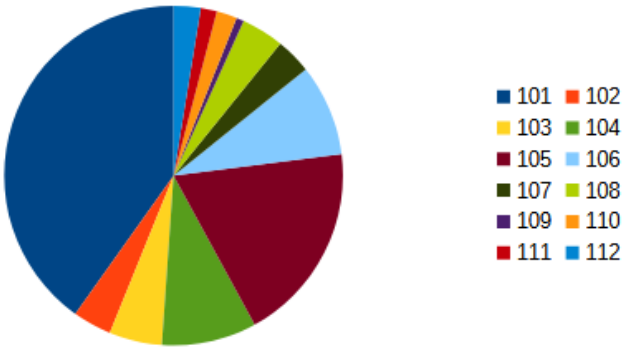

Figure 6.6 Sales statistics by product.

Sales statistics by product (Figure 6.6) determine the products that are most sought for or represent a significant share in the portfolio.

Date	Seller ID	Customer ID	Transaction ID	Product ID	Vsota - Produ
1.10.24	1	12	1	101	195,00 €
				102	45,00 €
				103	35,00 €
	2	14	2	104	55,00 €
			3	101	195,00 €
2.10.24	1	17	6	101	195,00 €
				102	45,00 €
				105	85,00 €
	3	15	4	101	195,00 €
				103	35,00 €
				105	85,00 €
		16	5	104	55,00 €
3.10.24	2	18	7	106	35,00 €
				107	65,00 €
				108	86,00 €
	4	19	8	101	195,00 €
				103	35,00 €
				105	85,00 €
			9	104	55,00 €

Figure 6.7 Transactions' overview.

Transaction's overview (Figure 6.7) by day, sales-office, transaction and buyer offer a structured insight into the data which is useful when solving inquiries on a certain product or sales transaction.

6.4 Decision support systems

A decision support system (DSS) is an interactive system, which assists the decision makers to use the data and models for solving unstructured or partly structured problems. An expert System (ES) is an application program or environment, which effectively supports problem solving in a specialized problem area, requiring expert knowledge and skills.

In addition to statistical and simulation methods used in BA, DSS often includes models that enable decision making, based on a set of distinctive criteria. These models may be simple (e.g., decision tables, trees, etc.) leading to a single correct solution. On the other hand, when dealing with multiple (possibly conflicting) criteria and multiple solutions, a more elaborate

model is necessary. An appropriate model often used in professional and personal life is the multi-criteria decision model.

Multi criteria decision making

Multi criteria decision making (MCDM), or multiple-criteria decision analysis (MCDA) is a sub-discipline of OR that explicitly evaluates multiple (possibly) conflicting criteria in decision making over a set of candidate solutions or variants.

A MCDM decision model is composed of:

- Criteria – the parameters of input variants, critical for our design.
- Weighs – the relative importance of selected criteria.
- Utility function – the function that combines the weighted parameters of variants into a fitness value.
- Data – the data representing the variants; input data to our MCDM model.

Data types in MCDM may be:

- Quantitative – representing values that can be compared as such.
- Qualitative – representing relative comparison values (e.g., high, comfortable, low temperature, etc.) that need to be quantified to provide unique values.
- Binary – representing binary criteria; a property being fulfilled (1) or not (0).

The procedure of multi-criteria decision making:

1. Representation of variants (V) by their characteristic parameters (P): $\{V_i (P_{i,1}; P_{i,2}; \dots P_{i,n}); i=1...m\}$.
2. Normalization of parameters by calculating the relative local grade $p_{i,j}$ for each $P_{i,j}$ (j=1...n), with reference to the j^{th} parameter maximum $P_{i,j}$ value from all i samples:
   a) $p_{i,j}=P_{i,j}/\max \{P_{i,j}\}$ if a greater value of $P_{i,j}$ is more beneficial.
   b) $p_{i,j}=1 - P_{i,j}/\max \{P_{i,j}\}$ if a smaller value of Pi,j is more beneficial.
3. The grades are weighted according to preferences: $x_{i,j}=p_{i,j}*U_j$ for each j=1...n, by weighs U_j which need to sum-up to 1, i.e. 100%.

4. The weighted grades of all variants are summed up: $X_i=\sum x_{i,j}$ for each i=1...m to obtain composite grades according to our utility function.

5. The best variant is chosen $Y=\max\{X_i\}$.

MCDM Example

When choosing new equipment in a company, we must often perform multi-criteria decision making. Let us consider an example of choosing the most cost-effective (below 300 EUR) mobile platform with the Android operating system for our company. In Table 6.1 the specimens, which have been chosen from an inquiry among employees to narrow down our assortment, are listed. For each of them the parameters which have been selected as most relevant, are given. In the sequel the selected parameter data are normalized to obtain comparable values, weighted to emphasize the values that are more significant and summed up to obtain grades for the selected specimens.

Table 6.1 MCDM for a cost-effective Android mobile platform.

PARAMETERS									
	Price (€)	Grade*	Performance			Properties			Camera
Model		(1-10)	proc.speed (GHz)	RAM (GB)	int.mem. (GB)	weight (g)	size (mm3)	bat.cap. (mAh)	(MP)
Honor Magic Lite 5	263 €	2	2,2	6	128	175	94344	5100	64
Honor X7a	210 €	2,5	2,3	4	128	196	106442	6000	50
Samsung A34	297 €	1,8	2,6	8	256	199	103300	5000	48
Redmi Note 12 Pro	240 €	1,9	2,6	6	128	187	99104	5000	50
Redmi Note 12 S	224 €	1,7	2,05	8	256	176	95715	5000	108

*Vir: www.testberichte.de

PARAMETER WEIGHS									
	Price (€)	Grade	Performance			Properties			Camera
			proc.speed (GHz)	RAM (GB)	int.mem. (GB)	weight (g)	size (mm3)	bat.cap. (mAh)	(MP)
			10 %	5 %	5 %	10 %	10 %	10 %	
Utež	20 %	10 %	20 %			30 %			20 %

NORMALIZED PARAMETERS									
	Price (€)	Grade*	Performance			Properties			Camera
Model		(1-10)	proc.speed (GHz)	RAM (GB)	int.mem. (GB)	weight (g)	size (mm3)	bat.cap. (mAh)	(MP)
Honor Magic Lite 5	0,11	0,20	0,85	0,75	0,50	0,12	0,11	0,85	0,59
Honor X7a	0,29	0,00	0,88	0,50	0,50	0,02	0,00	1,00	0,46
Samsung A34	0,00	0,28	1,00	1,00	1,00	0,00	0,03	0,83	0,44
Redmi Note 12 Pro	0,19	0,24	1,00	0,75	0,50	0,06	0,07	0,83	0,46
Redmi Note 12 S	0,25	0,32	0,79	1,00	1,00	0,12	0,10	0,83	1,00

FINAL PARAMETER ASSESSMENT									
	Price (€)	Grade*	Performance			Properties			Camera
Model		(1-10)	proc.speed (GHz)	RAM (GB)	int.mem. (GB)	weight (g)	size (mm3)	bat.cap. (mAh)	(MP)
Honor Magic Lite 5	0,02	0,02	0,08	0,04	0,03	0,01	0,01	0,09	0,12
Honor X7a	0,06	0,00	0,09	0,03	0,03	0,00	0,00	0,10	0,09
Samsung A34	0,00	0,03	0,10	0,05	0,05	0,00	0,00	0,08	0,09
Redmi Note 12 Pro	0,04	0,02	0,10	0,04	0,03	0,01	0,01	0,08	0,09
Redmi Note 12 S	0,05	0,03	0,08	0,05	0,05	0,01	0,01	0,08	0,20

The end result of our analysis is the summary table (Table 6.2) and possibly a graph (Figure 6.8) which summarizes the decision-making process and presents the best choice as well as the strengths and weaknesses of individual variants. Often-times the specimen with the best grade is not the one that performed best in all categories but the one that on average best

matches our selection criteria as well as weighing of the parameters. This is also the main strength of the MCDM method, since the choice by any single parameter might mislead us.

Table 6.2 MCDM summary for our mobile platform selection example.

FINAL PARAMETER ASSESSMENT						
Model	**Price**	**Grade**	**Performance**	**Properties**	**Camera**	**TOTAL:**
Honor Magic Lite 5	0,02	0,02	0,15	0,11	0,12	0,42
Honor X7a	0,06	0,00	0,14	0,10	0,09	0,39
Samsung A34	0,00	0,03	0,20	0,09	0,09	0,40
Redmi Note 12 Pro	0,04	0,02	0,16	0,10	0,09	0,41
Redmi Note 12 S	0,05	0,03	0,18	0,10	0,20	0,56

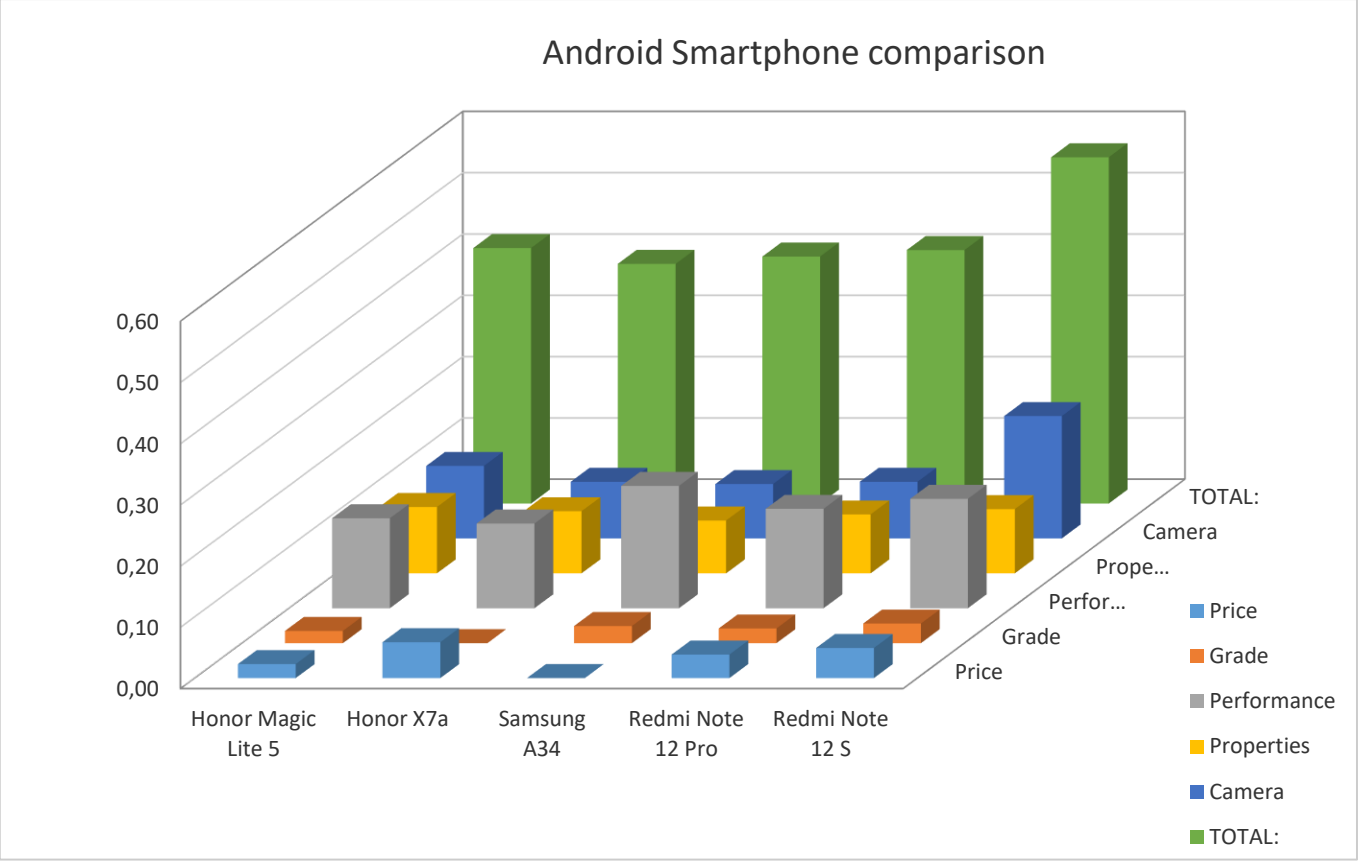

Figure 6.8 The choice of the best mobile platform.

Best choice: 0,56 (Redmi Note 12 S).

6.5 Knowledge based engineering

Knowledge Based Engineering (KBE) is an engineering methodology to integrate engineering knowledge systematically into the design system (Andersson & Larsson & Ola, 2011).

Lifecycle of experience management

The need to capture, manage, and utilize design knowledge and automate processes unique to a manufacturer's product development experience has led to the development of knowledge-based Engineering (KBE) technology (Prasad, 2005). KBE is meant to enrich institutional knowledge by experience management. The phases of experience management, according to (Andersson & Larsson & Ola, 2011) are:

1. Identify: a non-conformance with the desired state that appears in the manufacturing process due to an ill-defined product or process is selected.
2. Capture: the experience with its properties is captured.
3. Analyze: a root cause analysis of the captured experience is made to identify an appropriate remedy strategy and its re-use to prevent recurring anomalies.
4. Store: insights from the analysis are archived with the experience.
5. Search & Retrieve: the experience is searched for and retrieved.
6. Use: an element of the experience is used.
7. Reuse: concludes the cycle of knowledge management and starts a new one.

Knowledge based engineering is in general supplemented by further disciplines, whose closer consideration outreaches the scope of this chapter:

- Computer aided project management (PS).
- Computer aided design (CAD), production (CAM) and robotics (CIM).
- Computer simulation modeling and analysis (SMA).
- Computer aided detailed production planning (MPS / MRP).

6.6 Conclusion

As presented in this chapter, the main applications of BI in corporate management relate to business analytics (BA) and decision support systems (DSS). They are commonly termed operations research (OR). In addition to the underlying BI techniques, described in this chapter, in chapters 3 and 4 on data management and simulation modeling and analysis (SMA) some additional considerations on data collection, manipulation and presentation, which also support decision making, are given. In summary BI applications are found in Knowledge Management Systems (KMS), comprising:

- Decision support systems (DSS),
- Business analytics (BA) as an upgrade to Data Mining (DM) and
- Knowledge-based Engineering (KBE) as an upgrade to Computer-aided Engineering (CAE).

Based on BA, DSS and SMA results, experiences, which enhance institutional knowledge and constitute their knowledge-based expert systems (KBS), are devised. As demonstrated in (Gumzej et. al., 2023), they can be employed by KBE to introduce the "lessons-learned" principle into enterprise improvements management by strategic logistics planning.

References Chapter 6

- AIMS (2021). SCOR - Supply Chain Operations Model. [available at: https://aims.education/study-online/supply-chain-operations-reference-model-scor/, access June 20, 2023]
- Ciampa, D. (1992). Total Quality: A User's Guide for Implementation, Addison-Wesley.
- Britannica, T. Editors of Encyclopaedia (2023). Just-in-time manufacturing, Encyclopedia Britannica. [available at: https://www.britannica.com/topic/just-in-time-manufacturing, access June 20, 2023]
- Tennant, G. (2001). SIX SIGMA: SPC and TQM in Manufacturing and Services, Gower Publishing, Ltd.
- Wheat B. & Mills C. & Carnell M. (2003). Leaning into Six Sigma: a parable of the journey to Six Sigma and a lean enterprise, McGraw-Hill.

- Tague, N.R. (2005). Plan–Do–Study–Act cycle, The quality toolbox (2nd ed.), ASQ Quality Press, pp. 390–392.
- Chowdhury, S. (2002). Design for Six Sigma: The revolutionary process for achieving extraordinary profits, Prentice Hall,
- Ueda, M. (2010). How to Market OR/MS Decision Support. International Journal of Applied Logistics (IJAL), 1(2), 23-36.
- Tableau (2023). Comparing Business Intelligence, Business Analytics and Data Analytics. [Available from: https://www.tableau.com/learn/articles/business-intelligence/bi-business-analytics, access June 20, 2023]
- Prasad, B. (2005). Knowledge Technology, What Distinguishes KBE From Automation. COE NewsNet – June 2005. [available at: https://web.archive.org/web/20120324223130/http://legacy.coe.org/newsnet/Jun05/knowledge.cfm, access June 20, 2023]
- Robinson, S. (2004). Simulation: The Practice of Model Development and Use, Wiley.
- Gumzej, R., Kramberger, T., Dujak, D. (2023). A knowledge base for strategic logistics planning. In: Dujak, Davor (edt.). Proceedings of the 23rd International Scientific Conference Business Logistics in Modern Management: October 5-6, 2023, Osijek, Croatia. Osijek: Josip Juraj Strossmayer University of Osijek, Faculty of Economics and Business, pp. 317-330, illustr. Business logistics in modern management (Online). ISSN 1849-6148. https://blmm-conference.com/past-issues/.
- Andersson, P. & Larsson, T. & Ola, I. (2011). A case study of how knowledge based engineering tools support experience re-use. In Research into Design – Supporting Sustainable Product Development, Chakrabarti, A. (Edt.), Research Publishing, Indian Institute of Science, Bangalore, India.

7. Statistical data processing SPSS

By now, you have already acquired a fundamental understanding of statistics, data manipulation, simulation establishment, modelling, and analysis within logistics supply chains, along with straightforward linear regression methods. While statistics offers a diverse range of

models and techniques to enhance your optimization efforts, conduct analysis, and identify potential enhancements, you may have observed that as the complexity of the analyzed data and calculations grows, traditional approaches can become increasingly intricate and challenging to compute. As the intricacy of your data and computations escalates, conventional methods may become outdated and, in some cases, compromise the reliability of your results. To address this, statistics uses various software programs that automate the analysis and interpretation of collected data while also providing a multitude of models and functions to ensure reliable outcomes. One such software is IBM's SPSS, which will be a key tool in this chapter. In this chapter, we will provide a concise introduction to the primary usage of SPSS software, exploring its functionalities and practical applications. The initial introduction will be followed by the practical application of the program through four fundamental tests for result calculation: the T-test, correlations, Chi-Square, and ANOVA. To facilitate your learning, we will present simple problems and their solutions to help you become acquainted with these tests.

7.1 Basics of IBM`s SPSS

You may have already had some experience with SPSS software. However, if you still need to, let us offer a brief introduction to it. SPSS, much like its more widely recognized counterpart, Excel, facilitates data manipulation, analysis, and visualization. Nevertheless, unlike Excel, which can sometimes be laborious and complex in function programming, SPSS offers a user-friendly interface for statistical analysis (IBM, 2021). It offers a variety of functions and methodologies to handle your data provided efficiently. While the SPSS software excels in providing extensive statistical analysis capabilities, navigating data manipulation and configuring initial settings for analysis can sometimes be challenging (IBM, 2021). Therefore,

we will delve into the fundamentals of data importation and data preparation for subsequent statistical tests.

Given the widespread use of Excel for handling numerical data, your data may be obtained or prepared in an Excel spreadsheet. Fortunately, SPSS can import data from various file formats into its spreadsheets. Once you have your finalized Excel spreadsheets ready, open the SPSS software. On the initial screen, navigate to the "File" tab and select "Import Data". In the subsequent window, you can choose the data format you intend to import (refer to the figure 7.1). Following this step, locate your prepared file, select it, and proceed to the next window. This window will prompt you to configure additional settings. If you have already included column names in the first row of your data, opt for the "Read variable names from the first row of data" (refer to figure 7.1), and then click "Finish" to have the data appear in the spreadsheet (IBM, 2021).

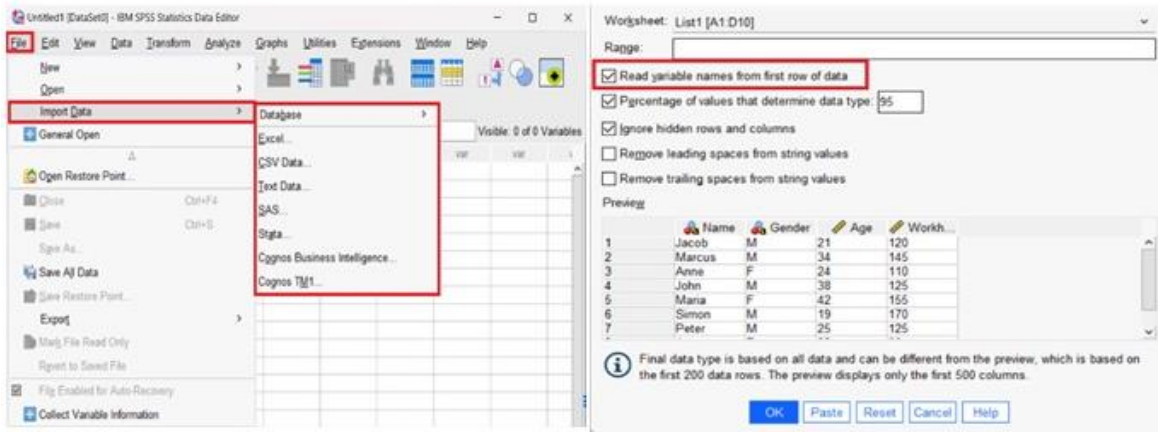

Figure 7.1 SPSS Import settings.

Now that we have the data in our spreadsheet, you will notice a distinct difference in presentation compared to Excel. SPSS categorizes data into two primary types, each with two additional sub-types. As illustrated in figure 7.2, data can be classified as numerical or categorical. Numerical data consists of numbers and can be categorized as discrete (with finite options) or continuous (offering infinite options). On the other hand, categorical data comprises words and can be further distinguished as ordinal (having a hierarchy) or nominal (lacking a hierarchy). Depending on the nature of your data, you may need to configure the variables to align with your desired analysis. In most cases, SPSS will automatically categorize variables appropriately. Suppose you wish to perform further manipulation of data types. In that case,

you can access the "View" option and, under "Variable View", adjust variable information such as Name, Type, Width, Measure, and more (IBM, 2021).

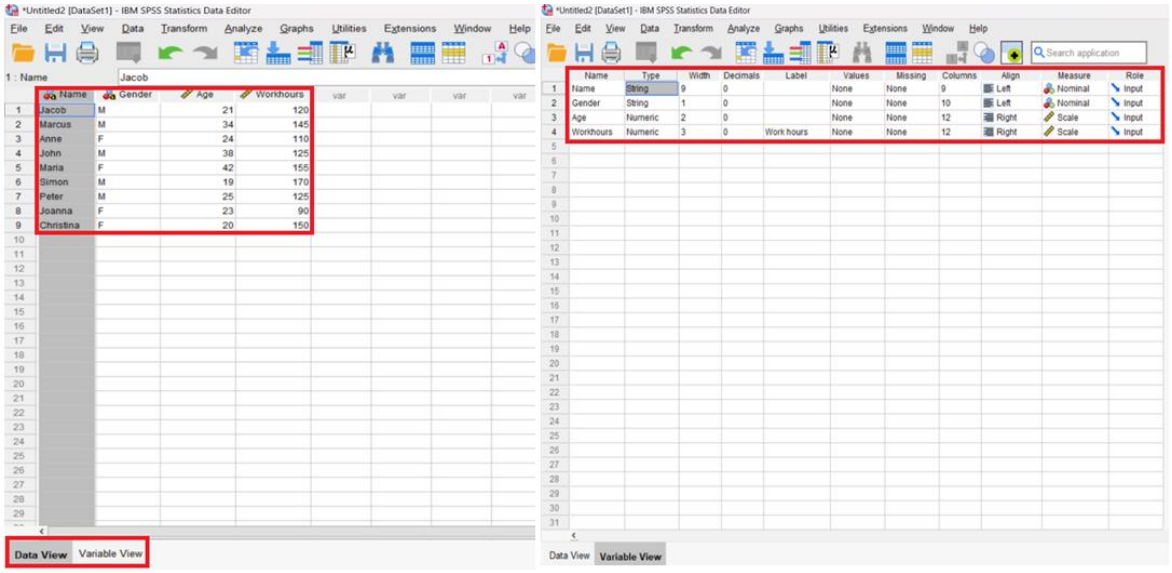

Figure 7.2 Data and variable view windows.

Once you have correctly set up your data, you can explore it within SPSS. SPSS allows users to perform fundamental statistical analysis without relying on predefined functions. On the initial screen (refer to figure 7.3), navigate to "Analyze", followed by "Descriptive Statistics", and then select "Explore". In the "Explore" section, you will find various options depending on the characteristics of the data you provided. In this mode, SPSS will provide you with "descriptive statistics" information about your data. While this is valuable for initial data analysis, it offers only fundamental insights and does not delve into more detailed statistical analysis, which will be covered in subsequent chapters. Before proceeding further, we will also explore another function in SPSS—graph visualization (IBM, 2021).

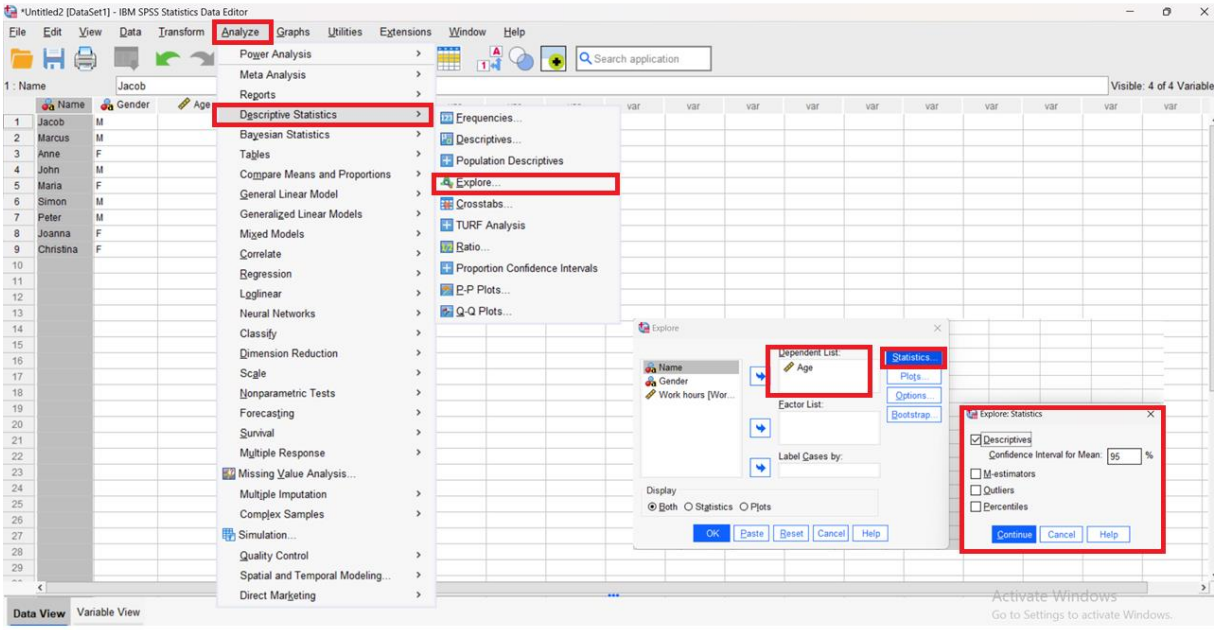

Figure 7.3 Descriptive statistics settings.

SPSS offers a range of data visualization options, including histograms, box plots, bar charts, scatterplots, line charts, pie graphs, and more. By this point, you should have a basic understanding of what each type of graph represents and how to interpret the results they provide. Therefore, we will focus on how to create these graphs within the SPSS software. To create graphs, select the "Graphs" tab on the initial screen, followed by "Chart Builder". In the new window, you can choose the type of graph you want to create and select the variables to be included. After selecting "Finish", a new window will appear with the results visualized in the chosen graph format. In this new window, you can interact with the graph actively, allowing you to modify variable colors and fonts, explore variable distributions across the graph, and more (IBM, 2021).

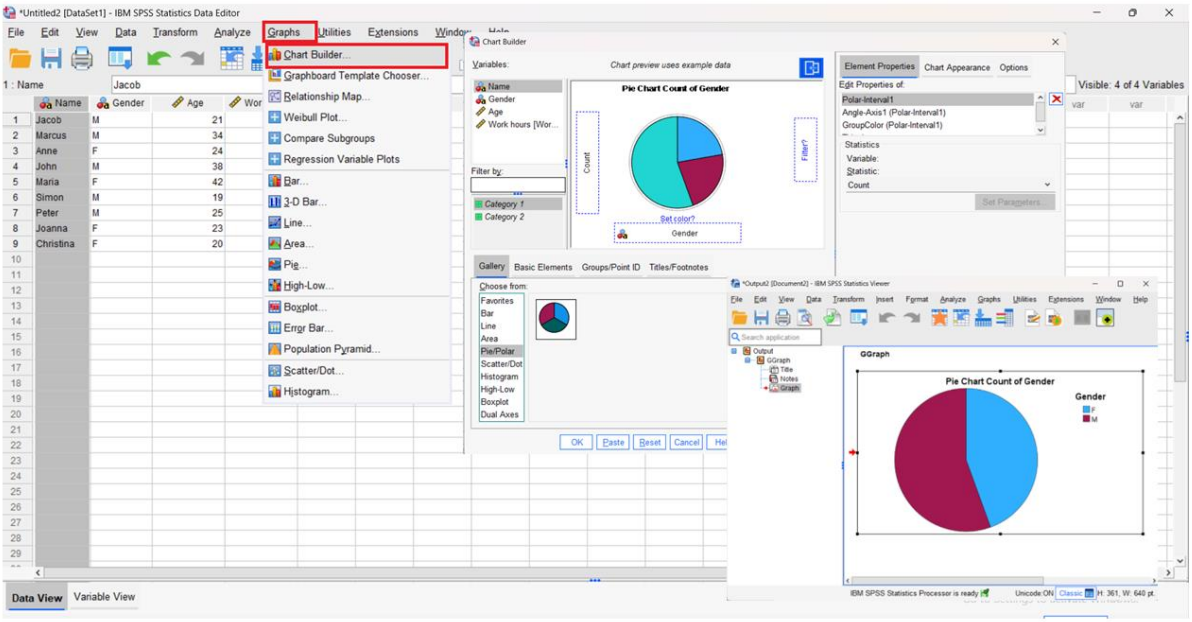

Figure 7.4 Chart Builder settings in SPSS.

Up to this point, we have covered three of the four rules for "Exploring data", which include looking at data (raw data exploration), identifying data (determining data types), and, to some extent, graphing and describing data through descriptive statistics and graph creation. The final rule is "Question Formulation", where we ask ourselves what we aim to achieve through data analysis and set up graphs and descriptive statistics accordingly to obtain answers to our specific questions. For instance, in our current example, the question could be, "Is our analyzed population predominantly female?" By employing both graphs and descriptive statistics, we can conclude that our population consists mainly of male individuals. When formulating your questions, always consider the available data and the variables you have identified (Garth, 2008). This concludes the first part of the SPSS data analysis, and we will now proceed with test preparation.

7.2 Data management

When engaging with crucial data in the SPSS software, it becomes crucial to comprehend the techniques for manipulating information across individual active datasets. SPSS provides functionalities facilitating the manipulation of existing data contained within active datasets. Occasionally, you may encounter two databases separately imported into datasets, yet the

preference is to merge them for enhanced analysis. Consider a logistic company with two branches, each contributing data on costs and transporting cargo in kilograms. The managerial objective is to analyse the overall efficiency of the company. In SPSS, accomplishing this involves navigating to "Data," selecting "Merge files," and having two distinct options. One involves selecting "Cases" and specifying the variable for merging, removing that variable while merging the others. Alternatively, opting for "Variable" retains the variable in the new dataset. A practical application is evident in our logistics scenario, where merging datasets simplifies the company's comprehensive performance analysis.

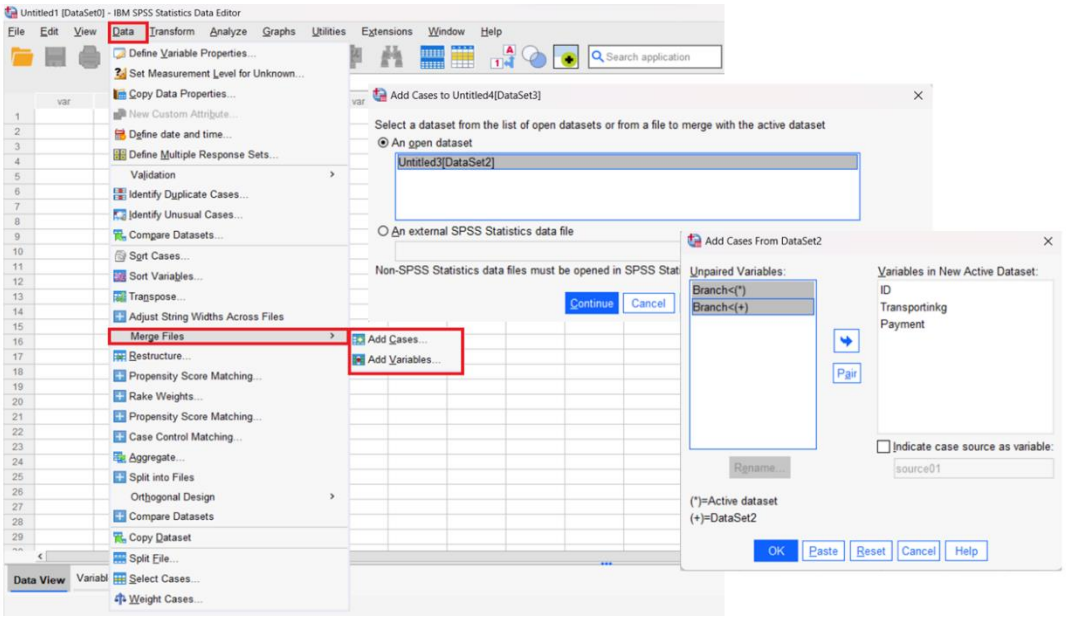

Figure 7.5 Merging file window.

While the merge and split functions enable specific data manipulation, the "Select cases" option offers distinct advantages. Imagine having data for shops B, C, and D in a single database, and the focus is solely on comparing Shop A and Shop C. By selecting "Data" and "Select Cases," one can specify the variables of interest, effectively filtering out unwanted data. For instance, setting Shop C as 2 instructs the software to concentrate solely on Shop C, generating output that is then available for subsequent analyses, such as descriptive statistics, focusing exclusively on the selected cases. Such an approach also enables comparative analysis between only Shop A and Shop C values.

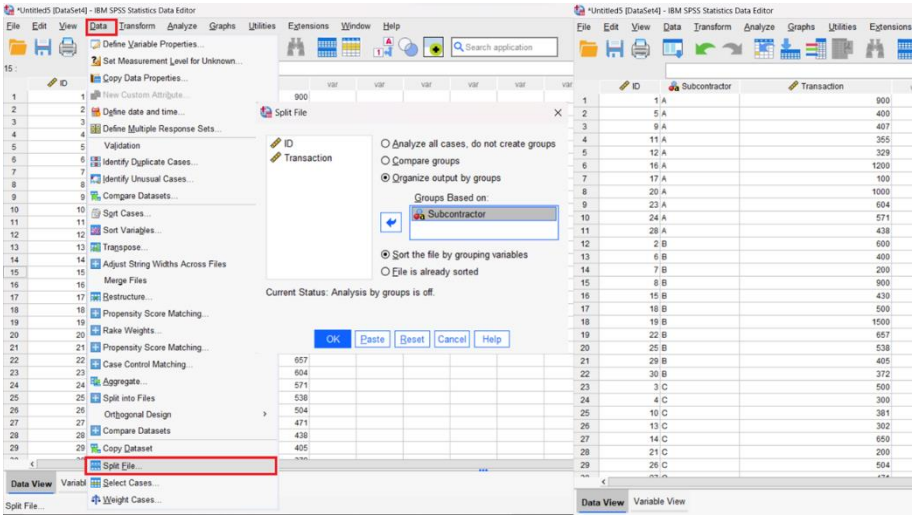

Figure 7.6 Splitting file window.

While merging and split functions enable certain data manipulation, there is also the option to "Select cases". Imagine that we know for certain that Shop A has on average 120 € profit and we want to compare that to Shop C. Unfortunately, in our database we have data for shops B, C and D in a single database and the analysis would include data from all three shops. By clicking "Data" and "Select Cases" we can select which variable we want to focus. In our cases we set that shop C should be set as 2 and then created the function for the software to focus only Shop C. The output can be then used for subsequent analysis by choosing this new column (e.g. descriptive statistics).

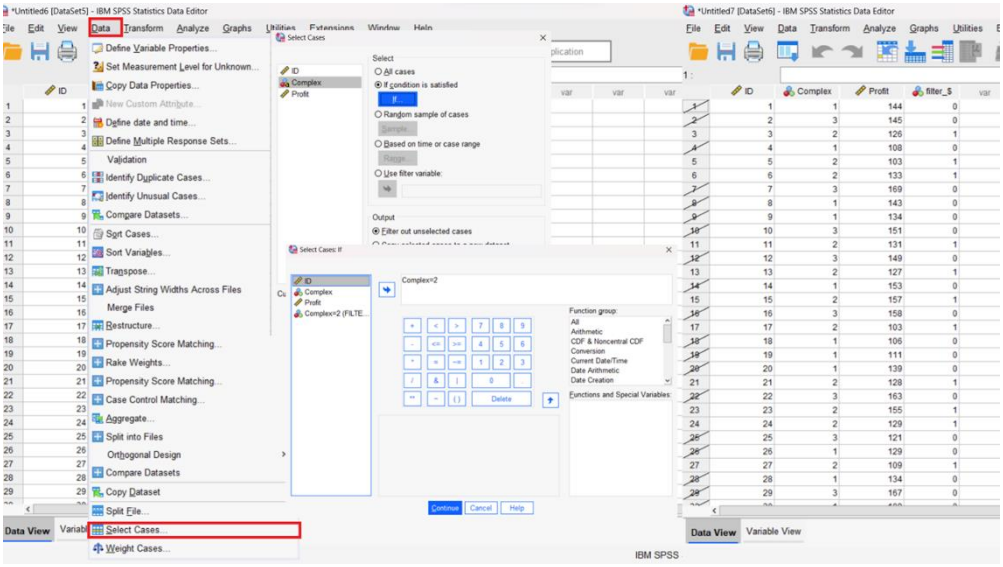

Figure 7.7 Selecting case procedure.

Occasionally, datasets may already contain variables, yet there is a need to introduce new variables based on existing ones. Take, for example, a logistic company manager who possesses data on the weight and distance travel for various products but requires delivery time for optimizing routes. In SPSS, achieving this involves clicking "Transform" and then "Compute Variables." A new variable, DeliveryTime, is created within the new window by setting numeric expressions. In this case, assigning a scale of 0.8 to distance and 0.2 to weight results in a new variable representing delivery time, a crucial addition to the dataset. The flexibility of computing additional variables exists, created for the needs of statistical tests.

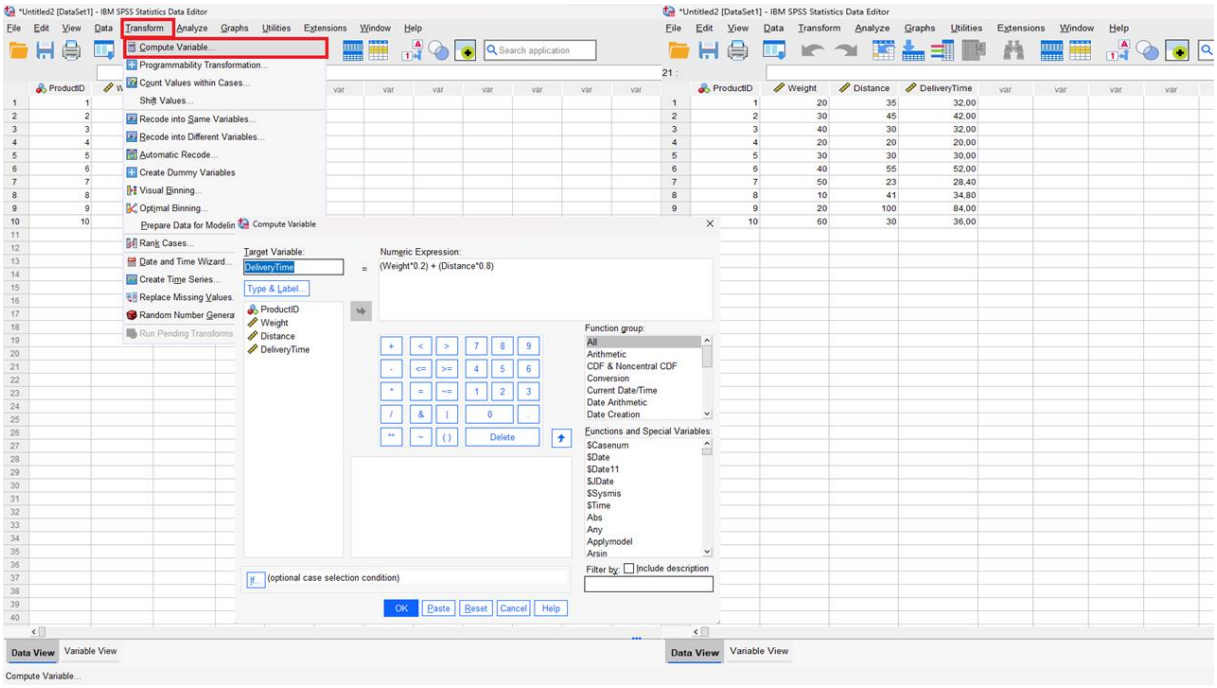

Figure 7.8 Computing variables procedure.

This concludes a small overview of data management functions that SPSS covers, which might be useful during the subsequent model tests covered in this chapter. We will continue with the phases needed before we can conduct a statistical test in the SPSS software.

7.3 Test preparation

Before proceeding with statistical tests, it is essential to adhere to a standard data analysis process flow, which includes data exploration (as covered in Chapter 7.1 and 7.2), data analysis, and results interpretation (Garth, 2008; George & Mallery, 2022). In this chapter, our focus is on data analysis using the SPSS software. Since hypotheses have already been

addressed in previous chapters, our primary focus will be on conducting normality tests within SPSS. There are three methods to assess normality: the histogram, QQ-plot, and the normality test. It is advisable to employ at least two, if not all three of these options, as they each provide distinct information (Ghasemi & Zadesiasl, 2012). To create a histogram, go to "Graphs", followed by "Chart Builder". In the new window, select "Histogram". If you have multiple variables, you must repeat this process for each one to obtain the results. A histogram validates the test for normal distribution if the bars representing variable values resemble a bell curve. If the bars lean more to the left or right side, it may indicate an exponential distribution. For example, we generated a database of 100 IDs, each with a variable representing weight in kilograms. Following the instructions, we created a histogram, as shown in the figure 7.9. As evident from the figure, the bars are distributed across the graph, and while they may not perfectly mirror the curve, they nonetheless suggest a normal distribution and a positive test result (George & Mallery, 2022; Goeman & Solari, 2021).

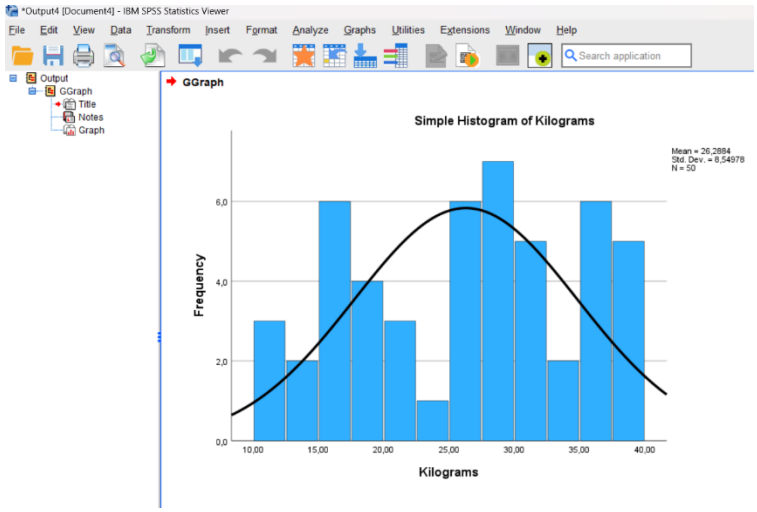

Figure 7.9 Histogram of normality test results.

Another option for conducting normality tests is the QQ-plot, which can be initiated by clicking "Analyze", followed by "Descriptive Statistics", and then selecting "Q-Q Plots". The advantage of this approach is that it allows for the assessment of multiple variables simultaneously (Williamson, b.d). The test is considered successful when the points on the plot cluster closely around a straight line, representing a normal distribution. If the points form "tails", it indicates a failed normality test (Andersen & Dennison, 2018). Using the same database from the histogram graph test, we conducted a Q-Q Plot test. In the figure 7.10 below, you can observe that most cluster points for our variable align with the straight line, indicating a normal

distribution of our data. While we could already conclude that the normality test is positive at this stage, we decided to seek confirmation from all three tests.

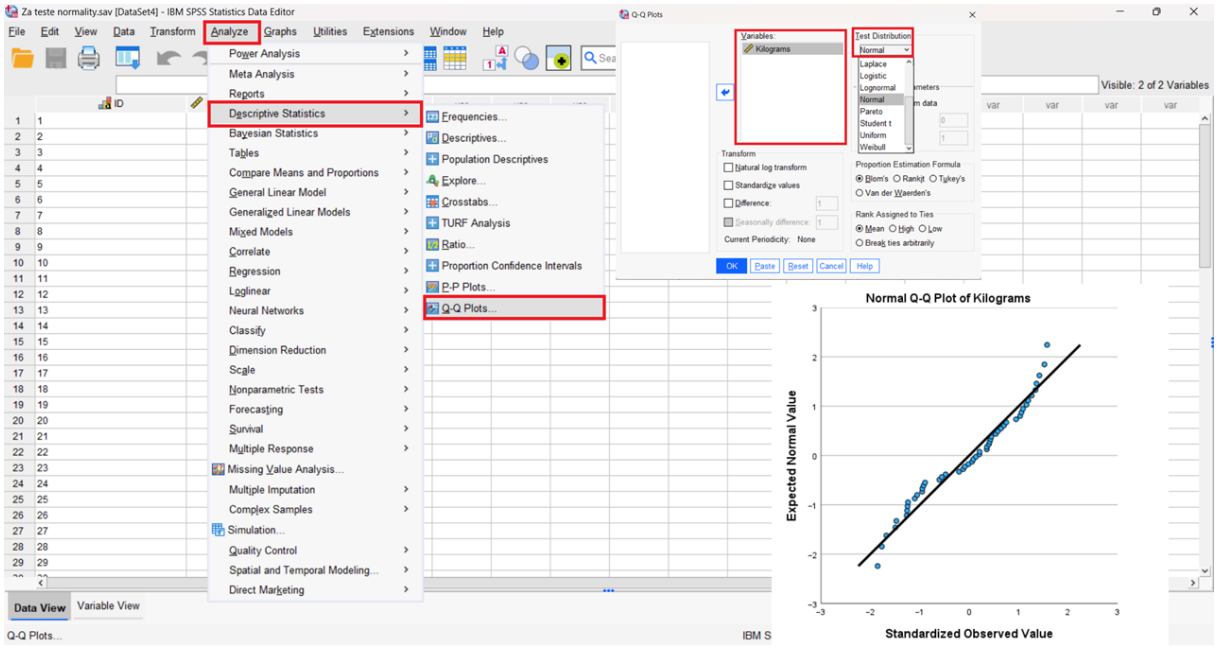

Figure 7.10 Q-Q plot normality test settings and results.

The final option for conducting a normality test is the so-called Test of Normality, considered a statistical test. Typically, it uses the Kolmogorov-Smirnov test, but for small sample sizes, the Shapiro-Wilk test can be employed (Goeaman & Solari, 2021). In SPSS, you can perform this test by clicking "Analyze", followed by "Descriptive Statistics", and then "Explore". You must set the variables you want to check under the "Dependent List" box. Then, under "Plots", select "Normality Plots with Tests". The test is considered successful if the Sig column (*p*-value) in the results is greater than 0.05, indicating a normal distribution. If the *p*-value is less than 0.05, it suggests a non-normal distribution, and the test is considered unsuccessful. We conducted this test once again using the same database as in the previous tests. From the results, we can conclude that according to the Kolmogorov-Smirnov standard, the test is positive as the *p*-value is higher than 0.05. However, for the Shapiro-Wilk test, the *p*-value is lower, indicating a negative test result. These differing results occur because both approaches have different sensitivity settings and power in detecting deviations (Ghasemi & Zahediasl, 2012). Since we have already conducted both Q-Q Plots and the histogram graph tests, the Test of Normality can be considered positive

overall. With the normality tests confirmed, we can conduct the main tests, such as the One-Sample Test.

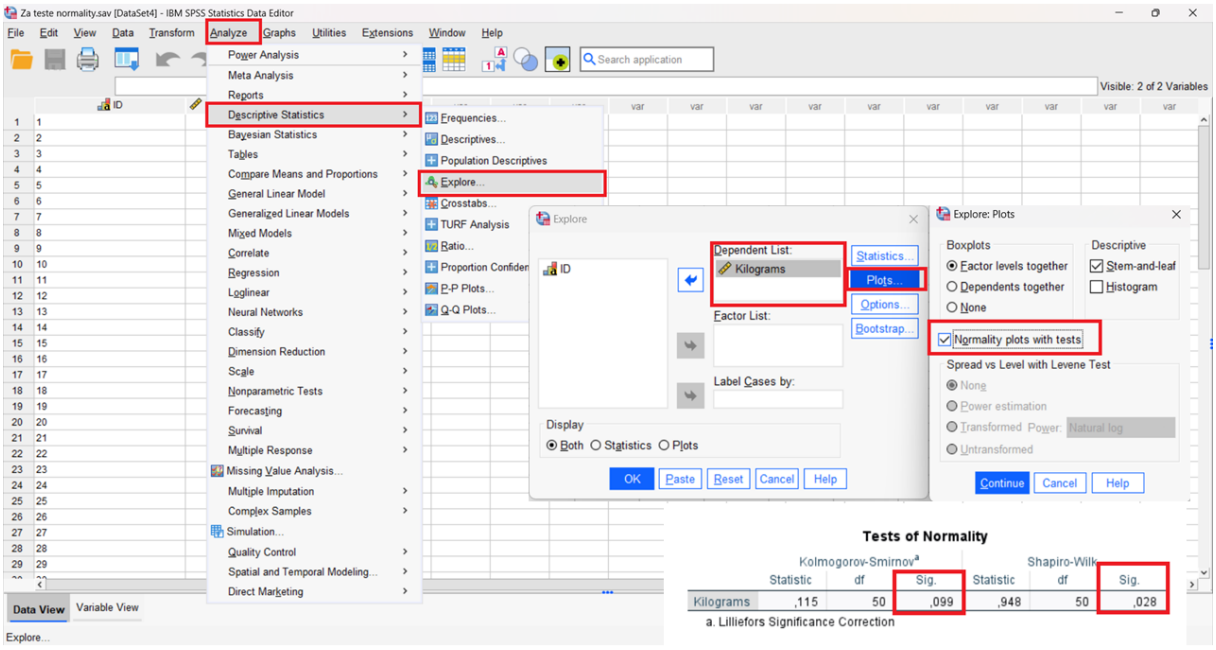

Figure 7.11 Test of Normality settings and results.

7.4 One Sample T-test

You have already covered the theory behind the One Sample T-test in the previous chapters; we will thus focus primarily on conducting a test with the SPSS software. For our One Sample T-test we have prepared a database with a sample of 200 inputs, which includes 1 categorical variable (Student ID) and 2 numerical variables (Weight and Age) (Kim, 2015). Following the guides from the previous subchapters we conduct:

- Explore the data, namely our **variables** and **descriptive statistics** and establish our **question**.
- Check the **normality**, since only one variable histogram and Q-Q plot should suffice.
- Set up hypothesis, where for **Null** the variable is no different from a certain value and **Alternative** where it is different.
- Conduct the **Students T-test**.

- Interpret results, focusing on **Null rejected** or **not**, answer the question and write a report on our test.

In our case, we decided that our question should be, "Is the average weight of students greater than 74 kilograms?". Following the question, we establish our hypothesis to the question, which is "Null = there is no difference" and "Alternative = there is a difference". We conducted the histogram and Q-Q plots to check for normality tests, and after their conclusion, we followed with the T-test. To run the T-test, we click "Analyze" and follow up with "Compare Means" and "One-Sample T-test". In the test variable box, we put our student ID, set the test value to 74 and start the test (refer to figure 7.12).

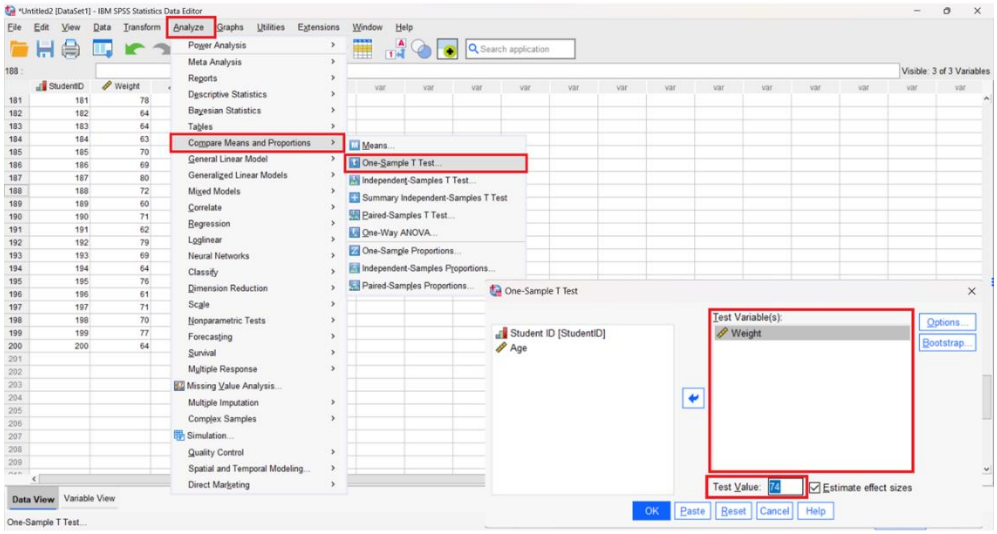

Figure 7.12 One Sample T-Test settings.

Upon confirming the test, another window will appear with the results of our analysis (refer to figure 7.13). This window provides several pieces of information regarding our analysis. In this case, both *p*-values are lower than 0.05, indicating the test's significance. Additionally, we check the *t* and *df* values, which, in our case, are -9.806 and 199, respectively. From these results, we can conclude that our null hypothesis is rejected. Therefore, the complete result report is as follows: "The average student weight is significantly lower (mean = 69.63) than the value of 74 kg (1-sample t-test, $t = -9.806$, $df = 199$, p-value < 0.001)".

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Weight	200	69,63	6,303	,446

One-Sample Test

Test Value = 74

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Weight	-9,806	199	<,001	<,001	-4,370	-5,25	-3,49

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Weight	Cohen's d	6,303	-,693	-,847	-,538
	Hedges' correction	6,326	-,691	-,844	-,536

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

Figure 7.13 One Sample T-Test results.

7.5 Correlation

Let us now move on to the second test, which is the correlation test. We will conduct it using the same database as in the one-sample t-test example. Similar to the one-sample t-test, we will follow the procedure with a few modifications. When performing a correlation between two variables, it is important to specify which one is the dependent variable, and which is the independent variable (Janse *et al.*, 2021; Mishra *et al.*, 2019). This selection can be made based on your research question. In our case, we want to investigate "Whether there is a correlation between a student's age and their weight?". Following the question, we consider weight as the dependent variable and age as the independent variable, as we want to explore if variations in age are related to variations in weight. We define our null and alternative hypotheses (see 7.3 and 7.4) and then run the test by clicking "Analyze", followed by "Correlate" and "Bivariate." Both variables should be placed in the "Variable" box. Ensure that "Pearson", "Two-Tailed", and "Flag Significant" are selected or set (refer to figure 7.14). In this case, we chose "Pearson" because our data indicated a normal distribution and could be analyzed using parametric methods. If normal distribution is not indicated, non-parametric methods should be used (in this case, you would select Spearman instead of Pearson) (George & Mallery, 2022; McClure, 2005).

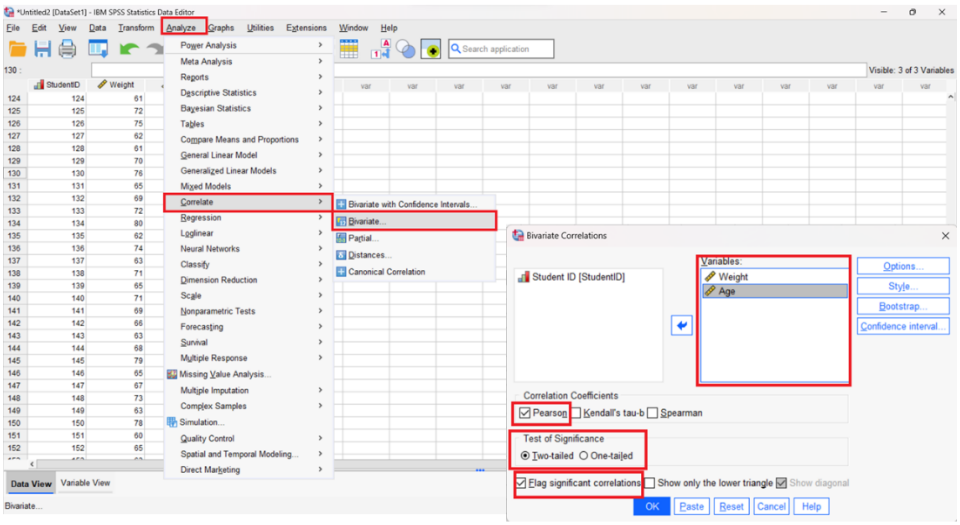

Figure 7.14 Correlation test settings.

Once again, we obtain the results in a new window (refer to figure 7.15). From the results, we can observe that our Pearson Correlation is -0.038, and our *p*-value is 0.596. In correlation analysis, the closer the correlation value is to zero, the weaker the correlation between the variables. In our case, the correlation is very close to zero, indicating no significant correlation between the two variables (McClure, 2005). Additionally, the high *p*-value (0.596) suggests that there is no substantial evidence to conclude that there is a meaningful correlation between the two selected variables (Williamson, b.d.). As a result, our null hypothesis is not rejected. Based on this, we can report that "There was no correlation between the students' age and weight".

Correlations

Correlations

		Weight	Age
Weight	Pearson Correlation	1	-,038
	Sig. (2-tailed)		,596
	N	200	200
Age	Pearson Correlation	-,038	1
	Sig. (2-tailed)	,596	
	N	200	200

Figure 7.15 Correlation test results.

7.6 Chi-Square

The third test we will perform in SPSS software is the Chi-Square test. Unlike the previous two tests, the Chi-Square test compares two categorical variables rather than numerical variables

(Turhan, 2020). Like the process in sections 7.4 and 7.5, we begin by exploring the data and formulating a research question. In our example, we have a logistics company with 200 customers, and we have data on the type of payment and the type of transportation chosen by each customer. The question we aim to answer is, "Do different payment types exhibit different preferences for transport types?" Since we are dealing with only categorical variables, there is no need for a normality test. We establish our Null hypothesis (The preferences for transport types are the same among all payment types) and the Alternative hypothesis. To conduct the Chi-Square analysis, click "Analyze", followed by "Descriptive Statistics", and select "Crosstabs". It is crucial to place the variables based on your research question in either the column or row box (refer to figure 7.16) (Garth, 2008).

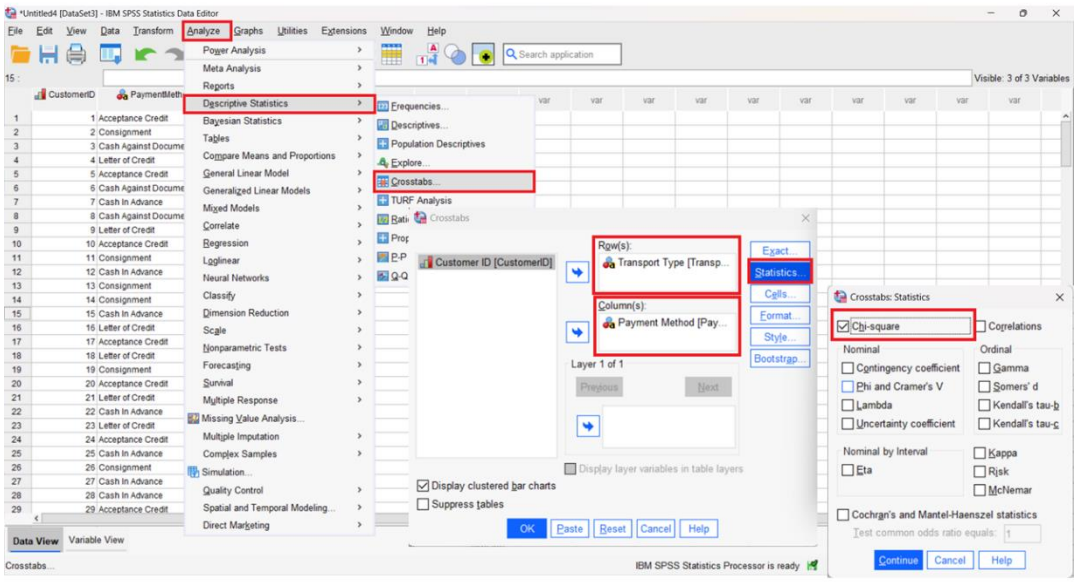

Figure 7.16 Chi-Square test settings.

After the analysis, a new window displays the results (refer to figure 7.17). In this window, you can observe that the Pearson Chi-Square value is 11.614, *df* value is 12, and the *p*-value (asymptotic significance) is 0.477. Based on these results, we can conclude that there is no significant association between the two variables, and the null hypothesis is not rejected. Therefore, the report follows: "There is no significant preference detected between different payment types for different transport types (2-tailed Chi-Square test, chi-sq = 11.614, *df* = 12, *p*-value = 0.477)."

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Transport Type * Payment Method	200	100,0%	0	0,0%	200	100,0%

Transport Type * Payment Method Crosstabulation

Count

		Payment Method					
		Acceptance Credit	Cash Against Documents	Cash In Advance	Consignment	Letter of Credit	Total
Transport Type	Airplane	11	9	12	13	6	51
	Ship	16	9	6	7	10	48
	Train	16	13	17	7	10	63
	Truck	7	6	11	9	5	38
Total		50	37	46	36	31	200

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	11,614[a]	12	,477
Likelihood Ratio	11,965	12	,448
N of Valid Cases	200		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,89.

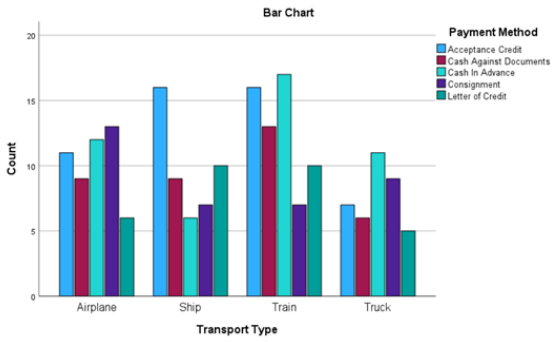

Figure 7.17 Chi-Square test results.

7.7 ANOVA

The final test we will cover is the ANOVA test, specifically focusing on the simpler model known as one-way ANOVA, which involves a categorical variable and a numerical variable (Goeman & Solari, 2021). As with the T-test, we will follow the same procedure: explore the data, formulate a research question, conduct a normality test, and set hypotheses. Let us consider a case study of a transport dispatcher working for a logistics company. The dispatcher closely collaborates with a partner company and regularly plans three different routes for the trucks to deliver their goods. Due to a "Just-in-time" policy emphasizing faster deliveries, the question arises: "Does the choice of delivery route impact on the delivery time for the company?" To run the ANOVA test in SPSS, go to "Analyze" followed by "Compare Means..." and then "One-way ANOVA". Place the dependent variable in the "Dependent List" box and the Factor variable in the "Factor" box (refer to figure 7.18). For thorough analysis, we have also included the Post Hoc setting. It is important to note that Post Hoc analysis should only be conducted if the initial ANOVA test is positive. By employing Post Hoc analysis, we can identify the most optimal choice (in our case, the route). The most reliable methods to use for Post Hoc analysis are either the Bonferroni correction or the Tukey HSD method (Goeman & Solari, 2021).

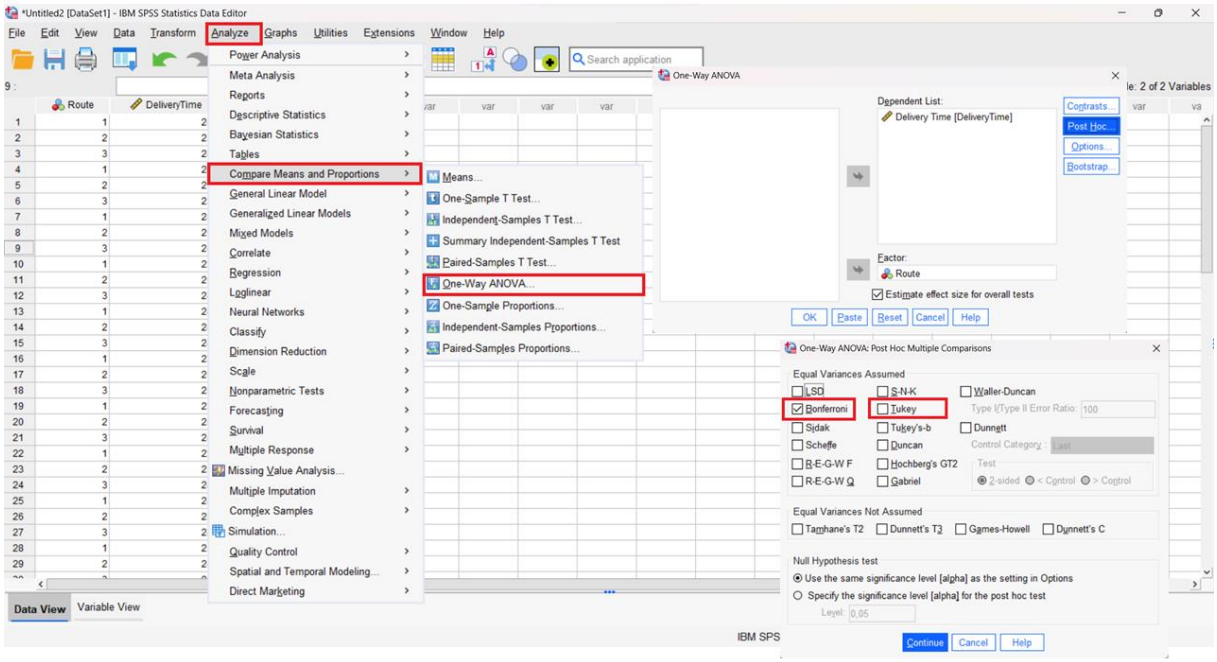

Figure 7.18 ANOVA settings.

The results of our analysis indicate that our *F*-statistic value is 11.173 (higher values indicate more variations between groups) and *p*-value <0.001, which means that our Null hypothesis is rejected (see figure 7.19). Since there is a significant difference between the three routes (<0.001), a post hoc test is also valid in our case (George & Mallery, 2022). After conducting the Bonferroni correction test, we can see that the best *p*-values are noted in the case of route 2 (refer to figure 7.19). In the report, we can conclude that "There was a significant difference in choosing a delivery route in correlation to delivery times (1-way ANOVA, *F*=11.173, *df* = 47, *p*-value = <0.001). Route 2 had the best delivery time results."

ANOVA

Delivery Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16,527	2	8,263	11,173	<,001
Within Groups	33,280	45	,740		
Total	49,807	47			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Delivery Time

Bonferroni

(I) Route	(J) Route	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	2	-1,4250*	,3040	<,001	-2,181	-,669
	3	-,5500	,3040	,231	-1,306	,206
2	1	1,4250*	,3040	<,001	,669	2,181
	3	,8750*	,3040	,018	,119	1,631
3	1	,5500	,3040	,231	-,206	1,306
	2	-,8750*	,3040	,018	-1,631	-,119

*. The mean difference is significant at the 0.05 level.

Figure 7.19 ANOVA initial results and Post Hoc Test results.

We conclude this chapter of the book with the understanding that we have covered some of the more common tests in this chapter. There are still other tests, such as Repeated Measures ANOVA, reliability tests, and sensitivity tests, which can also be modelled and analyzed using SPSS software. These additional tests provide a broader range of tools for data analysis and draw meaningful insights into various research and practical applications.

References Chapter 7

- Andersen, A.J. & Dennison, J.R. (2018). An Introduction to Quantile-Quantile Plots for the Experimental Physicist. Journal Articles, 51.
- Garth, A. (2008). Analysing data using SPSS [available at: https://students.shu.ac.uk/lits/it/documents/pdf/analysing_data_using_spss.pdf, access October 26, 2023]
- George, D. & Mallery, P. (2022). IBM SPSS Statistics 27 Stet by Step: A Simple Guide and Reference, 17TH edition, Abingdon: Routledge
- Ghasemi, A. & Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. International Journal of Endocrinology and Metabolism, 10(2), pp. 486-489.
- Goeman, J.J. & Solari, A. (2021). Comparing Three Groups. The American Statistician, 76(2), pp. 168-176
- IBM (2021). IBM SPSS Statistics 28 Brief [available at: https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/IBM_SPSS_Statistics_Brief_Guide.pdf, access October 26, 2023]
- Janse, R.J., Hoekstra, T., Jager, K.J., Zoccali, C., Tripepi, G., Dekker, F.W. & van Diepen, M. (2021). Conducting correlation analysis: important limitations and pitfalls, 14(11), pp. 2332-2337.
- Kim, T.K. (2015). T test as a parametric statistic. Korean Journal of Anesthesiology, 68(6), pp. 540-546
- Landau, S. & Everitt, B.S. (2004). A Handbook of Statistical Analyses using SPSS, 1st edition, London: Chapman & Hall/CRC
- McClure, P. (2005). Correlation Statistics Review of the Basics and Some Common Pitfalls. Journal of Hand Therapy, 18(3), pp. 378-380

- Mishra, P., Singh, U., Pandey, C.M., Mishra, P. & Pandey, G. (2019). Application of Student's t-test, Analysis of Variance, and Covariance. Annals of Cardiac Anesthesia, 22(4), pp. 407-411
- Turhan, N.S. (2020). Karl Pearson's chi-square tests. Educational Research and Reviews, 15(9), pp. 575-580
- Williamson, M. (b.d.). Data Analysis using SPSS [available at: https://med.und.edu/research/daccota/_files/pdfs/berdc_resource_pdfs/data_analysis_using_spss.pdf, access October 26, 2023]

8. Business analytics foundations including the R and SQL

What is business analytics (BA)? What problems does it solve and what tools does it use? What are R and SQL? How is BA related to R and SQL? Are there any examples of good practices where these software's are used to solve logistic business problems?

On these and similar questions, we will try to provide answers in the following chapter.

8.1 What is business analytics?

BA represents a holistic approach to data analysis and business decision-making. It is a data-driven environment with the goal of improving company business performance by providing a foundation for more informed decision-making. It is a systematic thinking process that applies qualitative, quantitative, and statistical computational tools and methods to analyse data, gain insights, inform, and support decision-making. Any particular analysis may use a variety of techniques including diagnostic, predictive, prescriptive, and optimisation models (Power et al., 2018). This Mikalef et al., (2019) provides a roadmap for both academic exploration and practical implementation, highlighting the transformative potential of analytics when properly integrated into organizational processes. Accordingly, the authors state that BA require organizations to radically redesign how such initiatives are approached, designed and refined, how resource planning and orchestration is executed and strategically aligned, as well as re-valuate their expected performance outcomes, their association with strategic objectives and, as a result, develop appropriate KPIs (Mikalef et al., 2019).

The main tasks of BA is to provide the knowledge pipeline to ensure a coherent link between raw data and business decisions. The overall goal is business effectiveness through 'verticalization,' usability, and integration with operational systems (Kohavi, Rothleder, & Simoudis, 2002). The BA has many application areas and related derivatives: Financial Analytics, Supply chain analytics, Crisis Analytics, Knowledge Analytics, Marketing Analytics, Customer Analytics, Service Analytics, Human Resource Analytics, Talent Analytics, Process Analytics, Risk Analytics (Holsapple, Lee-Post, & Pakath, 2014).

There are three types of Business Analytics platforms:

- **Descriptive** – They look at existing data and provide summary statistics and basic visualization.
- **Predictive** – They use existing data to estimate the most likely future scenarios.
- **Prescriptive** – They automatically process big data, business rules, market conditions, etc. These platforms utilize machine learning and artificial intelligence methods. The goal is fully automated decision-making on which actions a company should take considering the current situation to achieve the desired business objectives.

The interest in big data and business analytics has grown exponentially over the past decade (Mikalef et al., 2019). Modern-day BA is rooted in the ongoing advances of systems to support decision making. These advances include increasingly powerful mechanisms for acquiring, generating, assimilating, selecting, and emitting knowledge relevant to making decisions. Given its decision support heritage, business analytics necessarily partakes of and exploits these mechanisms. The knowledge that must be processed ranges from qualitative to quantitative and BA is concerned with operating on both knowledge types, as appropriate for the decision at hand (Kohavi, Rothleder, & Simoudis, 2002). The reason why should some organization apply BA is in the problems it solves. Problems which BA emphasise are the curtail problems for effective management of the company. Accordingly, there are several rationales for applying BA (Holsapple, Lee-Post, & Pakath, 2014):

- Achieving a competitive advantage
- Supporting of an organization's strategic and tactical goals
- Better organizational performance
- Better decision outcomes
- Better or more informed decision processes
- Knowledge production
- Obtaining value from data

Regardless, of the type of platform used in BA to solve and support the decision-making process in each company there are three key pillars of each BA solution (Figure 8.1). Generally, the BA occupies a place in the spectrum between Computer Science/Mathematics/Data Science

(on one hand) and Business and Management (on the other). Business analytics requires both technical and business knowledge. A major problem in designing a BA is that the boundaries are not distinct (Power et al., 2018). Accordingly, to perform the BA mathematical tools are needed to identify, extract and represent the insights in the right manner via tables, graphics, formulas, etc. Additionally, the programming tools serve as a support for this kind of activity and enable fast and error-less computations, compared to the traditional paper and pen approach. Last but not least, logistics expertise is needed in a given area or business problem to pinpoint the key influential factors and related ecosystem.

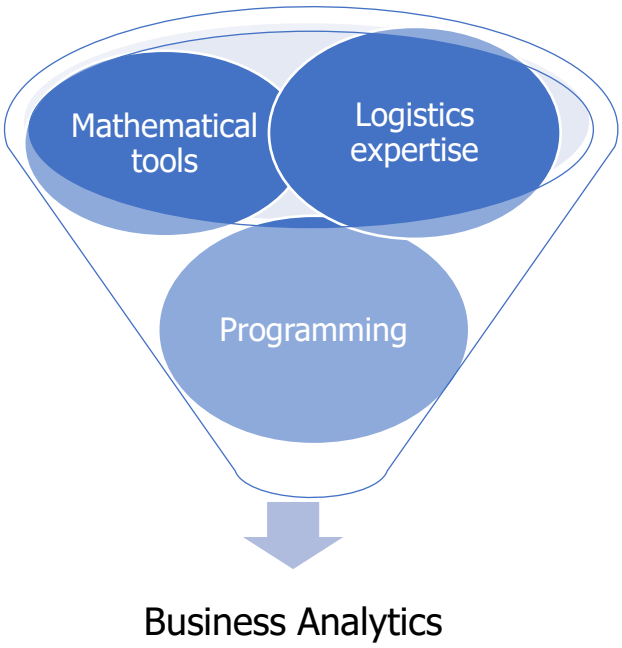

Figure 8.1 The key pillars of BA in the context of the supply chain and logistics.

The key consumer is the business user, whose job, possibly in merchandising, marketing, or sales, is not directly related to analytics per se, but who typically uses analytical tools to improve the results of some business process along one or more dimensions (such as profit and time to market). Business users do not want to deal with advanced statistical concepts; they want straightforward visualizations and task-relevant Outputs (Kohavi, Rothleder, & Simoudis, 2002).

8.2 What is R?

R is an integrated suite of software facilities for data manipulation, calculation and graphical

Display (R Core Team, 2019). Among other things it has:

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- Graphical facilities for data analysis and display either directly at the computer or on hard-copy, and a well-developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed, most of the system supplied functions are themselves written in the S language.)

The main advantages of R are the fact that R is freeware and that there is a lot of help available online. It is quite similar to other programming packages such as MatLab (not freeware), but more user-friendly than programming languages such as C++ or Fortran (Torfs & Brauer, 2014). R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis (R Core Team, 2019).

Instal R and R Studio

To install R, go to cran.r-project.org and click the downalod the R for a specific operating system on your machine (usually Windows) (Figure 8.2).

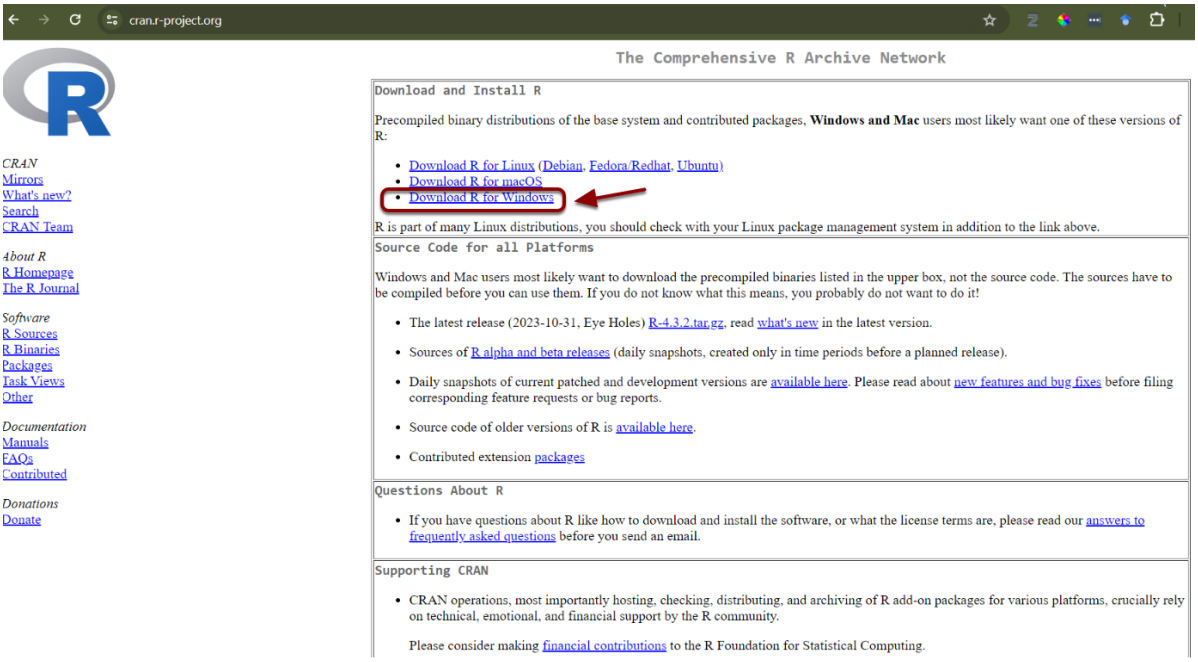

Figure 8.2 Download page for R software.

This will download the R software, and the installation procedure is the same as with any other software. When R software is installed, it will come without an advanced integrated development environment (IDE) to help and assist users in making different analyses. Although it is possible to do any kind of analysis with only R installed, it is preferable to pair it with some modern IDE, like RStudio, which is one of the most popular IDEs. The procedure on how to install the RStudio is similar to the core R software. Go to https://posit.co/download/rstudio-desktop/, search for RStudio Desktop open-source licence, download it, and install it. After installing the R and RStudio the user will have the following user interface screen (Figure 8.3).

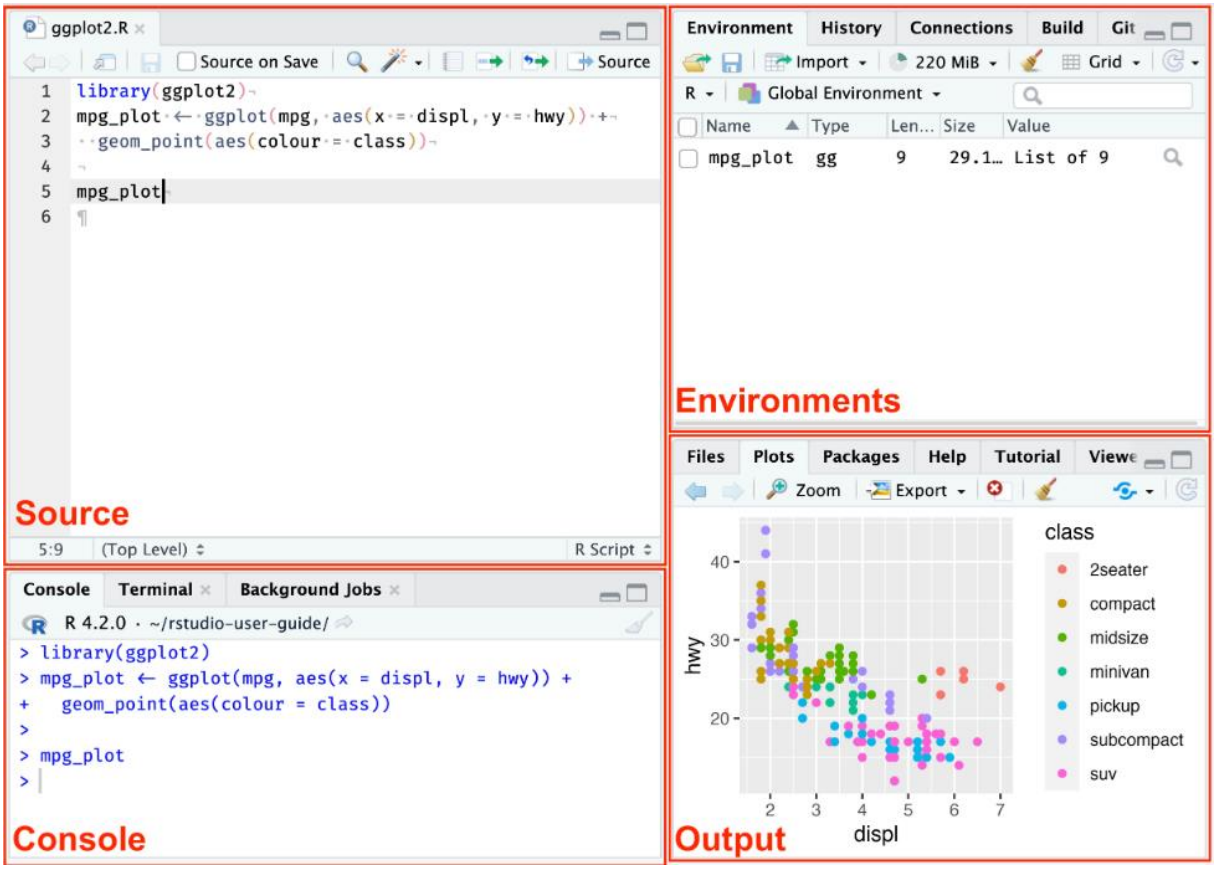

Figure 8.3 The user interface and the R & Rstudio (RStudio, 2024).

The RStudio user interface has 4 primary panes (RStudio, 2024):

- Source pane;
- Console pane;
- Environment pane, containing the Environment, History, Connections, Build, VCS, and Tutorial tabs;
- Output pane, containing the Files, Plots, Packages, Help, Viewer, and Presentation tabs.

Each of the panes and the subsection and options in it allow users to perform different operations, have control over some data analysis, or have a more structured and clear view of the data analytics process underway.

8.3 What is SQL and how is related to BA and R?

The Chinook Database is a sample database that is used for learning and demonstrating database management systems (DBMS) and SQL queries. It is designed as a digital media store using real data from an iTunes Library; fictitious names / addresses for customers and employees; and random data for the sales information's. The database contains a variety of tables that represent a music store's data, including information about artists, albums, tracks, customers, invoices, and more (Figure 8.4).

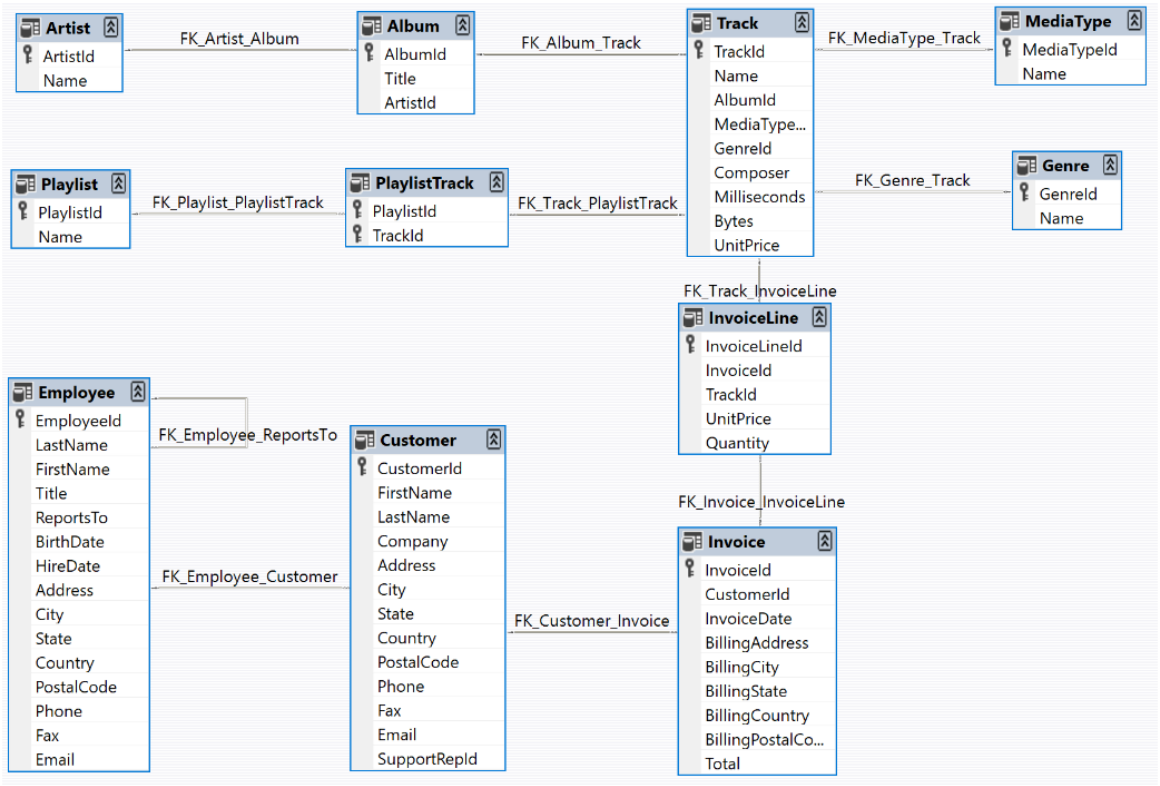

Figure 8.4 The data model of Chinook Database.

Figure 8.4 demonstrates the data model surrounding the Chinook database with the different data tables and their keys (unique identifiers) and joint tables like PlaylistTrack table. The tables convey different information's about the given digital store (Table 8.1).

Table 8.1 Information's contained in each of the tables of the Chinook Database.

Table Name	Description
Artist	Contains information about music artists.
Album	Contains information about music albums, each associated with an artist.

Track	Contains information about individual music tracks, including references to albums, media types, and genres.
Genre	Contains information about music genres.
MediaType	Contains information about different types of media (e.g., audio, video).
Customer	Contains information about customers, including contact details and support representative information.
Employee	Contains information about employees, including their roles, reports-to relationships, and contact details.
Invoice	Contains information about invoices, including customer details, billing information, and total amounts.
InvoiceLine	Contains detailed information about each item in an invoice, including references to tracks and quantities.
Playlist	Contains information about playlists.
PlaylistTrack	Links tracks to playlists, indicating which tracks are included in which playlists.

8.4 How are business analytics, SQL and R related?

The connection between BA, SQL and R is natural since all the business data should be stored in the SQL databases. This is still the idealistic goal since there is still poor data management in some fractions of the small and medium enterprises which still do not fully understand the power of the data. In large companies, this was recognized a long time ago and the data is properly structured in the databases (SQL or others, but usually in SQL). On the other hand, the analysis of the data can be performed in SQL, but for that purpose, it is better to use the statistically oriented software, where R comes in the focus, as one of the most popular statistical platforms to perform data analysis.

Accordingly, SQL and R can be seen as the perfect tool for collaboration when the problem at hand is from the BA area. There are several main reasons, and one of them is that the BA data is daily changing and updating according to the market reality and company activities: sales, employees, revenue, etc. The SQL databases are perfect for capturing those changes and updating existing data, while the R scripts are very good at automating tasks as well as designing new packages for analysing the given data. The reason for this is that the core R is more built around the concept of data analysing, than rather on general programming like Python for example.

Querying the SQL database with R

The R programming language and the SQL databases have a natural connection since the R is mainly built for statistical data analysis and the majority of the transactional data is found in databases. The "way" how the R operates to manage data manipulation from SQL databases is via DBI and RSQLite programming packages. The DBI package provides a standardized interface for interacting with various DBMS, allowing users to connect, query, and manage transactions consistently across different databases. The RSQLite package, which adheres to the DBI interface, specifically facilitates interaction with SQLite databases, enabling users to execute SQL queries, fetch data, and perform database operations directly from R. Together, these packages streamline the process of working with databases in R, offering a cohesive and efficient workflow.

To effectively demonstrate the performing the SQL operation from the R and generating desired insights from the data, regarding the problem on hand, we have provided several code snippets in Figures 8.5 and 8.6.

````
---
title: "BUSINES ANALYTICS FOUNDATINS INCLUDING THE R AND SQL"
format: html
editor: visual
---

# R & SQL

## Loading libraries

```{r setup, warning=FALSE, message=FALSE}
library(DBI)
library(RSQLite)
```

## Connect to the Chinook SQLite database

```{r}
con <- dbConnect(RSQLite::SQLite(), dbname = "Chinook_Sqlite.sqlite")
```

## List all tables in the database

```{r}
tables <- dbListTables(con)
print(tables)
```
````

Figure 8.5 The code snippet for establishing the connection between SQL & R and exploring the data tables contained in the SQL.

The first step in querying the SQL database via R is to establish the connection (Figure 8.5). The figure demonstrates the usage of DBI and RSQlite packages which enable establishing the connection via dbConect() function. The result of the connection and the data tables which are revealed via the aforementioned connection are then exported via dbListTables() functions which print a list of all the data found via the connection: Album, Artist, Customer, Employee, Genre, Invoice, InvoiceLine MediaType, Playlist, PlaylistTrack, Track.

After the connection is established, there are a number of possible analyses which can be performed, depending on a business goal and future usage of a given results. Here, due to the space restrictions, we will demonstrate only a fraction of possible data analysis, with a small code snippet and the set of code rules needed to extract the information from the SQL. The code performs database querying via R and displays the top-selling albums, their authors and the numbers sold (Figure 8.6). Table 8.2 represents the results of the data querying via code snippet in Figure 8.6.

```
## Choose a Album table from the database
```{r}
query_album <- "SELECT * FROM Album LIMIT 10"
data_album <- dbGetQuery(con, query_album)
print(data_album)
```

Query to get album details along with artist names
```{r}
query_album_artist <- "
SELECT Album.AlbumId, Album.Title AS AlbumTitle, Artist.Name AS ArtistName
FROM Album
JOIN Artist ON Album.ArtistId = Artist.ArtistId
LIMIT 10"

data_album_artist <- dbGetQuery(con, query_album_artist)
print(data_album_artist)
```

Query to get the top-selling albums along with artist names
```{r}
query_top_selling_albums <- "
SELECT
    Album.Title AS AlbumTitle,
    Artist.Name AS ArtistName,
    SUM(InvoiceLine.Quantity) AS TotalQuantitySold
FROM
    InvoiceLine
JOIN
    Track ON InvoiceLine.TrackId = Track.TrackId
JOIN
    Album ON Track.AlbumId = Album.AlbumId
JOIN
    Artist ON Album.ArtistId = Artist.ArtistId
GROUP BY
    Album.AlbumId, Album.Title, Artist.Name
ORDER BY
    TotalQuantitySold DESC
LIMIT 10"

# Execute the query
top_selling_albums <- dbGetQuery(con, query_top_selling_albums)
knitr::kable(top_selling_albums)
```
```

Figure 8.6 The code snippet for querying the SQL via R and determining the top 10 selling albums.

Table 8.2 The top 10 selling albums in the Chinook digital store.

Album title	Artist name	Quantity sold
Minha Historia	Chico Buarque	27
Greatest Hits	Lenny Kravitz	26
Unplugged	Eric Clapton	25
Acústico	Titãs	22
Greatest Kiss	Kiss	20
Prenda Minha	Caetano Veloso	19
Chronicle, Vol. 2	Creedence Clearwater Revival	19

Album title	Artist name	Quantity sold
My Generation - The Very Best Of The Who	The Who	19
International Superhits	Green Day	18
Chronicle, Vol. 1	Creedence Clearwater Revival	18

References Chapter 8

- Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics. Decision Support Systems, 64, 130-141.
- Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002). Emerging trends in business analytics. Communications of the ACM, 45(8), 45-48.
- Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data and business analytics: A research agenda for realizing business value. *Information & Management*. https://doi.org/10.1016/j.im.2019.103237
- Power, D. J., Heavin, C., McDermott, J., & Daly, M. (2018). Defining business analytics: An empirical approach. Journal of Decision Systems, 27(1), 40–53. https://doi.org/10.1080/2573234X.2018.1507605
- R Core Team. (2019). An introduction to R: Notes on R, a programming environment for data analysis and graphics. The R Foundation.
- RStudio. (2024). RStudio IDE cheatsheet: UI panes. Posit. https://docs.posit.co/ide/user/ide/guide/ui/ui-panes.html
- Torfs, P., & Brauer, C. (2014). A (very) short introduction to R. Hydrology and Quantitative Water Management Group, Wageningen University, The Netherlands, 1-12.

9. Demand forecasting, visualising and feature engineering of time series in supply chains

What is demand forecasting? How can we effectively visualize and make conclusions about the customer data? How to conduct the feature engineering of time series?

On these and similar questions, we will try to provide answers in the following chapter.

9.1 What is customer demand and demand forecasting?

The final customer's demand sets the entire supply chain in motion stakeholders (Syntetos et al., 2016). Accordingly, customer demand is a key component for planning all logistic processes in the supply chain, and therefore determining levels of customer demand is of great interest for supply chain managers. Complementary, demand forecasting is an essential activity for planning and scheduling logistic activities within the observed supply chain (Mircetic et al., 2017). Accurate demand forecasting models directly influence the decrease of logistics costs, since they provide an assessment of customer demand (Mircetic et al., 2016). Forecasting in supply chains goes beyond the operational task of extrapolating demand requirements at one echelon. It involves complex issues such as supply chain coordination and the sharing of information between multiple stakeholders (Syntetos et al., 2016).

Customer demand and accompanying forecasts are vital to SCs, as it provides the basic inputs for the planning and control of all functional areas, including logistics, marketing, production, etc (Mircetic, 2018). If the final consumers' demand were constant, or known with certainty well in advance, then the operation of a supply chain would be a straightforward (backwards) scheduling exercise. However, demand is not known and thus it needs to be forecasted. It is the uncertainty associated with this demand that makes supply chain management very difficult (Syntetos et al., 2016). The effectiveness of demand forecasting is influenced by the inherent uncertainties in the demand time series that supply chains have (Rostami-Tabar,

2013). Consequently, addressing and understanding these uncertainties is a major challenge for managers when coordinating and planning operations within supply chains (Mircetic, 2018).

Demand uncertainty is one of the most significant challenges for modern supply chains. The recent COVID-19 pandemic has further underscored this issue, causing widespread disruptions that have complicated supply chain planning and control (Nikolopoulos et al., 2020). Demand forecasting in supply chains often involves predicting the demand for numerous items. Forecasters in supply chains typically extrapolate time series data for each stock-keeping unit individually. For example, a retailer might use point-of-sale data to generate forecasts at the individual store level (Mircetic et al., 2022).

9.2 Demand forecasting steps in supply chains?

In accordance with the statements and conclusions made above, regarding the importance of demand and demand forecasting for supply chains, it is important to follow the specific steps when developing the forecasts in supply chains (Figure 9.1).

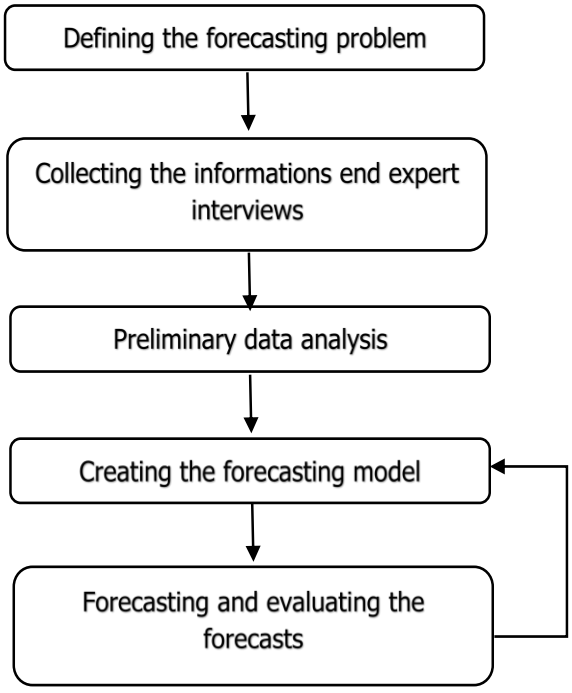

Figure 9.1 Thee basic steps for proper implementation of forecasts within a company (Makridakis et al., 1998; Makridakis et al., 1983).

Each of the steps in Figure 9.1 has its merits and contribution towards creating reliable and useful (business-oriented) forecasts. Accordingly, defining the problem is often the most difficult part of forecasting and requires an understanding of how the forecasts will be used, as well as the role of forecasting functions within the observed company. The forecaster should spend considerable time communicating with everyone involved in data collection, database maintenance, and using forecasts for future planning. One of the main aggravating factors in problem definition is how the final forecast will be utilized in everyday logistics operations (what platform, software design, user interface, etc).

For the information collection step, there are at least two types of information are always needed: statistical data and the accumulated expertise of the people who collect the data and use the forecasts. In practice, it is often difficult to obtain historical data to create a good statistical model. Also, there is a big misunderstanding of what is demand data and what can it be used as its proxy. There is a bad practice of using the shipment and delivery data as a proxy for demand data, which will only deteriorate the decision-making process based on forecasts made on the easy kind of data. Sales data is the only reliable proxy for demand data (Syntetos et al., 2016), although this simplification is not perfect, especially in supply chains with a lot of out-of-stock situations.

For the preliminary data analysis step, it is recommended to always start data analysis with graphical representations to answer the following questions. Are there consistent patterns? Is there a significant trend? Is there noticeable seasonality? Is there evidence of business cycles? How strong are the relationships between variables? These are the questions on which simple graphics can provide answers and allow further data analysis by narrowing the focus of which models to apply to discovered demand features. Usually, the simple model, determined in this way can beat the more sophisticated and complicated ones (Rostami-Tabar & Mircetic, 2023).

Selecting and creating forecasting models is the most important step when creating the forecasting model. Which model to use depends on several factors, the most important of which are the availability of historical data and the correlation between dependent and independent variables. It is common to compare two or three potential models when selecting a model. Each model is an artificial construct based on a set of assumptions (explicit and implicit) and generally involves one or more parameters that must be created using known

historical data. In the following chapter, the process of development and application of the ARIMA model will be represented, as one of the best and most popular models.

Evaluating the forecasting model is the step which measures the usability of the created model. After selecting the forecasting model and estimating its parameters, the model is used to create forecasts. The accuracy of the model is evaluated using various statistics, but it is also important to test forecasts through business implication measures (i.e., utility metrics).

9.3 Demand forecasting in the food industry

The consumption data for all products from the observed company in the food industry is presented as weekly demand spanning from January 2012 to December 2014 (Figure 9.2). The x axis represents the time, while the demand values are presented on y axis. This data is shown in weekly intervals, as this corresponds to the period during which the supply to final points of sale is conducted. Consequently, the company's management is focused on forecasting the market's weekly consumption.

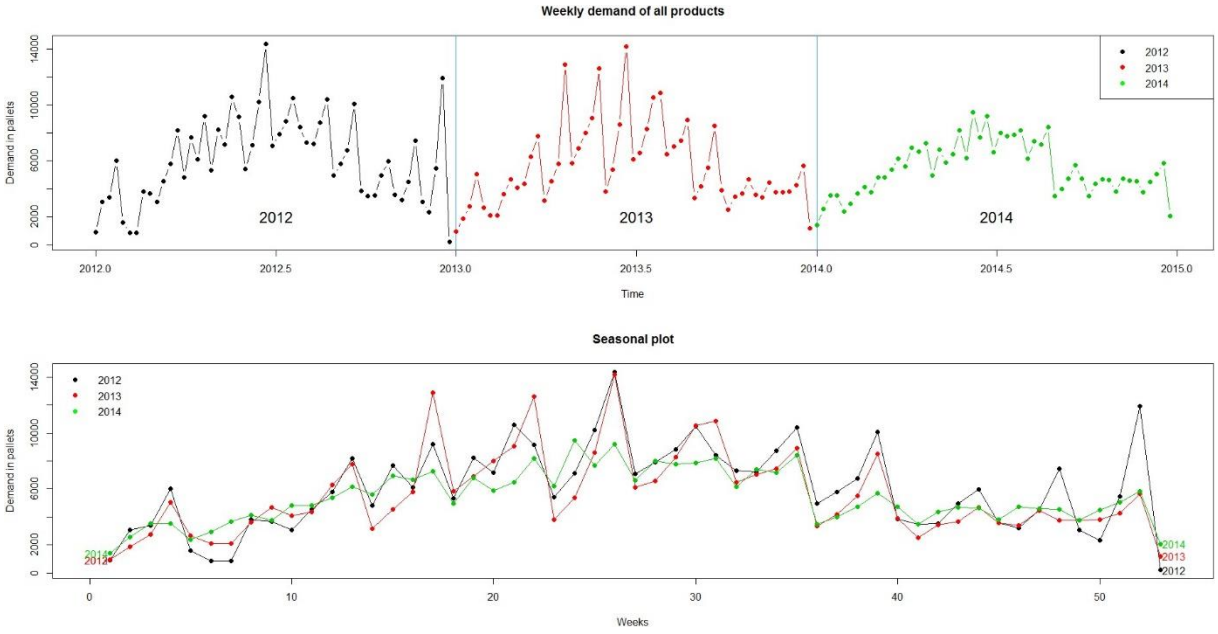

Figure 9.2 The demand data for observed food company.

Figure 9.2 demonstrates several important features and characteristics about the given data. First, the data has a strong seasonal character with peak sales occurring in the middle of the year (summer months). Secondly, the sub-seasonal plot (bottom plot in Figure 9.2) displays a significant pattern change of the downturn trend in 2014! These is very significant

characteristics for choosing the right forecasting model and for senior managers in the company since it reveals a significant drop in consumption and market loss. To further investigate the observed characteristics the seasonal trend decomposition (STL) methodology is used (Figure 9.3). The STL decomposition divides the original demand patterns in three components: seasonal, trend and remainder component.

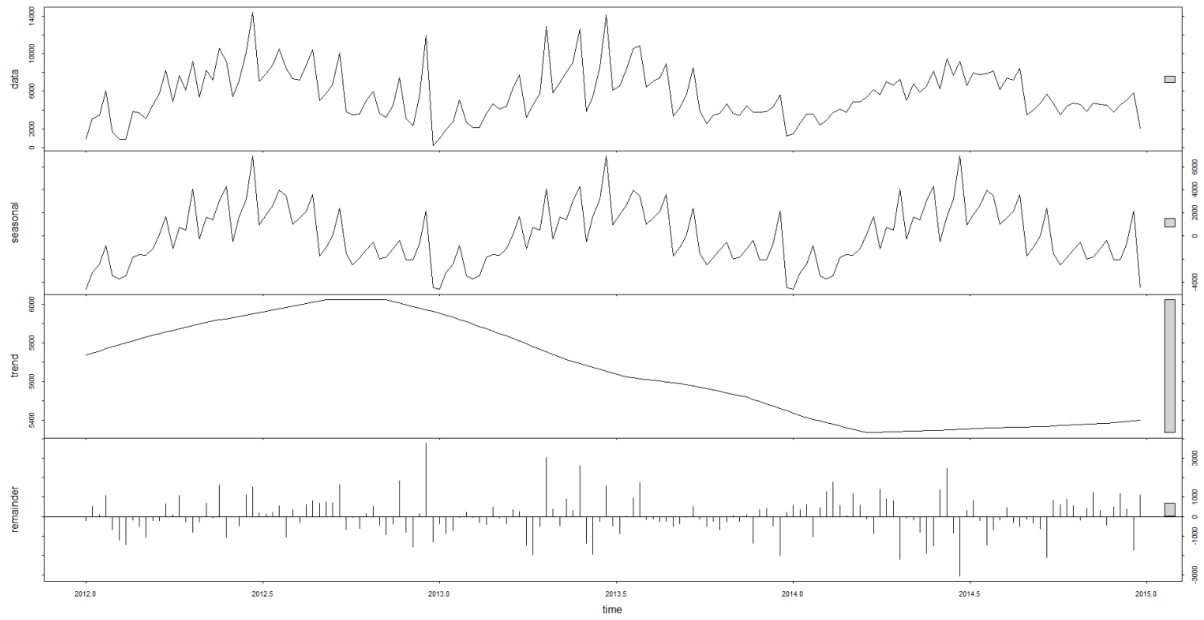

Figure 9.3 The STL decomposition of demand data.

The STL revealed that observed time series exhibit an additive nature, meaning that fluctuations around the trend-cycle curve do not significantly increase over time. As a result, Box-Cox transformations were not necessary for the raw time series. Decomposition revealed that the seasonal component is dominant in the observed series, showing high fluctuations within one year. Given the limited number of years of observations, it is difficult to identify a business cycle. Decomposition also indicated that the trend in the series is minimal, with a decreasing pattern starting from mid-2013.

These identified characteristics represented the important input during the process of design of the appropriate forecasting model. For this purpose, the S-ARIMA (Seasonal Autoregressive Integrated Moving Average) model is chosen. The S-ARIMA model is highly effective for forecasting because it combines both autoregressive and moving average components, along with differencing to make the data stationary. This model is particularly adept at capturing and modelling seasonal patterns in time series data, making it ideal for industries with cyclical

demand patterns, such as the food industry. Additionally, S-ARIMA's ability to handle complex seasonal structures and trends allows for more accurate and reliable forecasts, which are crucial for effective supply chain planning and inventory management. The S-ARIMA structural form is presented in Equation (1).

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = c + \Theta(B^m)\theta(B)\varepsilon_t, \quad (1)$$

where $\Phi(z)$ and $\Theta(z)$ are polynomials of order P and Q respectively, each containing no roots inside the unit circle. B is the backshift operator used for describing the process of differencing, i.e. $By_t=y_{t-1}$. If $c\neq 0$, there is an implied polynomial of order $d+D$ in the forecast function. Since the S-ARIMA is a highly parameterized model, the key question when using the S-ARIMA model is selecting the appropriate model order, which involves determining the values of p, q, P, Q, D, and d. If d and D are known, the orders p, q, P, and Q can be selected using an information criterion such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). The formulas for AIC and BIC are given by:

$$AIC = -2\log(L) + 2(k);$$

$$BIC = N\log\left(\frac{SSE}{N}\right) + (k+2)\log(N) \quad (2)$$

where $k=p+q+P+Q+1$ if a constant term is included and 0 otherwise, L is the maximized likelihood of the model fitted to the differenced data, SSE is the sum of squared errors, N is the number of observations used for estimation, and k is the number of predictors in the model.

For determining the optimal set of parameters Hyndman and Khandakar (2007) proposed the Canova-Hansen and KPSS unit root test in the following steps:

- Use the Canova-Hansen test for determining D in the ARIMA framework.
- Choose d by applying a successive KPSS unit root test on seasonally differenced data (if $D = 1$) or on the original data (if $D = 0$).
- Select the optimal values for p, q, P and Q by minimizing the AIC.

9.4 Developing the S-ARIMA forecasting model

Developing the S-ARIMA forecasting model

In accordance with the procedure mentioned above several parameter settings are tested and their performance is presented in Table 9.1. For testing the performances of different models several measures are used: mean absolute percentage error (MAPE), root mean square error (RMSE), mean absolute scaled error (MASE), AIC and BIC (Equations 2 and 3)

$$MAPE = \frac{1}{N}\sum_{1}^{N}|e_i|;$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}e_i};$$

$$MASE = \frac{1}{N}\sum_{1}^{N}|q_j|, \quad (3)$$

where e_i are residuals and $q_j = \frac{e_j}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|}$.

The most popular measures for the managers in the supply chains are RMSE and MAPE since they provide a "sense" of how good the model is performing in real numbers (RMSE) and percentages (MAPE).

Table 9.1 Performance of S-ARIMA models with different parameter settings.

Models[a]	RMSE	MAPE	MASE	AIC	BIC
S-ARIMA $(5,0,1)(1,0,0)_{52}$[b]	1023	14.18 %	0.60 %	86.62	110.50
S-ARIMA $(4,0,0)(1,0,0)_{52}$[c]	1041	14.53 %	0.62 %	86.73	105.31
S-ARIMA $(4,0,0)(0,1,1)_{52}$[d]	1882	22.12 %	1.05 %	41.74	53.560
S-ARIMA $(4,0,1)(1,0,0)_{52}$[e]	1050	14.18 %	0.60 %	88.23	109.47
S-ARIMA $(0,0,1)(0,1,0)_{52}$[f]	1797	21.42 %	1.02 %	36.52	40.460

[a] The model errors are calculated on the test data set.

[b] Detalis about S-ARIMA $(5,0,1)(1,0,0)_{52}$ model are provided bellow.

$_c\left(1-0.39B+0.06B+0.04B-0.46B\right)\left(1-0.65B^{52}\right)y_t=8.52$.

$_d\left(1-0.29B+0.19B-0.05B-0.19B\right)\left(1-B^{52}\right)y_t=\left(1+0.12B^{52}\right)e_t$.

$_e\left(1-0.29B+0.01B+0.05B-0.48B\right)\left(1-0.65B^{52}\right)y_t=8.52+\left(1+0.13B\right)e_t$.

$_f\left(1-B^{52}\right)y_t=\left(1+0.3B\right)e_t$.

Table 9.1 demonstrates that the S-ARIMA $(5,0,1)(1,0,0)_{52}$ model outperformed the competing models by achieving the lowest RMSE, MAPE, and MASE errors. The S-ARIMA $(5,0,1)(1,0,0)_{52}$ model form is presented in Equation (4), where $\phi_1=-0.5421$, $\phi_2=0.2962$, $\phi_3=-0.099$, $\phi_4=0.3974$, $\phi_5=0.4994$, $\Phi_1=0.9558$, $c=8.523$, and $\Theta_1=0.6345$.

$$\left(1-\phi_1B-\phi_2B-\phi_3B-\phi_4B-\phi_5B\right)\left(1-\Phi_1B^{52}\right)y_t=c+\left(1+\theta_1B\right)e_t\,, \tag{4}$$

Similar performance was observed with the S-ARIMA $(4,0,0)(1,0,0)_{52}$ and S-ARIMA $(4,0,1)(1,0,0)_{52}$ models. When performing forecasts the S-ARIMA $(5,0,1)(1,0,0)_{52}$ model on average produces an error of 1023 products, which translates to 14.18 %. These results highlight the importance of carefully selecting the terms to include in an S-ARIMA model since there are no significant differences between the performances of the model with different parameters. The evaluation indicated that models incorporating autoregressive (*p, P*) and moving average components (*q, Q*) were more effective in forecasting beverage consumption than those incorporating seasonal or non-seasonal differencing. Additionally, the comparative review revealed some unexpected findings, such as the models including seasonal differencing S-ARIMA $(4,0,0)(0,1,1)_{52}$ and S-ARIMA $(0,0,1)(0,1,0)_{52}$ performing worse than a simple average naive forecast model, as evidenced by their MASE errors exceeding one.

9.5 Forecasts of the future demand

Figure 9.4, demonstrates the performance of S-ARIMA $(5,0,1)(1,0,0)_{52}$. The top left panel represents the starting demand data, coloured for easier distinction of different marketing years. The top right panel is the input data to the S-ARIMA $(5,0,1)(1,0,0)_{52}$, representing the training data on which parameters in S-ARIMA are determined following the procedure described in subchapter 9.3. This is very important to understand how difficult the job of the

forecasting model is, since in this case it has two years of data as input and needs to forecast future demand one year ahead! This is quite a common case scenario in supply chains and logistics since there is an unwritten rule that companies keep a history of their data for three years after which they discard the data.

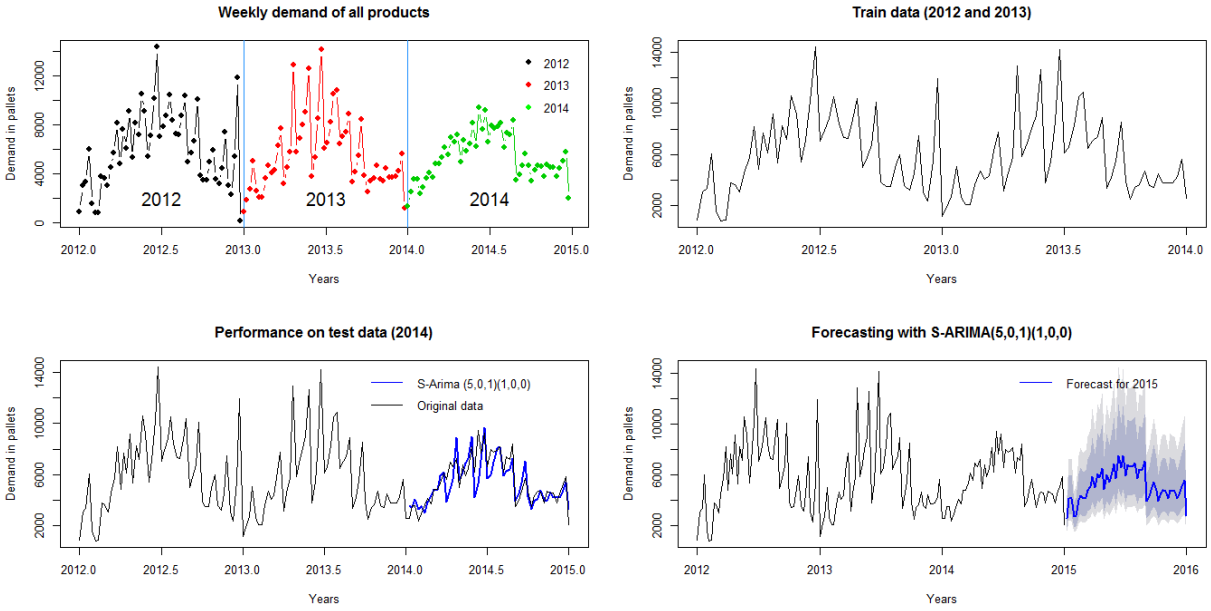

Figure 9.4 Train, test and forecasted demand of S-ARIMA $(5,0,1)(1,0,0)_{52}$ model.

The bottom left panel in Figure 9.4, presents the performance of an observed model. Model performance is already presented in Table 9.1 via different statistical measures, but for managers is usually hard to get the feeling of how good or bad the model is. For this purpose, the panel graphically demonstrates the performance of the model. It could be argued that the model follows the test data pretty well and in the majority of periods demonstrates excellent performance. In order to generate future forecasts, S-ARIMA $(5,0,1)(1,0,0)_{52}$ refitted by adding the data from 2014 to the train data. After that model produced 52- weekly steps ahead forecasts for 2015. The forecasts for 2015 are presented in the bottom left panel. The forecasts are accompanied by 80% and 95% prediction intervals, which demonstrate possible future forecasts disperse from the mean predicted values. The model forecasts a continued decline in demand that began in 2014. The causes of this decline could be diverse and should be further explored by managers at the strategic level of the company.

References Chapter 9

- Hyndman, R., & Khandakar, Y. (2007). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3). http://www.jstatsoft.org/v27/i03
- Makridakis, S., Wheelwright, S., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3rd ed.). New York: John Wiley & Sons.
- Makridakis, S., Wheelwright, S., & McGee, V. (1983). *Forecasting: Methods and applications* (2nd ed.). New York: John Wiley & Sons.
- Mircetic, D. (2018). *Boosting the performance of top-down methodology for forecasting in supply chains via a new approach for determining disaggregating proportions*. University of Novi Sad.
- Mircetic, D., Nikolicic, S., Maslaric, M., Ralevic, N., & Debelic, B. (2016). Development of S-ARIMA model for forecasting demand in a beverage supply chain. *Open Engineering, 6*(1).
- Mircetic, D., Rostami-Tabar, B., Nikolicic, S., & Maslaric, M. (2022). Forecasting hierarchical time series in supply chains: An empirical investigation. *International Journal of Production Research, 60*(8), 2514-2533.
- Mircetic, D., Nikolicic, S., Stojanovic, D., & Maslaric, M. (2017). Modified top-down approach for hierarchical forecasting in a beverage supply chain. *Transportation Research Procedia, 22*, 193–202.
- Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2020). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research, 290*(1), 99–115.
- Rostami-Tabar, B. (2013). *ARIMA demand forecasting by aggregation*. Université Sciences et Technologies Bordeaux I.
- Rostami-Tabar, B., & Mircetic, D. (2023). Exploring the association between time series features and forecasting by temporal aggregation using machine learning. *Neurocomputing, 548*, 126376.
- Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research, 252*(1), 1-26.

10. Artificial intelligence and machine learning in supply chains

What is artificial intelligence (AI)? Is it really an "alive" creature capable of thinking and making its own decisions based on one its mind, past experiences, ethics, beliefs, etc? How is it connected to machine learning (ML)? Are AI and ML the same thing?

What tools are used in AI & ML? What roles do AI and ML play in the business context and how can it be used in everyday business operations and optimizations? Are there specific architectures and examples of applying AI and ML to supply chains?

On these and similar questions, we will try to provide answers in the following chapter, closing with a real case study example of the application of AI & ML algorithms in the distribution warehouse.

10.1 What is artificial intelligence?

The field of AI began in the 1950s with computer scientists asking, "Can computers think like humans"? Researchers at that time were enthusiastic about the possibility of teaching computers to perform complex tasks and accordingly developed a set of different algorithms for that purpose. The definition of the field could be stated as the effort to automate intellectual tasks normally performed by humans (Chollet, 2021). The most notable algorithms in the field of AI today are coming from ML and deep learning, which are subsets of AI. Besides ML and deep learning AI includes a lot of non-learning algorithms. Moreover, we can argue that these kinds of algorithms were more dominant in the early phase of AI development. Accordingly, this is part of AI known as Symbolic AI which is based on the idea that human level of performance and intelligence can be achieved by programming computers with a large set of explicit rules for solving the observed problem. This approach provided excellent results in a logical problem which were well defined, like a computer playing a chess game, but it proved to be complicated for more complex problems. The real world has proven to be much more complicated than all explicit programming rules could be inserted into the computer. This approach is centred around an idea for a given situation do this or that (if-then rules). This is

an easily understandable approach, but on the other hand very time-consuming and sometimes very hard to determine all possible scenarios which need to be inserted into the program. As a funny example, but a good illustration of a given topic, please see Figure 10.1, which demonstrates several scenarios for determining the forecast based on the stone status.

Figure 10.1 If – Then programming rules (Gibbs, M. 2019).

The field of Symbolic AI has its biggest popularity in the 1980s with the emergence of expert systems (ES). ES represent a subset of decision support systems (DSS) (Turban, 1998), focused on delivering computerized decision-making abilities akin to those of a human specialist within a particular field. These systems are crafted to tackle intricate problems by employing a series of rules or algorithms that simulate human reasoning processes. Olson and Courtney describe expert systems (ES) as computer programs that simulate human thought processes to make decisions within a specific domain, incorporating a degree of artificial intelligence to match the conclusions a human expert would reach (Olson & Courtney, 1992). An ES component is particularly useful for supporting decision-makers in areas that require specialized knowledge (Turban et al., 2005). Essentially, an ES captures the expertise from a human expert (or another source) and transfers it to the computer. This technology can either aid decision-makers or fully substitute them, making it one of the most widely applied and commercially successful forms of artificial intelligence (Turban et al., 2005). One key reason for developing an ES is to distribute expert knowledge to a broader audience (Jackson, 1999). In the following subchapters, we will demonstrate the application of ES based on AI & ML algorithms, in the central warehouse as part of the overall DSS to managers.

Today, the field of AI consists of various approaches and algorithms, but the most used ones are described on Figure 10.2. It has diverged from the ES systems and the re-emergence of the field is mostly credited to the deep learning algorithms which had significant success in the last 12 years in the problems of image recognition, speech recognition, image segmentation, facial recognition, etc. Deep learning utilizes multiple layers of abstraction to identify complex patterns in high-dimensional data. This approach has achieved significant advancements in fields like speech and image recognition, drug discovery, and natural language processing. The deep learning's has ability to automatically discover relevant features and reduces the need for human intervention in feature design, making it highly efficient in leveraging large datasets and computational power (LeCun, Bengio, & Hinton, 2015).

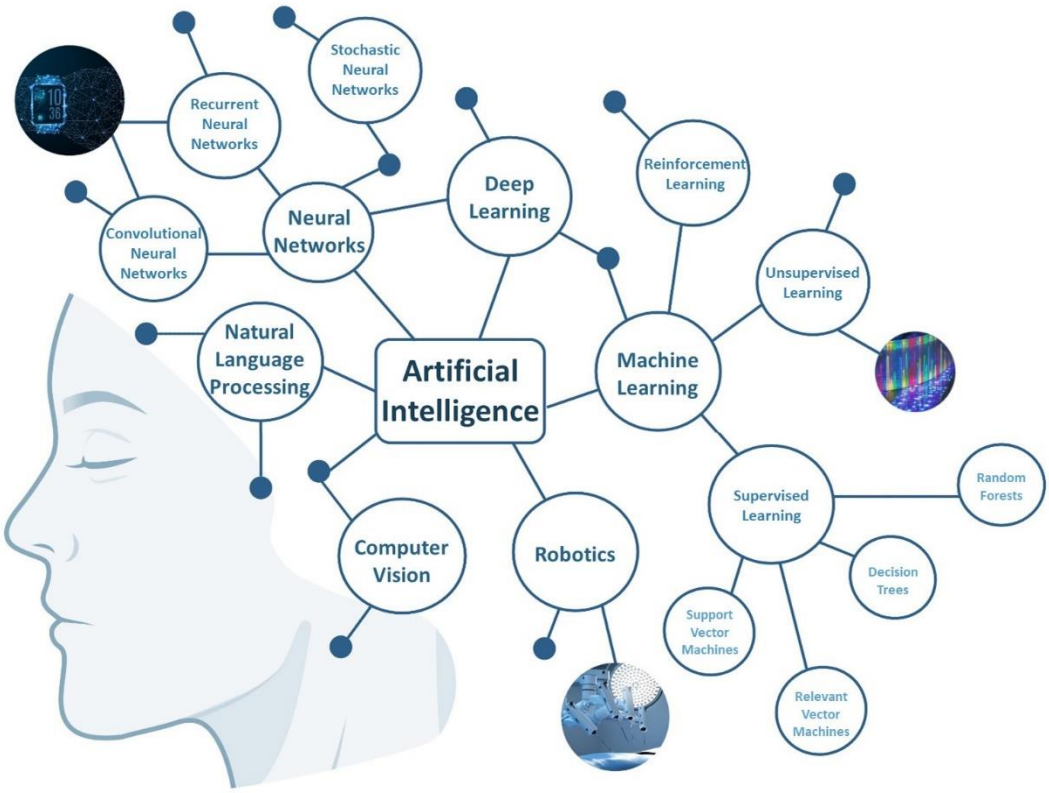

Figure 10.2 The main fields and subfields of AI (Athanasopoulou et al., 2022).

The big step towards today situation was research in the digit classification conducted by Hinton, Osindero, and Teh (2006), which manged to achieve more than 98% accuracy on classifying Modified National Institute of Standards and Technology (MNIST) data base. One way to think about how AI transitioned from Symbolic AI to ML, and what is the essence of the ML is to imagine ML algorithms as an amorphous mass that shapes itself according to the

desired outcomes. The system of rules that takes input to output changes from problem to problem and adapts to the existing situation. It aims to find rules that will automate the task by seeking statistical patterns within the data. This kind of approach for solving different problems significantly reduced the time for system set up (compared to the if – then rules), and make it more universal approach for tackling various problems.

Bengio, Lecun, and Hinton (2021) emphasised that the future of the AI is in the deep learning and the revolutionary impact of soft attention and transformer architectures in AI. These innovations allow neural networks to dynamically focus on important inputs and store information in differentiable memories, significantly improving sequential processing.

10.2 What is the ecosystem of AI & ML?

The ecosystem of AI & ML algorithms consists of three key pillars:

- Input data;
- Output data;
- Cost function.

The input data represents data recordings of specific feature or features (depending on an observed problem). The accuracy of the input data is crucial for building accurate algorithms. This is often not the case in real applications and usually significant time and effort are devoted to data collecting, cleaning, wrangling, unifying, checking on false inputs, etc. Besides, accurate data, another important aspect of input data characteristics is its representation and encoding. Different ways of encoding the data can reveal different features of the data and significantly "help" ML models in revealing the hidden patterns in the data. Here we see the Achilles' heel of ML algorithms. Often, too much attention is given to creating the methods for extracting information and intelligence from the data (i.e., the algorithms themselves), while insufficient attention is paid to the input data and its relationship with the output data. It is generally taken for granted that the input data has a causal relationship with the output data, which is sometimes not the case at all. Therefore, the next big step in ML development should be finding better ways to collect, represent and encode the data.

Output data represent the measurements of the particular problem that we are trying to solve. In a classification problem, it would be the label of the classes. In regression, it will be a real

number we are trying to predict. In a problem-specific context, for speech-recognition problems, the output data can be human-generated transcripts of the sound files. In an image recognition problem, the output can be image class labels, etc.

The cost function represents the way of measuring how the AI & ML is performing. Basically, we would ideally want answers from the algorithms to match the output data, for a given input data. The cost function is also a feedback signal to the set of parameters that guide the algorithm's work, i.e., it allows the optimization of overall algorithm performance via the process of learning (finding the optimal set of parameters). The learning process typically involves supervised learning, where a model is trained on labeled data to minimize prediction errors through techniques like stochastic gradient descent and backpropagation. This enables the model to adjust its internal parameters effectively, leading to improved performance on tasks such as object detection and classification (LeCun, Bengio, & Hinton, 2015).

10.3 What tools are used in ML?

Generally, ML algorithms can be classified into two main categories: supervised and unsupervised learning.

Supervised learning involves training algorithms on a labelled dataset, where each input data point is paired with the correct output. This clear "picture" of what the correct answer should be for a given input allows the algorithm to learn the mapping function from inputs to outputs. Accordingly, both the input and output data are known (Athanasopoulou et al., 2022). Common applications of supervised learning include classification tasks (e.g., determining whether an email is spam or not) and regression tasks (e.g., predicting house prices based on various features). Some of the most popular algorithms which have been proven by numerous applications are generalized additive models, random forests, boosting, classification and regression trees, support vector machines, extended linear regression, logistic regression, k-nearest neighbors, linear discriminant analysis, lasso, neural networks, adaptive neuro-fuzzy inference system, etc (Rostami-Tabar & Mircetic, 2023). Supervised learning is powerful because it leverages human-annotated data to achieve high accuracy in predictions. However, its effectiveness depends heavily on the quality and quantity of the labelled data.

In contrast, unsupervised learning deals with datasets that lack labelled responses. Accordingly, unsupervised machine learning algorithms use unlabelled datasets that include

only inputs (Athanasopoulou et al., 2022). Here, the algorithm is provided only with the input data, and its goal is to find underlying patterns, structures, or relationships within the data. Common techniques in unsupervised learning include clustering (e.g., grouping customers by purchasing behaviour) and dimensionality reduction (e.g., reducing the number of variables in a dataset while retaining important information). Unsupervised learning is valuable for exploratory data analysis and discovering hidden structures in data. It is often used when labelled data is scarce or unavailable.

Another important category, although distinct from supervised and unsupervised learning, is reinforcement learning. Here, the algorithm learns by interacting with an environment and receiving feedback in the form of rewards or penalties. This trial-and-error approach helps the algorithm learn optimal actions to maximize cumulative rewards. Reinforcement learning is widely used in fields such as robotics, game playing, and autonomous systems.

One of the most popular and successful algorithms for ML comes from the branch of neural networks. Neural networks exist since the 1950s but gained their popularity in the 1980s and in recent 12 years. They are built on the approximation of biological neurons and the way they share pieces of information in the brain, but beyond that, there are no significant connections between these two. Today, the most commonly used form of neural networks are in the form of deep learning, which represents several stack hidden layers between the input and the output features, which perform several nonlinear transformations of the input features. Since this has been proven very successful deep learning is today one of the most prominent subfields of ML (Chollet, 2021).

10.4 Case study?

The findings of Wenzel, Smit, and Sardesai (2019) on ML in supply chain management indicate a growing integration of ML applications across various SC tasks. Accordingly, observed case study represents the application of AI & ML, performed in the central warehouse of the food factory (Mirčetić et al., 2016; Mircetic et al., 2014). In the factory complex, there are 30 forklifts. Forklifts are engaged in various operations inside the complex, which are crucial for logistics operations in production, warehousing and dispatching the products. The central warehouse has the capacity of 11 100 pallet places and the annual output from 300 000 to

350 000 pallets. Currently, the factory is supplying around 20 000 supermarkets via direct delivery.

The problem of forklift engagement is related to the fact over or under-engagement of the forklifts in the different factory processes leads to significant financial and market losses. Currently, the process of decision-making making where and what will each forklift do is based on the expert (managers) decisions. Expert decisions are based on their experience, without the help of any decision-support system (DSS). Ample empirical evidence suggests that human intuitive judgment and decision-making often fall short of optimal, particularly under conditions of complexity and stress (Druzdzel & Flynn, 2002). This underscores the importance of incorporating decision support systems (DSS) to assist experts in the decision-making process.

In this application, we have chosen several ML algorithms to assist in optimizing the loading warehouse operation. The ML algorithms are assembled in a unique decision-making framework which serves as DSS for managers and experts in a given company. Moreover, the entire decision-making DSS can be observed as an AI platform, since it constantly recalculates the suggestions from several ML models (how many forklifts to use and which ones) and automatically chooses the best ones, regarding the provided operators' inputs.

Problem description

The loading process is crucial for warehouse logistics, impacting market service levels. During shipments, the warehouse expert determines the number and selection of forklifts for loading, guided by three factors: (1) completing loading within the specified timeframe, (2) minimizing disruption to other forklift tasks, and (3) aligning forklift use with maintenance capabilities, which can handle two overhauls simultaneously. Each forklift undergoes four to five maintenance overhauls annually.

Forklifts are vital for loading operations, which must support the company's marketing strategy while ensuring the smooth operation of other activities. Misallocating forklifts can lead to resource underutilization or harm the company's reputation and service levels. Delays in loading incur penalties. The manager must coordinate forklift use across all activities to avoid simultaneous overhauls and manage varying maintenance needs. Though managers typically make accurate decisions, high-stress environments can lead to errors. Therefore, a DSS is needed to enhance decision-making confidence and reliability.

AI & ML as DSS for central warehouse

The first step for generating the AI & ML systems is to acquire a stable and correct source of knowledge (database) and to shed light on the business roles it needs to support. Therefore, Figure 10.3 presents a methodology for building the AI & ML DSS system.

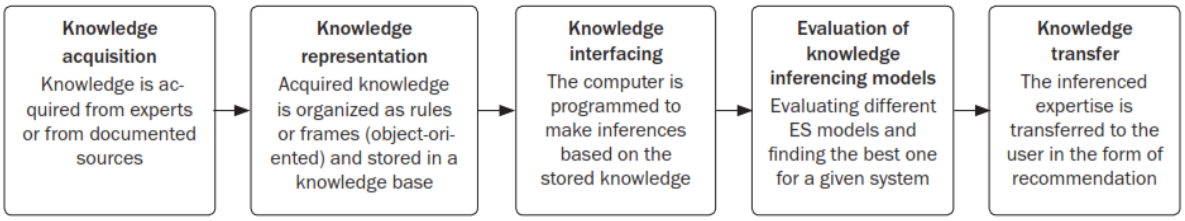

Figure 10.3 Methodology steps for building the AI & ML DSS system (Rainer & Turban, 2008; Turban, Aronson, & Liang, 2005).

Knowledge acquisition was achieved through manager interviews, observing their decision-making processes, and reviewing warehouse records. To develop the DSS for the given problem, two knowledge bases were established. The first knowledge base includes decisions on the number of forklifts deployed in the loading zone (434 expert decisions), while the second covers which forklifts were used (368 expert decisions) in various operational scenarios. During the knowledge inference stage, several ML algorithms were applied using Matlab software: Adaptive neuro-fuzzy inference system (ANFIS), generalized additive models (GAM), Random forests, Boosting, classification and regression trees (CART), Extended Linear Regression, Logistic Regression, k-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA). Various ML models were evaluated, and those with the best performance were identified. ANFIS and CART demonstrated superior results and were selected as the final DSSs for practical application in the company. Knowledge transfer was facilitated through the user interface of the final DSS models. The structure and logic of the DSS is illustrated in Figure 10.4.

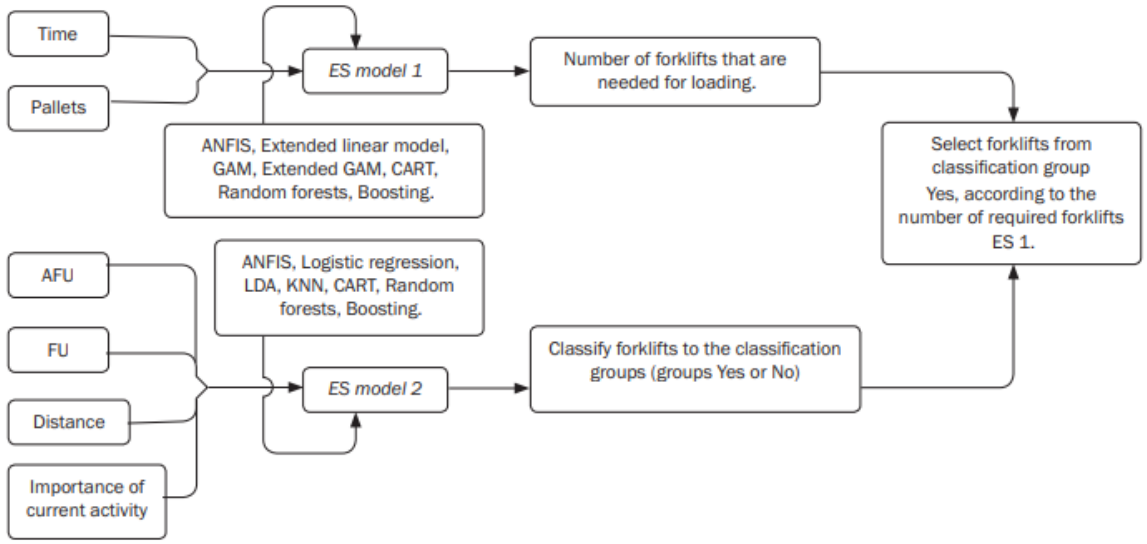

Figure 10.4 Building structure of warehouse DSS based on the AI & ML algorithms.

The DSS framework consists of an input layer, consisting of several key factors which influence forklift engagement. The ML layer has ML models that are recalculating the suggestion on how many and which forklifts to use in a given input scenario. The best-performing models are chosen as expert system models (ES models) since the knowledge base on which the ML models are created is extracted from the experts. The first model focuses on determining the number of forklifts required in a loading zone (ES model 1). The second model addresses the problem of selecting which specific forklifts should be engaged (ES model 2). Both models are developed using supervised machine learning techniques. According to Turban, Aronson, and Liang (2005), machine learning has demonstrated excellent results in designing intelligent decision support systems (DSS). The ES models send signals (ML suggestions and proposals) further to the sorting operation, where each forklift that is classified in the sorting group "Yes", can be engaged in a given loading operation.

The factors influencing the manager's decisions were identified through consultations. For determining the number of forklifts to deploy in the loading zone, the key factors are the specified loading time (15 to 135 minutes) and the amount of cargo (15 to 225 pallets). When choosing which forklifts to engage, the manager considers the importance of the current activity (rated 1 to 9 by company policy), the forklift's utilization rate, its distance from the loading dock, and the average utilization rate of all forklifts. Each forklift has a set number of working hours before an overhaul is needed, and its usage is restricted beyond this limit.

Forklift Utilization (FU) is the percentage of working hours used by an individual forklift, while Average Forklift Utilization (AFU) is the average working hours used by all forklifts. A higher AFU suggests that most forklifts will soon need an overhaul.

The user interface of DSS

In the majority of input situations, the best performance was demonstrated via ANFIS and CART. Accordingly, they were chosen as the engines of the given DSS and its ESs. The ES model 1 user interface is presented in Figure 10.5, and it allows operators to make decisions quickly and easily about the number of forklifts to deploy by simply moving a vertical line through the domain of input variables, based on the specified loading time and cargo amount.

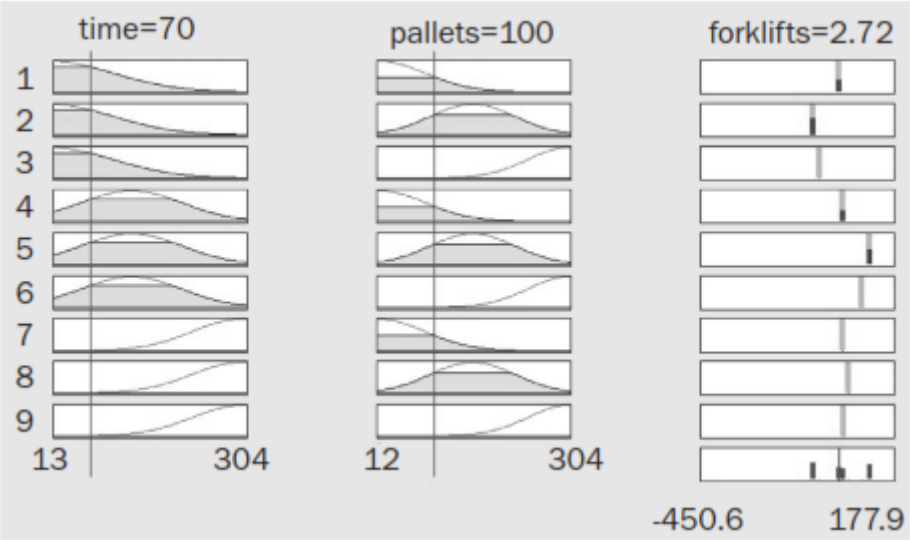

Figure 10.5 Fuzzy Inference System of ES model 1.

ES Model 2 serves as a supplementary tool to ES Model 1, enhancing decision-making by providing information on whether a specific forklift should be deployed in the loading zone (Figure 10.6). By considering the position of a forklift (distance from the loading zone), its current activity (importance of activity), its utilization of working hours (FU), and the average utilization of all forklifts (AFU), users can easily determine if a particular forklift is suitable for loading or if another should be selected. The CART decision tree is straightforward to interpret, eliminating the need to input values into software. Instead, the tree from Figure 10.6 can be printed and displayed prominently in the warehouse for quick reference.

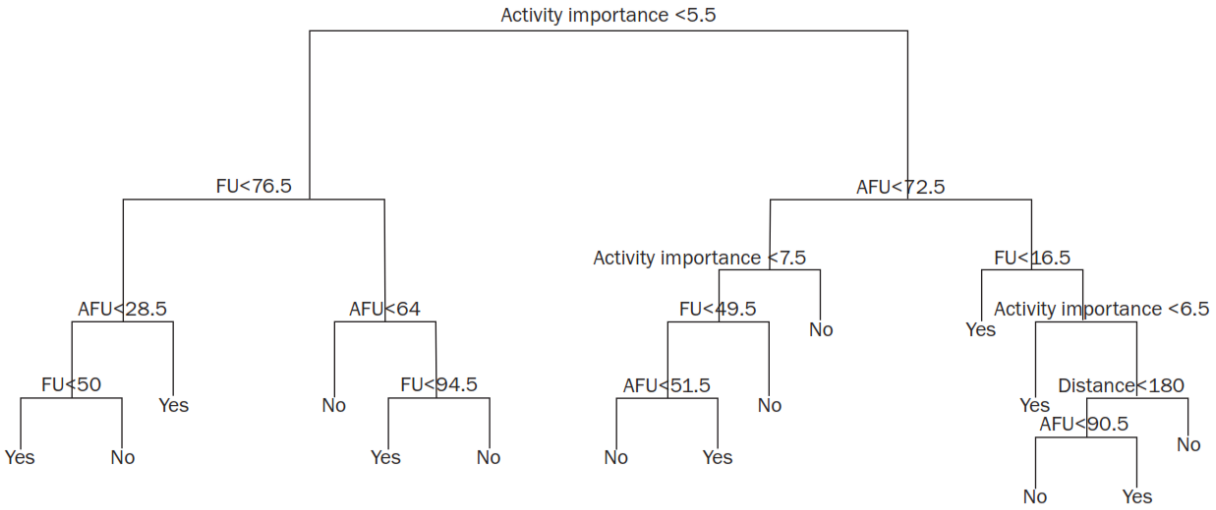

Figure 10.6 ES model 2 decision tree regarding the forklift engagement.

Managers can utilize the presented DSS daily, aiding in achieving higher supply chain responsiveness to customer demands and ensuring a high probability of on-time delivery. The proposed AI & ML DSS has demonstrated successful results in acquiring the expert's "know-how" knowledge and capturing their "inference logic." By using this method, managers' expertise can be extracted and applied to other warehouse operations. This is particularly valuable for practitioners since hiring warehouse experts is often expensive. Additionally, DSS can also serve as a training tool for novice managers, helping them gain experience and improve their decision-making skills over time. Therefore, AI and ML systems that can simulate a manager's decisions are essential tools, offering significant cost savings and increased efficiency in warehouse operations.

References Chapter 10

- Athanasopoulou, K., Daneva, G. N., Adamopoulos, P. G., & Scorilas, A. (2022). Artificial intelligence: The milestone in modern biomedical research. BioMedInformatics, 2(4), 727-744. https://doi.org/10.3390/biomedinformatics2040049
- Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, *64*(7), 58-65.
- Chollet, F. (2021). Deep learning with Python. Simon and Schuster.
- Druzdzel, M. J., & Flynn, R. R. (2002). Decision support systems. In A. Kent (Ed.), Encyclopedia of library and information science (2nd ed.). Marcel Dekker, Inc.
- Gibbs, M. (2019, December 17). Table-driven programming and the weather forecasting stone. Global Nerdy. https://www.globalnerdy.com/2019/12/17/table-driven-programming-and-the-weather-forecasting-stone/

- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.
- Jackson, P. (1999). Introduction to expert systems. Addison-Wesley.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.
- Mirčetić, D., Ralević, N., Nikoličić, S., Maslarić, M., & Stojanović, Đ. (2016). Expert system models for forecasting forklifts engagement in a warehouse loading operation: A case study. Promet-Traffic & Transportation, 28(4), 393-401.
- Mircetic, D., Lalwani, C., Lirn, T., Maslaric, M., & Nikolicic, S. (2014, July). ANFIS expert system for cargo loading as part of decision support system in warehouse. In 19th International Symposium on Logistics (ISL 2014).
- Rostami-Tabar, B., & Mircetic, D. (2023). Exploring the association between time series features and forecasting by temporal aggregation using machine learning. Neurocomputing, 548, 126376.
- Olson, D. L., & Courtney, J. F. (1992). Decision support models and expert systems. Macmillan.
- Rainer, R. K., & Turban, E. (2008). Introduction to information systems: Supporting and transforming business. John Wiley & Sons.
- Turban, E. (1998). Decision support and expert systems (2nd ed.). Macmillan.
- Turban, E., Aronson, J., & Liang, T.-P. (2005). Decision support systems and intelligent systems (7th ed.). Pearson Prentice Hall.
- Wenzel, H., Smit, D., & Sardesai, S. (2019). A literature review on machine learning in supply chain management. In W. Kersten, T. Blecker, & C. M. Ringle (Eds.), Artificial Intelligence and Digital Transformation in Supply Chain Management: Innovative Approaches for Supply Chains (Vol. 27, pp. 413-441). https://doi.org/10.15480/882.2478

LIST OF FIGURES

LIST OF TABLES

BUSINESS ANALYTICS SKILLS FOR THE FUTURE-PROOFS SUPPLY CHAINS -

BUSINESS ANALYTICS SKILLS FOR THE FUTURE-PROOFS SUPPLY CHAINS -