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.

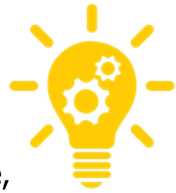
In addition, statistics are used to monitor and improve the quality of products and services in the supply chain, including quality control at suppliers.

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	𝍷𝍷𝍷	3
50 - 58	𝍸 𝍷	6
59 - 67	𝍸	5
68 - 76	𝍸 𝍷	6
77 - 85	𝍷𝍷	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.

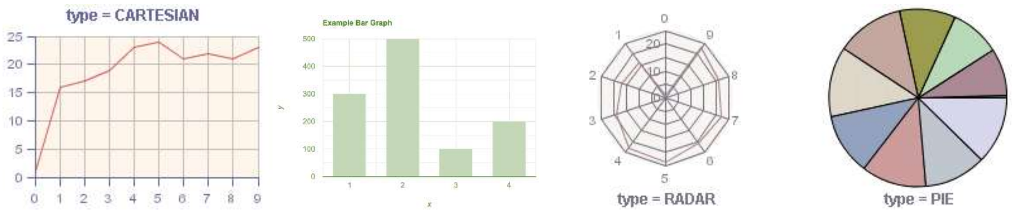


Figure 1.2 Examples of graphical representations of data.

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 analyse 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 analyse 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.

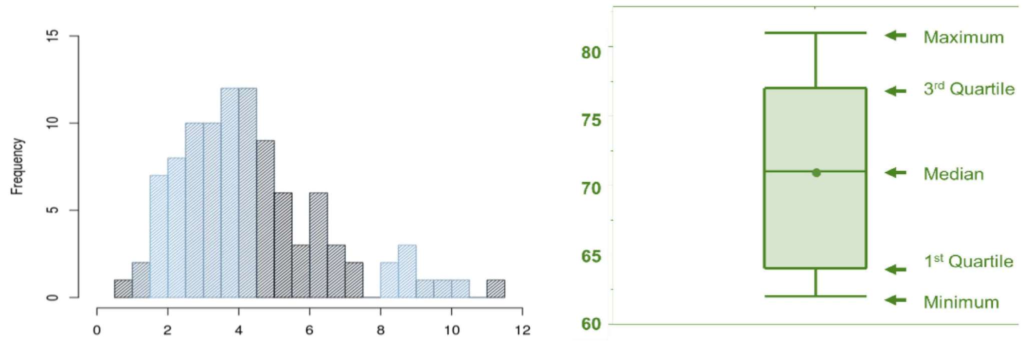


Figure 1.3 Histogram and Quantile chart (Box plot).

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	𝍷𝍷𝍷	3
50 - 58	𝍸 𝍷	6
59 - 67	𝍸	5
68 - 76	𝍸 𝍷	6
77 - 85	𝍷𝍷	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.

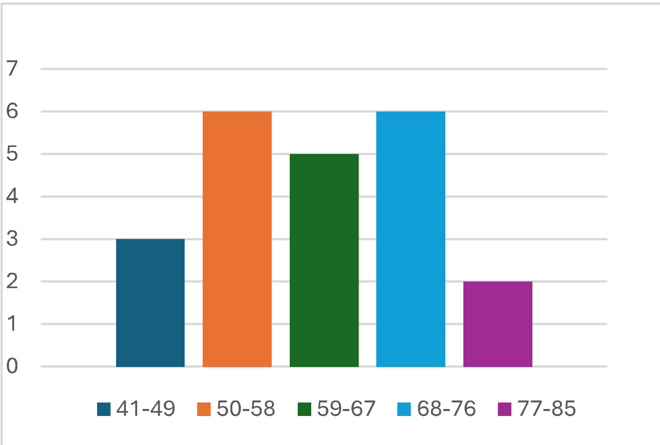


Figure 1.5 Frequency distribution graph.

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:

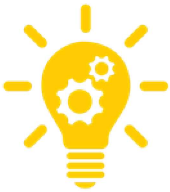
$$170\ cm\ \pm\ 1.96\ \times\ (\frac{5\ cm}{\sqrt{100}})\ =\ 170\ cm\ \pm\ 0.98\ cm$$

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).

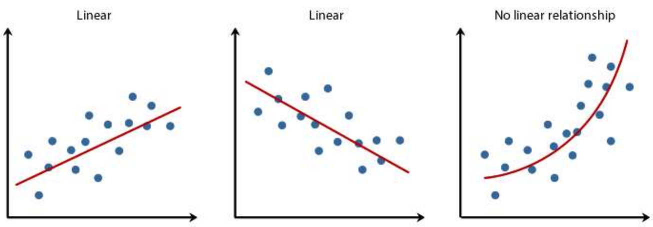


Figure 1.6 Simple linear regression graphs.

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.

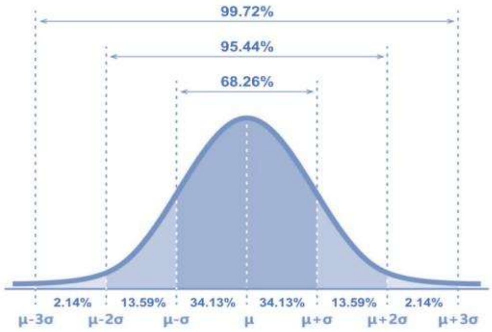


Figure 1.8 Normal distribution graph.

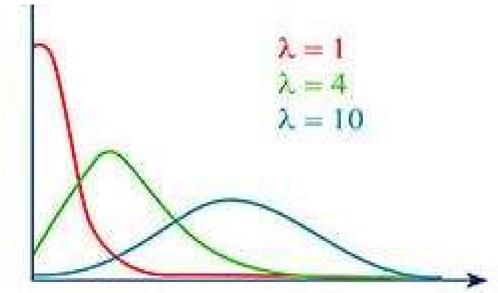


Figure 1.7 Poisson distribution graph.

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:10.2174/97898151231351230101

- 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);

- 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