TEXTBOOK OF AGRICULTURAL STATISTICS

A comprehensive guide in basic statistics for agricultural research

Dr. Pratheesh P. Gopinath Dr. Manju Mary Paul
Dr. Adarsh V.S. Mohammed Hisham M.

2024-02-12

Table of contents

Welcome

Welcome to the *Textbook of Agricultural Statistics* – a comprehensive resource thoughtfully crafted by the Department of Agricultural Statistics at the College of Agriculture, Vellayani, Kerala Agricultural University. This book was created to meet the need for a clear, accessible, and practical guide to statistics tailored for agricultural research.

Statistics is an essential tool in agriculture, enabling researchers and practitioners to uncover patterns, validate results, and make data-driven decisions. However, the subject is often perceived as complex and challenging to master. This textbook is designed to change that perception, offering straightforward explanations and practical guidance to make statistical concepts approachable and applicable.

As John Tukey once said, *"The greatest value of a picture is when it forces us to notice what we never expected to see."* We have embraced this philosophy by incorporating clear examples, practical applications, and data visualizations to illustrate concepts and deepen understanding.

While this book is primarily written with undergraduate students in mind, its simplicity and focus on real-world applications make it a valuable resource for a wide audience, including post-graduate students, researchers, and anyone seeking to build a strong foundation in statistics and experimental design.

You will find clear explanations, practical examples, and step-by-step instructions throughout the chapters, all tailored to the unique needs of agricultural studies. Our goal is to ensure that learners at all levels can confidently apply statistical methods to their work and research.

Whether you are new to statistics or looking to revisit the basics with a fresh perspective, we hope this book serves as a supportive companion in your journey to understanding and applying statistical tools effectively in agriculture.

Acknowledgements

This book is the result of collaborative efforts among dedicated teachers and statisticians, but the majority of the reviewing, editing, and refinement has been inspired and shaped by the students.
For two years, this book was made available online on the MeLON (Module for eLearning and Online Notes) platform of the College of Agriculture, Vellayani. During this time, students provided valuable feedback, pointed out areas for improvement, and offered insights that

greatly enhanced the quality and clarity of the content. We are sincerely grateful to all the students whose suggestions and input played a key role in the development of this textbook.

We would also like to thank the College of Agriculture, Vellayani, for fostering an environment that encourages learning, growth, and the sharing of ideas.

Note from the Publisher

This textbook is published by **PAPAYA**, the publication division of Statoberry LLP, which is committed to providing high-quality educational resources for agricultural research. The online version of this book is available for free, in line with our dedication to open access and knowledge sharing. Visit us at PAPAYA.

Copyright Information

Preface

For a long time, I have dreamed of writing a book that truly serves the needs of undergraduate students in agriculture—a book that demystifies statistics and makes it accessible and practical for their studies and research. Statistics, while being an essential tool in agricultural sciences, is often presented in ways that make it seem more complicated than it actually is. Textbooks in this field tend to delve into intricate details that go far beyond what most agricultural students require, leaving them overwhelmed and disconnected from the subject's practical relevance.

This book is my humble attempt to change that. It has been written with undergraduate students in mind, focusing on the basics of statistics and their direct applications in agricultural research. Each chapter is designed to simplify complex concepts, making them clear, relatable, and easy to understand. While the primary audience is undergraduate students, this book can also serve as a helpful resource for anyone looking to brush up on the fundamentals of statistics.

The journey of writing this book has been greatly enriched by the feedback and insights of the students at the College of Agriculture, Vellayani. For two years, an earlier version of this book was made available on MeLON (Module for eLearning and Online Notes), our college's online platform. The students, with their thoughtful suggestions and sharp observations, have helped refine the content and shape it into what it is today. Their enthusiasm and curiosity have been a constant source of inspiration throughout this process.

I would also like to express my heartfelt gratitude to my co-authors, **Dr. Manju Mary Paul**, **Dr. Adarsh V. S.**, and **Mohammed Hisham M.**, whose expertise, commitment, and contributions have been invaluable in bringing this book to life. Their collaboration and dedication have greatly enhanced the quality and depth of this work.

A special thanks to **Jithin Chandran**, **Gaatha Prasad**, **Anjana Biwas T.**, and **Varsha H.** for their valuable suggestions and minor corrections. They were postgraduate students in Agricultural Statistics at the time of writing this book, and their support has significantly increased the quality of this work.

I hope this textbook becomes a guiding light for students and researchers alike, helping them build a solid foundation in statistics while inspiring confidence in their ability to use these tools effectively. If this book makes statistics less intimidating and more approachable for even one reader, I will consider my efforts worthwhile.

With deep gratitude to my students, colleagues, and everyone who supported this work, I present this book as a tool to empower the next generation of agricultural scientists and researchers.

Dr. Pratheesh P. Gopinath
Head
Department of Agricultural Statistics
College of Agriculture, Vellayani
2 December 2024

1 Basics of statistics

Statistics is the science of understanding, analyzing, and interpreting data. It plays a crucial role in making informed decisions across various fields, from agriculture to medicine, economics to environmental studies. This chapter serves as an entry point into the fascinating world of statistics, introducing you to its basic concepts and practical applications.

We begin by exploring the origins and definitions of statistics, emphasizing its relationship with mathematics and its distinct role in solving real-world problems. From there, we focus on the importance of data—the raw material of statistics—examining its types and how it is collected, organized, and analyzed.

The chapter also covers essential concepts such as population and sample, variables and constants, and the different types of variables. These concepts form the building blocks for understanding how statistical studies are designed and conducted.

Finally, we introduce frequency distributions—an indispensable tool for summarizing and interpreting data. Topics such as construction of frequency distributions, grouped and cumulative frequency distributions, and relative frequency will help you make sense of data and uncover underlying patterns.

By the end of this chapter, you will have a comprehensive understanding of the core principles of statistics, setting the stage for deeper exploration and advanced applications in later chapters. The concepts presented here are largely based on the works of (Goon and Dasgupta 1983) and (Gupta and Kapoor 1997)

1.1 The word "statistics"

The term `statistics` originates from the Neo-Latin word `statisticum collegium`, meaning "council of state," and the Italian word `statista`, meaning "statesman" or "politician." The German term `Statistik` emerged in the early 18th century and initially referred to the "collection and classification of data," particularly data used by governments and administrative bodies. This usage was introduced by the German scholar Gottfried Achenwall in 1749, who is often credited as the founder of modern statistics.

In 1791, Sir John Sinclair introduced the term `Statistik` into English through his publication of the "Statistical Account of Scotland"(Ball 2004), a comprehensive 21-volume work. This marked the beginning of the use of the term statistics in English to describe the systematic

collection and analysis of data. Later, in 1845, Francis G.P. Neison an actuary[1] to the Medical Invalid and General Life Office published Contributions to Vital Statistics, the first book to include the word "statistics" in its title, focusing on actuarial and demographic data. These developments laid the foundation for statistics as a discipline, evolving from statecraft to a broader scientific approach to data analysis and interpretation.

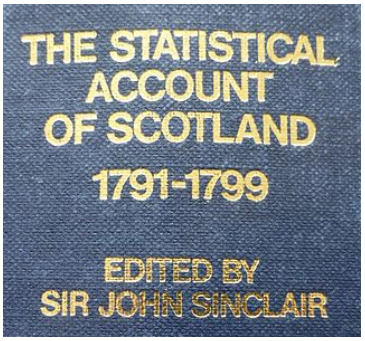


Figure 1.1: Statistical Account of Scotland by Sir John Sinclair (1791)

1.2 Statistics and mathematics

Mathematics and statistics, while closely related, serve distinct purposes and operate on fundamentally different principles. Mathematics can be thought of as a well-organized library, where everything follows strict rules and logical paths. Once a theorem is proven in mathematics, it remains universally true, leaving little room for ambiguity or change. It is a deductive science, relying on precise axioms and logical reasoning to arrive at exact and unchanging results.

Statistics, however, operates in a different realm. It deals with real-world data, which is often messy, unpredictable, and influenced by numerous uncontrolled factors. Statistics is more like an open field, where methods and approaches must adapt to the variability of data. Unlike the certainty of mathematics, statistics uses inductive reasoning to analyze data, account for randomness, and make decisions or predictions under uncertainty. This flexibility is essential because real-world phenomena, especially in fields like biology, are rarely as neat and predictable as mathematical constructs.

In biological sciences, we study complex systems such as plants, animals, and ecosystems, where exact outcomes are rarely achievable. These systems are influenced by a multitude of factors, many of which cannot be precisely measured or controlled. This is where the concept of the error term becomes important. The error term represents the difference between observed and predicted values in a statistical model, accounting for the inherent variability and uncertainty in biological phenomena.

[1]actuary: A person who compiles and analyses statistics and uses them to calculate insurance risks and premiums.

Statisticians embrace this uncertainty, developing mathematical models that approximate reality as closely as possible. Unlike mathematicians, whose focus is on achieving perfect precision, statisticians aim to draw meaningful insights from imperfect and variable data. In the study of biological systems, the goal is not to eliminate uncertainty but to understand patterns, relationships, and trends within the data.

Thus, while mathematics seeks absolute certainty, statistics accepts variability and uncertainty as fundamental characteristics of the real world. By acknowledging and incorporating these uncertainties, statisticians provide valuable tools to study and explain complex biological phenomena, making statistics an indispensable discipline for understanding the complexities of nature.

1.3 Definition of statistics

Statistics is the science which deals with the

- Collection of data
- Organization of data or classification of data
- Presentation of data
- Analysis of data
- Interpretation of data

> Just for Fun
>
> Let's give a definition to statistics using the words themselves:
> **S**trengthening **T**echnological **A**dvancement **T**hrough **I**mplementing **S**ystematic **T**echniques in **C**ontemporary **S**ciences

Two main branches of statistics are:

Descriptive statistics, which deals with summarizing data from a sample using indexes such as the mean or standard deviation etc.

Inferential statistics, use a random sample of data taken from a population to describe and make inferences about the population parameters.

1.4 Data

Data can be defined as individual pieces of factual information that are recorded and used to draw meaningful insights through the science of **statistics**. Think of data as the building blocks that form the foundation for understanding the world around us. It's the raw material from which we extract patterns, trends, and conclusions that help us make better decisions.

In today's fast-paced world, data is more important than ever. From predicting weather patterns to optimizing business strategies, data is at the heart of nearly every advancement. Without data, we're left with guesswork—making it impossible to understand complex systems or make informed decisions.

Here are some examples of data in action:

- **Number of farmers in a village**: Understanding this helps policymakers make decisions about agricultural development and rural economics.
- **Rainfall over a period of time**: This data is crucial for predicting crop yields, planning irrigation, and managing water resources.
- **Area under paddy crop in a state**: This informs agricultural policies, resource allocation, and even global food supply chains.

As you can see, data isn't just a collection of numbers; it's the key to solving real-world problems and shaping the future. In the hands of skilled statisticians, data has the power to unlock insights that can improve lives, drive innovation, and guide decisions at every level.

1.5 Scope and limits

Functions of statistics: Statistics plays a crucial role in simplifying complex data, transforming it into clear and meaningful information. It supports decision-making by presenting facts in an organized manner, aids in the formulation of effective policies, facilitates comparisons, and assists in making forecasts. By applying appropriate statistical methods, researchers can draw valid conclusions from experiments.

Applications of statistics: Statistics has become an integral part of almost every field of human activity. It is indispensable in areas such as administration, business, economics, research, banking, insurance, and more. Its ability to quantify and analyze data makes it an essential tool across industries.

Common limitations of statistics: Statistical methods are applicable only when there is variability in the data being studied. Statistics focuses on the analysis of groups or aggregates, rather than individual data points. The results derived from statistical analysis are often approximate and subject to uncertainty. Statistics is sometimes misapplied or misinterpreted, leading to erroneous conclusions.

As statisticians, we believe that the power of statistics knows no bounds. It's a tool that, when applied correctly, can unlock insights from any dataset. While the limitations listed above are commonly found in textbooks and curricula across SAUs (State Agricultural Universities), I believe these are more about guiding students on the appropriate use of statistics rather than presenting true constraints. With the right methodology and approach, statistics can be applied in any situation to derive valuable insights and support sound decision-making.

1.6 Population and sample

Consider the following example. Suppose we wish to study the height of all students in a college. It will take us a long time to measure the height of all students of the college, so we may select 20 of the students and measure their height (in cm). Suppose we obtain the measurements like this :

149, 156, 148, 161, 159, 143, 158, 152, 164, 171, 157, 152, 163,
158, 151, 147, 157, 146, 153, 159.

In this study, we are interested in the height of all students in the college. The set of height of all students in the college is called the **population** of this study. The set of 20 height, H = {149, 156, 148, ..., 153, 159}, is a **sample** from this population.

Population
In statistics, a *population* refers to the entire collection of elements, individuals, or objects that possess a particular characteristic and are the subject of a statistical study. It encompasses all possible observations or measurements that could be included in the analysis. For example, a population could be all the students in a university, all the trees in a forest, or all the farms in a region. The population provides the complete set of data from which conclusions can be drawn.

Sample
A *sample* is a subset of a population selected for the purpose of conducting a statistical analysis. It represents a smaller group drawn from the population, ideally chosen to reflect its characteristics. Samples are used to estimate population parameters when it is impractical or impossible to collect data from the entire population. The key to a good sample is that it should be representative of the population to allow valid inferences to be made.

Population parameter
A *population parameter* is a numerical characteristic or value that describes an aspect of an entire population. It is a fixed (constant), often unknown value that represents the true measurement of a specific attribute for every member of the population. Common population parameters include the population mean, population variance, and population proportion. Since it is usually impractical or impossible to measure the entire population, parameters are often estimated using sample data.

1.7 Variables and constants

Variables
A *variable* is a characteristic or attribute that can take different values for different individuals, at different times, or in different locations. In other words, variables are subject to change. Examples of variables include:

- The number of fruits on a branch, the number of plots in a field, or the number of schools in a country.
- Plant height, crop yield, panicle length, or temperature.

Variables can be classified into two broad categories: **quantitative** variables, which are measured on a numerical scale (such as height or yield), and **qualitative** (or categorical) variables, which describe categories or characteristics (such as plant species or color).

Constants
A *constant* refers to a value that does not change under any circumstances. Unlike variables, constants retain the same value throughout the study. Examples of constants include:

- Mathematical values such as pi (π), which is the ratio of the circumference of a circle to its diameter ($\pi = 3.14159\ldots$), and e, the base of the natural logarithms ($e = 2.71828$).

1.8 Types of variables

Quantitative variables
A *quantitative variable* is one that can be expressed in numerical terms and takes values that are measurable. Examples of quantitative variables include the number of fruits on a branch, the number of plots in a field, the number of schools in a country, plant height, crop yield, panicle length, and temperature. Quantitative variables can be further classified into two categories: **discrete** and **continuous**.

Discrete variables
Discrete variables are variables that can only take a finite or countable number of distinct values. They are often whole numbers and can be counted. For instance, the number of fruits on a branch, the number of plots in a field, or the number of schools in a country are all discrete variables.

Since discrete variables represent countable quantities, they can only take specific, separate values, such as 0, 1, 2, etc. For example, the number of daily hospital admissions is a discrete variable because it can only take whole number values like 0, 1, or 2, but not fractional values like 1.8 or 3.96.

Continuous variables
Continuous variables, on the other hand, are variables that can take any value within a given

range or interval and can be measured. These variables do not have distinct gaps or interruptions in their possible values. For example, plant height, yield, temperature, and panicle length are continuous variables because they can be measured to a high degree of precision, such as 5.5 cm, 5.8 cm, or any value within a relevant range. Continuous variables can assume an infinite number of possible values within a given range, making them different from discrete variables.

Categorical variables
A *categorical variable* is a type of variable where the data is divided into distinct categories that do not have a numerical value. For example, marital status (single, married, widowed), employment status (employed, unemployed), or religious affiliation (Protestant, Catholic, Jewish, Muslim, others) are examples of categorical variables. These variables are often referred to as *qualitative variables*, as they describe qualities or characteristics rather than measurable quantities.

Unlike quantitative variables, categorical variables cannot be measured or counted in the traditional sense. Instead, they classify data into specific groups or categories.

1.9 Measurement scales

Variables can be classified into four distinct levels of measurement scales each representing a different way of organizing and interpreting data. These four levels are **nominal**, **ordinal**, **interval**, and **ratio**.

Nominal scale
The **nominal scale** is the most basic level of measurement and applies to categorical (qualitative) variables. Data measured on the nominal scale consist of categories that are distinct but have no inherent order or ranking. The categories are simply used for labeling or naming objects or groups. For example, gender (male, female), blood group (A, B, AB, O), and marital status (single, married, divorced) are all nominal variables. In the nominal scale, arithmetic operations, such as addition or subtraction, cannot be performed on the data.

Ordinal scale
The **ordinal scale** also applies to qualitative data, but with an important distinction: the data on the ordinal scale are ordered. This means that the categories have a specific rank or order, but the differences between the categories are not necessarily uniform or meaningful. For example, the grades given in a class (excellent, good, fair, poor) are ordinal, where "excellent" is ranked higher than "good," and "good" is ranked higher than "fair," and so on. However, the difference between "excellent," and "good" is not numerically defined, making the exact magnitude of the difference unclear. Ordinal data allows us to say that one value is greater or lesser than another, but it does not allow for the measurement of exact differences.

Interval scale
The **interval scale** is used for quantitative (numerical) data, and it provides more information

than the nominal or ordinal scales. On the interval scale, the data points are ordered, and the differences between them are meaningful and measurable. However, the interval scale does not have a true zero point. This means that while we can measure the difference between values, we cannot make statements about ratios between them. An example of an interval scale is temperature measured in Celsius or Fahrenheit. For instance, if the temperature in two cities is 20°C and 30°C, we can say that the temperature in the second city is 10°C higher. However, we cannot say that the second city is "twice as hot" as the first city, because the zero point (0°C) does not represent the absence of temperature.

Ratio scale

The **ratio scale** is the highest level of measurement and applies to quantitative data. It shares the properties of the interval scale—ordered data with measurable differences between values—but it also has a meaningful zero point. This true zero point represents the total absence of the quantity being measured. With the ratio scale, not only can we measure differences between values, but we can also compute meaningful ratios. For example, weight is measured on the ratio scale. A weight of 60 kg is twice as much as a weight of 30 kg, and a weight of 0 kg indicates the complete absence of weight. Similarly, temperature measured on the Kelvin scale is an example of a ratio scale, where 0 Kelvin represents absolute zero, the complete absence of heat.

In summary, the key distinctions between these measurement scales are:

- **Nominal**: Categories without any order.
- **Ordinal**: Ordered categories without consistent differences.
- **Interval**: Ordered data with meaningful differences, but no true zero.
- **Ratio**: Ordered data with meaningful differences and a true zero point, allowing for meaningful ratios.

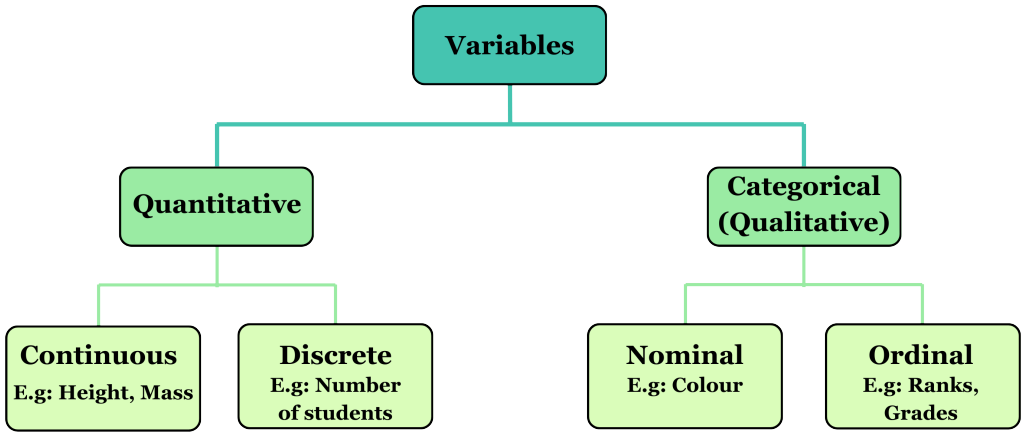


Figure 1.2: Classification of variables

1.10 Collection of data

The process of collecting data is the foundational step in any statistical investigation or research study. Data can be gathered for an entire population or for a sample drawn from it. Typically, data collection is performed on a sample basis, especially when studying large populations. Collecting data is a challenging task, requiring skill and precision. The person responsible for gathering the data, known as the **enumerator** or **investigator**, must be well-trained to ensure the accuracy and reliability of the data collected. The individuals or groups providing the information are referred to as the **respondents**.

1.10.1 Types of data

Data collection can be categorized into two main types based on the source from which the data is derived:

1. **Primary Data**
2. **Secondary Data**

Primary data
Primary data refer to first-hand, original data that are collected directly by the researcher or an organization for a specific purpose. These data have not been processed or analyzed previously and are considered the most authentic form of data. Primary data are typically gathered through surveys, interviews, experiments, or observations, and they represent a direct reflection of the phenomena being studied.

Example: Population census data collected by the government are considered primary data. These are collected directly from individuals by government authorities for the purpose of census enumeration and demographic analysis.

Secondary data
Secondary data refer to data that have already been collected, processed, and published by other organizations or researchers for a different purpose. These data may have undergone some degree of analysis or treatment before being made available for new studies. Secondary data are often more convenient to use, as they are readily accessible, but they may not always perfectly suit the specific needs of the researcher.

Example: An economic survey of a country, such as reports from the Bureau of Statistics or other governmental agencies, is an example of secondary data. These data were originally collected for purposes such as policy analysis or economic planning, and now can be used for additional research.

The distinction between primary and secondary data lies primarily in their origin and the process of collection. Primary data are first-hand, original data collected directly from a single source by the researcher for a specific purpose. These data are considered pure as they have

not undergone any prior statistical treatment. In contrast, secondary data are obtained from existing sources or agencies and have been previously collected and processed for different purposes. They are not considered pure as they have undergone some form of statistical treatment. While primary data are original and collected for the first time, secondary data are pre-existing and gathered from other sources. Both types of data have their respective advantages and limitations, and the choice between them depends on the research objectives, availability of resources, and the nature of the study.

1.11 Collecting primary data

Primary data can be collected using various methods, depending on the research requirements and resources available:

Personal investigation
In this method, the researcher directly conducts the survey and collects the data themselves. This approach often results in highly accurate and reliable data. It is best suited for small-scale research projects where direct involvement of the researcher is feasible.

Through investigation
In this method, trained investigators are employed to collect data. These investigators engage with individuals, asking questions and filling out questionnaires based on the responses. This method is widely used by organizations for larger data collection efforts.

Collection through questionnaire
Researchers distribute questionnaires to local representatives or agents who collect data based on their own experience and observations. While this method is relatively quick, it typically provides only a rough estimate of the information.

Through the phone
Data is gathered by contacting individuals via telephone/mobile phone. This method is fast and allows for accurate information to be collected efficiently, making it suitable for studies that require a broad reach but still need reliable data.

1.12 Collecting secondary data

Secondary data are collected through various established channels:

Official
Official sources include publications from government bodies such as the Statistical Division, Ministry of Finance, Federal Bureaus of Statistics, and various ministries (e.g., Agriculture, Food, Industry, Labor). These sources provide comprehensive and authoritative data.

Semi-Official
Semi-official sources include publications from institutions like the State Bank, Railway Board, Central Cotton Committee, and Boards of Economic Enquiry. It also encompasses reports from trade associations, chambers of commerce, technical journals, trade publications, and research organizations such as universities and other academic institutions. These sources provide valuable data, though they may not be as universally authoritative as official sources.

1.13 Frequency distribution

The data presented below shows the number of fruits per branch in a mango tree selected from a particular plot. The data, presented in this form in which it was collected, is called *raw data*.

0, 1, 0, 5, 2, 3, 2, 3, 1, 5,
5, 2, 3, 4, 4, 5, 4, 0, 5, 4,
2, 4, 4, 4, 1

It can be seen that, the minimum and the maximum numbers of fruits per branch are 0 and 5, respectively. Apart from these numbers, it is impossible, without further careful study, to extract any exact information from the data. But by breaking down the data into the form below

Number of fruits per branch	Tally	Frequency
0	|||	**3**
1	|||	**3**
2	||||	**4**
3	|||	**3**
4	𝍸 ||	**7**
5	𝍸	**5**
		Total = 25

Figure 1.3: Frequency distribution table

Now certain features of the data become apparent. For instance, it can easily be seen that, most of the branches selected have four fruits because number of branches having 4 fruits is 7. This information cannot easily be obtained from the raw data. The above table is called a **frequency table** or a **frequency distribution**. It is so called because it gives the frequency or number of times each observation occurs. Thus, by finding the frequency of each observation, a more intelligible picture is obtained.

1.13.1 Construction

In this section, we will discuss the process of constructing a frequency distribution. Follow the steps below. This method helps to clearly visualize the frequency of each observation, ensuring that the total frequency adds up to the total number of observations.

1. List all values of the variable in ascending order of magnitude.
2. Form a tally column, that is, for each value in the data, record a stroke in the tally column next to that value. In the tally, each fifth stroke is made across the first four. This makes it easy to count the entries and enter the frequency of each observation.
3. Check that the frequencies sum to the total number of observations.

1.14 Grouped frequency distribution

Data below gives the plant height of 20 paddy varieties, measured to the nearest centimeters.

109, 107, 129, 122, 118, 110, 124, 146, 138, 121,
115, 132, 131, 139, 142, 134, 143, 144, 127, 116

It can be seen that the minimum and the maximum plant height are 107 cm and 144 cm, respectively. A frequency distribution giving every plant height between 107 cm and 144 cm would be very long and would not be very informative. The problem is to overcome by grouping the data into classes.
If we choose the classes
100 – 109
110 – 119
120 – 129
130 – 139
140 – 149
we obtain the frequency distribution given below:

Mass (kg)	Tally	Frequency
101 – 109	||	2
110 – 119	||||	4
120 – 129	𝍸	5
130 – 139	𝍸	5
140 - 149	||	4
		Total = 20

Figure 1.4: Grouped Frequency distribution table

Above table gives the frequency of each group or class; it is therefore called a grouped frequency table or a grouped frequency distribution. Using this grouped frequency distribution, it is easier to obtain information about the data than using the raw data. For instance, it can be seen that 14 of the 20 paddy varieties have plant height between 110 cm and 139 cm (both inclusive). This information cannot easily be obtained from the raw data.
It should be noted that, even though above table is concise, some information is lost. For example, the grouped frequency distribution does not give us the exact plant height of the paddy varieties. Thus the individual plant height of the paddy varieties are lost in our effort to obtain an overall picture.

1.14.1 Terminologies

Class limits
The intervals into which the observations are put are called class intervals. The end points of the class intervals are called class limits. For example, the class interval 100 – 109, has lower class limit 100 and upper class limit 109.

Continuous classes
Continuous classes are intervals where the class limits represent a continuous range of values, with no gaps between the intervals.

Example: If the class intervals are 10 − 20, 20−30, 30−40, and so on, there are no gaps between them, and all values within these ranges are included seamlessly.

Discontinuous classes
Discontinuous classes are intervals where gaps exist between the class limits. In such cases, class boundaries are used to close the gaps and ensure continuity.

Example:
If the class intervals are 10 - 19, 20 - 29, 30 - 39, and so on, there is a gap between the end of one interval and the start of the next. The actual range of each interval is defined using class boundaries, which is explained below.

Class boundaries
The raw data in the above example were recorded to the nearest centimeters. Thus, a plant height of 109.5cm would have been recorded as 110cm, a plant height of 119.4 cm would have been recorded as 119cm, while a plant height of 119.5 cm would have been recorded as 120 cm. It can therefore be seen that, the class interval 110 – 119, consists of measurements greater than or equal to 109.5 cm and less than 119.5 cm. The numbers 109.5 and 119.5 are called the lower and upper boundaries of the class interval 110 – 120. The class boundaries of the other class intervals are given below:

Class interval	Class boundaries	Class mark	Frequency
101 – 109	100.5 – 109.5	105	2
110 – 119	109.5 – 119.5	114.5	4
120 – 129	119.5– 129.5	124.5	5
130 – 139	129.5– 139.5	134.5	5
140 - 149	139.5– 149.5	144.5	4

Figure 1.5: Class boundary and class limits

Note:
Notice that the lower class boundary of the i^{th} class interval is the mean of the lower class limit of the class interval and the upper class limit of the $(i-1)^{th}$ class interval (i = 2, 3, 4, …). For example, in the table above the lower class boundaries of the second and the fourth class intervals are (110 + 109) /2 = 109.5 and (130 + 129)/2 = 129.5 respectively.
It can also be seen that the upper class boundary of the i^{th} class interval is the mean of the upper class limit of the class interval and the lower class limit of the $(i+1)^{th}$ class interval (i = 1, 2, 3, …). Thus, in the above table the upper class boundary of the fourth class interval is (139 + 140)/2 = 139.5.

> **!** Note
>
> For continuous classes, class limits and boundaries are the same because there are no gaps between intervals. However, for discontinuous classes, boundaries are important as they close gaps and ensure every value belongs to one class.

Class mark
The mid-point of a class interval is called the class mark or class mid-point of the class interval. It is the average of the upper and lower class limits of the class interval. It is also the average of the upper and lower class boundaries of the class interval. For example, in the table, the class mark of the third class interval was found as follows: class mark = (120+129)/2 = (119.5 + 129.5)/2 = 124.5.

Class width
For continuous classes:
The class width is the difference between the upper and lower class limits of a class interval. Since the class limits and boundaries are the same for continuous classes, the width can also be determined by subtracting two consecutive lower or upper class limits.

For discontinuous classes:
The class width is the difference between the upper and lower class boundaries of a class interval. For discontinuous classes, class boundaries are used to account for gaps, and the width can also be determined by subtracting two consecutive lower or upper class boundaries.

Note:

In the grouped frequency table above with discontinuous classes, the width of the second class interval is calculated as |110 - 119| = 9. It can be observed that the width is the same for all classes. This result can also be obtained by taking the numerical difference between the lower class boundaries of the second and third class intervals.

1.14.2 Construction

Step 1. Decide how many classes you wish to use
Step 2. Determine the class width
Step 3. Set up the individual class limits
Step 4. Tally the items into the classes
Step 5. Count the number of items in each class

Consider the example where an agricultural student measured the lengths of leaves on an oak tree (to the nearest cm). Measurements on 38 leaves are as follows
9, 16, 13, 7, 8, 4, 18, 10, 17, 18,
9, 12, 5, 9, 9, 16, 1, 8, 17, 1, 10, 5, 9, 11, 15, 6, 14, 9, 1, 12,
5, 16, 4, 16, 8, 15, 14, 17

Step 1. Decide how many classes you wish to use

H.A. Sturges provides a formula for determining the approximation number of classes.

$$\mathbf{k = 1 + 3.322}.\log\mathbf{N}$$

Number of classes should be greater than calculated k.
In our example N=38, so k= (1+3.322)×log(38) = (1+3.322)×1.5797 = 6.24 = approx 7

So the approximated number of classes should be not less than 6.24 *i.e.* $k^{'}$ =7

Step 2. Determine the class width

Generally, the class width should be the same size for all classes. C= | max − min|/ k. Class width $C^{'}$should be greater than calculated C. For this example, C = | 18− 1|/**6.24** = 2.72, so approximately class width $C^{'}$ = 3 (Note that k used here is the calculated value using Sturges formula not the approximated).

Step 3. To set up the individual class limits, we need to find the lower limit only

$$L = min - \frac{C^{'} \times k^{'} - (max - min)}{2}$$

where, C and k here are final approximated class width and number of classes respectively in our example, $L = 1 - \frac{(3\times7)-(18-1)}{2}$ = 1-2 = -1; since there is no negative values in data = 0. Final frequency table will be as shown in Table 1.1

Table 1.1: Frequency distribution table

Class	Frequency
0-3	3
3-6	5
6-9	5
9-12	9
12-15	5
15-18	9
18-21	2

Even though the student only measured in whole numbers, the data is continuous, so "4 cm" means the actual value could have been anywhere from 3.5 cm to 4.5 cm.

1.15 Cumulative frequency

In many situations, we are not interested in the number of observations in a given class interval, but in the number of observations which are less than (or greater than) a specified value. For example, in the above table, it can be seen that 3 leaves have length less than 3.5 cm and 9 leaves (*i.e.* 3 + 6) have length less than 6.5 cm. These frequencies are called cumulative frequencies. A table of such cumulative frequencies is called a **cumulative frequency table** or **cumulative frequency distribution**.

Cumulative frequency is defined as a running total of frequencies. Cumulative frequency can also defined as the sum of all previous frequencies up to the current point. Notice that the last cumulative frequency is equal to the sum of all the frequencies. Two types of cumulative frequencies are **Less than Cumulative Frequency (LCF)** and **Greater than Cumulative Frequency(GCF)**. LCF is the number of values less than a specified value. GCF is the number of observations greater than a specified value.

The specified value for LCF in the case of grouped frequency distribution will be upper limits and for GCF will be the lower limits of the classes. LCF's are obtained by adding frequencies in the successive classes and GCF are obtained by subtracting the successive class frequencies from the total frequency. See calculated LCF and GCF in Table 1.2 below.

1.16 Relative frequency

It is sometimes useful to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the frequency of the

Table 1.2: LCF,GCF and Relative frequency

Class	Frequency	A	B	C
0.5 - 3.5	3	3	38	0.079
3.5 - 6.5	6	9	35	0.158
6.5 - 9.5	10	19	29	0.263
9.5 - 12.5	5	24	19	0.132
12.5 - 15.5	5	29	14	0.132
15.5 - 18.5	9	38	9	0.237

particular class interval by the total number of observations. **Relative frequency** of a class is the frequency of class divided by total observations. Relative frequencies all add up to 1. See relative frequency calculated in Table 1.2 .

Historical Insights

"The power of statistics: A wartime story"
During World War II, military engineers faced a daunting challenge: how to protect their planes and crews better during missions. When the planes returned from combat, they were covered in bullet holes. Naturally, the engineers assumed they should add armor to the most damaged areas to make the planes more resilient. It seemed like common sense—but then came a surprising twist. Abraham Wald, a brilliant statistician, was called in to analyze the problem. He noticed something others had missed. The planes being studied were the ones that survived the missions. The missing data—the planes that never made it back, held the key to solving the puzzle. Wald reasoned that the areas with no damage on the returning planes were likely the most critical: hits in these areas would have brought the planes down, leaving no data behind.
Wald's insight transformed how decisions were made, leading to smarter strategies that saved countless lives. This story is a powerful reminder of the importance of statistics not just in analyzing what's visible but in understanding what's missing. It's a perfect example of how data, combined with critical thinking, can uncover hidden truths and drive impactful decisions. This revolutionary application of statistics eventually led to the birth of operations research, a field that continues to solve complex problems in diverse areas today.

Quotes to Inspire

"Data is the sword of the 21st century, those who wield it well, the Samurai."
- Jonathan Rosenberg, former Google SVP

2 Statistics on agriculture

The agricultural sector accounts for approximately 18% of India's GDP and employs nearly half of its workforce. Reliable and timely information is vital for planners and policymakers to develop effective agricultural policies and make informed decisions on procurement, storage, public distribution, imports, exports, and other related matters. As such, the collection and management of agricultural statistics hold significant importance.

This chapter provides an overview of the system for collecting agricultural statistics in India. While agriculture is a state subject, agricultural statistics fall under the concurrent list, resulting in a decentralised system. State Governments, through their State Agricultural Statistics Authorities (SASAs), play a central role in collecting and compiling agricultural statistics at the state level. At the national level, the Directorate of Economics and Statistics, under the Ministry of Agriculture and Farmers Welfare, is responsible for compiling the data. Other key agencies involved include the National Statistical Office (NSO) and the State Directorates of Economics and Statistics (DESs).

2.1 Compiling crop statistics

Crop statistics comprise two key components: the area sown and the average yield.
While area estimates are derived from land revenue systems, yield estimates are obtained through crop estimation surveys.

2.1.1 Area statistics

The system for collecting area statistics across states and Union Territories (UTs) in India can be broadly classified into three categories:

1. **States with complete enumeration systems**
   These include states with land records or temporary settlement systems, covering 86% of the states (18 states and 3 UTs).
2. **States using sample surveys**
   These are states with no land record system or permanently settled states, representing 9% of the states.

3. **States with no developed system for area statistics**
   In these states, the collection is still based on conventional methods such as personal assessment, accounting for 5% of the states.

2.1.2 Crop yield estimation

Crop yields are estimated through **Crop Cutting Experiments (CCE)**, which are conducted extensively across the country. The General Crop Estimation Survey (GCES) covers 65 crops, including 51 food crops and 14 non-food crops. Approximately 9 lakh CCEs are carried out annually in India to estimate the yield of key crops such as rice, maize, bajra, groundnut, and sugarcane. These experiments are conducted systematically to ensure accurate and reliable yield data for principal crops.

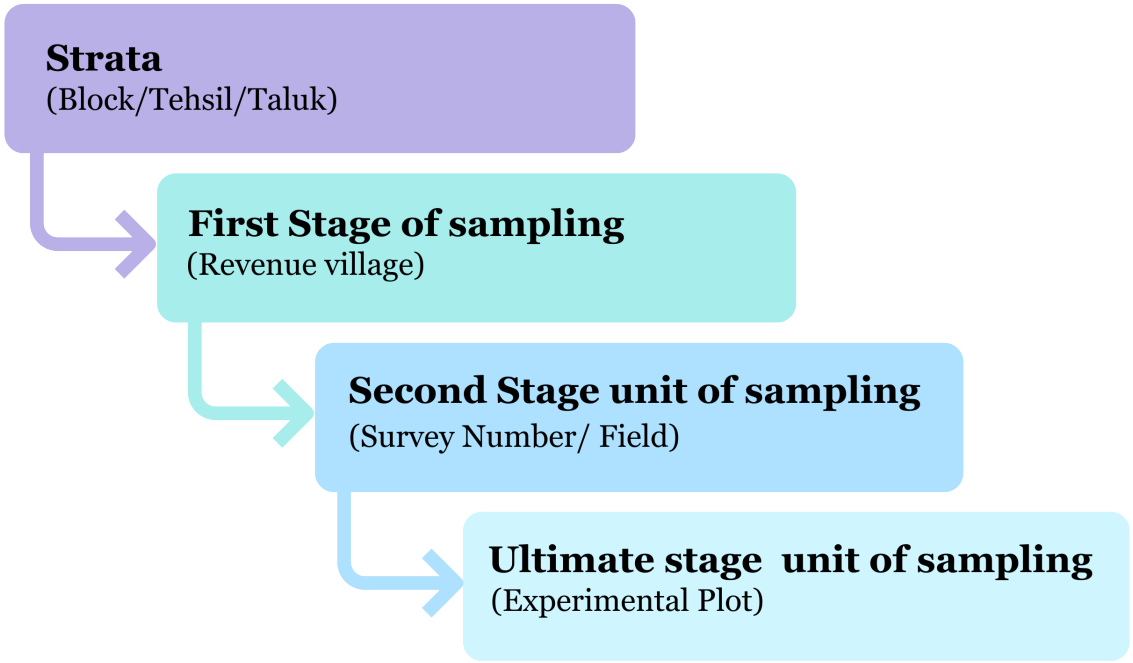


Figure 2.1: Sampling design under General Crop Estimation Survey

Final estimates of crop production are calculated using area figures obtained through complete enumeration and yield rates derived from crop-cutting experiments. These estimates become available only after the harvest. However, to support timely decision-making, the Government requires advance production estimates.

The Directorate of Economics and Statistics (DES), under the Ministry of Agriculture and Farmers Welfare, provides advance estimates of crop area and production for key food and non-food crops such as food grains, oilseeds, sugarcane, and fibres. These estimates are issued in four stages:

1. **First forecast**: Mid-September
2. **Second forecast**: January
3. **Third forecast**: Late March
4. **Fourth forecast**: Late May

In addition to these forecasts, **Final Estimates** of crop area and production are published in December. Subsequently, **Fully Revised Estimates** for all-India crop statistics are released in December of the following crop year.

Additionally, information on the structure and characteristics of the agricultural sector is gathered through the Agricultural Census.

2.2 Agricultural census

The Agricultural census is a comprehensive exercise conducted to gather and analyze data on the structure of the agricultural sector in India. It provides essential information about operational holdings, including their number, area, land use, cropping patterns, and input usage, down to the lowest geographical levels such as villages, tehsils (sub-districts), and districts. This census serves as a statistical framework for planning and conducting future agricultural surveys.

Initiated in **1970-71**, the Agricultural census is conducted every five years by the Department of Agriculture and Farmers Welfare in collaboration with State and Union Territory administrations. In states with land records, the number and area of operational holdings are collected through **complete enumeration**, while detailed data on the characteristics of operational holdings are gathered on a **sample basis**.

To date, **eleven Agricultural censuses** have been conducted, covering the reference years **1970-71, 1976-77, 1980-81, 1985-86, 1990-91, 1995-96, 2000-01, 2005-06, 2010-11, 2015-16 and 2020-21**. The reference period for each census corresponds to the agricultural year, spanning from **July to June**.

The data derived from the Agricultural census plays a crucial role in policy formulation, resource allocation, and the overall development of the agricultural sector in India.

Additional data pertaining to various sectors can be obtained from the sources listed in the Appendix 1

Historical Insights

"Mahalanobis and statistical planning in India"
In post-independence India, statistician *Prasanta Chandra Mahalanobis* played a key role in shaping the country's economic planning. He developed statistical methods to guide the allocation of resources in India's *Five-Year Plans*. Mahalanobis introduced the *Mahalanobis Distance*, a method for analyzing multi-dimensional data, and created the *Input-Output Model* to understand how different sectors of the economy interacted. His work, especially in the first Five-Year Plan (1951-1956), helped optimize resource use and industrial growth. Mahalanobis' contributions laid the foundation for using statistics in economic policy-making, which continues to guide India's development. He is known as the 'Father of Indian statistics'.

Quotes to Inspire

"Statistics is the art of never having to say you're certain." – W. Edwards Deming

3 Graphical representation

Graphs and diagrams play a vital role in statistics by transforming complex data into clear, visual formats that are easier to interpret and analyze. While frequency distributions in tabular form help organize raw data, graphical representations provide a more intuitive way to understand patterns, trends, and relationships within the data. By converting numbers into visual elements, graphs make it simpler to convey information effectively, making them indispensable tools in research, analysis, and communication. Depending on the nature of the data and the intended purpose, various types of graphs and diagrams can be employed to illustrate key insights. This chapter focuses on the fundamental graphs and charts used in statistics to visually represent data.

3.1 Histogram

A histogram is a graphical representation used to display the frequency distribution of continuous data. It consists of adjacent rectangles, where:

- The **base** of each rectangle lies along the horizontal axis, with the width determined by the class intervals.
- The **height** of each rectangle is proportional to the frequency of the corresponding class.

Unlike bar charts, histograms have no gaps between the rectangles, emphasizing the continuity of the data. The height of each rectangle represents the frequency for equal-width classes. Histograms are effective tools for visualizing data distribution, identifying patterns, and highlighting skewness or outliers.

> **!** Note
>
> If the class intervals are of equal width, the height of each rectangle in a histogram is directly proportional to the class frequency. In such cases, the class frequencies can be used as the heights of the rectangles.
> However, when class intervals have varying widths, the height of each rectangle should be proportional to the **frequency density**, which is calculated as:

$$\text{Frequency Density} = \frac{\text{Class Frequency}}{\text{Class Width}}$$

In these cases, the frequency density is plotted on the y-axis to ensure that the *area of each rectangle* accurately represents the frequency of the class. This approach maintains the correct visual representation of the data distribution regardless of the class interval widths.

Example 3.1 Table 4.1 displays the frequency distribution of plant heights for a sample of 50 plants. This data can be visualized effectively using a histogram, as shown in Figure 3.1.

Table 3.1: Grouped frequency table of plant heights

Plant height (cm)	Frequency
130 – 140	3
140 – 150	6
150 – 160	17
160 – 170	13
170 – 180	8
180 – 190	3

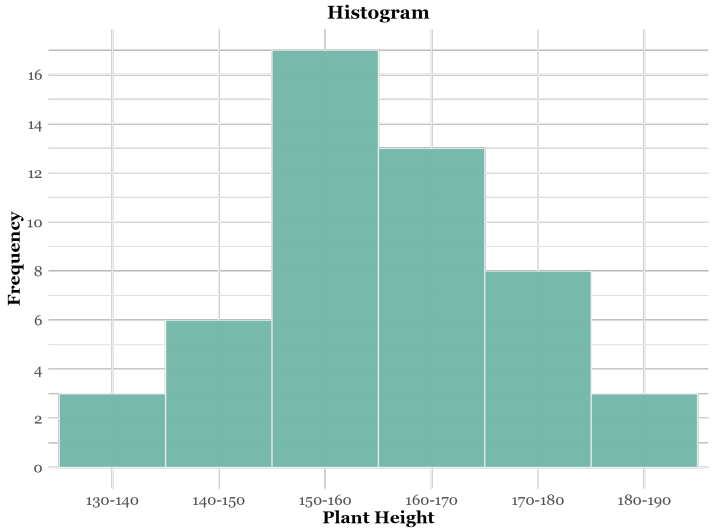


Figure 3.1: Histogram

3.2 Ogive

Ogive, also known as the cumulative frequency curve, is a graphical representation that plots cumulative frequencies against class boundaries. The points are typically connected using straight lines, forming a continuous curve. This visualization effectively illustrates the accumulation of frequencies, making it useful for understanding data distribution and determining percentiles or the median.

3.3 Types of ogives

There are two main types of cumulative frequency curves:
1. **Less than ogive**
2. **Greater than ogive**

Less than ogive

The less than ogive, also known as the **less than type cumulative frequency curve**, is created by plotting the less than cumulative frequencies against the upper class boundaries. For example, consider the plant height data for 50 plants. By using the upper class limits and their cumulative frequencies, we can construct a smooth curve that provides insights into the data distribution. See Table 3.2, which is constructed from Table 4.1. The less than ogive, shown in Figure 3.2, is drawn using Table 3.2.

Table 3.2: Upper limit and LCF of plant heights

Upper limit	140	150	160	170	180	190
LCF	3	9	26	39	47	50

Note: LCF denotes less than cumulative frequency.

Greater than ogive

The **greater than ogive**, also known as the **greater than type cumulative frequency curve**, is constructed by plotting the greater than cumulative frequencies against the lower class boundaries. In this case, instead of using the upper limits like in the "Less than ogive", we use the lower class limits and their corresponding cumulative frequencies. This curve helps visualize the cumulative frequency distribution from the highest class down to the lowest, providing insights into the number of observations greater than a specific value. See Table 3.3 constructed from Table 4.1. The greater than ogive, shown in Figure 3.3, is drawn using Table 3.3.

Figure 3.2: Less than ogive

Table 3.3: Lower limit and GCF of plant heights

Lower limit	130	140	150	160	170	180
GCF	50	47	41	24	11	3

Note: GCF denotes greater than cumulative frequency.

Figure 3.3: Greater than ogive

> **Note**
>
> Intersection of both less than and greater than ogives gives the median.

3.4 Frequency polygon

A grouped frequency table can also be represented by a frequency polygon, a special type of line graph. To construct it, plot the class frequencies against the corresponding class midpoints and connect successive points with straight lines. The frequency polygon can also be derived by joining the midpoints of a histogram. See Table 3.4, constructed from Table 4.1. The frequency polygon, created using Table 3.4, is shown in Figure 3.4. The relation between

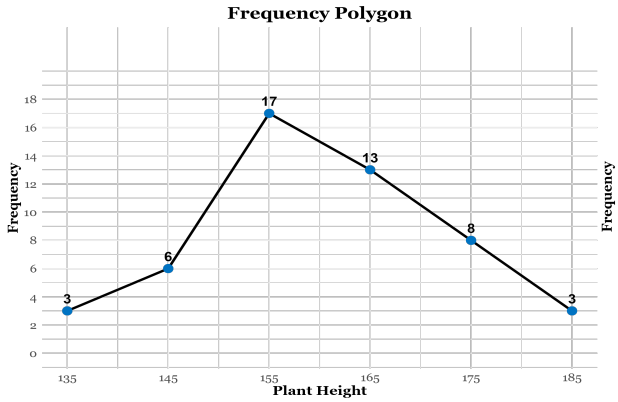


Figure 3.4: Frequency Polygon

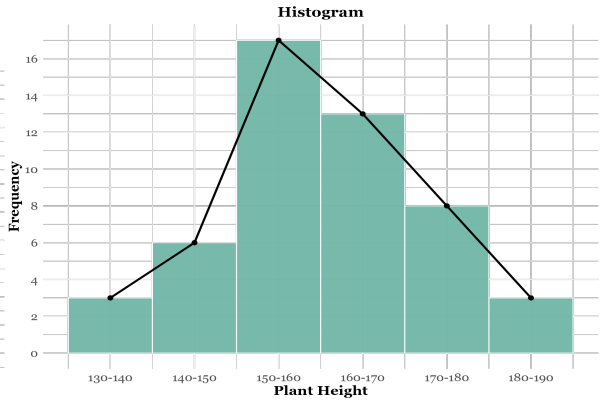


Figure 3.5: Frequency Polygon and Histogram

3.5 Stem and leaf plot

A stem and leaf plot is a graphical device useful for representing a relatively small set of data that takes numerical values. To construct a stem and leaf plot, we partition each measurement into two parts: the **stem** (the leading digits) and the **leaf** (the trailing digits). This method retains the exact value of each observation, unlike a frequency distribution. It also clearly shows the distribution of data within each group. A stem and leaf plot conveys similar information as a histogram, with the added benefit of retaining individual data points. It provides insights into the range, concentration of measurements, and symmetry of the data.

Consider the example:
12, 16, 21, 25, 29, 26, 30, 31, 37, 42, 45.

The stem and leaf plot for this data is shown Figure 3.6

1	2 6
2	1 5 6 9
3	0 1 7
4	2 5

Figure 3.6: Stem and leaf plot

A stem-and-leaf plot is not only useful for small data sets but can also effectively represent

larger sets of numerical data. For instance, consider the monthly income of 50 employees in a company:

19710, 24096, 23618, 26490, 25626, 24653, 24297, 23609, 19120, 25942, 23591, 27302, 29569, 25332, 29396, 20725, 25202, 20763, 30556, 21961, 22910, 21826, 21547, 21015, 19825, 24124, 22275, 26127, 24297, 20564, 26943, 26627, 23602, 24585, 25725, 24322, 23198, 25590, 23366, 23313, 22840, 25514, 24959, 23194, 21337, 26030, 27215, 19260, 27467, 29737.

The corresponding stem-and-leaf plot for this data, shown in Figure 3.7, lists the leaves in increasing order under their respective stems. The proper choice of stems is crucial as it organizes the data effectively, revealing patterns and distribution with clarity.

19	120, 260, 710, 825
20	564, 725, 763
21	015, 337, 547, 826, 961
22	275, 840, 910
23	194, 198, 313, 366, 591, 602, 609, 618
24	096, 124, 297, 297, 322, 585, 653, 959
25	202, 332, 514, 590, 626, 725, 942
26	030, 127, 490, 627, 943
27	215, 302, 467
29	396, 569, 737
30	556

Figure 3.7: Stem and leaf plot of 5 digit data

Consider a different dataset representing the percentage of adults with a college degree in 20 cities.

48.5, 53.2, 42.1, 65.4, 70.3, 38.7, 55.9, 47.3, 59.2, 33.5, 45.6, 62.8, 50.1, 41.3, 36.2, 43.7, 39.8, 66.4, 58.1, 31.2.

The stem and leaf plot for this data is shown in Figure 3.8. Here, the tens digit serves as the stem, and the decimal values form the leaves.

3	1.2, 3.5, 6.2, 8.7, 9.8
4	1.3, 2.1, 3.7, 5.6, 7.3, 8.5
5	0.1, 3.2, 5.9, 8.1
6	2.8, 4.2, 5.4, 6.4
7	0.3

Figure 3.8: Stem and leaf plot of decimal data

3.6 Bar chart

A bar chart or bar graph is a diagram consisting of a series of horizontal or vertical bars of equal width. The bars represent various categories of the data. There are three types of bar charts, and these are simple bar charts, component bar charts and grouped bar charts.

Simple bar chart

In a simple bar chart, the height (or length) of each bar is equal to the value of category in the y-axis it represents. Table 3.5 presents hypothetical data on coconut production across five districts of Kerala for a specific year. The data represented using barchart is shown in Figure 3.9

Table 3.5: hypothetical data on coconut production

District	Production (million nuts)
Alappuzha	700
Kannur	800
Thrissur	980
Ernakulam	1100
Wayanad	1400

Component bar chart

In a component bar chart, the bar for each category is subdivided into component parts; hence its name. Component bar charts are therefore used to show the division of items into components. Component bar chart is also known as *stacked barchart*.

Figure 3.10 shows the distribution of sales of agricultural produce from a farm in 1995, 1996 and 1997 and its corresponding component barchart in Figure 3.11.

The component bar chart shows the changes of each component over the years as well as the comparison of the total sales between different years.

Grouped bar chart

Figure 3.10 can also be represented using a grouped bar chart shown in Figure 3.12. For a grouped bar chart, each category within a group is represented by a bar with a distinct shade or color, allowing for clear comparisons of both within and across groups.

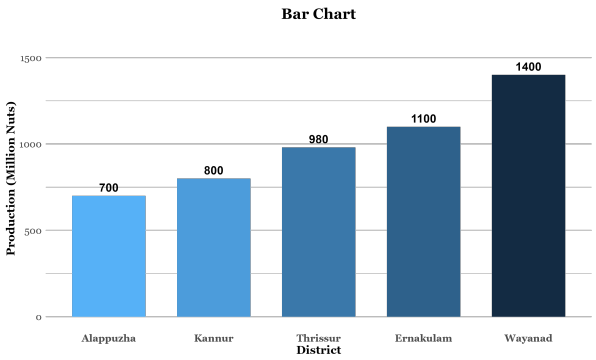


Figure 3.9: Barchart

Agricultural Produce Sales

Sales in Million Dollars (1995-1997)[1]

Agricultural Produce	1995	1996	1997
Coffee	90	120	180
Cocoa	180	140	220
Palm oil	30	30	20

[1] Data represents agricultural sales trends across three years.

Figure 3.10: Sales data of agricultural produce

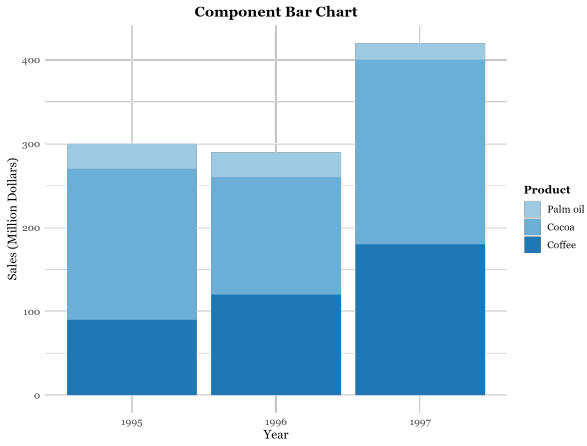


Figure 3.11: Component bar chart

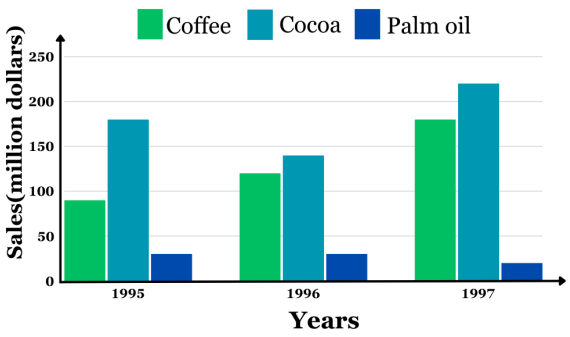


Figure 3.12: Grouped bar chart

3.7 Histogram versus bar chart

Table 3.6 highlights the key differences between histograms and bar charts, two commonly used graphical tools in data visualization. While both employ bars to represent data, they serve

distinct purposes and are applied to different types of data. Understanding these differences ensures the correct choice of graph for effectively presenting and interpreting data.

Table 3.6: Comparison between histogram and bar chart

Feature	Histogram	Bar chart
Meaning	A graphical representation using bars to display the frequency of numerical data.	A pictorial representation using bars to compare different categories of data.
Purpose	Depicts the distribution of continuous (non-discrete) data.	Compares discrete (categorical) data.
Type of data	Quantitative data.	Categorical data.
Bar spacing	Bars are adjacent with no gaps.	Bars are separated by spaces.
Grouping of elements	Data is grouped into ranges or intervals (bins).	Data is represented as individual categories.
Bar order	Bars cannot be reordered.	Bars can be reordered.
Bar width	Bar widths may vary.	Bar widths are uniform.

3.8 Pie chart

A pie chart is a circular graph divided into sectors, each sector representing a different value or category. The angle of each sector of a pie chart is proportional to the value of the part of the data it represents. The bar chart is more precise than the pie chart for visual comparison of categories with similar relative frequencies.

Steps for constructing a pie chart

1. Find the sum of the category values.

2. Calculate the angle of the sector for each category, using the following formula. Angle of the sector for category A $= \frac{\text{value of category A}}{\text{sum of category values}} \times 360$

3. Construct a circle and mark the centre.

4. Use a protractor to divide the circle into sectors, using the angles obtained in step 2.

5. Label each sector clearly.

Table 3.7 presents hypothetical data on the production of different commodities in India during a particular year. Pie chart based on this data is shown in Figure 3.13

Table 3.7: Hypothetical data on the production of different commodities

Commodities	Production(tonnes)	Angle
Wheat	27000	(27000/81000)×360= 120
Grams	22500	100
Maize	13500	60
Rice	6750	30
Sugar	11250	50
Total	**81000**	**360**

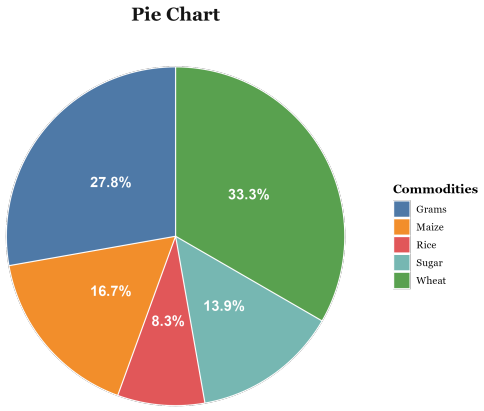


Figure 3.13: Pie chart

3.9 Boxplot

A boxplot, also known as a **box-and-whisker plot**, visually represents the five-number summary of a dataset: the minimum value, first quartile, median, third quartile, and maximum value. These key statistics provide insights into the dataset's central tendency, spread, and potential outliers. Quartiles and the median, explained in detail in Section 5.5, are critical components of this summary.

In a boxplot, a rectangular box spans from the first quartile (Q1) to the third quartile (Q3), with a vertical line inside the box indicating the median. Whiskers extend from each end of the box to the dataset's minimum and maximum values, providing a clear picture of the range and variability.

Figure 3.14 below shows the parts of a box plot.

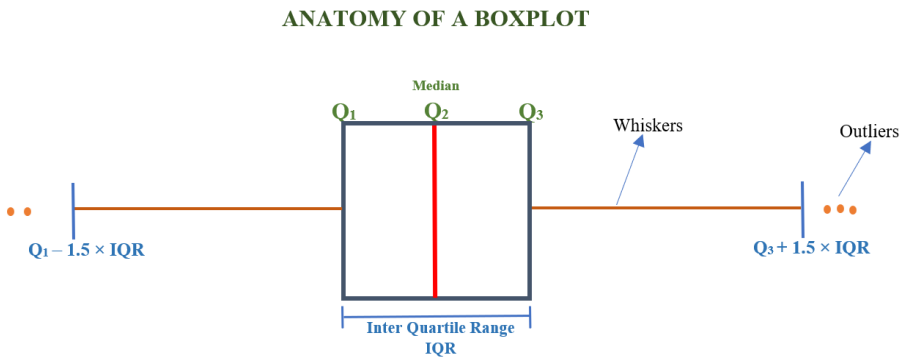


Figure 3.14: Anatomy of box plot

In a boxplot, the minimum value is defined as $Q_1 - 1.5 \times IQR$, and the maximum value is $Q_3 + 1.5 \times IQR$, where, Q_1 and Q_3 represent the first and third quartiles, and IQR stands for the interquartile range. Any data points falling below the minimum or above the maximum are considered outliers.

3.10 Advanced visualization

While this book focuses on basic plots and charts, significant advancements have been made in the field of data visualization. New types of graphs and charts have been developed to help in more effective representation and communication of data. Although a detailed discussion of these advanced graphs is beyond the scope of this book, we provide an overview of some common and recently developed types for reference. For more detailed information, you can explore resources such as The R Graph Gallery.

It is important to be aware of the wide variety of visualization tools available, as they can enhance your understanding of data and improve your ability to communicate insights clearly. From Figure 3.15 to Figure 3.26 you can see a few popular and advanced graph types widely used in modern data analysis.

Historical Insights

The cholera map: A life-saving visualization
In 1854, during a devastating cholera outbreak in London, physician John Snow transformed public health through the power of visualization. By mapping the locations of cholera cases and water pumps, he revealed a striking correlation: cases clustered around a single contaminated pump on Broad Street. Snow's map not only pinpointed the outbreak's source but also challenged prevailing beliefs that diseases spread through "miasma" (bad air). His groundbreaking work demonstrated how data visualization could

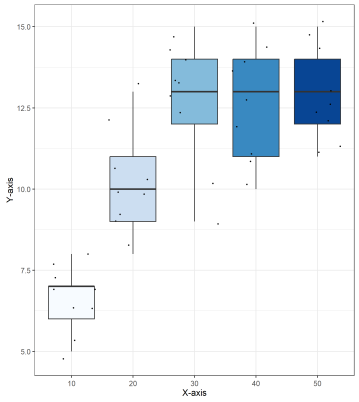


Figure 3.15: Box Plot

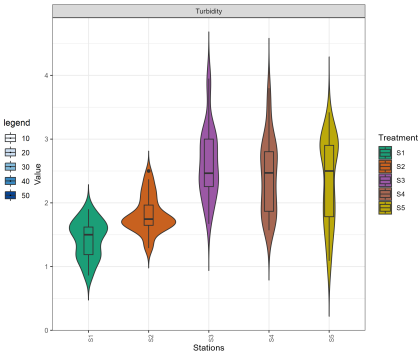


Figure 3.16: Violin Plot

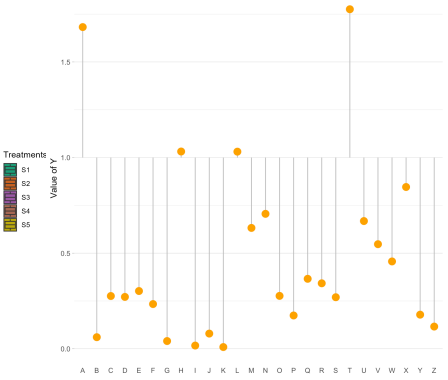


Figure 3.17: Lollipop Plot

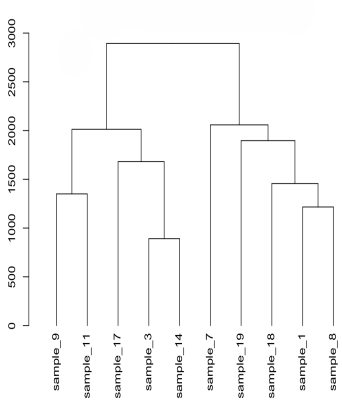


Figure 3.18: Dendrogram

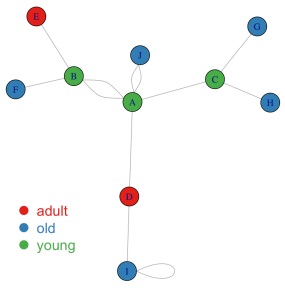


Figure 3.19: Network Graph

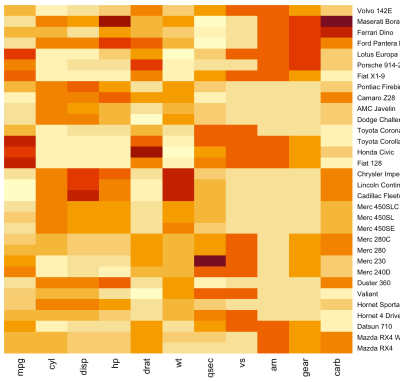


Figure 3.20: Heat Map

Figure 3.21: Circular Bar Plot

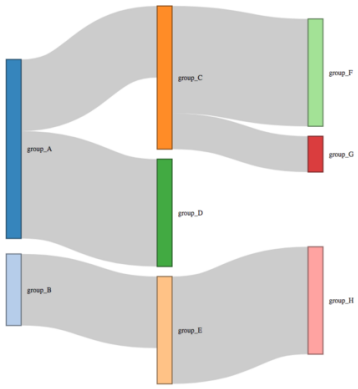


Figure 3.22: Sankey Diagram

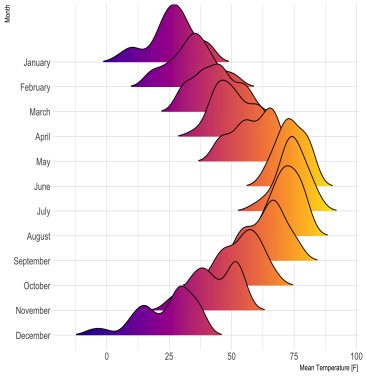


Figure 3.23: Ridgeline Plot

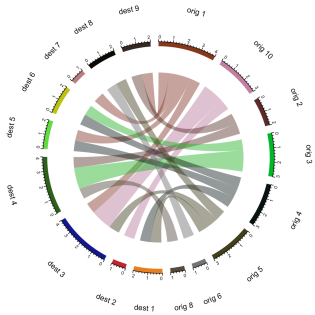


Figure 3.24: Chord Diagram

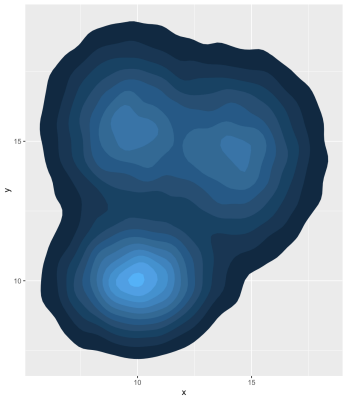


Figure 3.25: Density Plot

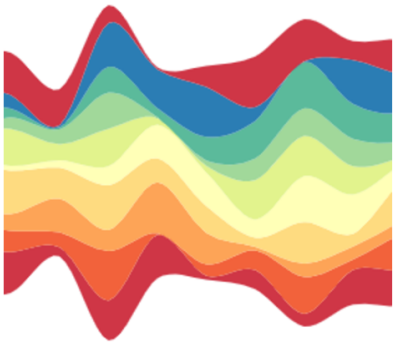


Figure 3.26: Stream Graph

provide critical insights, save lives, and shape public health policy, marking a pivotal moment in the history of epidemiology.

Quotes to Inspire

“Statistics is the grammer of science”
- Karl Pearson

4 Central tendency I

In the previous chapter, you explored how data can be summarized using tables and visually presented through graphs, enabling important features to be highlighted effectively. In this chapter, we shift our focus to **statistical measures** that describe the characteristics of a dataset.

One key aspect of data analysis is identifying a single value that represents the overall dataset. This is where **measures of central tendency** come into play. These are summary statistics that capture the center or typical value of a dataset, providing a concise numerical summary.

There are five commonly used averages: **mean**, **median**, and **mode**, collectively referred to as **simple averages**, and **geometric mean** and **harmonic mean**, known as **special averages**. In addition to these, there are **positional averages**, such as **quartiles**, **deciles**, and **percentiles**, which are determined based on the position of values within an ordered dataset. These measures provide insights into the central value and distribution of the data, making them fundamental tools for understanding and interpreting data patterns.

In this section, we will focus on **simple averages**, with a detailed discussion on positional averages and special averages following later.

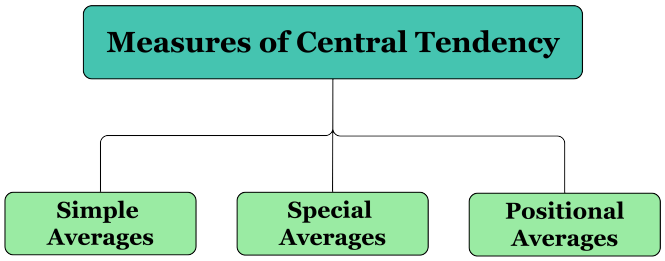


Figure 4.1: Measures of central tendency

Requisites of a good measure of central tendency:

- It should be rigidly defined.
- It should be simple to understand & easy to calculate.

- It should be based upon all values of given data.
- It should be capable of further mathematical treatment.
- It should have sampling stability.
- It should be not be unduly affected by extreme values.

The main objectives of measure of central tendency:

- To condense data in a single value.
- To facilitate comparisons between data.

4.1 Arithmetic mean

This is what people usually intend when they say "average". Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. Mean of set of numbers $x_1, x_2, ..., x_n$ is denoted as $\overline{x}$. It is given by the formula

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.1}$$

Example 4.1: Find the mean of the numbers 2, 4, 7, 8, 11, 12

$$\overline{x} = \frac{2+4+7+8+11+12}{6} = \frac{44}{6} = 7.33$$

4.1.1 Mean of ungrouped frequency distribution

Direct method

If the numbers $x_1, x_2, ..., x_n$ occur with frequencies $f_1, f_2, ..., f_n$ respectively then

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 \quad + ... + x_n f_n}{f_1 + f_2 + ... + f_n}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \tag{4.2}$$

Example 4.2: Table 4.1 below shows the plant height of 50 plants. Find the mean plant height.

Table 4.1: Plant height of 50 plants

Plant height(cm)	159	160	161	162	163
Frequency	3	9	23	11	4

Solution 4.2

The calculation can be arranged as shown

Table 4.2: Solution using direct method

Plant height(x)	Frequency(f)	fx
159	3	477
160	9	1440
161	23	3703
162	11	1782
163	4	652
	$\sum_{i=1}^{n} f_i$= 50	$\sum_{i=1}^{n} f_i x_i$= 8054

$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{8054}{50}$= 161.08 cm

Assumed mean method (Indirect method)

The amount of computation involved above can be reduced by using the following formula:

$$\overline{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i} \tag{4.3}$$

where, A is the assumed mean, which can be any value in x. $d_i = x_i - A$, f_i is the frequency of x_i

Consider the Table 4.1

let $A = 161$; it can be any number in x

Table 4.3: Solution using assumed mean method

Plant height(x)	Frequency(f)	$d_i = x_i - 161$	$f_i d_i$
159	3	-2	-6

Plant height(x)	Frequency(f)	$d_i = x_i - 161$	$f_i d_i$
160	9	-1	-9
161	23	0	0
162	11	1	11
163	4	2	8
	$\sum_{i=1}^{n} f_i = 50$		$\sum_{i=1}^{n} f_i d_i = 4$

using Equation 4.3, $\overline{x} = 161 + \frac{4}{50} = 161.08$ cm

The mean plant height is 161.08 cm.

4.1.2 Mean of grouped frequency distribution

Direct method

The mean for grouped data is obtained from the following formula:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n} \tag{4.4}$$

Where x_i = the mid-point of i^{th} class (i^{th} class mark); f_i= the frequency of i^{th} class; n = the sum of the frequencies or total frequencies in a sample. Note that i =1, 2,..., k, *i.e.* there are k classes.

Example 4.3: Table 4.4 shows the distribution of the marks scored by 60 students in a Maths examination. Find the mean mark.

Table 4.4: Distribution of the marks scored by 60 students

Mark (%)	60-65	65-70	70-75	75-80	80-85
Number of students	2	15	25	14	4

Solution 4.3

The solution can be arranged as shown

Table 4.5: Mean of grouped frequency table using direct method

Marks	Class mark(x_i)	Frequency(f_i)	$f_i x_i$
60-65	62.5	2	125
65-70	67.5	15	1012.5

Marks	Class mark(x_i)	Frequency(f_i)	$f_i x_i$
70-75	72.5	25	1812.5
75-80	77.5	14	1085
80-85	82.5	4	330
		$\sum_{i=1}^{n} f_i$= 60	$\sum_{i=1}^{n} f_i x_i$= 4365

$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{4365}{60}$= 72.75

The mean mark is 72.75%.

Coding method or Indirect method

If all the class intervals of a grouped frequency distribution have equal size C (class width); then the following formula can be used instead of direct method above. This formula makes calculations easier.

$$\overline{x} = A + C\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} \tag{4.5}$$

where, A is the class mark with the highest frequency, $u_i = \frac{x_i - A}{C}$, f_i is the frequency of x_i, C is the class width.

This is called the "coding" method for computing the mean. It is a very short method and should always be used for finding the mean of a grouped frequency distribution with equal class widths.

Consider the Table 4.4 of the Example 4.3.

A = 72.5, class mark with highest frequency; C = 5

Table 4.6: Solution using coding method

Marks	Class mark(x_i)	Frequency(f_i)	$u_i = \frac{x_i - 72.5}{5}$	$f_i u_i$
60-65	62.5	2	-2	-4
65-70	67.5	15	-1	-15
70-75	72.5	25	0	0
75-80	77.5	14	1	14
80-85	82.5	4	2	8
		$\sum_{i=1}^{k} f_i$= 60		$\sum_{i=1}^{k} f_i u_i$=3

using Equation 4.5, $\overline{x} = 72.5 + 5 \times \left(\frac{3}{60}\right)$= 72.75

The mean mark is 72.75%.

Merits and demerits of arithmetic mean

Merits

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement *i.e.* Intelligence, beauty, honesty etc.
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

4.1.3 Weighted average

A weighted average is a method of computing an average where some data points contribute more than others.

> **!** Note
>
> If all the weights of the data point are equal then the weighted average is the same as the simple mean.

The formula for the weighted average is:

$$\text{Weighted average} = \frac{\sum w_i x_i}{\sum w_i} \tag{4.6}$$

where, x_i = individual data values; w_i = corresponding weights

Example 4.4: If $x = [10, 20, 30]$ and its corresponding weights are $w = [1, 2, 3]$ then calculate its weighted average
Solution 4.4

Using the Equation 4.6

$$\text{Weighted average} = \frac{(1 \times 10) + (2 \times 20) + (3 \times 30)}{1 + 2 + 3} = 23.33$$

Example 4.5: What is the weighted average of the first n natural numbers if the weights assigned to each number are equal to the numbers themselves?
Solution 4.5

Using the Equation 4.6

$$\begin{aligned}\text{Weighted average} &= \frac{(1 \times 1) + (2 \times 2) + ... + (n \times n)}{1 + 2 + ... + n} \\ &= \frac{1^2 + 2^2 + ... + n^2}{1 + 2 + ... + n} \\ &= \frac{n(2n+1)(n+1)/6}{n(n+1)/2} \\ &= \frac{2n+1}{3}\end{aligned}$$

4.2 The median

The **median** is the middle value in a set of data when the values are arranged in order from smallest to largest. If there is an **odd** number of values, the median is the one in the middle, with half of the values smaller and half larger. If there is an **even** number of values, the median is the average of the two middle values. The median is a **positional measure**, which means it depends on the order of the data, not the actual values. It helps to find the central point of the data, especially when there are extreme values or outliers that could affect the average.

4.3 Median of ungrouped or raw data

Arrange the given n observations $x_1, x_2, \ldots, x_n$ in ascending order. If the number of values is odd, median is the middle value. If the number of values is even, median is the mean of middle two values.

Arrange data in ascending then use the following formula

When n is odd, Median = Md $= \left(\frac{n+1}{2}\right)^{\text{th}}$ value

When n is even, Median = Md =Average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ value

Example 4.6: Find the median of each of the following sets of numbers.

a) 12, 15, 22, 17, 20, 26, 22, 26, 12

b) 4, 7, 9, 10, 5, 1, 3, 4, 12, 10

Solution 4.6

a) Arranging the data in an increasing order of magnitude, we obtain 12, 12, 15, 17, 20, 22, 22, 26, 26. Here, N = 9 is odd, and so, median $= \left(\frac{9+1}{2}\right)^{\text{th}} = 5^{\text{th}}$ ordered observation = 20.

> **!** Note
>
> If a number is repeated, we still count it the number of times it appears when we calculate the median.

b) Arranging the data in an increasing order of magnitude, we obtain 1, 3, 4, 4, 5, 7, 9, 10, 10, 12. Here, N = 10 is an even number and so median $= \frac{1}{2}\{5^{\text{th}}$ ordered observation $+ 6^{\text{th}}$ ordered observation$\} = \frac{1}{2}(5+7) = 6$.

> **!** Note
>
> You can see in each case, the median divides the distribution into two equal parts, with 50% of the observations greater than it and the other 50% less than it.

4.4 Median of ungrouped frequency distribution

The median is the middle number in an ordered set of data. In a frequency table, the observations are already arranged in an ascending order. We can obtain the median by looking for the value in the middle position.

Odd number of observations

When the number of observations (n) is odd, then the median is the value at the $\left(\frac{n+1}{2}\right)^{\text{th}}$ positional value. For that we use less than cumulative frequency.

Example 4.7: The Table 4.7 shows the frequency of the score obtained in a mathematics quiz. Find the median score.

Table 4.7: Score obtained in a mathematics quiz

Score	0	1	2	3	4
Frequency	3	4	7	6	3

Solution 4.7

Total frequency = 3 + 4 + 7 + 6 + 3 = 23 (odd number). Since the number of scores is odd, the median is at $\left(\frac{23+1}{2}\right)^{\text{th}} = 12^{\text{th}}$ position.

Table 4.8: Less than cumulative frequency of the scores in mathematics quiz

Score	0	1	**2**	3	4
Frequency	3	4	**7**	6	3
Less than cumulative frequency	3	7	**14**	20	23

To find out the 12^{th} position, we use less than cumulative frequencies in Table 4.8 which helps us track how many values are less than or equal to each score. In the table, the less than cumulative frequency for the score 0 is 3 (meaning the first 3 values are 0), for the score 1 is 7 (meaning the first 7 values are 0 and 1), and for the score 2 is 14 (meaning the first 14 values are 0, 1, and 2). If we list the data in order, it would look like this: 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2. (no need to list, this is just for reader's understanding, less than cumulative frequency is enough).

Now, we need to find where the 12^{th} value falls. The 12^{th} value is between the 7^{th} and 14^{th} values, so based on less than cumulative frequency it is clear that 12^{th} value is 2. So the median is 2.

Even number of observations

When the number of observations is even, then the median is the average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ position values.

Example 4.8: The Table 4.9 is a frequency table of the marks obtained in a competition. Find the median score.

Table 4.9: Distribution of marks obtained in a competition.

Mark	0	1	2	3	4
Frequency	11	9	5	10	15

Solution 4.8

Total frequency = 11 + 9 + 5 + 10 + 15 = 50 (even number). Since the number of scores is even, the median is at the average of the values in $\left(\frac{n}{2}\right)^{th} = 25$ *and* $\left(\frac{n}{2}+1\right)^{\text{th}} = 26$ positions. To find out the 25^{th} position and 26^{th} position, we add up the frequencies as shown:

Table 4.10: Less than cumulative frequency of marks obtained.

Mark	0	1	2	3	4
Frequency	11	9	5	10	15
Less than cumulative frequency	11	20	25	35	50

The mark at the 25^{th} position is 2 and the mark at the 26^{th} position is 3. The median is the average of the scores at 25^{th} and 26^{th} positions = $\frac{2+3}{2} = 2.5$

4.5 Median of grouped frequency distribution

The exact value of the median of a grouped data cannot be obtained because the actual values of a grouped data are not known. For a grouped frequency distribution, the median is in the class interval which contains the $\left(\frac{N}{2}\right)^{\text{th}}$ ordered observation, where N is the total number of observations. This class interval is called the **median class**. The median of a grouped frequency distribution can be estimated by either of the following two methods:

Linear interpolation method

The median of a grouped frequency distribution can be estimated by linear interpolation. We assume that the observations are evenly spread through the median class. The median can then be computed by using the following formula:

$$Median = L + \left(\frac{\frac{1}{2}N - F}{f_m}\right) C \tag{4.7}$$

where, N = total number of observations, L = lower limit of the median class, F = sum of all frequencies below L(cumulative frequency), f_m = frequency of the median class, C = class width of the median class.

Estimation from cumulative frequency curve

The median of a grouped frequency distribution can be estimated from a cumulative frequency curve. A horizontal line is drawn from the point $\frac{N}{2}$ on the vertical axis to meet the cumulative frequency curve. From the point of intersection, a vertical line is dropped to the horizontal axis. The value on the horizontal axis is equal to the median.

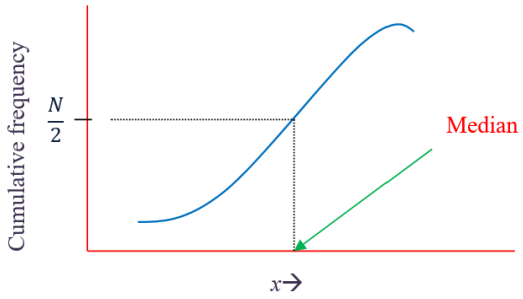


Figure 4.2: Median from a cumulative frequency curve

Example 4.9 Table 4.11 below gives the distribution of the heights of 60 students in a senior high school. Find the median height of the students

Table 4.11: Distribution of heights of 60 students

Height(cm)	145-150	150-155	155-160	160-165	165-170	170-175
Number of students	3	9	16	18	10	4

Solution 4.9

(i) Linear interpolation method

$N = 60$ (Sum of frequencies)

Median class= class interval which contains the $\left(\frac{N}{2}\right)^{\text{th}}$ ordered observation; here $\left(\frac{60}{2}\right)^{\text{th}} = 30^{\text{th}}$ observation. Before the class 160-165 there are 3+9+16 = 28 observations so 30^{th} observation will be in the class 160-165, therefore it is the median class.

L = lower limit of the median class = 160

F = sum of all frequencies below 160(cumulative frequency) = 16+9+3 = 28

f_m = frequency of the median class = 18

C = class width of the median class = 5

using Equation 4.7, $median = 160 + \left(\frac{\frac{1}{2}60-28}{18}\right)5 = 160.56$

(ii) From a cumulative frequency curve

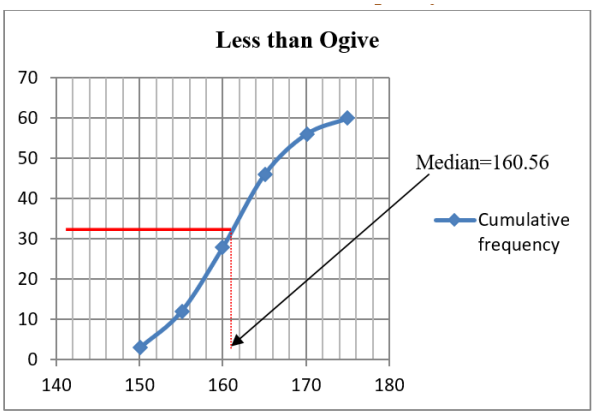


Figure 4.3: Median from a cumulative frequency curve Example 4.7

Merits and demerits of median

Merits

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals.
3. Median can be located even if the data are incomplete.

Demerits

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in calculating mean deviation.
4. It does not take into account all the observations.

4.6 The mode

The mode of a set of data is the value which occurs with the greatest frequency. The mode is therefore the most frequently occuring value in a dataset. The mode is an important measure in case of qualitative data. The mode can be used to describe both quantitative and qualitative data.

4.6.1 Mode of ungrouped or raw data

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example 4.10:

a) The modes of 1, 2, 2, 2, 3 is 2.

b) The modes of 2, 3, 4, 4, 5, 5 are 4 and 5.

c) The mode does not exist when every observation has the same frequency. For example, the following sets of data have no modes: (i) 3, 6, 8, 9; (ii) 4, 4, 4, 7, 7, 7, 9, 9, 9.

> **!** Note
>
> It can be seen that the mode of a distribution may not exist, and even if it exists, it may not be unique. Distributions with a single mode are referred to as *unimodal*. Distributions with two modes are referred to as *bimodal*. Distributions may have several modes, in which case they are referred to as *multimodal*.

Example 4.11: 20 patients selected at random had their blood groups determined. The results are given in the Table 4.12

Table 4.12: Blood group of 20 patients

Blood group	A	AB	B	O
No. of patients	2	4	6	8

Solution 4.11

The blood group with the highest frequency is O. The mode of the data is therefore blood group O. We can say that most of the patients selected have blood group O. Notice that the mean and the median cannot be applied to the data. This is because the variable "blood group" cannot take numerical values. However, it can be seen that the mode can be used to describe both quantitative and qualitative data.

4.7 Mode of grouped data

Mode of a grouped frequency distribution can be found out using the formula below.

$$mode = L + \left(\frac{f_m - f_p}{2f_m - f_p - f_s}\right) C \tag{4.8}$$

Locate the highest frequency the class corresponding to that frequency is called the **modal class**.

where, L = lower limit of the modal class; f_m = the frequency of modal class; f_p= the frequency of the class preceding the modal class; f_s= the frequency of the class succeeding the modal class and C = class interval

Example 4.12: For the frequency distribution of weights of sorghum ear-heads given in Table 4.13 below. Calculate the mode.

Table 4.13: Frequency distribution of weights of sorghum ear heads

Weights of ear heads (g)	No of ear heads (f)
60-80	22
80-100	38
100-120	45
120-140	35
140-160	20

Solution 4.12

Modal class is **100-120**, since it is the class with highest frequency.

Using Equation 4.8 mode is calculated as below.

$mode = 100 + \left(\frac{45-35}{90-38-35}\right) 20 = 111.76$

Mode using histogram

Consider the figure below. The modal class is the class interval which corresponds to rectangle ABCD. An estimate of the mode of the distribution is the abscissa of the point of intersection of the line segments $\overline{\text{AE}}$ and $\overline{\text{BF}}$ in the figure.

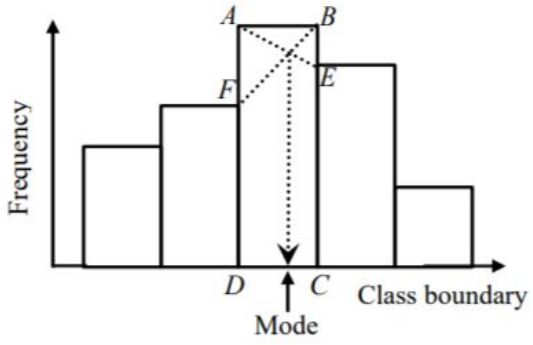


Figure 4.4: Median from a cumulative frequency curve for Example 4.10

Merits and demerits of mode

Merits

1. It is readily comprehensible and easy to compute. In some case it can be computed merely by inspection.
2. It is not affected by extreme values. It can be obtained even if the extreme values are not known.
3. Mode can be determined in distributions with open classes.
4. Mode can be located on the graph also.
5. Mode can be used to describe both quantitative and qualitative data.

Demerits

1. The mode is not unique *i.e.* there can be more than one mode for a given set of data.
2. The mode of a set of data may not exist.
3. It is not based upon all the observation.

Historical Insights

Ancient wall measuring with the mode!
Back in the 5^{th} century BCE, the Athenians used a clever "statistical hack" to plan their siege of Platea. Soldiers counted bricks in an unplastered section of the wall multiple times, and the most frequent count (what we now call the mode) was taken as the best estimate. They then multiplied this by the height of a brick to calculate the wall's height and build ladders tall enough to scale it. Problem-solving with statistics!
Astronomy and the mean
Although the Greeks knew the concept of the arithmetic mean, it wasn't generalized for multiple values until the 16^{th} century. Simon Stevin's invention of the decimal system in 1585 made it much easier to calculate. Astronomer Tycho Brahe was one of the first to use the mean, reducing errors in his estimates of celestial body locations.
Navigation and the median
The concept of the median first appeared in 1599 in Edward Wright's book on navigation, *Certaine errors in navigation*. Wright used it to determine the most likely value in a series of compass readings. Later, in 1669, Christiaan Huygens noticed the difference between the mean and the median while working with Graunt's tables. It's amazing how these early navigators and mathematicians paved the way for the stats we use today.

Quotes to Inspire

“If the statistics are boring, you’ve got the wrong numbers”:- Edward R. Tufte

5 Central tendency II

While simple averages like mean, median, and mode are widely used to summarize data, certain situations call for more specialized measures to capture the essence of a dataset. **Special averages**, including the **geometric mean** and **harmonic mean**, are tailored for specific contexts where the nature of the data or the relationships between data points require a different approach.

5.1 Geometric mean

The **geometric mean (GM)** is a specialized measure of central tendency, particularly suited for datasets involving growth rates, ratios, or percentages, such as population growth, investment returns, or interest rates. Unlike the arithmetic mean, which calculates the average by summing values, the geometric mean finds the average by multiplying values and then taking the root (typically the n^{th} root for n values).

This approach captures the compounding effects present in the data, making the geometric mean an essential tool for accurately summarizing proportional changes or rates over time. Its utility lies in providing a more representative measure for datasets where changes are multiplicative rather than additive.

The geometric mean of a series containing n observations is the n^{th} root of the product of the values. If $x_1, x_2, \dots, x_n$ are observations then

$$GM = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{5.1}$$

where, $\prod_{i=1}^{n} x_i$ means the product of $x_1, x_2, \dots, x_n$

$$= (x_1 x_2 \cdots x_n)^{\frac{1}{n}}$$

$$\log \mathrm{GM} = \frac{1}{n} \log (x_1 x_2 \cdots x_n)$$

$$= \frac{1}{n}\left(\log x_1 + \log x_2 + \cdots + \log x_n\right)$$

$$= \frac{\sum_{i=1}^{n} \log x_i}{n} \tag{5.2}$$

$$\text{GM} = \text{Antilog}\left(\frac{\sum_{i=1}^{n} \log x_i}{n}\right) \tag{5.3}$$

5.1.1 Geometric mean for frequency table

$$\text{GM} = \text{Antilog}\left(\frac{\sum_{i=1}^{k} f_i \log x_i}{n}\right) \tag{5.4}$$

where, x_i is the i^{th} value in the dataset and for a grouped frequency table x_i will be the midpoint of the i^{th} class interval calculated as the average of upper and lower limit, f_i is the frequency of the i^{th} value or class , k is the number of classes.

Example 5.1: If the weight of sorghum ear heads are 45, 60, 48, 100, 65 gms. Find the geometric mean?

Table 5.1: Log values of sorghum ear head weights

Weight of ear head (x)	$\log(x)$
45	1.653
60	1.778
48	1.681
100	2.000
65	1.813
Total	**8.926**

Solution 5.1

Here n =5

using Equation 5.3

$$= \text{Antilog}\left(\frac{8.926}{5}\right)$$

$$= \text{Antilog}(1.785) = 60.95$$

note: here $\text{Antilog}\,(x) = 10^x$ *i.e.*

$$\text{Antilog}\,(1.785) = \ 10^{1.785} = 60.95$$

Example 5.2: Geometric mean of a frequency distribution

Table 5.2: Frequency distribution and log for GM calculation

Weight of ear head (x)	Frequency(f)	$\log(x)$	$f.\log(x)$
45	5	1.653	8.266
60	4	1.778	7.113
48	6	1.681	10.087
100	8	2.000	16.000
65	9	1.813	16.316
Total	**32**		**57.782**

Solution 5.2

Here n =32

using Equation 5.4

$$\sum_{i=1}^{k} f_i \log x_i = 57.782$$

$$\text{GM} = \ \textit{Antilog}\left(\frac{57.782}{32}\right)$$

$$= \textit{Antilog}\,(1.8056) = 10^{1.8056} = 63.92$$

Example 5.3: Geometric mean of a grouped frequency distribution

Table 5.3: Geometric mean calculation for grouped frequency table

Class	Mid value (x)	Frequency(f)	$\log(x)$	$f.\log(\text{x})$
60-80	70	5	1.845	9.225
80-100	90	4	1.954	7.817

Class	Mid value (x)	Frequency(f)	log(x)	f.log(x)
100-120	110	6	2.041	12.248
120-140	130	8	2.114	16.912
140-160	150	9	2.176	19.585
	Total	**32**		**65.787**

Solution 5.3
Here $n = 32$

using Equation 5.4

$$\sum_{i=1}^{k} f_i \log x_i = 65.787$$

$$\text{GM} = \ Antilog\left(\frac{65.787}{32}\right)$$

$$= Antilog\,(2.0558) = 10^{2.0558} = 113.71$$

Merits and demerits of geometric mean

Merits

- It is rigidly defined.
- It is based on all the observations of the series.
- It is suitable for measuring the relative changes.
- It gives more weights to the small values and less weight to the large values.
- It is used in averaging the ratios, percentages and in determining the rate gradual increase and decrease.
- It is capable of further algebraic treatment.

Demerits

- It is not easy to understand.
- It is difficult to calculate.
- It cannot be calculated, if the number of negative values is odd.
- It cannot be calculated, if any value of a series is zero.
- At times it gives a value which may not be found in the series or impractical.

5.2 Harmonic mean

Harmonic means are often used in averaging things like rates (e.g. the average travel speed given duration of several trips). Harmonic mean (HM) of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given value.

> **!** Note
>
> Harmonic mean can be easily remembered as "reciprocal of the mean of the reciprocals".

If $x_1,\ x_2, \ldots,\ x_n$ are n observations then

$$\text{H.M} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{5.5}$$

Steps in calculating Harmonic Mean (H.M)

1. Calculate the reciprocal (1/value) for every value.
2. Find the average of those reciprocals (just add them and divide by how many there are).
3. Then do the reciprocal of that average (=1/average).

Example 5.4: From the given data 5, 10, 17, 24, 30. calculate H.M

Solution 5.4

Here $n = 5$

Table 5.4: Inverse of numbers to calculate harmonic mean

x	$1/x$
5	0.2
10	0.1
17	0.059
24	0.042
30	0.033
Total	**0.434**

using Equation 5.5

$$\text{H.M} = \frac{5}{0.434} = 11.525$$

5.2.1 Harmonic mean for frequency table

$$\text{H.M} = \frac{n}{\sum_{i=1}^{k} f_i \frac{1}{x_i}} \tag{5.6}$$

where, x_i is the i^{th} value in the dataset and for a grouped frequency table x_i will be the midpoint of the i^{th} class interval calculated as the average of upper and lower limit, f_i is the frequency of the i^{th} value or class , k is the number of classes.

Example 5.5: For the given data calculate the harmonic mean.

Table 5.5: Model data for harmonic mean calculation

x	20	21	22	23	24	25
frequency (f)	4	2	7	1	3	1

Solution 5.5

Table 5.6: Calculation of harmonic mean from frequency table

x	f	$1/x$	f . (1/x)
20	4	0.050	0.200
21	2	0.048	0.095
22	7	0.045	0.318
23	1	0.043	0.043
24	3	0.042	0.125
25	1	0.040	0.04
	18		**0.822**

Here n =18

using Equation 5.6

$$\text{H.M} = \frac{18}{0.822} = 21.90$$

Example 5.6: *Cistern Problem*-Two pipes are used to fill a cistern. Pipe A can fill the cistern in 3 hours. Pipe B can fill the cistern in 5 hours. If both pipes are opened at the same time, how long will it take to fill the cistern completely?

Solution 5.6

To solve the cistern problem, we first calculate the rate at which each pipe fills the cistern. Pipe A fills the cistern in 3 hours, so its rate is $\frac{1}{3}$ of the cistern per hour. Pipe B fills the cistern

in 5 hours, so its rate is $\frac{1}{5}$ per hour. To find the combined rate, we add these rates together; $\frac{1}{3} + \frac{1}{5} = \frac{8}{15}$. This means the two pipes together fill $\frac{8}{15}$ of the cistern each hour. To find the total time, we take the reciprocal of the combined rate *i.e.* Total time $= \frac{1}{\frac{8}{15}} = \frac{15}{8} = 1.875$. To convert 1.875 hours into hours and minutes, we first separate the whole number from the decimal part. 1.875 hours consists of 1 hour (the whole number) and 0.875 hours (the decimal part). Next, we convert the decimal part into minutes. Since 1 hour is equal to 60 minutes, we multiply 0.875 by 60 *i.e.* $0.875 \times 60 = 52.5$ minutes. Rounding 52.5 minutes gives approximately 53 minutes. Thus, 1.875 hours is equivalent to **1 hour and 53 minutes**.
The problem can be easily solved using the harmonic mean, the harmonic mean of two pipes is

$$\text{H.M} = \frac{2}{\frac{1}{3} + \frac{1}{5}}$$

$$= \frac{2 \cdot 3 \cdot 5}{3 + 5} = \frac{30}{8} = 3.75 \,\text{hours}$$

the harmonic mean of the pipes is 3.75 hours, and since both pipes are working together, the total time is half of that, which confirms the answer of 1 hour and 53 minutes.

Example 5.7: A car travels a certain distance from City A to City B at a speed of 60 km/h, and returns the same distance from City B to City A at a speed of 90 km/h. What is the average speed for the entire trip?

Solution 5.7

To find the average speed when traveling the same distance at two different speeds, we use the harmonic mean. The harmonic mean for the speed is $\frac{2 \cdot S_1 \cdot S_2}{S_1 + S_2}$, where $S_1 = 60\,\text{km/h}$ is the speed from City A to City B, $S_2 = 90\,\text{km/h}$ is the speed from City B to City A.

In this case, we are calculating the average speed over the entire round trip. The harmonic mean is used because it accounts for the fact that traveling at different speeds over the same distance results in an overall average speed that is closer to the lower of the two speeds, rather than simply averaging the two speeds.

Now, applying the harmonic mean formula:

$S_{\text{avg}} = \frac{2 \cdot 60 \cdot 90}{60 + 90} = \frac{10800}{150} = 72\,\text{km/h}$

So, the average speed for the entire trip is 72 km/h.

Merits and demerits of harmonic mean

Merits

- It is rigidly defined.
- It is defined on all observations.

- It is amenable to further algebraic treatment.
- It is the most suitable average when it is desired to give greater weight to smaller and less weight to the larger ones.

Demerits

- It is not easily understood.
- It is difficult to compute.
- It is only a summary figure and may not be the actual item in the series.
- It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.
- It is rarely used in grouped data.

5.3 AM, GM or HM ?

The arithmetic mean (AM), geometric mean (GM), and harmonic mean (HM) are termed as **Pythagorean means**. The Pythagorean means refer to three specific types of means that were known to the ancient Greek mathematicians, particularly to Pythagoras and his followers. Choosing the right average depends on what you're measuring and how the data behaves. Let's break it down step by step in a simple way:

Arithmetic Mean (AM)
The arithmetic mean is the most common type of average. You use it when the values in your data add together in a straightforward way. This is suitable for quantities like:
- Heights or weights
- Lengths or distances
- Marks in exams

For example, if you want to find the average height of students in a class, you add up all the heights and divide by the number of students. The arithmetic mean gives a meaningful average because height or weight adds linearly.

Harmonic Mean (HM)
The harmonic mean is useful when you are working with rates, ratios, or situations where quantities add up as reciprocals. Some examples include:
- Speeds (distance per unit time)
- Capacitors in a series circuit
- Rates like fuel efficiency or cost per unit

For instance, imagine you are driving the same distance at different speeds. If you want to find the average speed for the entire trip, the harmonic mean is the best choice. This is because speeds relate inversely to time when you go faster, you take less time, and vice versa.

Geometric Mean (GM)
The geometric mean is the right choice when your data involves multiplication or compounding, such as:
- Growth rates (like population growth or interest rates)
- Percentages (like inflation rates)
- Ratios

For example, if you have annual interest rates for 10 years and want to find a single rate that represents the same total growth over that period, the geometric mean gives you the answer. It works by multiplying the rates and taking the root, which accounts for the compounding effect.

Here's a simple example to understand how the geometric mean works. Suppose you invest Rs.100, and your investment changes over three years as follows: in the first year, it grows by 10%, represented by a growth factor of 1.10; in the second year, it grows by 20%, represented by 1.20; and in the third year, it drops by 10%, represented by 0.90. To find a single rate that would give the same overall growth over the three years, you multiply the growth factors: $1.10 \times 1.20 \times 0.90 = 1.188$. Then, take the cube root (since there are three years): $\sqrt[3]{1.188} \approx 1.059$. This gives a geometric mean rate of approximately 1.059, or 5.9% per year. In other words, if your Rs. 100 investment grew steadily at a rate of 5.9% each year, it would result in the same total growth over three years, leaving you with about Rs. 118.80 at the end.

Key Takeaways
1. Use **Arithmetic Mean** when values combine directly, like adding lengths or weights.
2. Use **Harmonic Mean** for rates or quantities that work reciprocally, like speed or resistance.
3. Use **Geometric Mean** for data involving multiplication or compounding, like growth rates or percentages.

By understanding the relationship between the data and the type of average, you can choose the most meaningful measure for your analysis.

5.4 Relation between AM, GM and HM

The formula for the relation between AM, GM, HM is the product of arithmetic mean and harmonic mean is equal to the square of the geometric mean. This can be presented here in the form of Equation 5.7

$$\mathrm{AM} \times \mathrm{HM} = \mathrm{GM}^2 \tag{5.7}$$

also

$$\mathbf{AM} \geq \mathbf{GM} \geq \mathbf{HM} \tag{5.8}$$

5.4.1 Geometric illustration

Consider two numbers a and b. See Figure 5.1 a semi circle can be drawn with diameter a+b. Then its radius is half the diameter, which will be the arithmetic mean $\frac{a+b}{2}$.

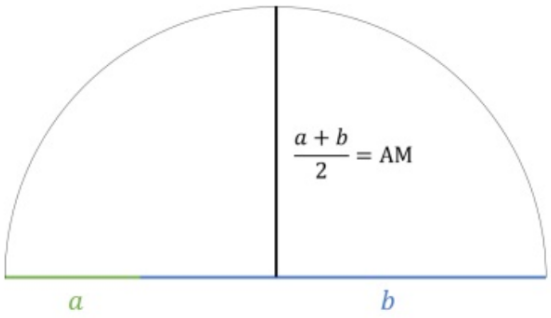


Figure 5.1: Arithmetic mean on a semi circle

The geometric mean is the length of the perpendicular where a and b meet, which is never larger than the radius of the circle as illustrated in Figure 5.2

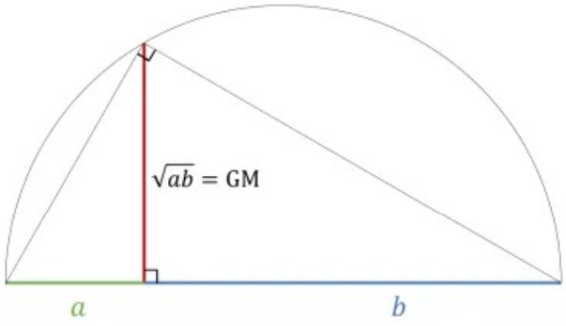


Figure 5.2: Geometric mean on a semi circle

Now draw a line from the top of red line which is the GM in Figure 5.2 to the center of the circle. Now that line is the radius of the circle, so it is equal to AM, which is now the hypotenuse of the newly formed triangle with GM as the leg of the triangle. So from Figure 5.3 it is clear that

$$\mathbf{AM} \geq \mathbf{GM} \tag{5.9}$$

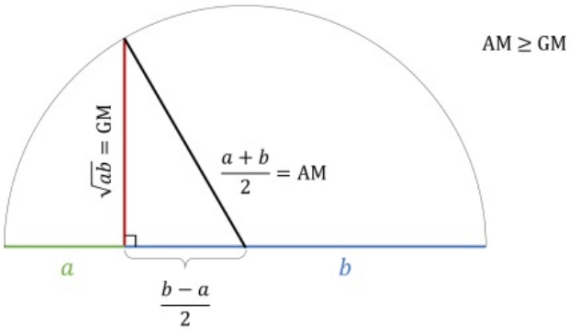


Figure 5.3: Geometric mean and aritmetic mean on a semi circle

Now if we draw an altitude to the hypotenuse as shown in Figure 5.3, the upper length on the hypotenuse is the harmonic mean . We can now consider another triangle where HM is a leg and the GM is the hypotenuse, this shows the GM is never smaller than the HM. So

$$\mathbf{GM} \geq \mathbf{HM} \tag{5.10}$$

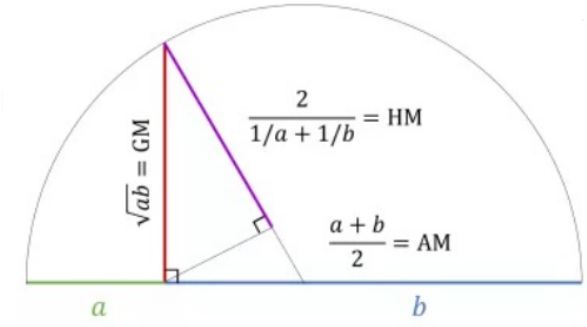


Figure 5.4: AM, GM and HM on a semi circle

Now from Equation 5.9 and Equation 5.10 it is clear that $\mathbf{AM} \geq \mathbf{GM} \geq \mathbf{HM}$

5.5 Positional averages

Positional averages are measures derived directly from the values in a dataset. These averages are based on the position of the values within the series and are used to represent the overall dataset or highlight specific positional characteristics.

> **!** Note
>
> The **median** although a simple average is also a positional average that represents the middle value of an ordered dataset, making it a central point of reference. Similarly, the **mode**, which identifies the most frequently occurring value in the dataset, is also a positional average since it is directly taken from the series.

The other common positional averages include **percentiles**, **quartiles**, and **deciles**, which divide the data into equal parts to analyze its distribution.

In contrast, measures like the **arithmetic mean**, **geometric mean**, and **harmonic mean** are referred to as **mathematical averages**, as they are calculated through specific mathematical operations rather than being derived from the data's positional properties.

5.6 Quartiles

The **median** divides a dataset into two equal halves. Similarly, it is possible to divide a dataset into more than two parts. When an **ordered** dataset is divided into four equal sections, the points that mark these divisions are called **quartiles**.

The **first or lower quartile ($\mathbf{Q_1}$)** is a value that has one fourth, or 25% of the observations below its value.

The **second quartile ($\mathbf{Q_2}$)**, has one-half, or 50% of the observations below its value. The second quartile is equal to the **median**.

The **third or upper quartile**, **($\mathbf{Q_3}$)**, is a value that has three-fourths, or 75% of the observations below it. If there are n items in a dataset then

$$Q_1 = \left(\frac{n+1}{4}\right)^{\text{th}} \text{item} \tag{5.11}$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{\text{th}} \text{item} \tag{5.12}$$

Calculations of quartiles are explained using the example below. See in the example the procedure followed when a fraction appear in the calculation.

Example 5.8: Compute quartiles for the data 25, 18, 30, 8, 15, 5, 10, 35, 40, 45.

Solution 5.8

First arrange the data in ascending order

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

here $n = 10$

using Equation 5.11 and Equation 5.12

$Q_1 = \left(\frac{10+1}{4}\right)^{th} = 2.75^{\text{th}}$ item; when such a fraction appears we use the following procedure

$Q_{1} = $2.75^{th} item = 2^{nd} item + 0.75(3^{rd} item - 2^{nd} item)

So from the given data $Q_1 = 8+0.75(10-8) = 9.5$, also Q_2 is the median, here $Q_2 = (18+25)/2 = 21.5$

$Q_3 = \left(\frac{3(10+1)}{4}\right)^{th} = 8.25^{\text{th}}$ item $= 8^{\text{th}}$ item + 0.25(9^{th} item - 8^{th} item) = 35+0.25(40-35) = 36.25

Quartiles of a discrete frequency data

The following steps explain how to calculate quartiles for discrete frequency data

1. Find cumulative frequencies.
2. Find $\left(\frac{n+1}{4}\right)$.
3. See in the cumulative frequencies, the value just greater than $\left(\frac{n+1}{4}\right)$, then the corresponding value of x is Q_1.
4. Find $\left(\frac{3(n+1)}{4}\right)$.
5. See in the cumulative frequencies, the value just greater than $\left(\frac{3(n+1)}{4}\right)$, then the corresponding value of x is Q_3.

Example 5.9: Compute quartiles for the data given below

Table 5.7: A model frequency distribution

x	5	8	12	15	19	24	30
f	4	3	2	4	5	2	4

Solution 5.9

Table 5.8: Cumulative frequency data for quartile calculation

x	f	cf
5	4	4
8	3	7
12	2	9
15	4	13

x	f	cf
19	5	18
24	2	20
30	4	24

Here n =24

$\left(\frac{n+1}{4}\right) = \left(\frac{25}{4}\right) = 6.25$

The cumulative frequency value just greater than 6.25 is 7, the x value corresponding to cumulative frequency 7 is 8. So $Q_1 = 8$

$\left(\frac{3(n+1)}{4}\right) = \left(\frac{3\times25}{4}\right) = 18.75$

The cumulative frequency value just greater than 18.75 is 20, the **x** value corresponding to cumulative frequency 20 is 24. So $Q_3 = 24$

Quartiles of a grouped frequency data

Below it is explained steps in calculating quartiles for a continuous frequency data

1. Find cumulative frequencies.
2. Find $\left(\frac{n}{4}\right)$.
3. See in the cumulative frequencies, the value just greater than$\left(\frac{n}{4}\right)$, and then the corresponding class interval is called **first quartile class**.
4. Find $3\left(\frac{n}{4}\right)$.
5. See in the cumulative frequencies the value just greater than $33\left(\frac{n}{4}\right)$then the corresponding class interval is called **3rd quartile class**. Then apply the respective formulae

$$Q_1 = l_1 + \frac{\frac{n}{4} - m_1}{f_1} \times c_1 \tag{5.13}$$

$$Q_3 = l_3 + \frac{3\left(\frac{n}{4}\right) - m_3}{f_3} \times c_3 \tag{5.14}$$

where, l_1 = lower limit of the first quartile class

f_1 = frequency of the first quartile class

c_1 = width of the first quartile class

m_1 = cumulative frequency preceding the first quartile class

l_3= 1ower limit of the 3rd quartile class

f_3= frequency of the 3rd quartile class

c_3= width of the 3rd quartile class

m_3 = cumulative frequency preceding the 3rd quartile class

Example 5.10: Find the quartiles for the grouped frequency data given

Table 5.9: A model grouped frequency data

Class	**frequency**	**cumulative frequency**
0–10	11	11
10–20	18	29
20–30	25	54
30–40	28	82
40–50	30	112
50–60	33	145
60–70	22	167
70–80	15	182
80–90	12	194
90–100	10	204

Solution 5.10

$(\frac{n}{4}) = \frac{204}{4} = 51$

The cumulative frequency value just greater than 51 is 54 so the class 20-30 is the 1st quartile class

using Equation 5.13

$$Q_1 = 20 + \frac{51 - 29}{25} \times 10 = 28.8$$

$3\,(\frac{n}{4})= 3 \times \frac{204}{4} = 153$

The cumulative frequency value just greater than 153 is 167 so the class 60-70 is the 3rd quartile class.

using Equation 5.14

$$Q_1 = 60 + \frac{153 - 145}{22} \times 10 = 63.63$$

5.7 Percentiles

Percentiles divide an ordered dataset into 100 equal parts, with each part containing 1% of the observations. The p^{th} percentile, denoted as P_p, is the value below which x percent of the data falls.

For example:
- The $\mathbf{P_{50}}$, $\mathbf{50^{th}}$ **percentile** is equivalent to the **median**, representing the middle value of the dataset.
- The $\mathbf{P_{25}}$, $\mathbf{25^{th}}$ **percentile** corresponds to the first quartile (Q_1), which marks the lower 25% of the data.
- The $\mathbf{P_{75}}$, $\mathbf{75^{th}}$ **percentile** is the third quartile (Q_3), indicating that 75% of the data falls below this value.

For raw data, first arrange the n observations in increasing order. Then the x^{th} percentile is given by

$$P_p = \left(\frac{p(n+1)}{100}\right)^{\text{th}} \text{ item} \tag{5.15}$$

Percentiles of discrete frequency data

To calculate percentiles for discrete frequency data, follow these steps which is similar to that of quartiles:

1. Find the cumulative frequencies.
2. Find the position of the percentile. For the p-th percentile, calculate the position using the formula:
$$P_p = \left(\frac{p(n+1)}{100}\right)$$
where p is the percentile and n is the total number of data points.
3. Identify the value in the cumulative frequencies.

find the cumulative frequency that is just greater than or equal to the calculated position P_p. The corresponding value of x is the p-th percentile.

For example, to find the first percentile (P_1): - Calculate $P_1 = \left(\frac{1(n+1)}{100}\right)$. - Locate the cumulative frequency that is just greater than P_1, and the corresponding value of x is P_1.

Percentiles of a grouped frequency data

Calculation of percentile is very much similar to that of quartile. For a frequency distribution the p^{th} percentile is given by following steps

1. Find cumulative frequencies.

2. Find $\left(\frac{p.n}{100}\right)$.
3. See in the cumulative frequencies, the value just greater than $\left(\frac{p.n}{100}\right)$ and then the corresponding class interval is called **Percentile class**.
4. Use the following formula

$$P_p = l + \frac{\left(\frac{p \times n}{100}\right) - cf}{f} \times c \tag{5.16}$$

where,

l = lower limit of the percentile class

cf = cumulative frequency preceding the percentile class

f = frequency of the percentile class

c = class interval

n = total number of observations

Example 5.11: Compute P_{25} and P_{75} for the data 25, 18, 30, 8, 15, 5, 10, 35, 40, 45.

Solution 5.11

First arrange the data in ascending order

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

Here $n = 10$

$P_{25} = \left(\frac{25(10+1)}{100}\right)^{\text{th}} = 2.75^{\text{th}}$ item

$P_{25} = 2.75^{\text{th}}$ item = 2^{nd} item + 0.75(3^{rd} item – 2^{nd} item)

So from the given data P_{25} = 8+0.75(10– 8) = 9.5

$P_{75} = \left(\frac{75(10+1)}{100}\right)^{\text{th}} = 8.25^{\text{th}}$ item

i.e. $P_{75} = \left(75 \times \frac{10+1}{100}\right)^{th} = 8.25^{\text{th}}$ item = 8^{th} item + 0.25(9^{th} item –8^{th} item) = 35+0.25(40-35) = 36.25

> **!** Note
>
> Data in Example 5.11 is same as Example 5.8; it can be seen that $P_{25} = Q_1$ & $P_{75} = Q_3$ always

Try yourself
Find P_{25}, P_{50} & P_{75} for Example 5.9 & 5.10; verify that $P_{50} = Q_2$, $P_{25} = Q_1$ & $P_{75} = Q_3$

5.8 Deciles

Deciles consist of 9 points that divide an ordered dataset into ten equal parts. The d^{th} decile is denoted as D_d. It is important to note that the **median** is the **5th decile**.

$$D_d = \left(\frac{d(n+1)}{10}\right)^{\text{th}} \text{ item} \tag{5.17}$$

where, d is the decile number (e.g., $d = 1$ for the first decile, $d = 9$ for the ninth decile), and n is the total number of data points.

Deciles of discrete frequency data

To calculate deciles for discrete frequency data, follow these steps which are similar to that of percentiles:

1. Find the cumulative frequencies.
2. Find the position of the decile.
   For the d^{th} decile, calculate the position using the formula:
$$D_d = \left(\frac{d(n+1)}{10}\right)$$
3. Identify the value in the cumulative frequencies.
   Find the cumulative frequency that is just greater than or equal to the calculated position D_d. The corresponding value of x is the d^{th} decile.

For example, to find the first decile (D_1): - Calculate $D_1 = \left(\frac{1(n+1)}{10}\right)$. - Locate the cumulative frequency that is just greater than D_1, and the corresponding value of x is D_1.

Deciles of a grouped frequency data

For a frequency distribution the d^{th} decile is given by following steps

1. Find cumulative frequencies.
2. Find $\left(\frac{d.n}{10}\right)$.

3. See in the cumulative frequencies, the value just greater than$\left(\frac{d.n}{10}\right)$and then the corresponding class interval is called **decile class**.

4. Use the following formula

$$D_d = l + \frac{\left(\frac{d \times n}{10}\right) - cf}{f} \times c \tag{5.18}$$

where,

l = lower limit of the decile class

cf = cumulative frequency preceding the decile class

f = frequency of the decile class

c = class interval

n = total number of observations

Try yourself
Find D_5 for Example 5.9, 5.10 & 5.11; verify that $D_5 = Q_2 = P_{50} = median$

Historical Insights

The harmonic mean and the perfect fourth
The harmonic mean a term derived from the ancient Greeks, particularly associated with Pythagoras or his followers. The harmonic mean is closely related to musical intervals, specifically the *perfect fourth*. In music theory, an octave change upwards corresponds to a doubling of the frequency (a 1:2 ratio). The harmonic mean of 1 and 2, which is $\frac{4}{3}$, defines the frequency ratio for the perfect fourth, making it a crucial concept in understanding musical harmony and acoustics.

Quotes to Inspire

"The best thing about being a statistician is that you get to play in everybody else's backyard". – John Tukey

6 Measures of dispersion

In the previous chapters, we have seen how a set of data can be summarized by a single representative value that describes the central tendency of the data. Consider the two sets of data, A and B, in Table 6.1.

Table 6.1: Model dataset to demonstrate dispersion

A	1	2	3	3	4	5
B	-1	0	3	3	5	8

You can see mean, median and mode for both the sets **A** & **B** in Table 6.1 is 3.

The plot of values of **A** and **B** in Table 6.1 can be seen in Figure 6.1. The figure is known as dot plot.

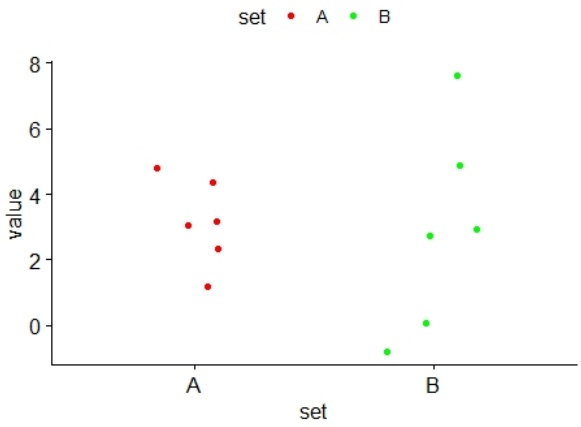


Figure 6.1: Dot plot of datasets A & B

It can be seen in Figure 6.1 that, while values of data set **A** are grouped close to their mean, while the values of data set **B** are more spread out. We say that values of data set **B** are more dispersed (or scattered) than those of data set **A.** This example shows that the measures of central tendency are not enough in describing a set of data. In addition to using these measures, we need numerical measures of dispersion (or variation) of a set of data.

> **!** Note
>
> **Dispersion** refers to the extent to which numerical data values deviate from an average or central value. The statistical measures calculated from the data to quantify this dispersion are known as measures of dispersion.

6.1 Characteristics of a good measure of dispersion

An ideal measure of dispersion is expected to possess the following properties

1. It should be rigidly defined.
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.
4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to calculate.

The most important measures of dispersion are **range**, **quartile deviation**, **variance**, **inter-quartile range**, **mean absolute deviation** and **standard deviation**.

6.2 The range

This is the simplest possible measure of dispersion. The range of a set of data is defined as the difference between the largest observation and the smallest observation in the set of data.

Thus,

Range = largest observation − smallest observation.

It can be denoted as, Range = L – S.

where, L = Largest value; S = Smallest value.

Example 6.1: The marks obtained by 8 students in mathematics and physics examinations are as follows:

Table 6.2: Model dataset of marks obtained in two subjects

Student	Mathematics	Physics
1	35	50
2	60	55

Student	Mathematics	Physics
3	70	70
4	40	65
5	85	89
6	96	68
7	55	72
8	65	80

Find the ranges of the two sets of data. Are the physics marks more dispersed than the mathematics marks?

Solution 6.1

For mathematics,

Highest mark = 96, lowest mark = 35, range = 96 – 35 = 61

For physics,

Highest mark = 89, lowest mark = 50, range = 89 – 50 = 39.

The mathematics marks have a wider range than the physics marks. The mathematics marks are therefore more dispersed than the physics marks.

In individual observations and discrete series, L and S are easily identified. In case of grouped frequency distribution, the following method is employed.

L = Upper boundary of the highest class

S = Lower boundary of the lowest class.

$$Range = L - S \tag{6.1}$$

Example 6.2: Calculate range from the following distribution

Table 6.3: Model frequency distribution table for range calculation

Size	60–63	63–66	66–69	69–72	72–75
Number	5	18	42	27	8

Solution 6.2

L = Upper boundary of the highest class = 75

S = Lower boundary of the lowest class = 60

Range = L – S = 75 – 60 = 15

Merits and demerits of range

Merits

1. It is simple to understand.
2. It is easy to calculate.
3. In certain types of problems like quality control, weather forecasts, share price analysis, etc.

Demerits

1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.
3. It cannot be calculated from open-end class intervals.
4. It is not suitable for mathematical treatment.
5. It is a very rarely used measure.

6.3 The inter-quartile range (IQR)

The range is a simple and quick measure to calculate. However, because it relies solely on the maximum and minimum values in a data set, it does not provide information about how the data is distributed between these two values. As a result, the range may not be an effective measure of dispersion, especially if one or both of these values are significantly different from the rest of the data. To address this limitation, the interquartile range is often used. The interquartile range is a more robust measure of dispersion, defined as the difference between the upper and lower quartiles of the data. IQR is also known as midspread. Thus,

$$IQR = Q_3 - Q_1 \tag{6.2}$$

The inter-quartile range of a set of data is therefore not affected by values of the data outside Q_1 and Q_3 making it a more reliable measure of spread for skewed or non-normal distributions.

Example 6.3: Consider the two sets of data A & B below, find IQR

Table 6.4: Model dataset for IQR calculation

A	3	4	5	6	8	9	10	12	15
B	3	8	8	9	9	9	10	10	15

For data set A, $Q_1 = 4.5$, $Q_3 = 11$; so Inter-Quartile Range = 11 – 4.5 = 6.5

For data set B, $Q_1 = 8$, $Q_3 = 10$; so Inter-Quartile Range = 10 – 8 = 2

Since the interquartile range (IQR) of data set A is greater than that of data set B, these results indicate that data set A is more dispersed than data set B. It is also noticeable that the range is the same for both sets.

Merits and demerits of IQR

Merits

1. It is simple to calculate and easy to understand.
2. It is not affected by extreme values (outliers) in the data, making it a more reliable measure of spread than the range.
3. IQR provides a clear measure of the spread of the middle 50% of the data, giving a better representation of variability when data is skewed.
4. It can be used for skewed distributions, where the range and standard deviation may not be as useful.
5. IQR is particularly useful in identifying outliers, as data points outside 1.5 times the IQR from the quartiles are often considered outliers.

Demerits

1. The IQR does not use all the data points, which means it may not represent the variability of the entire dataset.
2. It may not be as intuitive as the range or standard deviation for some users, particularly in more complex datasets.
3. The IQR is less sensitive to variations in the data outside of the interquartile range, meaning it might not fully reflect extreme values or trends.
4. It is not as effective when comparing datasets with significantly different shapes or distributions.

6.4 Mean absolute deviation (MAD)

The mean absolute deviation (MAD) is a measure of variability that indicates the average distance between observations and their mean. MAD uses the original units of the data, which simplifies interpretation. Larger values signify that the data points spread out further from the average. Conversely, lower values correspond to data points bunching closer to it. The mean absolute deviation is also known as the mean deviation and average absolute deviation.

Here is how to calculate the mean absolute deviation.

1. Calculate the mean.
2. Calculate the difference of each observation from mean and take absolute value *i.e.* ignore the sign. This difference is known as absolute deviation.
3. Add those deviations together.
4. Divide the sum by the number of data points.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n} \tag{6.3}$$

Example 6.4: Find the mean absolute deviation of the following 10, 15, 15, 17, 18, 21

Table 6.5: Calculation of mean absolute deviation

x_i	$x_i - \overline{x}$	$\lvert x_i - \overline{x}\rvert$
10	-6	6
15	-1	1
15	-1	1
17	1	1
18	2	2
21	5	5
$\overline{x} = 16$		$\sum_{i=1}^{n} \lvert x_i - \overline{x}\rvert = 16$

Here $n = 6$ and $\sum_{i=1}^{n} |x_i - \overline{x}| = 16$ therefore MAD $= \frac{16}{6} = 2.67$

Merits and demerits of MAD

Merits

1. Mean deviation is simple and easy.
2. Different items of observations can be easily compared with mean deviation.

3. Mean deviation is better than quartile deviation and range because it is based on all the observations of the series.
4. Mean deviation is less affected by the extreme values in the series while comparing to standard deviation.
5. Mean deviation is rigidly defined. So, it has fixed value.
6. Mean deviation about median will be least.

Demerits

1. Mean deviation becomes difficult to compute mean deviation in case of fractions.
2. It is not applicable for algebraic calculations.
3. It cannot be calculated from open-end class intervals.
4. Mean deviation is not a good measure as it ignores negative signs of deviations.

6.5 The variance and standard deviation

The most important measures of variability are the sample variance and the sample standard deviation. If x_1, x_2, ...,x_n is a sample of n observations, then the **sample variance** is denoted by s^2 and is defined by the equation.

$$\text{sample variance, } s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{6.4}$$

The sample **standard deviation**, s, is the positive square root of the sample variance.

$$\text{standard deviation, } s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{6.5}$$

> **!** Note
>
> **Why standard deviation?**
> While both variance and standard deviation measure data dispersion, standard deviation is preferred for practical interpretation. Variance is expressed in squared units, making it harder to interpret. For example, if the data represents lengths in meters, the variance is in square meters (m^2), which complicates understanding variability. In contrast, standard deviation is the square root of variance, preserving the original unit (e.g., meters), making it more intuitive. Thus, standard deviation is preferred for its clarity and ease of interpretation, especially when analyzing how data points deviate from the mean.

If the standard deviation of data set A is greater than that of data set B, it indicates that data set A is more dispersed than data set B. A higher standard deviation means that the values in data set A are more spread out from the mean compared to the values in data set B. It's important to note that the standard deviation of any data set is always a non-negative number, as it represents the square root of the variance, which is always non-negative. Variance and standard deviation can never be negative values.

Example 6.5: Consider the Table 6.1 discussed earlier, find the standard deviation?

Solution 6.5

Table 6.6: Calculation of standard deviation of set A

Set A	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
	1	-2	4
	2	-1	1
	3	0	0
	3	0	0
	4	1	1
	5	2	4
Sum	**18**	**0**	**10**

Mean $(\overline{x}) = \frac{18}{6} = 3$

$$\textit{Sample variance, } s_A^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{10}{5} = 2$$

$$\text{Sample standard deviation, } s_A = \sqrt{s_A^2} = \sqrt{2} = 1.414$$

Table 6.7: Calculation of standard deviation of set B

Set B	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
	-1	-4	16
	0	-3	9
	3	0	0
	3	0	0
	5	2	4
	8	5	25
Sum	**18**	**0**	**54**

Mean $(\overline{x}) = \frac{18}{6} = 3$

$$\textit{Sample variance, } s_B^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{54}{5} = 10.8$$

$$\textit{Sample standard deviation, } s_B = \sqrt{s_B^2} = \sqrt{10.8} = 3.29$$

It can be seen that $s_B > s_A$, confirming that data set B is more dispersed than data set A as shown in Figure 6.1.

An alternative formula for computing the variance

The computation of s^2 requires calculations of $\overline{x}$, n subtractions and n squaring and adding operations. If the original observations or the deviations $(x_i - \overline{x})$ are not integers, the deviations $(x_i - \overline{x})$ may be difficult to work with, and several decimals may have to be carried to ensure numerical accuracy. A more efficient computational formula for s^2 is given by

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right\} \tag{6.6}$$

Example 6.6: Consider the data set below; find standard deviation?

Table 6.8: Model dataset for standard deviation calculation

3	4	5	6	8	9	10	12	15

Solution 6.6

Table 6.9: Calculation of sd using alternate formula

x_i	x_i^2
3	9
4	16
5	25
6	36
8	64
9	81
10	100
12	144
15	225
$\sum_{i=1}^{9} x_i = \mathbf{72}$	$\sum_{i=1}^{9} x_i^2 = \mathbf{700}$

using Equation 6.6 ; here $n = 9$

$s^2 = \frac{1}{8}\left\{700 - \frac{1}{9}(72)^2\right\} = 15.5$

$s = \sqrt{15.5} = 3.94$

6.5.1 Standard deviation for frequency table

For discrete frequency table

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} f_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} f_i x_i\right)^2\right\} \tag{6.7}$$

where, x_i is the i^{th} observation and f_i is the corresponding frequency

Example 6.7: The frequency distributions of seed yield of 50 sesamum plants are given below. Find the standard deviation.

Table 6.10: A model frequency distributions of seed yield

Seed yield in gms (x)	3	4	5	6	7
Frequency (f)	4	6	15	15	10

Solution 6.7

Table 6.11: Calculation of standard deviation for frequency table

x_i	f_i	$f_i.x_i$	$f_i.x_i^2$
3	4	12	36
4	6	24	96
5	15	75	375
6	15	90	540
7	10	70	490
Total	**50**	**271**	**1537**

using Equation 6.7

$$\textit{sample variance},\ s^2 = \frac{1}{50-1}\left\{1537 - \frac{271^2}{50}\right\} = 1.3914$$

$$\textit{standard deviation},\ s = \sqrt{1.3914} = 1.179$$

For grouped frequency table

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} f_i d_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} f_i d_i\right)^2\right\} \tag{6.8}$$

where, f_i is the frequency of i^{th} class, $d_i = \frac{x_i - A}{c}$, where x_i is the class mark, A is the class mark with the highest frequency and c is the class interval.

Example 6.8: The frequency distributions of seed yield of 50 sesamum plants are given below. Find the standard deviation

Table 6.12: A model gropued frequency distributions of seed yield

Seed yield in gms (x)	2.5–3.5	3.5–4.5	4.5–5.5	5.5–6.5	6.5-7.5
Frequency (f)	4	6	15	15	10

Solution 6.8

Here $n = 50$; $c = 1$

Table 6.13: Calculation of sd for grouped frequency distribution

Seed yield	f_i	x_i	$d_i = \frac{x_i - A}{c}$	$f_i.d_i$	$f_i.d_i^2$
2.5–3.5	4	3	-2	-8	16
3.5–4.5	6	4	-1	-6	6
4.5–5.5	15	5	0	0	0
5.5–6.5	15	6	1	15	15
6.5–7.5	10	7	2	20	40
Total	**50**	**25**	**0**	**21**	**77**

$A = 5$

using Equation 6.8

$$\textit{sample variance},\ s^2 = \frac{1}{49}\left(77 - \frac{(21)^2}{50}\right) = 1.3914$$

$$\textit{standard deviation},\ s = \sqrt{1.3914} = 1.179$$

6.5.2 Merits and demerits of standard deviation

Merits

1. It is rigidly defined and its value is always definite and based on all the observations.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.
4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling and hence stable.
6. It is the basis for measuring the coefficient of correlation and other measures.

Demerits

1. It is not easy to understand and it is difficult to calculate.
2. It gives more weight to extreme values because the values are squared up.
3. As it is an absolute measure of variability, it cannot be used for the purpose of comparison.

6.6 Coefficient of variation

The standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of plants cannot be compared with the standard deviation of weights of the grains, as both are expressed in different units, *i.e.* heights in centimetre and weights in kilograms.

Therefore, the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the **coefficient of variation**. The coefficient of variation is obtained by dividing the standard deviation by the mean and expressed in percentage.

$$\text{Coefficient of variation}\,(C.V) = \frac{\text{standard deviation}}{\text{mean}} \times 100 \tag{6.9}$$

A higher C.V. indicates greater variability in the dataset, meaning the data values are more dispersed relative to the mean. In contrast, a lower C.V. signifies lower variability, indicating that the data values are more closely clustered around the mean. This measure is particularly useful when comparing datasets with different units or scales.

Example 6.9: Consider the measurement on yield and plant height of a paddy variety. The mean and standard deviation for yield are 50 kg and 10 kg respectively. The mean and standard deviation for plant height are 55 cm and 5 cm respectively. Compare the variability.

Solution 6.9

Here, the measurements for yield and plant height are in different units. Hence the variability can be compared only by using coefficient of variation.

For yield, CV = $\frac{10}{50} \times 100 = 20\%$

For plant height, CV = $\frac{5}{55} \times 100 = 9.1\%$

The yield is subject to more variation than the plant height.

Historical Insights

Exploring variability

The term "standard deviation" was first introduced in writing by Karl Pearson in 1894 in his paper "Contributions to the Mathematical Theory of Evolution." Prior to this, the concept was referred to by other names, including "mean error," "mean square error," and "error of mean square," reflecting its origins in the study of measurement errors and variability.(Pearson 1894)

The concept of variance was formalized in 1918 by Sir Ronald Aylmer Fisher in his seminal paper "The Correlation Between Relatives on the Supposition of Mendelian Inheritance." While earlier mathematicians like Carl Friedrich Gauss made significant contributions to the development of probability and error theory, which influenced the understanding of variability, the term "variance" as we know it today was introduced by Fisher.(Fisher 1918)

Quotes to Inspire

"The object of our being statistical is to learn how to improve the whole health of humanity" – Florence nightingale

7 Skewness and kurtosis

In the previous chapter, we explored numerical measures of central tendency and dispersion. Together, these measures give us insights into the location and spread of our data. However, they don't fully describe the data distribution. What about its shape?

The shape of a distribution helps us understand the symmetry, peakedness, and presence of tails in the data. While a histogram provides a visual summary of the shape, we often need numerical measures for precise analysis. These measures include:

- **Skewness**, which quantifies the degree of asymmetry in the data distribution.
- **Kurtosis**, which measures the "tailedness" or peakedness of the distribution.

In this chapter, we will have a detailed discussion on these two important measures of shape, understanding how they are calculated and interpreted. By the end, you'll be able to evaluate whether a distribution is symmetric, positively or negatively skewed, and whether it has light or heavy tails.

7.1 Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. Then you may ask, what will a symmetric distribution looks like. Histogram of a symmetric distribution is showed in Figure 7.1.

A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point. In our discussion we are including only unimodal cases.

> **!** Note
>
> For a symmetric distribution skewness = 0; mean = median = mode. Figure 7.2 shows how a symmetric distribution looks like.

Figure 7.3 shows a model data set with skewness = 0 (symmetric distribution)

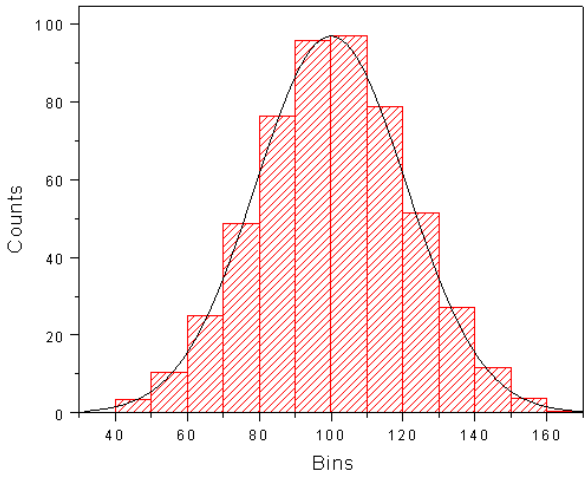


Figure 7.1: Histogram of a symmetric distribution

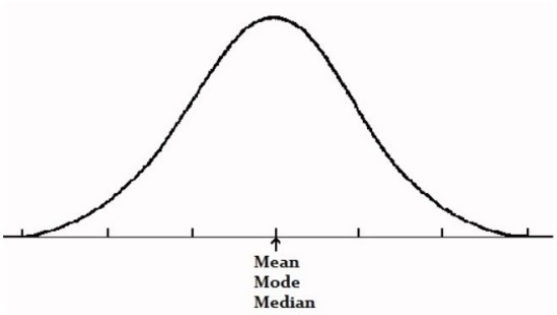


Figure 7.2: Symmetric distribution

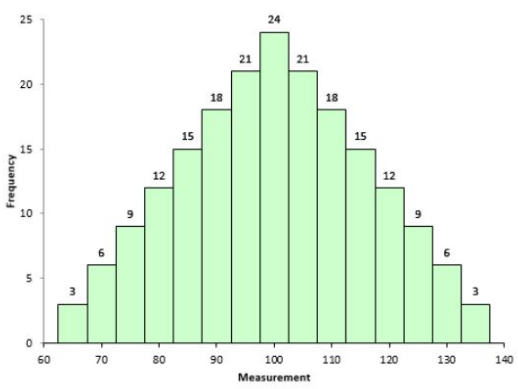


Figure 7.3: Data set with skewness = 0

7.1.1 Negatively skewed

A negatively skewed distribution, also known as a left-skewed distribution, is characterized by a longer tail on the left side of the distribution. The bulk of the data values, or the "mass" of the distribution, is concentrated on the right, as shown in Figure 7.4.

This type of distribution is referred to as left-skewed, left-tailed, or skewed to the left because of the extended left tail. In such cases, the numerical relationship between the mean, median, and mode typically follows this pattern:

Mean < Median < Mode

This occurs because the mean is pulled towards the longer tail, while the median and mode remain closer to the center of the data's bulk.

Figure 7.4: Left skewed or negatively skewed distribution

Figure 7.5 shows a model dataset with negative skewness.

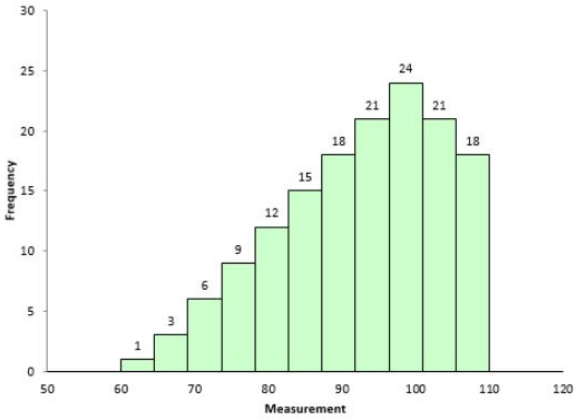


Figure 7.5: Negatively skewed data set

7.1.2 Positively skewed

A positively skewed distribution, also known as a right-skewed distribution, is characterized by a longer tail on the right side. The bulk of the data values, or the "mass" of the distribution, is concentrated on the left, as illustrated in Figure 7.6.

This type of distribution is referred to as right-skewed, right-tailed, or skewed to the right, due to the extended tail on the right. In such cases, the relationship between the mean, median, and mode typically follows this pattern:

Mean > Median > Mode

This occurs because the mean is influenced by the extreme values in the longer right tail, while the median and mode remain closer to the center of the data's bulk.

Figure 7.6: Right skewed or positively skewed distribution

Figure 7.7 shows a model dataset with positive skewness.

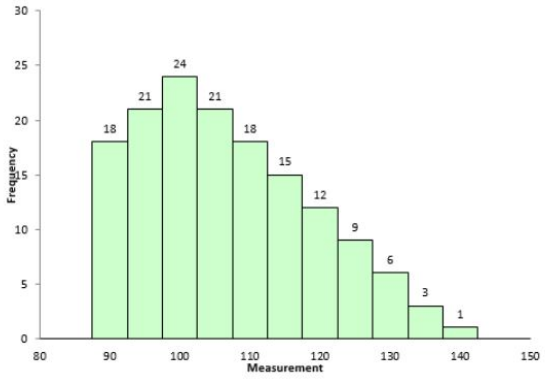


Figure 7.7: Data set with positive skewness or right skewed

7.2 Measures of skewness

The direction and extent of skewness can be measured in various ways. We shall discuss four measures.

7.2.1 Karl Pearson's coefficient of skewness (S_k)

You have noticed that the mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the divergence of mean from mode in a skewed distribution.

$$S_k = \frac{mean - mode}{\text{standard deviation}} \tag{7.1}$$

The sign of S_k gives the direction of skewness and its magnitude gives the extent of skewness. If $S_k > 0$, the distribution is positively skewed, and if $S_k < 0$ it is negatively skewed.

In Equation 7.1 since mode is used, there is a problem that if mode is not defined for a distribution we cannot find S_k. But empirical relation between mean, median and mode states that, for a moderately symmetrical distribution $mean - mode \approx 3(mean - median)$. So Equation 7.1 can be written as

$$S_k = \frac{3(mean - median)}{\text{standard deviation}} \tag{7.2}$$

Example 7.1: Compute the Karl Pearson's coefficient of skewness from the following data:

Table 7.1: Model dataset for skewness calculation

Height (x)	frequency (f)
58	10
59	18
60	30
61	42
62	35
63	28
64	16
65	8

Solution 7.1

Mean, $\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{11482}{187} = 61.40$

Table 7.2: Karl Pearson's coefficient of skewness

Height (x_i)	frequency (f_i)	$f_i x_i$	$f_i x_i^2$
58	10	580	33640
59	18	1062	62658
60	30	1800	108000
61	42	2562	156282
62	35	2170	134540
63	28	1764	111132
64	16	1024	65536
65	8	520	33800
Sum	187	11482	705588

Sample variance, s^2 using Equation 6.7 $= \frac{705588 - \frac{(11482)^2}{187}}{186} = 3.123$

Standard deviation, $s = \sqrt{3.123} = 1.76$

To calculate the median, refer to the Table 7.3. Locate the cumulative frequency just greater than $\frac{n+1}{2}$, and the corresponding value of x will be the median (Q_2).

Here, $\frac{n+1}{2} = \frac{187+1}{2} = \frac{188}{2} = 94$.
From the Table 7.3, it is evident that the median is 61.

Table 7.3: Cumulative frequency for skewness calculation

Height (x_i)	frequency (f_i)	cumulative frequency
58	10	10
59	18	28
60	30	58
61	42	100
62	35	135
63	28	163
64	16	179
65	8	187

using Equation 7.2

$$S_k = \frac{3(61.40 - 61)}{1.76} = \frac{1.2}{1.76} = 0.68$$

Hence, the Karl Pearson's coefficient of skewness $S_k = 0.68$, Thus the distribution is positively skewed.

7.2.2 Bowley's measure of skewness (S_Q)

Karl Pearson's coefficient of skewness is most commonly used skewness measure. However, in order to use it you must know the mean, mode (or median) and standard deviation for your data. Sometimes you might not have that information; instead you might have information about quartiles. If that's the case, you can use Bowley's measure of skewness as an alternative to find out more about the asymmetry of your distribution. It's very useful if you have extreme data values (outliers) or if you have an open-ended distribution.

$$\textit{Bowley's measure of Skewness, } S_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \tag{7.3}$$

where, $Q_1 = 1^{\text{st}}$ quartile; $Q_2 =$ median; $Q_3 = 3^{\text{rd}}$ quartile

Equation can be further modified into

$$S_Q = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \tag{7.4}$$

- $S_Q = 0$ means that the curve is symmetrical.
- $S_Q > 0$ means the curve is positively skewed.
- $S_Q < 0$ means the curve is negatively skewed.

Lets find Bowley's measure of skewness for Table 7.1 in Example 7.1 from the cumulative frequency in Table 7.3, quartiles can be calculated. Calculation of Q_1, Q_2, Q_3 is given in Section 5.6.

$$Q_1 = 60$$

$$Q_2 = 61$$

$$Q_3 = 63$$

$$S_Q = \frac{63 - (2 \times 61) + 60}{63 - 60} = \frac{1}{3} = 0.33$$

Since $S_Q > 0$ means the curve is positively skewed.

7.2.3 Kelly's measure of skewness (S_p)

Bowley's measure of skewness is based on the middle 50% of the observations; it leaves 25% of the observations on each extreme of the distribution. As an improvement over Bowley's measure, Kelly has suggested a measure based on Percentiles, including P_{10} and P_{90} so that only 10% of the observations on each extreme are ignored.

$$Kelly's\ Measure\ of\ Skewness,\ S_p = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})} \tag{7.5}$$

> **Try yourself**
> Try to find Kelly's measure of skewness for Table 7.1

7.2.4 Measure based on moments

Before going into measuring skewness using moments, one should know what a moment is:

Moments

The r^{th} **moment about mean** of a distribution, denoted by μ_r is given by

$$\mu_r = \frac{\sum_{i=1}^{N} f_i \left(x_i - \overline{x}\right)^r}{N} \tag{7.6}$$

where, f_i is the frequency of i^{th} observation or class mark x_i, $N = \sum f_i$, number of observations

Moment about mean is also called as **central moment**.

If $r = 0$, $\mu_0 = \frac{\sum_{i=1}^{N} f_i (x_i - \overline{x})^0}{N} = 1$

If $r = 1$, $\mu_1 = \frac{\sum_{i=1}^{N} f_i (x_i - \overline{x})^1}{N} = 0$ (sum of deviation about mean is zero)

If $r = 2$, $\mu_2 = \frac{\sum_{i=1}^{N} f_i (x_i - \overline{x})^2}{N} = \sigma^2$, Population variance

> **Note**
>
> It should be remembered that first moment about mean is 0 and second moment about mean is **variance**.

For Table 7.1 in Example 7.1 given above, calculate third central moment, μ_3

Mean = 61.40

Table 7.4: Third central moment calculation

Height (x_i)	frequency (f_i)	$(x_i - \overline{x})^3$	$f_i (x_i - \overline{x})^3$
58	10	-39.304	-393.040
59	18	-13.824	-248.832
60	30	-2.744	-82.320
61	42	-0.064	-2.688
62	35	0.216	7.560
63	28	4.096	114.688
64	16	17.576	281.216
65	8	46.656	373.248
Sum	187	12.608	49.832

$$\mu_3 = \frac{\sum_{i=1}^{N} f_i (x_i - \overline{x})^3}{N} = \frac{49.832}{187} = 0.266$$

In short values of following moments about mean are

Table 7.5: Moments about mean

Moments about mean	Value
μ_0	1
μ_1	0
μ_2	σ^2

7.2.4.1 Beta one and gamma one

The moment measure of skewness is based on the property that, for a symmetrical distribution, all odd ordered central moments are equal to zero. We note that $\mu_1 = 0$, for every distribution, therefore, the lowest order moment that can provide an absolute measure of skewness is μ_3. So measures of skewness are based on μ_3.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \tag{7.7}$$

Pronounced as 'beta one'.

β_1= 0 means that the curve is symmetrical. The greater the value of β_1 the more skewed the distribution. One serious limitation of β_1 is that it cannot tell the direction of skewness

i.e. whether it is positive or negative. Since μ_2 is always positive (as it is variance) and μ_3^2 is positive, β_1 will be positive always. This drawback is removed by calculating $_1$, called as Karl Pearson's $_1$, pronounced as 'gamma one'.

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} \tag{7.8}$$

If μ_3 is positive γ_1 is positive, If μ_3 is negative γ_1 is negative

- γ_1= 0 means that the curve is symmetrical.
- $\gamma_1 > 0$ means the curve is positively skewed.
- $\gamma_1 < 0$ means the curve is negatively skewed.

For Table 7.1 in Example 7.1, β_1 and γ_1 can be calculated as follows

$\mu_3 = 0.226$

$\mu_2 = 3.123$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.226)^2}{(3.123)^3} = \frac{0.051}{30.46} = 0.0016$

$\gamma_1 = \sqrt{\beta_1} = \sqrt{0.0016} = +0.04$

Since μ_3 is positive $\gamma_1$is positive. Since γ_1 is slightly greater than 0, distribution is a slightly skewed to right.

7.3 Kurtosis

Kurtosis is a statistical measure that describes the shape of a distribution's frequency curve, focusing on its relative peakedness. While skewness measures the asymmetry or lack of symmetry in a distribution, kurtosis evaluates how sharp or flat the peak of the curve is. There are three categories of frequency curves depending upon the shape of their peak as shown in Figure 7.8.

Kurtosis refers to degree of flatness or peakedness of the curve. It is measured relative to the peakedness of normal curve. The normal curve is considered as *mesokurtic*. If a curve is more peaked than normal curve, it is called *leptokurtic*. If a curve is more flat-topped than normal curve, it is called *platykurtic*. The condition of peakedness (leptokurtic) or flatness (platykurtic) is called **kurtosis of excess**.

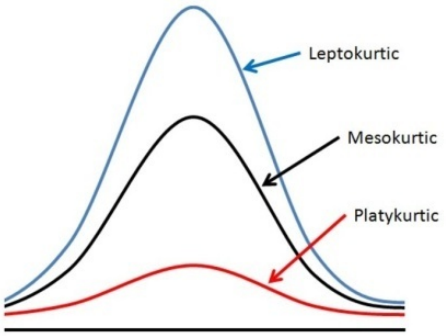


Figure 7.8: Three categories of frequency curves

7.3.1 Measure of kurtosis

Kurtosis is measured using β_2 'beta two' and γ_2 'gamma two' given by Karl Pearson

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \tag{7.9}$$

where, μ_4 is the 4^{th} central moment, μ_2 is the 2^{nd} central moment

- $\beta_2 = 3$ means that the curve is mesokurtic.
- $\beta_2 > 3$ means the curve is leptokurtic.
- $\beta_2 < 3$ means the curve is platykurtic.

$$\gamma_2 = \beta_2 - 3 \tag{7.10}$$

- $\gamma_2 = 0$ means that the curve is mesokurtic.
- $\gamma_2 > 0$ means the curve is leptokurtic.
- $\gamma_2 < 0$ means the curve is platykurtic.

For Table 7.1 in Example 7.1, kurtosis can be examined as follows

Mean, $\overline{x} = 61.40$

$\mu_2 = 3.123$ (calculation shown in previous example)

$\mu_4 = \frac{\sum_{i=1}^{N} f_i(x_i - \overline{x})^4}{N} = \frac{4312.747}{187} = 23.062$

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{23.062}{(3.123)^2} = 2.364$

Table 7.6: Measures of kurtosis

Height (x_i)	frequency (f_i)	$(x_i - \overline{x})^4$	$f_i\,(x_i - \overline{x})^4$
58	10	133.634	1336.336
59	18	33.178	597.197
60	30	3.842	115.248
61	42	0.026	1.075
62	35	0.130	4.536
63	28	6.554	183.501
64	16	45.698	731.162
65	8	167.962	1343.693
Sum	187	391.021	4312.747

β_2 is 2.364, which is close to 3, distribution can be considered slightly platykurtic close to symmetric.

You can verify the frequency curve of Example 7.1 Figure 7.9, it can be seen that it is slightly right tailed (positively skewed).

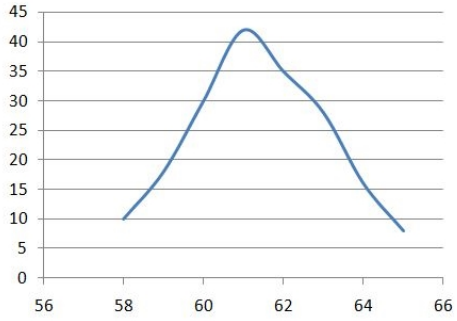


Figure 7.9: Frequency curve of Example 7.1

Historical Insights

"Crabs and kurtosis"
The story of **kurtosis** and **skewness** begins with a fascinating scientific journey involving **crabs**! In the late 1800s, Karl Pearson, a pioneering statistician, worked with biologist Walter Weldon to study variations in the size of crustaceans, like crabs. They noticed that the data didn't follow the usual normal pattern, so Pearson developed new tools to

better understand the shapes of these unusual data distributions.
He created the concept of **skewness** to measure whether the data was symmetrical or had long tails on one side. Then, he developed **kurtosis**, a measure of how "peaked" or "flat" the data distribution was compared to the normal curve. These ideas helped statisticians better analyze data that didn't fit the typical patterns, paving the way for modern statistical tools we still use today! (Fiori and Zenga 2009)

Quotes to Inspire

"We are just statistics, born to consume resources" – Horace

8 Measures of association

In the previous chapters, we examined measures of central tendency, which summarize the location of data, and measures of dispersion, which describe the spread around that location. Together, these tools provided us with a foundation for understanding and summarizing a single dataset. However, in many real-world scenarios, we are interested in understanding the relationships between variables rather than focusing on one variable in isolation.

Measures of association allow us to explore and quantify the connections between two or more variables. For example, does fertilizer use influence crop yield? Is there a relationship between farm size and agricultural income? Do education levels impact the adoption of new farming technologies? By studying association, we can answer such questions and uncover meaningful patterns in data.

In this chapter, we will introduce key measures of association, such as **covariance**, **correlation coefficients**, and other related metrics. We will discuss how these measures are calculated, interpreted, and applied to analyze relationships between variables. By the end of this chapter, you will be equipped to assess both the strength and direction of relationships, providing deeper insights into agricultural and social science research.

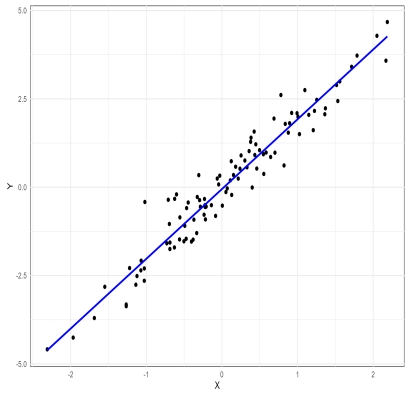
8.1 Linear and monotonic relationship

Linear relationship

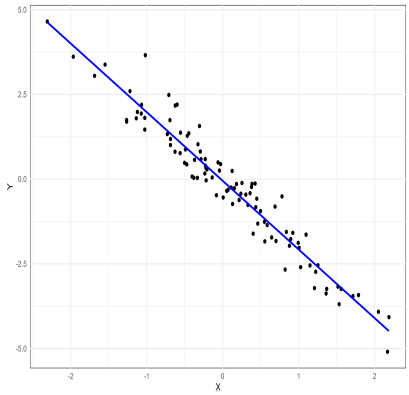
A linear relationship (or linear association) is a statistical term used to describe a straight-line relationship between variables. Linear relationships can be expressed either in a graphical format where the variable plotted on X-Y plane gives a straight line or relation between two variables (consider x and y) can be expressed with an equation of a straight line ($y = a + bx$) (will be more clear when we discuss regression in Chapter 9).

Monotonic relation

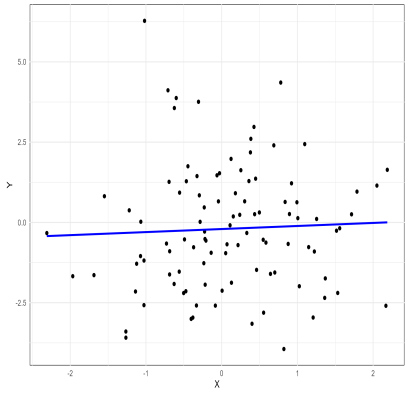
A monotonic relationship between two variables means that as one variable increases or decreases, the other tends to move in the same direction, but not necessarily at a constant rate. In contrast, a linear relationship implies that the variables move in the same direction at a constant rate. While all linear relationships are monotonic, not all monotonic relationships are linear. Refer to the illustration Figure 8.1 for a clearer distinction.



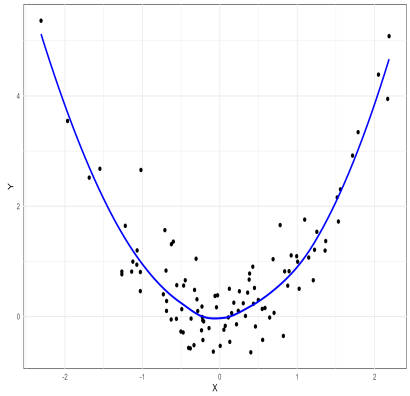
(a) Strong positive linear relationship



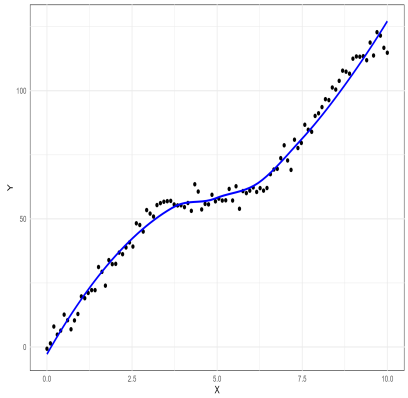
(b) Strong negative linear relationship



(c) Weak linear relationship



(d) Non-linear relationship



(e) Monotonic relationship

Figure 8.1: Linear and monotonic relationship

8.2 Scatter diagram

Consider two variables x and y, a scatter diagram is used to visually investigate whether there is any relationship between them. This graphical method helps explore whether an association exists between the two variables. If the variables x and y are plotted along the X-axis and Y-axis respectively in the X-Y plane of a graph sheet the resultant diagram of dots is known as **scatter diagram**. From the scatter diagram we can say whether there is any association between x and y. Figure 8.2 gives the scatter diagram of Example 8.1

Example 8.1: Consider the data on sepal length (x) and sepal width (y) of *Iris setosa.*

Table 8.1: Data on sepal length (x) and sepal width (y) of *Iris setosa.*

Sepal length (x)	Sepal width (y)
5.1	3.5
4.9	3
4.7	3.2
4.6	3.1
5	3.6
7	3.2
6.4	3.2
6.9	3.1
5.5	2.3
6.5	2.8
6.3	3.3
5.8	2.7
7.1	3
6.3	2.9
6.5	3

Example 8.2: A research station investigates the relationship between the average daily soil moisture content (as influenced by irrigation levels, in percentage) and the corresponding monetary yield (in Rs.) of a crop. The data collected over different periods are as follows:

Table 8.2: Model data on crop yield influenced by soil moisture

Soil moisture (%)	Crop yield (in Rs./cent)
14.2	215
16.4	325
11.9	185
15.2	332

Soil moisture (%)	Crop yield (in Rs./cent)
18.5	406
22.1	522
19.4	412
25.1	614
23.4	544
18.1	421
22.6	445
17.2	408

Scatter diagram for the Example 8.2 is given in Figure 8.3. You can see a linear association between the two variables *i.e.* between soil moisture and crop yield in rupees. It can be shown using a line as in Figure 8.4. It is clear that as soil moisture percentage increases crop yield in rupees increases, indicating a positive correlation.

From the examples above it is clear that scatter diagram gives an idea on linear association between variables, so it can also used as a graphical tool to see whether there any association is present or not. But we cannot quantify the association using scatter diagram.

8.3 Correlation

Correlation is a statistical technique used to examine the relationship between two or more variables. It quantifies the degree and strength of the *linear association* between two variables, expressed as a single numerical value.

This measure allows us to summarize the extent to which changes in one variable are associated with changes in another. When two or more quantities vary in a related manner, such that movements in one variable are consistently accompanied by movements in the other, the variables are said to be correlated. Based on the nature of relationship, correlation can be classified into three categories *positive correlation*, *negative correlation* and *no correlation*.

Positive correlation

Positive correlation refers to a relationship between two variables in which both move in the same direction. A positive correlation exists when an increase in one variable is accompanied by an increase in the other, or when a decrease in one variable corresponds with a decrease in the other.

Examples of positive correlation:

1. *Fertilizer use and crop yield*: The more fertilizer (x) a farmer applies to a field, the higher the crop yield (y). Here, as x increases, y also increases.

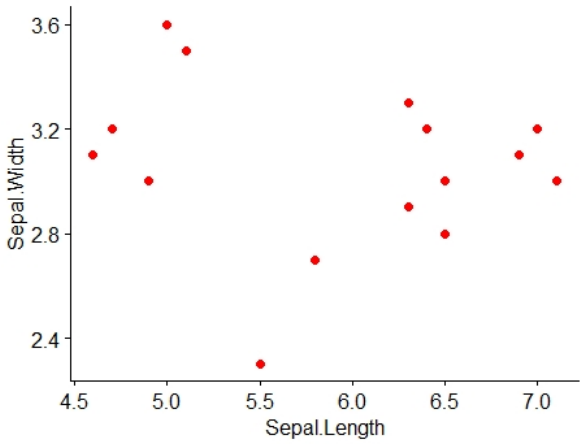


Figure 8.2: Scatter diagram of data in Example 8.1

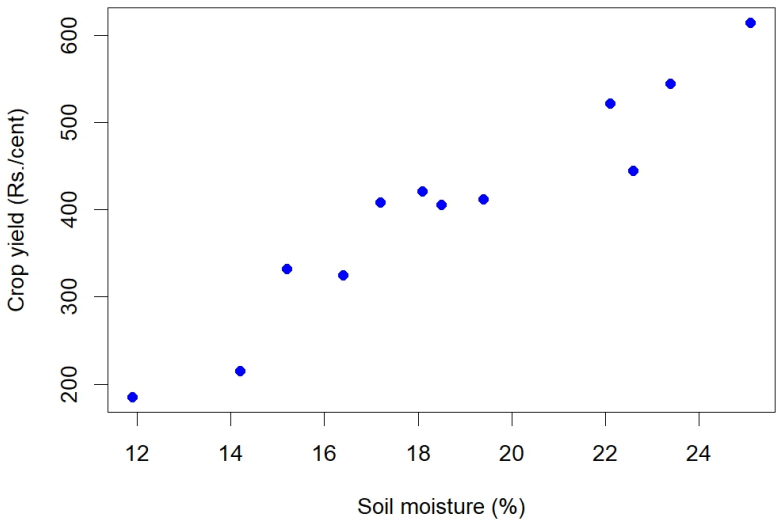


Figure 8.3: Scatter diagram of data in Example 8.2

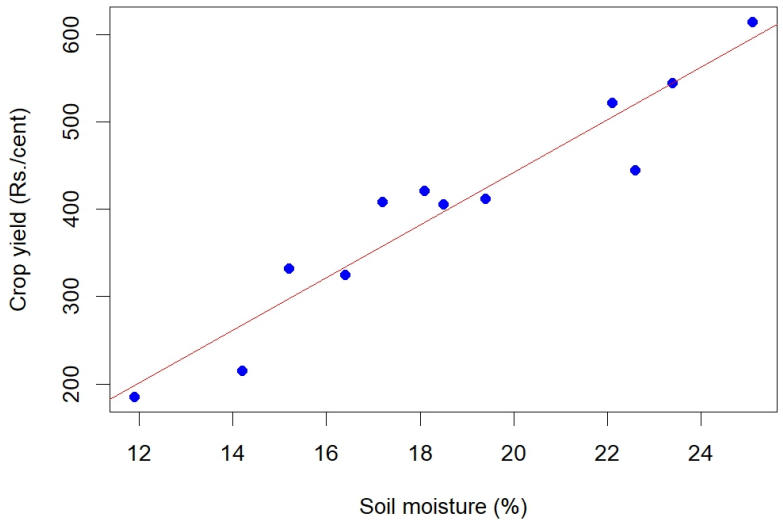


Figure 8.4: Linear relationship between variables

2. *Rainfall and crop growth*: Increased rainfall (x) often leads to better crop growth (y).
3. *Farm size and agricultural output*: Larger farm sizes (x) are associated with greater total agricultural output (y).
4. *Labor hours and harvest quantity*: The more hours spent harvesting (x), the higher the quantity of crops harvested (y).

Negative correlation

Negative correlation refers to a relationship between two variables in which one variable increases as the other decreases, and vice versa.

Examples of negative correlation:

1. *Weed density and crop yield*: As the density of weeds in a field (x) increases, the crop yield (y) decreases. Here, as x increases, y decreases.
2. *Soil salinity and plant growth*: Higher soil salinity levels (x) result in reduced plant growth (y).
3. *Age of livestock and milk production*: As a cow's age (x) increases, the amount of milk it produces (y) decreases. Here, as x increases, y decreases.

4. *Pesticide application and pest population*: As the amount of pesticide applied (x) increases, the pest population (y) decreases. Here, x increases while y decreases.

No correlation refers to a statistical relationship where two variables exhibit no apparent association with each other. In other words, changes in one variable do not systematically correspond to changes in the other.

Direction of correlation can be identified using a scatter diagram as shown below in Figure 8.5

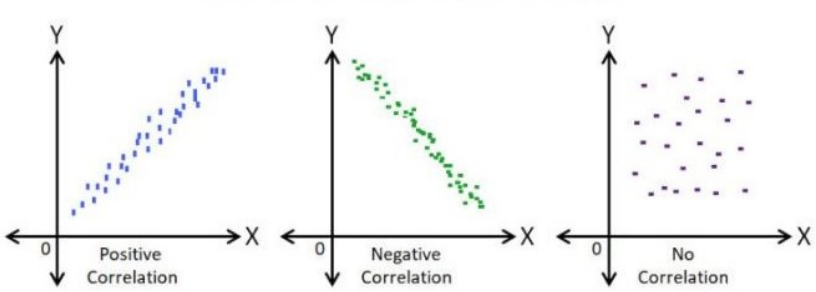


Figure 8.5: Scatter plot and nature of relationship

8.4 Correlation types

Simple and multiple

In simple correlation the relationship is confined to two variables only. In multiple correlation the relationship between more than two variables is studied.

Linear and non-linear correlation

Linear correlation: A type of correlation in which the relationship between two variables can be represented by a straight line. In this case, a change in one variable corresponds to a proportional change in the other, either in a positive or negative direction.

Non-linear correlation: A type of correlation where the relationship between two variables cannot be represented by a straight line. Instead, the relationship follows a curved pattern, indicating that the variables do not change at a constant rate relative to each other. This is also referred to as *curvilinear correlation*.

Partial and total correlation

Partial correlation: In multiple correlation analysis, partial correlation examines the relationship between two variables after controlling for or eliminating the linear effect of other correlated variables.

Total correlation: Total correlation considers the relationship between variables based on all relevant variables without controlling for the influence of any specific variable.

Partial and multiple correlation are discussed in detail in Section 8.8.

> **!** Note
>
> **Perfect Correlation**: If there is any change in the value of one variable, the value of the other variable is changed in a fixed proportion then the correlation between them is said to be in perfect correlation. If there is a perfect correlation, the points will lie in the straight line. If there was a perfect correlation the data will look like in Figure 8.6 below.

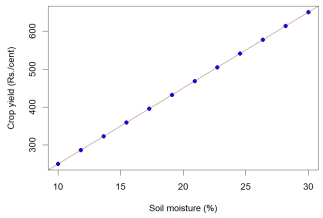


Figure 8.6: Perfect correlation

8.5 Measuring correlation

While a scatter diagram provides a visual representation to examine whether there is an association between two variables, it does not give a precise measure of the strength or direction of the relationship. To understand the degree and nature of the correlation more quantitatively, we use numerical measures. In this section, we discuss various methods to quantify correlation, enabling a more comprehensive and objective analysis of the relationship between variables.

8.5.1 Karl Pearson's coefficient of correlation

It is the most important and widely used measure of correlation. A measure of the intensity or degree of *linear relationship* between two variables is developed by Karl Pearson, a British Biometrician - known as the **Pearson's correlation coefficient** denoted by 'r' which is expressed as the ratio of the **covariance** to the product of the standard deviations of the two variables.

Covariance

Covariance is a measure of the joint linear variability of the two variables. Covariance of two variables x and y is denoted as $cov(x, y)$. Covariance measure is used to find correlation coefficient. Consider two variables x and y with n observations each, then covariance is given by Equation 8.1

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y}) \tag{8.1}$$

When covariance = 0 there is no joint variability or there is no linear relationship. The unit of covariance is the product of the units of the two variables.

Pearson's correlation coefficient

Assumptions of Pearson's correlation coefficient

To ensure the validity of Pearson's correlation coefficient, the following assumptions must be met:

1. **Linearity**: The relationship between the two variables must be linear. Pearson's correlation only measures the strength of linear associations, so it may not accurately describe relationships that are non-linear.
2. **Continuous data**: Both variables should be measured on continuous scales (interval or ratio).
3. **Normality**: Both variables should be approximately normally distributed, especially for small sample sizes. This assumption is less critical for large sample sizes due to the Central Limit Theorem.
4. **Homoscedasticity**: The variability in one variable should remain constant across the range of the other variable. In other words, the scatter of points around the regression line should be uniform.
5. **Independence**: Observations in the data should be independent of each other.
6. **No significant outliers**: Outliers can have a disproportionate effect on Pearson's correlation, potentially distorting the results. An *outlier* is an observation in a dataset that is significantly different from other observations. It deviates markedly from the overall pattern of the data and may occur due to variability in the data, measurement errors, or rare events.

Violations of these assumptions may lead to incorrect or misleading interpretations of the correlation coefficient. In such cases, alternative methods like Spearman's correlation discussed in Section 8.5.2 may be more appropriate.

The Pearson's correlation coefficient between the two variables (x and y) is calculated using Equation 8.2. Equation 8.2 can be also written as in Equation 8.3

$$r = \frac{cov(x, y)}{sd(x)sd(y)} \tag{8.2}$$

where sd is the standard deviation.

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{8.3}$$

> **!** Note
>
> **Product-moment correlation**
> Karl Pearson's correlation coefficient (r) is also called the **product-moment correlation** coefficient because it is calculated using the product of the deviations of the two variables from their respective means. In other words, the Equation 8.3 involves multiplying the deviations of each data point from the mean of its variable and then averaging those products. The term product-moment refers to the multiplication (product) of the deviations (moments) of the variables from their means.

A simplified formula for by hand computation of correlation coefficient can be derived by modifying Equation 8.3

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \tag{8.4}$$

Properties of the correlation coefficient (r)

1. It is a pure number independent of both origin and scale of the units of the observations.
2. It always lies between -1 and $+1$ (absolute value cannot exceed unity).*i.e.* $-1 \quad r \quad +1$
3. $r = +1$, indicates perfect positive correlation. $r = -1$, indicates perfect negative correlation. $r = 0$, indicates no correlation.
4. When the correlation is zero then there is no linear relationship between the variables.
5. Karl Pearson's correlation coefficient is also called as *product-moment correlation* coefficient.

6. Two independent variables are always uncorrelated, meaning that there is no relationship between them. However, the reverse is not always true. Just because two variables are uncorrelated doesn't mean they are independent. Uncorrelated variables may still have a relationship, but the relationship might be non-linear, which correlation cannot capture. Independence implies no relationship in any form (linear or non-linear), but uncorrelated means there's no linear relationship between them.

> **!** Note
>
> **Spurious correlation**
> When we calculate the correlation between two variables, we often get a numerical value that quantifies the strength and direction of the relationship between them. However, if there is no actual meaningful relationship between these variables, the correlation value obtained may be misleading. For example, even if there's no practical or causal link, we might find a correlation between variables like "fertilizer price" and "Kohli's batting average." Despite the absence of any real-world connection between these two variables, we can still compute a correlation value. This type of correlation is known as a spurious correlation, which means it exists purely due to random chance or statistical artifacts rather than any practical or causal relationship.

Example 8.3: Consider the Table 8.2 in Example 8.2; find correlation coefficient (r)

Table 8.3: Calculation of Pearson's correlation coefficient for Table 8.2

Sl. No.	Moisture (x_i)	Crop yield (y_i)	$x_i - \overline{x}(1)$	$y_i - \overline{y}(2)$	(1).(2)	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	14.2	215	-4.48	-187.42	838.69	20.03	35125.01
2	16.4	325	-2.28	-77.42	176.12	5.18	5993.34
3	11.9	185	-6.78	-217.42	1473	45.90	47270.01
4	15.2	332	-3.48	-70.42	244.70	12.08	4958.51
5	18.5	406	-0.18	3.58	-0.63	0.031	12.84
6	22.1	522	3.43	119.58	409.57	11.73	14300.17
7	19.4	412	0.73	9.58	6.9479	0.53	91.84
8	25.1	614	6.43	211.58	1359.42	41.28	44767.51
9	23.4	544	4.73	141.58	668.98	22.33	20045.84
10	18.1	421	-0.58	18.58	-10.69	0.33	345.34
11	22.6	445	3.93	42.58	167.14	15.41	1813.34
12	17.2	408	-1.48	5.58	-8.24	2.16	31.17
Total	**224.1**	**4829**	0	0	**5325.03**	**176.98**	**174754.9**

n =12

$$mean, \overline{x} = \frac{224.1}{12} = 18.68$$

$$mean, \overline{y} = \frac{4829}{12} = 402.42$$

$cov\ (x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

$\sum_{i=1}^{12}(x_i - \overline{x})(y_i - \overline{y}) = 5325.03$

$Cov\ (x,y) = \frac{5325.03}{12} = 443.75$

$$Standard\ deviation,\ S.D\,(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{176.98}{12}} = 3.84$$

$$Standard\ deviation,\ S.D\,(y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{174754.9}{12}} = 120.68$$

Using Equation 8.2, $r = \frac{443.75}{3.84 \times 120.68} = 0.96$, which indicates a strong positive correlation

8.5.2 Spearman's rank order correlation coefficient

The Spearman's correlation coefficient evaluates the strength and direction of a monotonic relationship between two variables, whether they are continuous or ordinal. Spearman's correlation is denoted by a Greek letter ρ pronounced as "*rho*". The range of Spearman's rank correlation also lies between -1 and $+1$ always, *i.e.* $-1 \qquad +1$

Unlike Pearson's correlation, which measures linear relationships, Spearman's correlation is based on the ranked values of the variables rather than their raw data. This makes it particularly useful in situations where:

- The relationship between variables is non-linear but monotonic.
- The data contains outliers or is not normally distributed, as ranking reduces the influence of extreme values.
- Variables are measured on an ordinal, interval, or ratio scale.

Spearman's correlation is a more robust alternative to Pearson's when the assumptions of Pearson's correlation coefficient are violated.

There are two cases in calculating ρ :

1. No tied rank case

2. Tied rank case

No tied rank case

When two or more distinct observations have the same value, thus being given the same rank, they are said to be tied. The formula for the Spearman rank correlation coefficient when there are *no tied ranks* is:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \tag{8.5}$$

where d_i is the difference between ranks of i^{th} pair of observation

Example 8.4: Calculation of Spearman's rank correlation when there is no tied rank is explained step by step by using the simple example below

The scores for nine students in physics and mathematics are as follows:

Physics: 35, 23, 47, 17, 10, 43, 9, 6, 28

Mathematics: 30, 33, 45, 23, 8, 49, 12, 4, 31

Compute the student's ranks in the two subjects and compute the Spearman rank correlation.

Table 8.4: Non tied rank case example dataset

Physics	Mathematics
35	30
23	33
47	45
17	23
10	8
43	49
9	12
6	4
28	31

Step 1: Find the ranks for each individual subject. Rank the scores from greatest to smallest; assign the rank 1 to the highest score, 2 to the next highest and so on:

Table 8.5: Non tied rank case calculation table1

Physics (x)	Rank$_x$	Mathematics (y)	Rank$_y$
35	3	30	5
23	5	33	3
47	1	45	2
17	6	23	6
10	7	8	8
43	2	49	1
9	8	12	7
6	9	4	9
28	4	31	4

Step 2: Add a column d, to your data. The d is the difference between ranks.

d = Rank$_x$ – Rank$_y$

For example, the first student's physics rank is 3 and math rank is 5, so the difference is -2. In the next column, square your d values.

Table 8.6: Non tied rank case calculation table2

Physics (x)	Rank$_x$	Mathematics (y)	Rank$_y$	d	d^2
35	3	30	5	-2	4
23	5	33	3	2	4
47	1	45	2	-1	1
17	6	23	6	0	0
10	7	8	8	-1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0
Total					12

Step 4: Sum (add up) all of your d^2 values. $\sum_{i=1}^{n} d_i^2 = 4 + 4 + 1 + 0 + 1 + 1 + 1 + 0 + 0 = 12$.

Step 5: Insert the values into Equation 8.5

$$\rho = 1 - \frac{6 \times 12}{9\,(81 - 1)} = 0.90$$

The Spearman's rank correlation for this set of data is 0.90. This indicates there is a high correlation between the marks of physics and mathematics in the sample.

Tied rank case

When two or more data points have the same value, same ranks were given to these data points and a tied rank case occurs. When there are tied ranks, the formula for calculating ρ is given below

$$\rho = 1 - \frac{6\left(\sum_{i=1}^{n} d_i^2 + T_x + T_y\right)}{n\left(n^2 - 1\right)} \tag{8.6}$$

If there are m individuals tied (having same rank), and s such sets of ranks are there in X-series then,

$$T_x = \frac{1}{12}\sum_{i=1}^{s} m_i\left(m_i^2 - 1\right) \tag{8.7}$$

If there are w individuals tied (having same rank), and s' such sets of ranks are there in Y-series then,

$$T_y = \frac{1}{12}\sum_{i=1}^{s'} w_i\left(w_i^2 - 1\right) \tag{8.8}$$

Calculation of Spearman's rank correlation when there is tied rank is explained step by step by using the example below

Example 8.5: The scores for nine students in physics and mathematics are as follows:

Table 8.7: Tied rank case example dataset

Physics (x)	Mathematics (y)
35	30
23	33
47	45
23	23
10	8
43	49
9	12
6	33
28	33

Step 1: Consider the marks in Physics, ranked as usual without considering the repeated value. Here you can see 23 is repeated but first value is given rank 5 and second repeated value is given the next rank 6.

Table 8.8: Tied rank case calculation table-1

Physics (x)	Rank
35	3
23	5
47	1
23	6
10	7
43	2
9	8
6	9
28	4

Then the average of two ranks 5 and 6 is assigned to both the values; $\left(\frac{5+6}{2}\right) = 5.5$

Table 8.9: Tied rank case calculation table-2

Physics (x)	Rank
35	3
23	5.5
47	1
23	5.5
10	7
43	2
9	8
6	9
28	4

Similarly for marks in mathematics you can see 33 is repeated thrice.

Table 8.10: Tied rank case calculation table-3

Mathematics (y)	Rank
30	6
33	3
45	2
23	7
8	9
49	1
12	8

Mathematics (y)	Rank
33	4
33	5

You can see the value 33 is repeated thrice, so the average of three ranks 3, 4 and 5 is given $\left(\frac{3+4+5}{3}\right) = 4$

Table 8.11: Tied rank case calculation table-4

Mathematics (y)	Rank
30	6
33	4
45	2
23	7
8	9
49	1
12	8
33	4
33	4

Step 2: Calculate T_x and T_y

In our example marks in Physics (x) there are two 23 values tied therefore $m = 2$; since only one such a set is there $s = 1$. Applying these values in Equation 8.7 $T_x = \frac{1}{12}(2 \times (2^2 - 1) = 0.5$

In our example marks in Mathematics (y) there are three 33 values tied therefore $w = 3$; since only one such a set is there $s = 1$. Applying these values in Equation 8.8 $T_y = \frac{1}{12}(3 \times (3^2 - 1) = 2$

Step 2: Calculate d and then use the Equation 8.6

Table 8.12: Tied rank case calculation table-5

Physics (x)	Rank	Mathematics (y)	Rank	d	d^2
35	3	30	6	-3	9
23	5.5	33	4	1.5	2.25
47	1	45	2	-1	1
23	5.5	23	7	-1.5	2.25
10	7	8	9	-2	4
43	2	49	1	1	1

Physics (x)	Rank	Mathematics (y)	Rank	d	d^2
9	8	12	8	0	0
6	9	33	4	5	25
28	4	33	4	0	0
Total				0	44.5

using Equation 8.6, $\rho = 1 - \frac{6\times(44.5+0.5+2)}{9(9^2-1)} = 1 - \frac{282}{720} = 0.61$

A Spearman rank correlation of 0.61 indicates a moderately strong positive monotonic relationship between the two variables.

> **!** Note
>
> Spearman's rank correlation is a non-parametric method, meaning it does not rely on assumptions about the data's distribution and is suitable for ordinal data or non-linear relationships. In contrast, Karl Pearson's correlation is a parametric method that assumes the data follows a normal distribution and is used for continuous variables with a linear relationship. Parametric methods are more sensitive to outliers and deviations from assumptions, while non-parametric methods are more robust and versatile for different types of data.

8.5.3 Kendall's Rank Correlation Coefficient

Kendall's rank correlation coefficient, also known as Kendall's τ or the coefficient of concordance, is a measure of association that evaluates the degree of agreement between ranked variables. It ranges from 0 to 1 *i.e.* $0 \leq \tau \leq 1$, where $\tau = 0$: indicates no agreement. $\tau = 1$: indicates perfect agreement.

When there are multiple sets of rankings (k sets), Kendall's coefficient of concordance (τ) can be used to assess the association among them. This measure is particularly useful in evaluating the reliability or consistency of scores assigned by multiple judges or raters. By quantifying the agreement across different rankings, it provides a robust way to analyze consensus in subjective evaluations or repeated assessments.

To calculate Kendall's coefficient of concordance (τ), the data is organized into a table where each row represents the ranks assigned by a different judge to the same set of n objects. Each column in the table corresponds to the ranks given by a judge to n objects.

For k judges, there will be k sets of rankings for each object. The table should look like this:

Table 8.13: Model table for the calculation of Kendall's τ

Object	Judge 1	Judge 2	...	Judge k
1	R_{11}	R_{12}	...	R_{1k}
2	R_{21}	R_{22}	...	R_{2k}
...	...	...	...	...
n	R_{n1}	R_{n2}	...	R_{nk}

In Table 8.13 each row represents the ranks assigned to an object by all (k) judges. Each column represents the ranks for each object assigned by a particular judge. R_{ij} is the rank assigned to the i^{th} object by the j^{th} object, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., k$.

Kendall's coefficient of concordance (τ) is calculated using Equation 8.9.

$$\tau = \frac{12\left[\sum_{i=1}^{n} R_i^2 - \frac{\left(\sum_{i=1}^{n} R_i\right)^2}{n}\right]}{k^2 n\left(n^2-1\right)} \tag{8.9}$$

where $R_i = \sum_{j=1}^{k} R_{ij}$, *i.e.* the sum of ranks obtained by each object. Calculation of Kendall's τ is explained in Example 8.6

Example 8.6: In a crop production competition, 10 entries of farmers were ranked by agricultural scientists (judges). Find the degree of agreement among the scientist for the competition result given below.

Table 8.14: Rankings of 10 farmers by agricultural scientists

Farmers	Scientist 1	Scientist 2	Scientist 3	Scientist 4
1	4	5	3	7
2	10	9	8	6
3	8	6	10	9
4	3	4	2	1
5	1	3	4	2
6	2	1	1	4
7	5	7	6	5
8	6	2	5	3
9	7	8	9	10
10	9	10	7	8

Solution 8.6

Table 8.15: Calculation of Kendall's τ

Farmers	S1	S2	S3	S4	R_i (sum of ranks)	R_i^2
1	4	5	3	7	19	361
2	10	9	8	6	33	1089
3	8	6	10	9	33	1089
4	3	4	2	1	10	100
5	1	3	4	2	10	100
6	2	1	1	4	8	64
7	5	7	6	5	23	529
8	6	2	5	3	16	256
9	7	8	9	10	34	1156
10	9	10	7	8	34	1156
Total					220	5900

here k = number of judges = 4; n = number of farmers = 10; $\sum_{i=1}^{10} R_i^2 = (220)^2 = 48400$; $\sum_{i=1}^{10} R_i^2 = 5900$. Using Equation 8.9

$$\tau = \frac{12\left[5900 - \frac{48400}{10}\right]}{16 \times 10\,(100-1)} = 0.80$$

A Kendall's τ of 0.80 suggests that the judges' rankings are highly consistent. This level of concordance implies that the rankings are strongly aligned, indicating a high level of agreement across the judges' evaluations.

8.6 Correlation matrix

A correlation matrix is a table that displays the correlation coefficients between multiple variables. It is particularly useful for presenting the correlation values of several variables at the same time. The matrix can be computed using either Pearson's correlation or Spearman's correlation. Each cell in the matrix represents the correlation coefficient between the two variables at the intersecting row and column. These values range from -1 to 1. The diagonal elements are always equal to 1, as a variable is perfectly correlated with itself. Additionally, the matrix is symmetrical across the diagonal, since the correlation between x and y is the same as the correlation between y and x.

Example 8.7: Data below gives the measurement on several plant growth characters. Create a correlation matrix.

Correlation matrix for Table 8.16 can be created as shown in Table 8.17 below. Here the diagonal elements are 1 because correlation of any variable to itself is 1. Also from the matrix

Table 8.16: Plant growth characters

Growth	Water	Sunlight	Fertilizer	Nutrient
4.37	167.55	11.08	56.60	243.43
9.56	325.36	6.71	40.68	519.04
7.59	231.40	5.65	35.86	455.06
6.39	209.62	14.49	78.06	342.38
2.40	77.79	14.66	81.01	127.70
2.40	115.16	13.08	70.34	148.98
1.52	76.76	8.05	45.48	92.12
8.80	280.54	5.98	34.18	532.97
6.41	195.48	11.84	59.45	401.31
7.37	236.65	9.40	48.08	431.84
1.19	51.96	6.22	31.41	146.65
9.73	328.38	9.95	56.11	566.87
8.49	286.58	5.34	29.84	443.16
2.91	131.66	14.09	75.54	234.76
2.64	102.81	7.59	47.03	185.93

below it is easy to find the correlation between any two variables, for example correlation between growth and water is 0.988. Correlation matrix is an effective way of presenting correlation results of several variables.

Table 8.17: Correlation matrix of data in Table 8.16

	Growth	Water	Sunlight	Fertilizer	Nutrient
Growth	1	0.988	-0.343	-0.332	0.983
Water	0.988	1	-0.328	-0.314	0.967
Sunlight	-0.343	-0.328	1	0.986	-0.371
Fertilizer	-0.332	-0.314	0.986	1	-0.366
Nutrient	0.983	0.967	-0.371	-0.366	1

8.7 Correlogram

A correlogram is a graphical representation of a correlation matrix. It is used to visualize the pairwise correlation coefficients between multiple variables in a dataset. The correlogram uses color and size to represent the strength and direction of the correlation, helping to quickly identify relationships among variables. The most common way to display a correlogram is through a matrix of circles or squares, where each shape represents the correlation between

two variables. Correlogram can be drawn using packages like `corrplot` (Wei and Simko 2021) in R software (Team 2024). Different colours and styles are available to make the presentation attractive. Correlation matrix and correlogram can also be generated using our online platform available at www.kaugrapes.com (Pratheesh P. Gopinath 2020).

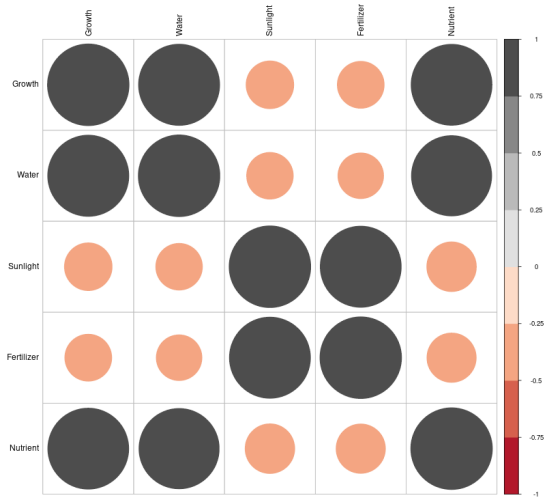


Figure 8.7: Correlogram of growth characters in Example 8.7

8.8 Partial and multiple correlation

Historical Insights

Correlation and Sir Francis Galton
The concept of correlation dates back to Sir Francis Galton, who introduced the idea of *"co-relation"* in the 19th century while studying the relationship between physical traits, such as the height of parents and their children. Galton's cousin, Karl Pearson, further developed this idea by formalizing the calculation of correlation and introducing the formula we use today in terms of Pearson's correlation coefficient. Pearson's work in the late 1800s provided a mathematical framework for understanding the strength and direction of relationships between two variables, a concept that has since become fundamental in statistics, especially in the fields of genetics, psychology, and social sciences.

Quotes to Inspire

"Statistics is like a high-caliber weapon: helpful when used correctly and potentially disastrous in the wrong hands".- Herman Chernoff

9 Regression analysis

Regression analysis is one of the most important tools in statistics, used to understand and quantify the relationships between variables. In essence, regression helps us answer questions such as, "How does a change in one factor (like fertilizer usage) affect another factor (like crop yield)?" It provides a mathematical framework to explore these relationships based on observed data.

Regression analysis involves two types of variables:

- **Dependent variable** (y): This is the main outcome or the variable you are trying to predict or explain. For example, crop yield might be the dependent variable in an agricultural study. We usually denote dependent variable as y. The dependent variable is also known by various names such as the target variable, response variable, outcome variable, predicted variable, explained variable, **regressand**, y-variable, criterion variable, or output variable.
- **Independent variables** (x)*:* These are the factors that are thought to influence or predict changes in the dependent variable, such as the amount of water, fertilizer, or sunlight received by plants. The independent variable is also known as the predictor variable, explanatory variable, input variable, **regressor**, feature, covariate, x-variable, or control variable.

Why use regression analysis?

Regression analysis is particularly useful because it allows you to:

1. **Quantify relationships**: It measures how strongly one or more independent variables are associated with the dependent variable.
2. **Predict outcomes**: Once the relationship is understood, regression can be used to predict the dependent variable for new values of the independent variables.
3. **Identify key factors**: It can highlight which variables have the most significant impact on the dependent variable, guiding decision-making.
4. **Control for multiple factors:** By including several independent variables, regression helps isolate the effect of each variable while controlling for others.

Types of regression

There are different types of regression techniques, depending on the nature of the data and the relationship between variables:

- **Simple linear regression**: Examines the relationship between one independent variable and one dependent variable, assuming a straight-line relationship.
- **Multiple linear regression**: It is an extension of simple linear regression, allowing for the analysis of the relationship between one dependent variable (y) and multiple independent variables (more than one x). It is used when the outcome (dependent variable) is influenced by more than one factor (independent variables).
- **Nonlinear regression:** Deals with situations where the relationship between variables is not a straight line.
- **Logistic regression:** Used when the dependent variable is categorical, such as predicting whether a plant will survive (yes or no) based on environmental factors.

Practical applications

Regression analysis has a wide range of applications across fields:

- In **agriculture**, it can be used to study the effect of irrigation, soil nutrients, and weather on crop yield.
- In **economics**, it helps analyze the impact of income, education, and employment on consumer behavior.
- In **medicine**, regression can predict health outcomes based on patient characteristics.

By the end of this chapter, you will learn how to perform regression analysis, interpret its results, and understand its assumptions and limitations. This will enable you to use regression as a powerful tool for making informed decisions and predictions.

9.1 Simple linear regression

Regression can be simply defined as a technique of fitting best line or line of best fit to estimate value of one variable on the basis of another variable. Now what is a best line? or line of best fit?. To understand this concept better, consider the data presented in Table 8.2, which shows the average daily soil moisture content and the corresponding monetary yield from crops in Example 8.2 of Section 8.2. This example helps visualize how the relationship between two variables—soil moisture content (independent variable) and crop yield (dependent variable). We can use regression analysis to answer the following questions. What will be the crop yield in rupees when soil moisture content is maintained at 20%?. What is the functional form of relationship between soil moisture content and monetary crop yield?.

Refer to the scatter diagram of Table 8.2 in Figure 8.3. To represent the relationship between soil moisture content and monetary crop yield, we might attempt to draw a line, as illustrated in Figure 9.1. However, as shown in Figure 9.1, it's possible to draw numerous lines through the data points. This raises the question: *which line is the best fit*?

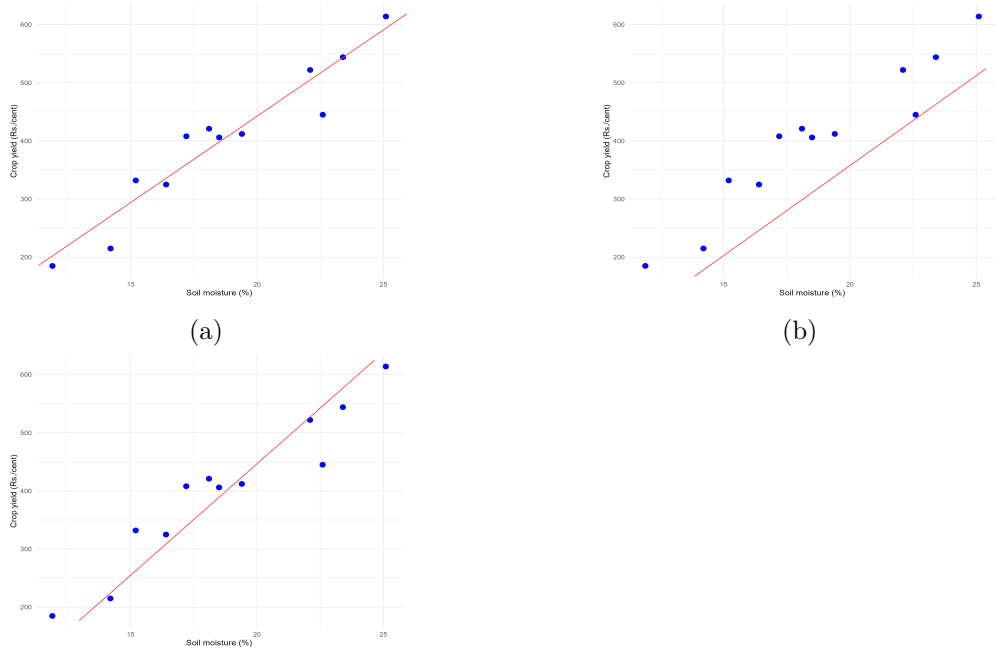


Figure 9.1: lines drawn to show functional relationship between soil moisture and yield

The best fit line is defined as the one that minimizes the distances between the observed data points and the line itself. These distances, which represent how far each data point is from the line, are minimized collectively using a specific criterion. The regression technique provides a systematic approach to determine and draw this best fit line. Before diving further into regression, it is essential to understand the concepts of error and residuals, which play a critical role in determining the best fit line. The entire topic of regression is on *how to draw a best fit line*?.

9.1.1 Error and residual

In regression analysis, an error represents the difference between an observed value (a data point) and the true regression line, which reflects the actual relationship between the dependent and independent variables in the population. Since the true regression line is based on the entire population and is usually unknown, the error is a theoretical concept that cannot be directly calculated.

A residual is the difference between an observed value and the value predicted by a regression line based on sample data. Specifically, for a given data point, the residual is calculated as the observed value minus the predicted value from the regression line. Residuals are measurable because they are derived from the observed data and the regression line that is obtained using the sample. Unlike errors, which are theoretical and represent deviations from the true underlying model, residuals provide a practical estimate of these deviations, allowing us to assess the goodness of fit and identify any patterns or discrepancies in the model.

In essence, a residual serves as an estimate of the error. While errors represent the deviation from the actual true value, residuals reflect the discrepancy between observed data and the fitted model. From Figure 9.2 you can see the residual e_i of an $\hat{i}$th observation in a fitted regression line.

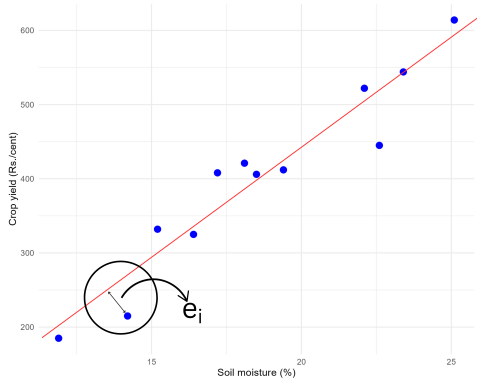


Figure 9.2: Residual and a best fit line

The distance of i^{th} observation (e_i) from the fitted line can be considered as the residual (error). Best fit line can be obtained by minimizing this distance. This can be achieved using the mathematical technique "**principle of least squares**" discussed in Section 9.1.3. Before going to identify a best fit line on should know the concept of a straight line.

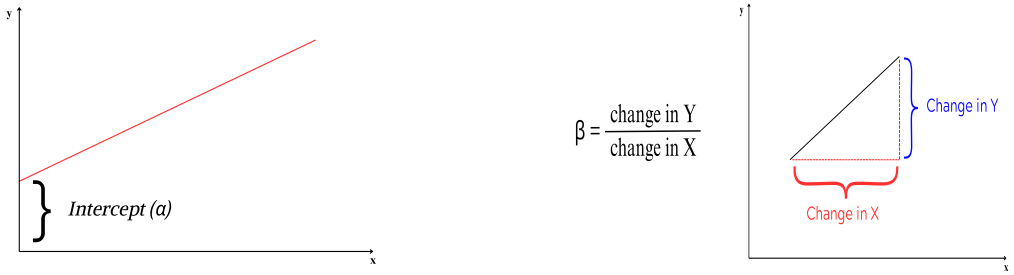
9.1.2 Straight lines

A straight line is the simplest figure in geometry. Mathematical equation of a straight line is

$$Y = \alpha + \beta X \tag{9.1}$$

Two important features of a line **intercept** (α) and **slope**(β). α is the Y-intercept, the intercept of a line is the y-value of the point where it crosses the y-axis. β is the **slope of a line,** which is a number that measures its "steepness". It is the change in Y for a unit change in X along the line. In regression β is called as **regression coefficient** explained in Section 9.1.4.

Intercept and slope



(a) Intercept of a straight line

(b) Slope of a straight line

Figure 9.3: Intercept and slope

α and β can be considered as a finger print of a line; with these values we can easily identify the line. So now our problem is simple, to find a line of best, estimate α and β, such that error e_i of each observation is minimized. For that we use the *method of least squares*.

9.1.3 Method of least squares

On considering the error term e_i; equation of a straight line is

$$y_i = \alpha + \beta x_i + e_i \tag{9.2}$$

Where e_i is the i^{th} error term corresponding to y_i, i =1, 2, ..., n

> **Note**
>
> One way to obtain *line of best fit* is by estimating α and β by minimizing error sum $\sum_{i=1}^{n} e_i$. By theorem $\sum_{i=1}^{n} e_i = 0$. So α and β are estimated by minimizing $\sum_{i=1}^{n} {e_i}^2$. The term $\sum_{i=1}^{n} e_i^2$ is called as error sum of squares. As we are minimizing the sum of the squares of error term the process is known by the name `principle of least squares`.

Principle of least squares

`Principle of least squares` is the statistical method used to determine a line of best fit by minimizing the sum of squares of the error term *i.e* minimizing $\sum_{i=1}^{n} {e_i}^2$.

Consider Equation 9.2

$$y_i = \alpha + \beta x_i + e_i$$

$$e_i = y_i - (\alpha + \beta x_i) \tag{9.3}$$

$$e_i^2 = [y_i - (\alpha + \beta x_i)]^2 \tag{9.4}$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 \tag{9.5}$$

we want to minimize Equation 9.5 and estimate α and β. $\sum_{i=1}^{n} e_i^2$ can be minimized by taking derivative with respect to α and β and equating to zero. On doing so we will get two equations, these equations are termed as *normal equations* and solving those normal equations will give the formulas for estimating α and β.

Differentiating $E = \sum_{i=1}^{n} e_i^2$ with respect to α and equating to 0.

$$\frac{\partial E}{\partial \alpha} = \sum_{i=1}^{n} 2\left[y_i - (\alpha + \beta x_i)\right](-1) \tag{9.6}$$

$$= -2 \sum_{i=1}^{n} [y_i - \alpha - \beta x_i] \tag{9.7}$$

equating the derivative in Equation 9.7 to 0 and on simplifying:

$$\sum_{i=1}^{n} [y_i - \alpha - \beta x_i] = 0 \tag{9.8}$$

expand the summation in Equation 9.8:

$$= \sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} x_i = 0 \tag{9.9}$$

on rearranging Equation 9.9 you will get the first normal equation.

$$\sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i \tag{9.10}$$

Differentiating $E = \sum_{i=1}^{n} e_i^2$ with respect to β:

$$\frac{\partial E}{\partial \beta} = \sum_{i=1}^{n} 2\left[y_i - (\alpha + \beta x_i)\right](-x_i) \tag{9.11}$$

$$= -2\sum_{i=1}^{n} x_i\left[y_i - \alpha - \beta x_i\right] \tag{9.12}$$

equating the derivative in Equation 9.12 to 0 and on simplifying:

$$\sum_{i=1}^{n} x_i\left[y_i - \alpha - \beta x_i\right] = 0 \tag{9.13}$$

expand the summation in Equation 9.13:

$$= \sum_{i=1}^{n} x_i y_i - \alpha\sum_{i=1}^{n} x_i - \beta\sum_{i=1}^{n} x_i^2 = 0 \tag{9.14}$$

on rearranging Equation 9.14 you will get the second normal equation.

$$\sum_{i=1}^{n} y_i x_i = \alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 \tag{9.15}$$

On solving normal equations Equation 9.10 and Equation 9.15, we derive the equations to estimate α and β, which are considered population parameters. Since these parameters are usually unknown, we estimate them using equations derived from sample data. The estimated values of α and β are denoted as $\hat{\alpha}$ and $\widehat{\beta}$, where the hats indicate that they are sample-based estimates. These are pronounced as "alpha cap" and "beta cap," respectively, and are used as approximations of the true population parameters.

9.1.4 Regression coefficient

The regression coefficient, β in linear regression, represents the slope of the regression line. It quantifies the relationship between the independent variable (x) and the dependent variable (y). Specifically, β indicates the expected change in y for a one-unit increase in x, holding other factors constant. Regression coefficients can take any real value, ranging from $-\infty$ to $+\infty$, depending on the nature of the relationship between the variables. A positive β implies a direct relationship (as x increases, y increases), while a negative β implies an inverse relationship (as x increases, y decreases). A coefficient of 0 suggests no linear relationship between the variables. The magnitude of β reflects the strength of the association, with larger absolute values indicating stronger relationships. Equation 9.16 and Equation 9.16 for the calculation

of estimate of β is obtained by solving normal equations Equation 9.10 and Equation 9.15. Equation 9.16 can be used for hand calculations.

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{9.16}$$

Equation 9.16 can be written as

$$\hat{\beta} = \frac{cov(x, y)}{var(x)} \tag{9.17}$$

9.1.5 Intercept

The intercept, often denoted as α in linear regression, represents the value of the dependent variable y when the independent variable x is equal to zero. It is the point at which the regression line crosses the y-axis. The intercept provides a baseline value for the dependent variable before any influence from the independent variable is considered. The intercept can take any real value ($-\infty$ to $+\infty$), and its meaning depends on the specific context of the data. In some cases, it may not have a practical interpretation, especially if $x = 0$ is not within the range of observed data. Equation 9.18 for estimating α is obtained by solving normal equations Equation 9.10 and Equation 9.15.

$$\hat{\alpha} = \overline{y} - \hat{\beta}.\overline{x} \tag{9.18}$$

where $\overline{y}$ = mean of y; $\overline{x}$ = mean of x

> **!** Note
>
> Once the estimates of α and β, which are denoted as $\hat{\alpha}$ and $\hat{\beta}$ respectively are obtained using Equation 9.17 and Equation 9.18. The estimated regression line can be written as:
>
> $$y = \hat{\alpha} + \hat{\beta}x \tag{9.19}$$

9.1.6 Assumptions

The goal of linear regression is to estimate the coefficients of the regression equation, which help explain how changes in the independent variables affect the dependent variable. However, for the results of a regression analysis to be reliable and meaningful, certain underlying assumptions must be met. These assumptions ensure that the estimates are accurate, the

predictions are unbiased, and the conclusions drawn from the model are valid. Before conducting a regression analysis, it is crucial to understand and verify these assumptions to avoid misleading results.These assumptions are:

1. **Linearity**
   The relationship between the independent variable(s) and the dependent variable is linear. This means that the changes in the dependent variable are proportional to changes in the independent variable(s).

2. **Independence**
   The observations in the dataset are independent of each other. Additionally, the residuals (errors) are assumed to be independent.

3. **Homoscedasticity**
   The variance of the residuals is constant across all levels of the independent variable(s). In other words, the spread of the residuals should remain consistent and not show patterns of increasing or decreasing variance.

4. **Normality of residuals**
   The residuals (errors) are normally distributed. This is particularly important for hypothesis testing and constructing confidence intervals. Normality assumption doesnot much influence the estimation of regression coefficient.

5. **No multicollinearity**
   In the case of multiple regression, the independent variables should not be highly correlated with each other, as multicollinearity can distort the estimates of regression coefficients.

6. **No autocorrelation**
   There should be no autocorrelation in the residuals. This means that the residuals of one observation should not be correlated with the residuals of another.

7. **Correct model specification**
   The model should include all relevant variables and exclude irrelevant ones. The functional form of the relationship between variables should be correctly specified.

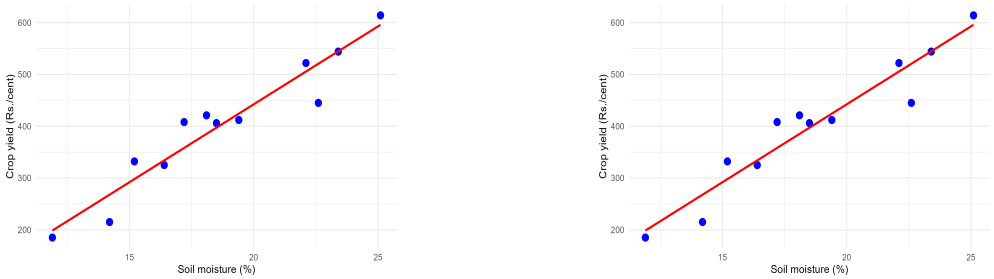
Violations of these assumptions can lead to biased, inconsistent, or inefficient estimates, affecting the validity of the regression analysis.

> **!** Note
>
> The essence of the assumptions in linear regression can be summarized as $e \sim \text{i.i.d.}(0, \sigma^2)$. This denotes that the errors are *independent and identically distributed (i.i.d.)*, with a mean of zero and a constant variance σ^2. Independence ensures that the error for one observation neither depends on nor influences the error for another. Identically distributed means that all errors are drawn from the same probability distribution, without variation

across observations. A mean of zero ensures that the errors do not introduce systematic bias into the model's predictions. Additionally, the constant variance (homoscedasticity) implies that the errors maintain a consistent level of variability across all values of the independent variable(s).

9.2 Two lines of regression

Consider the data presented in Table 8.2, which shows the average daily soil moisture content and the corresponding monetary yield from crops in Example 8.2 of Section 8.2. For the data we can draw two lines of regression interchanging variables in X and Y axis as shown in Figure 9.4.



(a) Moisture in X-axis and yield in Y-axis

(b) Yield in X-axis and moisture in Y-axis

Figure 9.4: Two lines of regression

From Figure 9.4 it is clear that two lines of regression that of y on x and x on y is possible.

Regression of y on x

Consider the two variables x and y, if you are considering y as dependent variable and x as independent variable then your equation is:

$$y = \alpha + \beta_{yx} x \tag{9.20}$$

This is used to predict the unknown value of variable y when value of variable x is known. Usually β here is denoted as β_{yx} and it is obtained using Equation 9.21.

$$\beta_{yx} = \frac{cov(x, y)}{var(x)} \tag{9.21}$$

Regression of x on y

Consider the two variables x and y, if you are considering x as dependent variable and y as independent variable then your equation is:

$$x = \alpha_1 + \beta_{xy}.x \tag{9.22}$$

This is used to predict the unknown value of variable x when value of variable y is known. Usually β here is denoted as β_{xy} and it is obtained using Equation 9.23.

$$\beta_{xy} = \frac{cov(x, y)}{var(y)} \tag{9.23}$$

You can see from Equation 9.21 and Equation 9.23 both the regression coefficients were different. It depends on the experimenter to choose dependent and independent variable. In the Example 8.2 there may be situation that considering moisture as dependent variable is meaningless, *i.e.* it depends on the fact that what is the usefulness in predicting soil moisture based on monetary crop yield?. So the selection of dependent and independent variable is entirely the discretion of experimenter based on the objective of his study.

9.2.1 Properties of regression coefficients

1. The correlation coefficient between x and y denoted as r_{xy} is the geometric mean of the two regression coefficients β_{yx} and β_{xy}

$$r_{xy} = \sqrt{\beta_{yx}.\beta_{xy}} \tag{9.24}$$

2. Regression coefficients are independent of change of origin but not of scale. Regression coefficients exhibit specific behaviors under transformations of the variables x or y, particularly when there are changes in origin or scale. They are independent of a change in origin for both x and y. This means that adding or subtracting a constant to either variable (e.g., transforming x to $x' = x + c$ or y to $y' = y + c$, where c is a constant) does not affect the slope β of the regression line. The slope depends on the relative differences between values, which remain unchanged by shifts in origin. However, changes in origin will affect the intercept α by adjusting it to accommodate the shift in y. In contrast, regression coefficients are not independent of changes in scale. When either x or y is multiplied or divided by a constant (e.g., $x' = kx$ or $y' = ky$, where k is a non-zero constant), the slope β changes proportionally. Specifically, the regression coefficient is inversely proportional to the scale factor for x because scaling affects the covariance and variance of the variables. Similarly, scaling y affects both the slope and the intercept, as the entire regression equation is scaled by the factor k. Understanding these effects is crucial for accurately interpreting regression results, especially when transformations are applied during data preprocessing.

3. If one regression coefficient is greater than unity, then the other must be less than unity but not vice versa. *i.e.* both the regression coefficients can be less than unity but both cannot be greater than unity, *i.e.* if $\beta_{yx} > 1$ then $\beta_{xy} < 1$ and if $\beta_{xy} > 1$, then $\beta_{yx} < 1$.
4. Also if one regression coefficient is positive the other must be positive (in this case the correlation coefficient is positive) and if one regression coefficient is negative the other must be negative (in this case the correlation coefficient is negative). This relationship arises because the regression coefficients and the correlation coefficient share the same sign, reflecting the direction of the association between the two variables.
5. The range of regression coefficients is $-\infty$ to $+\infty$.
6. If the variables (x) and (y) are independent, the regression coefficients are zero. This is referred to as the independence property of regression coefficients.

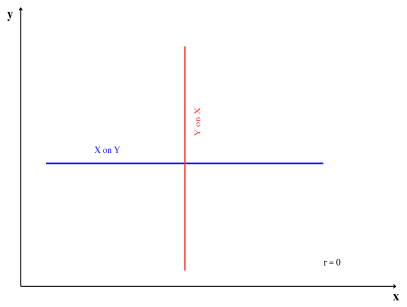
9.2.2 Properties of regression lines

1. Regression lines minimize the sum of squared deviations of observed values from the predicted values, ensuring the best possible fit.
2. The regression lines intersect at the mean values of x and y *i.e.,* at ($\overline{x}$,$\overline{y}$)
3. The slopes of the regression lines are related to the correlation coefficient r. If r=0, the lines are perpendicular, indicating no linear relationship.

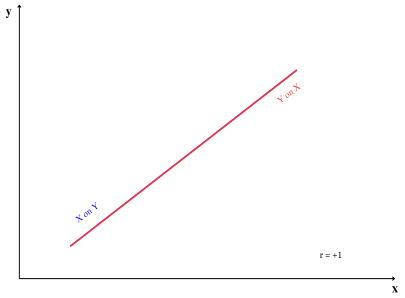
The position of regression lines is closely related to the strength of the correlation between x and y. As shown in Figure 9.5, the placement of the two lines changes with the correlation value, demonstrating how the relationship between x and y influences the regression line's position.

9.3 Uses of regression

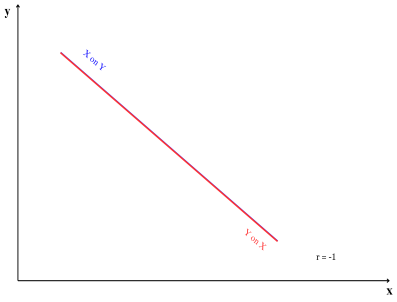
- **Prediction**
  Regression analysis is used to predict the value of a dependent variable (y) based on one or more independent variables (x). Examples in agricultural research include:
  - Predicting crop yield based on weather parameters such as temperature, rainfall, and humidity.
  - Estimating soil nutrient levels using remote sensing data or environmental variables.
  - Forecasting pest or disease outbreaks based on climatic and ecological conditions.



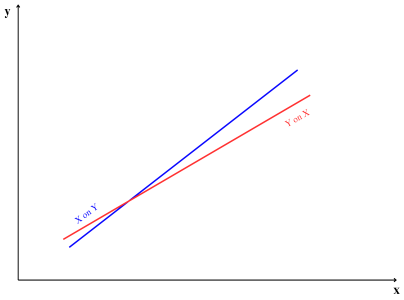
(a) When correlation = 0



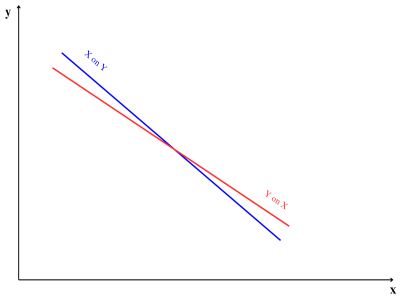
(b) Perfect positive correlation



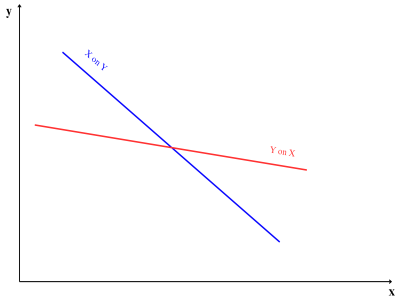
(c) Perfect negative correlation



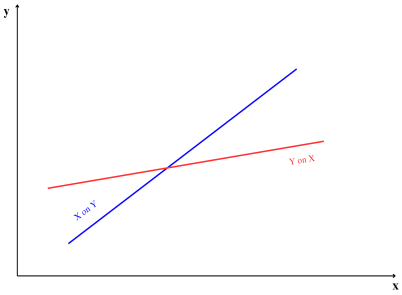
(d) High degree positive correlation



(e) High degree negative correlation



(f) Low degree negative correlation



(g) Low degree positive correlation

Figure 9.5: Effect of correlation strength on the position of regression lines

- **Identify the strength of relationships**
  Regression helps quantify the strength of the relationship between variables. This is essential in agricultural research to identify influential factors. Examples include:
  - Determining the effect of fertilizer dosage on crop yield.
  - Analyzing the relationship between irrigation frequency and plant growth.
  - Understanding the impact of livestock feed composition on milk production.
- **Forecast effects or impact of changes**
  Regression models allow researchers to evaluate how changes in one or more independent variables affect the dependent variable. For example:
  - Assessing how seed quality impacts overall harvest productivity.
  - Analyzing the effects of varying water availability on crop output in drought-prone areas.
  - Estimating the economic benefits of adopting precision farming techniques.
- **Predict Trends and Future Values**
  Regression is valuable for modeling trends and forecasting future values, aiding in strategic planning and policy-making. Applications include:
  - Predicting future crop yields under different climate change scenarios.
  - Estimating long-term price trends for agricultural commodities such as rice, wheat, or coffee.
  - Forecasting the adoption rates of new agricultural technologies among farmers.

! Note

Multiple regression is an extension of simple linear regression that models the relationship between a dependent variable and two or more independent variables. It allows researchers to account for the combined effect of multiple factors on an outcome, making it particularly useful in agricultural research. For instance, crop yield can be predicted based on a combination of variables such as soil nutrients, rainfall, temperature, and fertilizer application.
In a multiple regression model, the relationship between the dependent variable y and independent variables $x_1, x_2, \ldots, x_n$ is expressed as:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + e \tag{9.25}$$

Where:

Table 9.1: Calculation table for regression

Sl No.	Soil moisture (x)	Crop yield in Rs(y)	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
1.00	14.2	215	-4.48	-187.42	838.69	20.03
2.00	16.4	325	-2.28	-77.42	176.12	5.18
3.00	11.9	185	-6.78	-217.42	1473.00	45.90
4.00	15.2	332	-3.48	-70.42	244.70	12.08
5.00	18.5	406	-0.18	3.58	-0.63	0.03
6.00	22.1	522	3.43	119.58	409.57	11.73
7.00	19.4	412	0.73	9.58	6.95	0.53
8.00	25.1	614	6.43	211.58	1359.42	41.28
9.00	23.4	544	4.73	141.58	668.98	22.33
10.00	18.1	421	-0.58	18.58	-10.69	0.33
11.00	22.6	445	3.93	42.58	167.14	15.41
12.00	17.2	408	-1.48	5.58	-8.24	2.18
SUM	224.1	4829	0.00	0.00	5325.03	176.98

- y: Dependent variable (response).
- α: Intercept (value of Y when all X values are zero).
- $\beta_1, \beta_2, \ldots, \beta_n$: Coefficients representing the effect of each independent variable on y.
- $x_1, x_2, \ldots, x_n$: Independent variables (predictors).
- e: Error term accounting for variability not explained by the predictors.

Example 9.1: We will be using the data presented in Table 8.2, which shows the average daily soil moisture content and the corresponding monetary yield from crops in Example 8.2 of Section 8.2 to demonsrtate how regression analysis can be used to answer the following questions.

1. What is the functional form of relationship between soil moisture and monetary crop yield?
2. What will be the estimated monetary crop yield when average daily soil moisture is maintained around 20%?

Solution 9.1:

1. Fit a model considering monetary crop yield as dependent variable (y) and average soil moisture as independent variable (x). Fitting a model means estimating β and α using equation.
2. After fitting the model put 20 in the x value you will get the predicted y value

n =12

$$mean, \overline{x} = \frac{224.1}{12} = 18.675$$

$$mean, \overline{y} = \frac{4829}{12} = 402.416$$

$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{5325.03}{12} = 443.752$$

variance of x:

$$var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{176.983}{12} = 14.7485$$

now using Equation 9.17

$$\hat{\beta} = \frac{cov(x, y)}{var(x)}$$

$$\hat{\beta} = \frac{443.752}{14.7485} = 30.088$$

using Equation 9.18

$$\hat{\alpha} = \overline{y} - \hat{\beta}.\overline{x}$$

$$\hat{\alpha} = 402.416 - 30.088\,(18.675) = -159.477$$

So our estimated model is

$$y = -159.477 + 30.088x$$

$$\text{crop yield in Rs} = -159.477 + 30.088(\text{soil moisture})$$

for soil moisture content at 20%, *i.e.* $x = 20$

$$y = -159.477 + 30.088(20)$$

= 442.28

So the predicted monetary yield in rupees at a average soil moisture of 20% is 442.283

9.4 Correlation and regression

Correlation and regression are fundamental concepts in statistics, often used to explore and model relationships between variables. While both techniques examine how variables relate to one another, they differ in their purpose, interpretation, and methodology. Correlation focuses on measuring the strength and direction of an association between two variables, without assuming causation. In contrast, regression goes a step further by modeling the relationship, enabling predictions of one variable based on the other(s). The Table 9.2 below provides a detailed comparison of these two approaches, helping to clarify their unique characteristics and applications.

Table 9.2: Correlation versus regression

Item	**Correlation**	**Regression**
Definition	Measures the strength and direction of the relationship between two variables.	Models the relationship between a dependent variable and one or more independent variables.
Objective	To quantify the degree of association between variables.	To predict the value of the dependent variable based on the independent variable(s).
Causation	Does not imply causation; it only measures association. Causation means changes in one variable cause changes in another.	Can imply causation if assumptions are met and the model is well-specified.
Equation	No equation is derived.	Derives an equation: $y = \alpha + \beta x + e$.
Variables involved	Considers two variables at a time.	Can involve one or multiple independent variables to predict a dependent variable.
Symmetry	Correlation between (x) and (y) is the same as (y) and (x).	The regression coefficient of y on x is different from x on y.
Range	The correlation coefficient (r) ranges from -1 to 1.	Regression coefficients (α, β) ranges from $-\infty$ to $+\infty$.
Units	Unit less measure.	Dependent on the units of the variables involved.
Purpose	To understand the strength of the relationship.	To predict outcomes or explain variability in the dependent variable.

Historical Insights

"Regression and study of heights"

The story of *regression* begins with Sir Francis Galton's groundbreaking work on heredity in the late 19th century. While studying the heights of parents and their children, Galton noticed a fascinating pattern: tall parents tended to have slightly shorter children, and shorter parents tended to have slightly taller children. He called this phenomenon "regression toward the mean", as the offspring's heights seemed to move closer to the population average. This observation not only introduced the term "regression" but also inspired the development of statistical tools for studying relationships between variables. Galton's work, later expanded by Karl Pearson, laid the foundation for modern regression analysis, which remains an essential technique in several fields ranging from agriculture to space exploration.

Quotes to Inspire

"Statistics is the art of never having to say you're wrong":-Robert P. Abelson

10 Probability

Probability is a cornerstone of statistical methods, providing the mathematical framework to quantify and analyze uncertainty. In the diverse and dynamic field of agricultural research, uncertainty often arises in areas such as yield predictions, pest control strategies, and the impact of environmental variables on crop growth. Understanding probability equips researchers and students with the tools to make informed decisions in the face of variability and randomness.

This chapter introduces fundamental probability concepts. We begin with the basic definitions and axioms of probability. The chapter progresses to explore key topics such as conditional probability, Bayes' theorem, and independence of events.

10.1 Random experiment

A random experiment is a process or procedure that produces an outcome which cannot be predicted with certainty beforehand. The outcome is determined by chance, and the set of all possible outcomes is known as the `sample space`. Random experiments form the basis of probability theory and are critical for understanding uncertainty in various contexts.

Characteristics of a random experiment

- Uncertainty of outcome: The result of the experiment cannot be predetermined.
- Repeatability: The experiment can be repeated under the same conditions.
- Defined sample space: All possible outcomes are known and well-defined.

Some common examples of random experiments used to illustrate probability theory were discussed below:

Tossing a coin

Tossing a coin is one of the simplest and most fundamental random experiments used to illustrate the principles of probability. Despite its simplicity, this seemingly mundane activity provides profound insights into random processes and lays the groundwork for understanding more complex probabilistic systems.

When a coin is tossed, there are two possible outcomes:

- Heads (H)

- Tails (T)

We say that for an unbiassed coin the probability of the coin landing H is $\frac{1}{2}$ and the probability of the coin landing T is $\frac{1}{2}$

> **!** Note
>
> An unbiased coin is a theoretical concept in probability, referring to a coin designed or assumed to have no preference for either side when tossed, such that the likelihood of landing on heads (H) is exactly the same as landing on tails (T), with each outcome having a probability of $\frac{1}{2}$ or 50%. Its characteristics include perfect symmetry, ensuring the coin is balanced without physical imperfections or uneven weight distribution that could influence the result, and equal probability, where each side is equally likely to face upward after a toss. This assumption generally holds under ideal experimental conditions, free from external factors like uneven surfaces or human bias in the tossing technique.

Throwing dice

A standard die has six faces, numbered from 1 to 6. Under fair conditions, each face has an equal probability of appearing when the die is rolled. Thus, the probability of getting any one face is $\frac{1}{6}$.

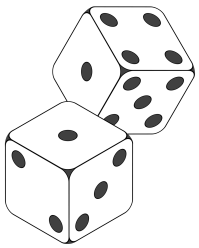


Figure 10.1: Two dice

Playing cards

A standard deck of playing cards consists of 52 cards, divided into four suits: spades, hearts, diamonds, and clubs, with each suit containing 13 cards. Four suits in playing cards were given in the Figure 10.2. The spades suit includes 9 numbered cards from 2 to 10, along with the picture cards ace, king, queen, and jack. Similarly, the hearts, diamonds, and clubs suits each contain 13 cards. Of these 52 cards, 26 are red, consisting of the hearts and diamonds suits, while the other 26 cards are black, comprising the spades and clubs suits. Playing cards can be used to create examples for random experiments, such as drawing a single card from a shuffled deck to determine its specific identity. Other experiments include drawing a card and checking its suit etc. These simple experiments can be easily modified with conditions or repetition, making playing cards a versatile tool for exploring probability and random events.

(a) Spades

(b) Hearts

(c) Diamonds

(d) Clubs

Figure 10.2: Four suits in playing cards

10.2 Random variable

A **random variable** is a numerical value that represents the outcome of a random experiment. It is a function that assigns a real number to each possible outcome in the sample space of the experiment.

Random variables can be classified into two types:

1. **Discrete random variables**: These take on a countable set of values, such as integers (e.g., the number of heads in three coin tosses).

2. **Continuous random variables**: These can take on any value within a given range, which is typically uncountable (e.g., the height of individuals in a population).

In mathematical terms, if S is the sample space of a random experiment, a random variable X is a function that maps sample space to real number set, which can be represented as $X : S \to \mathbb{R}$, where $\mathbb{R}$ represents the set of real numbers.

The probability of a random variable can be represented by $p(X = x)$ or $p(x)$, where the small letter x denotes the value taken by the random variable X.

Consider the example of throwing a die once.Here, you can define a random variable of your interest. Defining a random variable involves assigning a numerical value to the outcomes of the experiment. Let X represent the number appearing on the die. The possible values of X are: $x = \{1, 2, 3, 4, 5, 6\}$.

Consider another example of even number appearing on the die. Let X represent the even number appearing on the die. The possible values of X are: $x = \{2, 4, 6\}$.

In each of the above case, you can see that the random variable X assigns a value to the outcomes of the experiment.

10.3 Probability

Probability is a measure of the likelihood or chance that a particular event will occur. It quantifies uncertainty and is expressed as a number between 0 and 1, where 0 indicates an impossible event and 1 represents a certain event. Mathematically, the probability of an event

A is defined as the ratio of the number of favorable outcomes to the total number of possible outcomes, assuming all outcomes are equally likely, and is given by

Probability of an event A happening:

$$p(A) = \frac{\text{Number of favourable outcomes}}{\text{Total possible outcomes}} \tag{10.1}$$

here favourable outcome means outcomes favouring the happening of event A.

Example 10.1: What is the chances of rolling a "4" with a die ?

Solution 10.1: Let the event A = rolling a "4" with a die. Number of ways event A can happen is one, as there is only 1 face with a "4" on it. Total number of outcomes is 6 (there are 6 faces altogether). So using Equation 10.1; probability,$p(A) = \frac{1}{6}$.

Example 10.2: There are 5 marbles in a bag: 4 are blue, and 1 is red. What is the probability that a blue marble gets picked?

Solution 10.2: Let the event A = a blue marble gets picked. Number of ways event A can happen is 4 (there are 4 blues). Total number of possible outcomes is 5 (there are 5 marbles in total). So using Equation 10.1; probability,$p(A) = \frac{4}{5} = 0.8$

Common terms used

Outcome: An outcome is the result of a random experiment, representing a single occurrence or observation. For instance, when rolling a die, the outcomes could include the appearance of 1 dot, 2 dots, and so on.

Sample space: The sample space (denoted as S) of a random experiment is the set that contains all possible outcomes of that experiment. For example, when tossing a coin, the sample space is S = {H, T}, and when throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}.

Sample point: A sample point is an individual element of the sample space, representing a single possible outcome. For example, "heads" in a coin toss or the "5 of clubs" in a deck of cards are sample points. In the case of rolling a die, there are six distinct sample points within the sample space.

Try yourself
If a die is tossed

1. What is the sample space?
2. What is the probability of getting a 1?
3. What is the probability of obtaining a even number?
4. What is the probability of getting a 7?

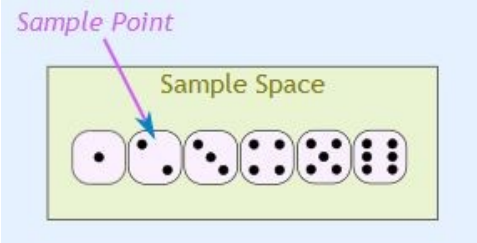


Figure 10.3: Sample space and sample point

10.4 Event

An event refers to one or more outcomes of a random experiment. It can represent any subset of the sample space, where the event consists of specific outcomes that are of interest in the context of the experiment. An event can be a single outcome. For example, when tossing a coin, getting a "Tail" is an event that represents just one outcome from the sample space {H, T}. Similarly, rolling a "5" on a die is an event that includes only one specific outcome, 5, from the sample space {1, 2, 3, 4, 5, 6}.

On the other hand, an event can also include multiple outcomes. For instance, when selecting a "king" from a deck of cards, the event includes any one of the four kings (king of hearts, king of diamonds, king of clubs, king of spades), forming a set of four possible outcomes. Similarly, when rolling an "even number" on a die, the event would encompass the outcomes {2, 4, 6}, which are all the even numbers in the sample space of a die roll. Thus, an event can be a single outcome or a combination of multiple outcomes depending on the context of the experiment.

Example 10.3: Ram wants to see how many times a "double" (both dice have same number) comes up when throwing 2 dice.

Solution 10.3: The *sample space* is all possible outcomes (36 Sample Points): {1,1} {1,2} {1,3} {1,4} … {6,3} {6,4} {6,5} {6,6}. The *event* Ram is looking for is a "double", where both dice have the same number. It is made up of these 6 sample points: {1,1} {2,2} {3,3} {4,4} {5,5} and {6,6}. You can calculate the probability of the event using Equation 10.1. Probability = $\frac{6}{36} = 0.17$.

10.4.1 Types of events

Independent events

Independent events are events where the occurrence of one does not influence the occurrence of another. For instance, consider tossing a coin. If it lands on "heads" three times in a row, the outcome of the next toss remains unaffected by the previous results. The probability of getting "heads" on the next toss is still $\frac{1}{2}$ (or 0.5), just as it is for any fair coin toss.

This demonstrates that past outcomes do not influence the probabilities of future events in independent scenarios.

Example 10.4: A coin is tossed 100 times, and heads appear in 99 of those tosses. What is the probability that heads will appear on the 100^{th} toss?

Solution 10.4: Answer is $\frac{1}{2}$ as each toss is independent of previous toss.

Dependent events

Dependent events are events where the outcome of one event influences the probabilities of subsequent events. For example, consider drawing cards from a deck. The probabilities change as cards are removed from the deck, altering the available outcomes. Initially, the chance of drawing a king on the first card is $\frac{4}{52}$. However, for the second draw, the probabilities depend on the outcome of the first draw. If the first card is a king, only 3 kings remain among the 51 remaining cards, reducing the likelihood of drawing another king to $\frac{3}{51}$. Conversely, if the first card is not a king, all 4 kings remain, making the probability $\frac{4}{51}$. This illustrates how previous outcomes affect probabilities in dependent events.

> **!** Note
>
> When cards are drawn `with replacement`, each card is returned to the deck after being drawn, so the total number of cards remains constant. This keeps the probabilities unchanged, making the events independent. However, when cards are drawn `without replacement`, the total number of cards decreases after each draw, which alters the probabilities and makes the events dependent.

Mutually exclusive events

Mutually exclusive events are events that cannot occur at the same time. It is either one event or the other, but not both. For example, turning left or right is mutually exclusive because you cannot do both simultaneously. Similarly, heads and tails in a coin toss is mutually exclusive. Drawing a king and drawing an ace are mutually exclusive events because you can't draw both in a single card. However, not all events are mutually exclusive. For instance, kings and hearts are not mutually exclusive because it is possible to have a king of hearts, an outcome that belongs to both categories.

Exhaustive events

A set of events is called exhaustive if, together, they encompass the entire sample space. For example, when tossing a die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Consider the events: event A, which is getting an even number $\{2, 4, 6\}$, and event B, which is getting an odd number $\{1, 3, 5\}$. These events are exhaustive because, when combined, they include all possible outcomes in the sample space.

Equally likely events

Equally likely events are events that have the same theoretical probability of occurring. For example, when tossing a coin, event A (getting a head) and event B (getting a tail) are equally likely events because both have an equal probability of occurring.

10.5 Definitions of probability

In probability theory, different approaches are used to define and calculate the likelihood of events occurring. These approaches vary depending on the nature of the experiment and the available information. The three primary definitions of probability are mathematical (classical), statistical (empirical), and axiomatic, each offer unique perspectives and methods for determining probabilities. The classical approach is based on equally likely outcomes, the empirical approach relies on experimental data, and the axiomatic approach, developed by A.N. Kolmogorov, is based on a set of foundational principles. Understanding these definitions helps in applying probability theory to various real-world situations, from simple experiments to more complex scenarios.

10.5.1 Mathematical approach

It is also known as classical, theoretical or a priori approach to probability. If an experiment with n `exhaustive`, `mutually exclusive` and `equally likely` outcomes, m outcomes are favourable to the happening of an event E, the probability p of happening of E is given by

$$p\,(E) = \frac{m}{n} \tag{10.2}$$

p is termed as probability of success.

Example 10.5: When a coin is tossed, there are two possible outcomes head or tail. Outcomes are exhaustive, mutually exclusive and equally likely. What is the probability of getting head?

Solution 10.5: Solution using mathematical approach, consider the event E : getting a head; probability of event E, $p(E)$ can be determined using Equation 10.2.

here, number of outcomes are favourable to the happening of event E, $m = 1$; n = total number of possible outcomes (head and tail) = 2

$$p\,(E) = \frac{1}{2}$$

i.e. probability of getting a head is $\frac{1}{2}$

Limitations

The mathematical approach has certain limitations. For instance, it cannot account for situations where the outcomes are not equally likely, such as tossing a biased die. Additionally, this approach cannot define probability when the total number of possible outcomes n is unknown or tends to infinity, as in determining the probability of raining tomorrow. To address these limitations, other definitions of probability have been developed.

10.5.2 Statistical approach

The statistical approach to probability, also known as the *empirical approach*, is based on observing and recording outcomes from repeated experiments. The probability of an event is determined as the ratio of the number of times the event occurs m to the total number of trials n as when n approaches infinity. This method is useful when theoretical probabilities cannot be calculated but relies on the practicality of conducting a large number of trials and assumes consistency in experimental conditions.

The probability p of happening of E is given by

$$p\left(E\right) = \lim_{n\to\infty} \frac{m}{n} \tag{10.3}$$

where n is the number of times the process is performed which tends to infinity, and m is the number of times the outcome 'E' happens.

Limitations

The statistical approach also has limitations. In some cases, the experiment may not be practically repeatable, making it impossible to rely on repeated trials. Additionally, it raises the question of how large (n) must be to provide a good approximation of the probability. To address these issues, Russian mathematician A.N. Kolmogorov introduced the axiomatic approach, which does not rely on precise definitions but instead establishes probability based on a set of fundamental axioms or postulates.

10.5.3 Axiomatic Approach

The axiomatic approach to probability, introduced by A.N. Kolmogorov, provides a formal framework based on a set of foundational axioms rather than specific definitions. This approach overcomes the limitations of earlier methods by establishing universally accepted rules that apply to all probability calculations, making it applicable to a wide range of theoretical and practical scenarios.

Whole field of probability theory is based on the following three axioms

1. Probability of an event, $p(E)$ lies between 0 and 1. That is $0 \leq p(E) \leq 1$

2. Probability of entire sample space is 1. That is $p(S) = 1$
3. If A and B are mutually exclusive events then probability of occurrence of either A or B is denoted by $p(A \cup B)$ shall be given by $p(A \cup B) = p(A) + p(B)$

> **Note**
>
> The probability p of the occurrence of an event is referred to as the *probability of success*, while the probability q of its non-occurrence is known as the *probability of failure*. Both p and q are non-negative and cannot exceed unity, *i.e.*, $0 \leq p \leq 1$ and $0 \leq q \leq 1$. This means the probability of an event always lies between 0 and 1, inclusive. The probability of an impossible event is 0, and the probability of a certain event is 1. For instance, if $p(A) = 1$, event A is guaranteed to occur, and if $p(A) = 0$, event A cannot occur. Additionally, the number of favorable outcomes m for an event cannot exceed the total number of outcomes n.

Figure 10.4: Probability always lies between 0 and 1

Example 10.6: In a simultaneous toss of two coins, find the probability of (i) getting 2 heads. (ii) exactly 1 head?

Solution 10.6: Here, the possible outcomes are HH, HT, TH, TT. *i.e.*, Total number of possible outcomes = 4.

(i) Number of outcomes favorable to the event (2 heads *i.e.*,HH) = 1. p(2 heads) = $\frac{1}{4}$.

(ii) Now the event consisting of exactly one head has two favourable cases, namely HT and TH. p(exactly one head) = $\frac{2}{4} = \frac{1}{2}$

Example 10.7: In a single throw of two dice, what is the probability that the sum is 9?

Solution 10.7: The number of possible outcomes is $6 \times 6 = 36$.

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)

(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)

(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)

(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)

(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)

(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

Let the event A be sum the is 9. Four outcomes are there with sum 9, they are (5,4), (6,3), (3,6), (4,5). $p(A) = \frac{4}{36} = \frac{1}{9}$

Example 10.8: From a bag containing 10 red, 4 blue and 6 black balls, a ball is drawn at random. What is the probability of drawing (i) a red ball? (ii) a blue ball? (iii) not a black ball?

Solution 10.8: There are 20 balls in all. So, the total number of possible outcomes is 20

(i) Number of red balls = 10, p(getting a red ball) $= \frac{10}{20} = \frac{1}{2}$

(ii) Number of blue balls = 4, p(getting a blue ball) $= \frac{4}{20} = \frac{1}{5}$

(iii) Number of balls which are not black = 14, p(not a black ball) $= \frac{14}{20} = \frac{7}{10}$

10.6 Event relations

In probability theory, relationships between events help in analyzing their combined or individual outcomes within a sample space. These relations include concepts such as union, intersection, complement and mutual exclusivity, which are fundamental for understanding how events interact and influence probabilities.

Complement of an event

In probability, the complement of an event refers to all outcomes in the sample space that are not favorable to the given event. If A is an event, its complement is denoted by A^c, and it represents the occurrence of all outcomes where A does not happen.

For example, consider the experiment of tossing a die. Let A be the event of getting an even number. The favorable outcomes for A are $\{2, 4, 6\}$. The remaining outcomes, $\{1, 3, 5\}$, do not satisfy A and are therefore the complement of A. These outcomes represent the occurrence of A^c, where A does not occur, therefore in this case $A^c = \{1, 3, 5\}$.

The complement of an event is essential in probability calculations and is related to the event by the formula:

$$P(A^c) = 1 - P(A) \tag{10.4}$$

The relationship in Equation 10.4 highlights that the probability of an event and its complement together always sum to 1. *i.e.* $P(A) + P(A^c) = 1$

Event A or B

Denoted as '$A \cup B$', spelled as A union B, represents the occurrence of either event A, event B or both. Let us consider the example of throwing a die. Suppose A is an event of getting a multiple of 2 and B be another event of getting a multiple of 3. The outcomes 2, 4 and 6 are favourable to the event A and the outcomes 3 and 6 are favourable to the event B *i.e.* $A = \{2, 4, 6\}$, $B = \{3, 6\}$ then $A \cup B = \{2, 3, 4, 6\}$.

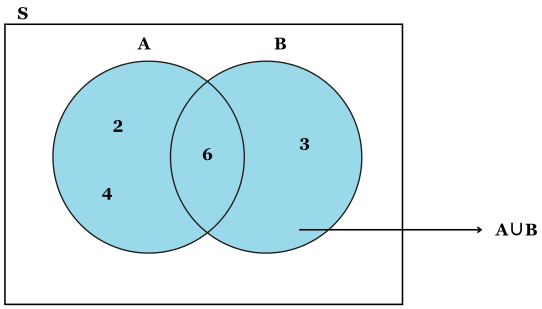


Figure 10.5: Venn diagram showing union of two events

Event A and B

Denoted as ' $A \cap B$ ' spelled as A intersection B, represents the occurrence of both events A and B simultaneously. It includes only those outcomes that are common to both events. For example, on throwing a die in which A is the event of getting a multiple of 2 and B is the event of getting a multiple of 3. The outcomes favorable to A are 2, 4, 6 and the outcomes favorable to B are 3 and 6. Here 6 is present in both A and B so here $A \cap B = 6$. Figure 10.6 below shows the venn diagram of intersection of two events.

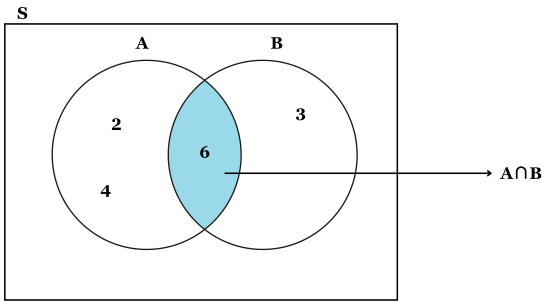


Figure 10.6: Venn diagram showing intersection of two events

10.7 Additive law of probability

According to additive law of probability, for any two events A and B of a sample space S, the probability of the union of two events A and B is equal to the sum of their individual probabilities, minus the probability of their intersection.

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \tag{10.5}$$

For mutually exclusive case $p(AB) = 0$; in that case:

$$p(A \cup B) = p(A) + p(B) \tag{10.6}$$

Example 10.9: A card is drawn from a well-shuffled deck of 52 cards. What is the probability that it is either a spade or a king?

Solution 10.9: If A denotes the event of drawing a 'spade card'. B denotes the events of drawing a 'king' respectively. The event A consists of 13 sample points, whereas the event B consists of 4 sample points. $p(A) = \frac{13}{52}$, $p(B) = \frac{4}{52}$, $p(A \cap B) = \frac{1}{52}$, $p(A \cup B) = p(A) + p(B) - p(AB) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}$

Example 10.10: In a single throw of two dice, find the probability of a total of 9 or 11.

Solution 10.10: Let the events A = a total of 9 and B = a total of 11. Events are mutually exclusive $A \cap B = 0$. Now there are four such cases were sum = 9, such as (3, 6), (4, 5), (5, 4), (6, 3); therefore $p(A) = \frac{4}{36}$. Similarly there are two cases were sum = 11, such as (5, 6), (6, 5); therefore $p(B) = \frac{2}{36}$. Thus from Equation 10.6, $p(A \cup B) = \frac{4}{36} + \frac{2}{36} = \frac{6}{36} = \frac{1}{6}$

10.8 Conditional probability

Conditional probability measures the likelihood of an event occurring, given that another event has already occurred. If A and B are two events, the conditional probability of A given B is denoted by $P(A \mid B)$ and is defined as:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}, \quad \text{provided } p(B) > 0 \tag{10.7}$$

Equation 10.7 gives the probability of A happening under the condition that B has occurred.

Example 10.11: If a card is drawn from a standard deck of 52 cards, what is the probability that it is a king, given that the card drawn is a face card?

Solution 10.11: Let A be the event of drawing a king, B be the event of drawing a face card (king, queen, or jack). The probability of B, drawing a face card, is $p(B) =$

$\frac{\text{Number of face cards}}{\text{Total cards}} = \frac{12}{52}$. The probability of $A \cap B$, drawing a King that is also a face card, is:$p(A \cap B) = \frac{\text{Number of Kings}}{\text{Total cards}} = \frac{4}{52}$. Using Equation 10.7 the conditional probability of drawing a king given that the card is a face card is $p(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{4}{52}}{\frac{12}{52}} = \frac{4}{12} = \frac{1}{3}$.

10.9 Multiplication law of probability

The multiplication law of probability states that if A and B are two events, the probability of both events occurring (*i.e.*, $A \cap B$) is given by:

$$p(A \cap B) = p(A) \cdot p(B \mid A) \tag{10.8}$$

This formula expresses the relationship between the joint probability of two events and their conditional probability. If A and B are independent events, then $p(B \mid A) = p(B)$, and the formula simplifies to:

$$p(A \cap B) = p(A) \cdot p(B) \tag{10.9}$$

Example 10.12: What is the probability of drawing two aces in succession from a standard deck of 52 cards?

Solution 10.12: Let A be the event that the first card drawn is an ace. B be the event that the second card drawn is an ace. The probability of A (drawing an ace on the first draw) is $p(A) = \frac{4}{52}$. If the first card is an ace, there are only 3 aces left in a deck of 51 cards. The probability of B (drawing an ace on the second draw given that the first card was an ace) is $p(B \mid A) = \frac{3}{51}$. Now using Equation 10.8 the probability of drawing two aces in succession is given by $p(A \cap B) = p(A) \cdot p(B \mid A) = \frac{4}{52} \cdot \frac{3}{51} = \frac{12}{2652} = \frac{1}{221}$

Example 10.13: A die is tossed twice. Find the probability of a number greater than 4 on each throw?

Solution 10.13: Let A be the event that 'a number greater than 4' on first throw. B be the event that 'a number greater than 4' in the second throw. Clearly A and B are independent events. In the first throw, there are two outcomes, namely, 5 and 6 favourable to the event A. $p(A) = \frac{2}{6} = \frac{1}{3}$. Similarly in the second throw, there are two outcomes, namely, 5 and 6 favourable to the event B, therefore $p(B) = \frac{2}{6} = \frac{1}{3}$. Now by Equation 10.6, probability to get a number greater than 4 on each throw is given by $p(A \cap B) = p(A).p(B) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$

10.10 Probability using combinations

Combinations can be used to calculate the total number of possible outcomes in a probability problem. The formula for combinations is given by:

$$^{n}C_{r} = \frac{n!}{r!(n-r)!} \tag{10.10}$$

Example 10.14: Calculate $^{3}C_{2}$

Solution 10.14: $^{3}C_{2} = \frac{3!}{2!(3-2)!} = \frac{3\times2\times1}{2\times1} = 3$

Example 10.15: A bag contains 3 red, 6 white, and 7 blue balls. What is the probability that two balls drawn are white and blue?

Solution 10.15: Total number of balls $= 3 + 6 + 7 = 16$. The number of ways to draw 2 balls from 16 is given by $^{16}C_{2} = 120$. The number of ways to draw one white ball from 6 white balls is $^{6}C_{1} = 6$, and the number of ways to draw one blue ball from 7 blue balls is $^{7}C_{1} = 7$. Since the events are independent, the total number of favorable outcomes is $^{6}C_{1} \times ^{7}C_{1} = 6 \times 7 = 42$. Thus, the required probability is:p(one white and one blue) $= \frac{^{6}C_{1}\times^{7}C_{1}}{^{16}C_{2}} = \frac{42}{120} = \frac{7}{20}$

Example 10.16: A bag contains 5 red and 4 black balls. What is the probability that both balls drawn are red?

Solution 10.16:Total number of balls $= 5 + 4 = 9$. The number of ways to draw 2 balls from 9 is given by $^{9}C_{2} = \frac{9\times8}{2\times1} = 36$. The number of ways to draw 2 red balls from 5 red balls is $^{5}C_{2} = \frac{5\times4}{2\times1} = 10$. Thus, the required probability is: p(both red balls) $= \frac{^{5}C_{2}}{^{9}C_{2}} = \frac{10}{36} = \frac{5}{18}$

10.11 Bayes' theorem

Bayes' theorem gives a mathematical rule for inverting conditional probabilities, allowing one to find the probability of a cause given its effect.

Let $E_1, E_2, \ldots, E_n$ be a set of events associated with a sample space S, where all the events $E_1, E_2, \ldots, E_n$ have non-zero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem,

$$p(E_i \mid A) = \frac{p(E_i) \cdot P(A \mid E_i)}{\sum_{k=1}^{n} p(E_k) \cdot p(A \mid E_k)} \tag{10.11}$$

The following terminologies are commonly used when applying Bayes' theorem:

Hypotheses: The events $E_1, E_2, \ldots, E_n$ are called the hypotheses.

Prior probability: The probability $p(E_i)$ is considered the prior probability of hypothesis E_i.

Posterior probability: The probability $p(E_i \mid A)$ is considered the posterior probability of hypothesis E_i.

For any two events A and B, the formula for the Bayes theorem is given by:

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B)}, p(B) \neq 0 \tag{10.12}$$

Where $p(A)$ and $p(B)$ are the probabilities of events A and B. $p(A \mid B)$ is the probability of event A given B. $p(B \mid A)$ is the probability of event B given A.

Example 10.17: A bag I contains 4 white and 6 black balls while another bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags, and it is found to be black. Find the probability that it was drawn from bag I.

Solution 10.17: Let E_1 be the event of choosing bag I, E_2 the event of choosing bag II, and A be the event of drawing a black ball. $p(E_1) = p(E_2) = \frac{1}{2}$. Also,probability of drawing a black ball from bag I is given by $p(A \mid E_1) = \frac{6}{10} = \frac{3}{5}$. Probability of drawing a black ball from bag II $p(A \mid E_2 =) \frac{3}{7}$.

By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

$$p(E_1 \mid A) = \frac{p(E_1) \cdot P(A \mid E_1)}{p(E_1) \cdot P(A \mid E_1) + p(E_2) \cdot P(A \mid E_2)}$$

$$= \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{1}{2} \times \frac{3}{5} + \frac{1}{2} \times \frac{3}{7}} = \frac{7}{12}$$

Example 10.18: A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the number obtained is a four. Find the probability that the number obtained is actually a four.

Solution 10.18: Let A be the event that the man reports that number four is obtained. Let E_1 be the event that four is obtained and E_2 be its complementary event. Then, $p(E_1) =$ probability that four occurs $= \frac{1}{6}$. $p(E_2) =$ probability that four does not occur $= 1 - p(E_1) = 1 - \frac{1}{6} = \frac{5}{6}$. Also, $p(A|E_1) =$ probability that man reports four and it is actually a four $= \frac{2}{3}$ (because man speaks truth 2 out of 3 times). $p(A|E_2) =$Probability that man reports four and it is not a four $= \frac{1}{3}$.

By using Bayes' theorem, probability that number obtained is actually a four, $p(E_1 \mid A) = \frac{p(E_1) \cdot P(A|E_1)}{p(E_1) \cdot P(A|E_1) + p(E_2) \cdot P(A|E_2)}$.

$$p(E_1 \mid A) = \frac{\frac{1}{6} \times \frac{2}{3}}{\frac{1}{6} \times \frac{2}{3} + \frac{5}{6} \times \frac{1}{3}} = \frac{2}{7}$$

Try yourself

Six cards are drawn at random from a pack of 52 cards. What is the probability that 3 will be red and 3 black?

Historical Insights

The gambler's fallacy

In 1913, a group of gamblers gathered at a casino in Monte Carlo, where they witnessed a highly unusual event at the roulette table. The ball landed on black 26 times in a row—an incredibly rare streak. As the streak continued, more and more gamblers bet heavily on red, convinced that red was "due" to appear. But to their shock, black kept winning, and many lost fortunes that night. This event is now a famous example of the *gambler's fallacy* also known as the *Monte Carlo fallacy* or *the fallacy of the maturity of chances*, the mistaken belief that past events influence future outcomes in independent events, like the spin of a roulette wheel. In reality, the probability of the ball landing on black or red remains the same on every spin, regardless of previous results.

Quotes to Inspire

"The theory of probabilities is at bottom nothing but common sense reduced to calculus." – Pierre-Simon Laplace

11 Appendix 1

Source of data related to different sectors

Sl No	Data Particulars	Source	Organisation
1.	Estimates of area, production of important crops in India	https://agriwelfare.gov.in/en/AgricultureEstimates	MoAFW, GOI
2.	State level, district level aggregates of Area, production and productivity of principal crops in Kerala	https://www.ecostat.kerala.gov.in	Department of Economics & Statistics, Govt of Kerala
3.	District Wise Birth & Death Data, State Level Birth & Death Data	https://www.ecostat.kerala.gov.in	Department of Economics & Statistics, Govt of Kerala
4.	Minimum Support Price (MSP) Statement	https://desagri.gov.in/statistics-type/latest-minimum-support-price-msp-statement/	MoAFW, GOI
5.	Annual Survey of Industries, Index of Industrial Production, Household Consumer Expenditure, Economic Census, Enterprises Surveys, Periodic Labour Force Survey, CPI, etc	https://www.mospi.gov.in/	Ministry of Statistics and Programme Implementation

Sl No	Data Particulars	Source	Organisation
6.	State/UT wise estimates on population, health, family planning and nutrition related key indicators like fertility, mortality, maternal, child and adult health, women and child nutrition, domestic violence, etc	https://main.mohfw.gov.in/	Ministry of Health & Family Welfare
7.	Commodity wise export import data	https://tradestat.commerce.gov.in, https://ftddp.dgciskol.gov.in	Ministry of Commerce and Industry
8.	Data on various aspects of Indian economy, banking and finance	https://www.rbi.org.in	Reserve Bank of India
9.	Sustainable Development Goal report	https://www.niti.gov.in/	NITI Aayog

12 References

Ball, Philip. 2004. *Critical Mass.* Farrar, Straus; Giroux.

Fiori, Anna M., and Michele Zenga. 2009. "Karl Pearson and the Origin of Kurtosis." *International Statistical Review* 77 (1): 40–50. https://doi.org/10.1111/j.1751-5823.2009.00076.x.

Fisher, Ronald A. 1918. "The Correlation Between Relatives on the Supposition of Mendelian Inheritance." *Transactions of the Royal Society of Edinburgh* 52 (2): 399–433. https://doi.org/10.1017/S0080456800012163.

Goon, Gupta, A. M., and B Dasgupta. 1983. *Fundamentals of Statistics. Vol. I.* TheWorld Press.

Gupta, S. C., and V. K. Kapoor. 1997. *Fundamentals of Mathematical Statistics.* Sulthan Chand Publications, New Delhi.

Pearson, Karl. 1894. "Contributions to the Mathematical Theory of Evolution." *Philosophical Transactions of the Royal Society of London. A* 185: 71–110. https://doi.org/10.1098/rsta.1894.0003.

Pratheesh P. Gopinath, Brigit Joseph, Rajender Parsad. 2020. *GRAPES: General r-Shiny Based Analysis Platform Empowered by Statistics* (version 1.0.0). https://doi.org/10.5281/zenodo.4923220.

Team, R Core. 2024. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing. https://www.R-project.org/.

Wei, Taiyun, and Viliam Simko. 2021. "Corrplot: Visualization of a Correlation Matrix." https://cran.r-project.org/web/packages/corrplot/corrplot.pdf.