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 postgraduate 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.

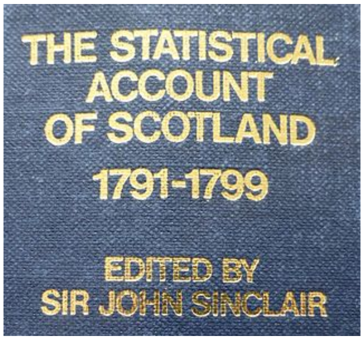


Figure 1.1: Statistical Account of Scotland by Sir John Sinclair (1791)

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

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

predicted values in a statistical model, accounting for the inherent variability and uncertainty in biological phenomena.

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 ($\pi = 3.14159...$), 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.

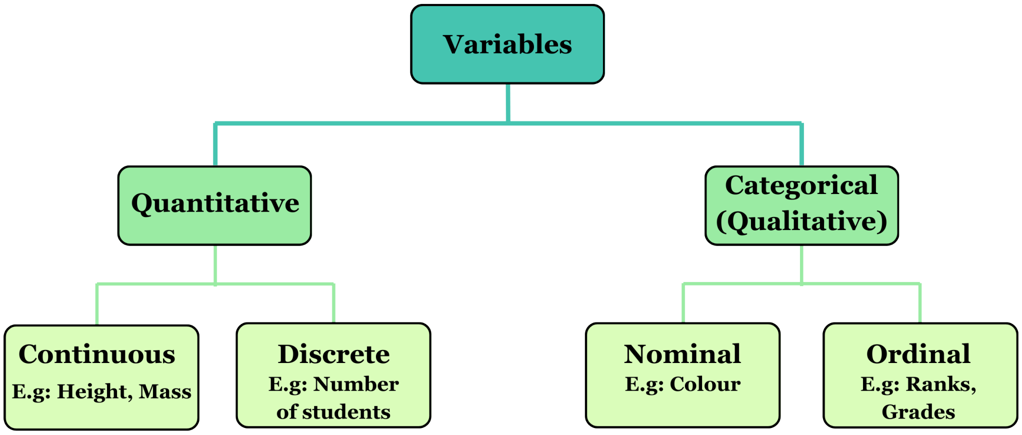


Figure 1.2: Classification of variables

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

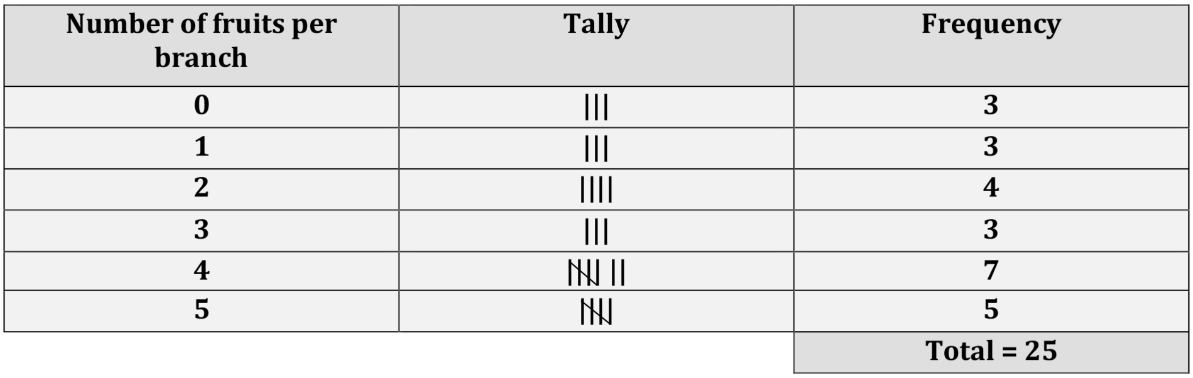
Semi-official sources include publications from institutions like the State Bank, Railway Board, Central Cotton Committee, and Boards of Economic Enquiry. It also encompasses reports from trade associations, chambers of commerce, technical journals, trade publications, and research organizations such 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.

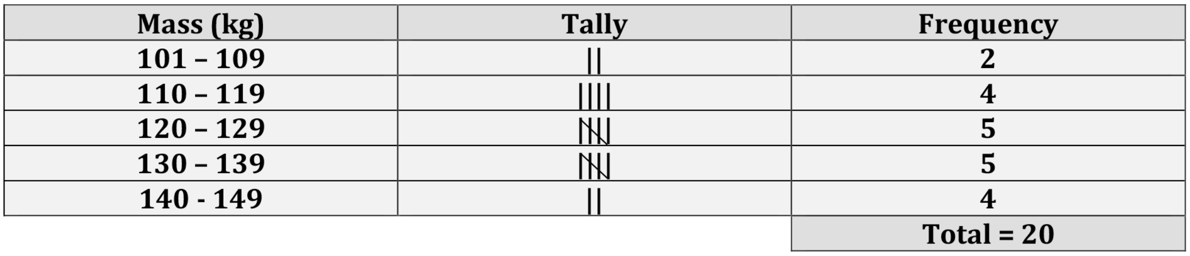
1. List all values of the variable in ascending order of magnitude.
2. Form a tally column, that is, for each value in the data, record a stroke in the tally column next to that value. In the tally, each fifth stroke is made across the first four. This makes it easy to count the entries and enter the 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	卌	5
130 – 139	卌	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.

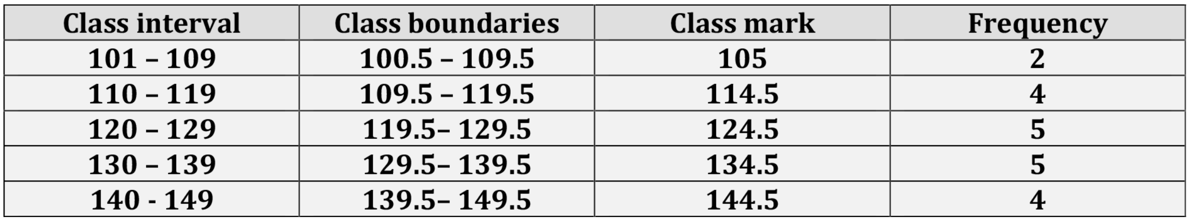
Continuous classes
Continuous classes are intervals where the class limits represent a continuous range of values, with no gaps between the intervals.

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.

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

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 .

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.

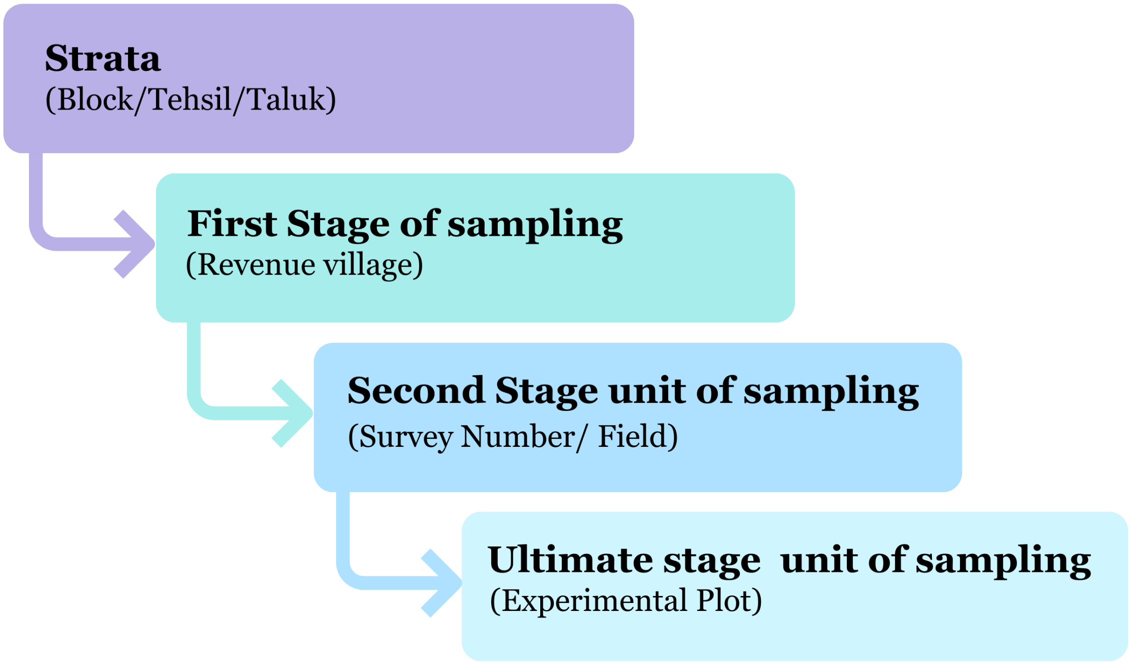

Strata
(Block/Tehsil/Taluk)
First Stage of sampling
(Revenue village)
Second Stage unit of sampling
(Survey Number/ Field)
Ultimate stage unit of sampling
(Experimental Plot)

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

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.

Table 3.1 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

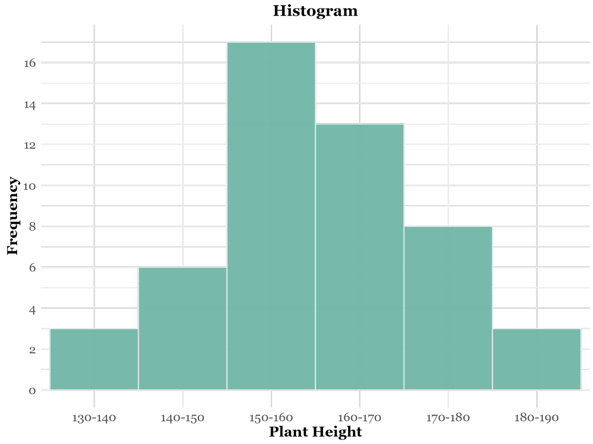


Figure 3.1: Histogram

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 3.1. 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

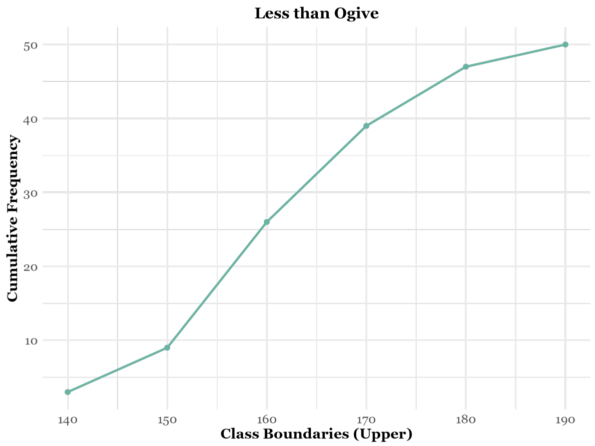


Figure 3.2: Less than Ogive

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 3.1. The greater than ogive, shown in Figure 3.3, is drawn using Table 3.3.

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

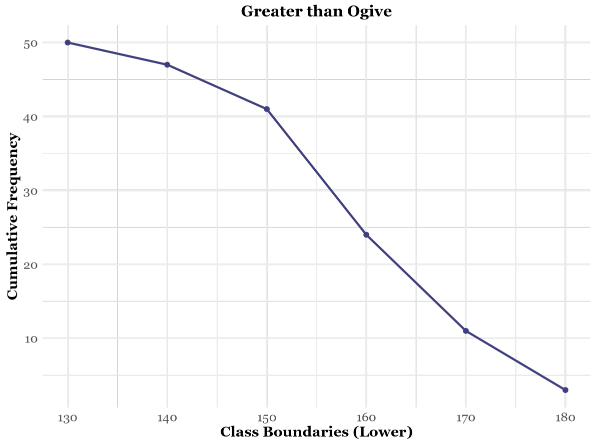


Figure 3.3: Greater Than Ogive

! 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 3.1. The frequency polygon, created using Table 3.4, is shown in Figure 3.4. The relation between frequency polygon and histogram can be seen in Figure 3.5

Table 3.4: Midpoints and frequencies

Class Midpoints	135	145	155	165	175	185
Frequencies	3	6	17	13	8	3

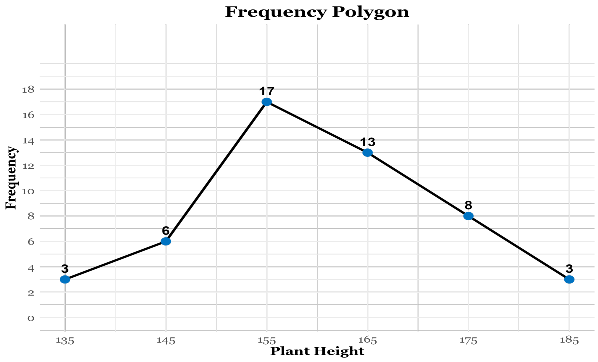


Figure 3.4: Frequency Polygon

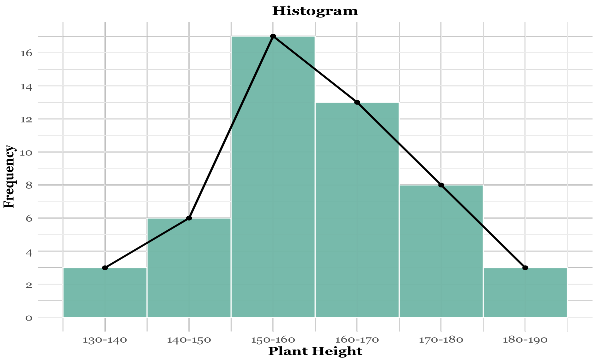


Figure 3.5: Frequency Polygon and Histogram

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

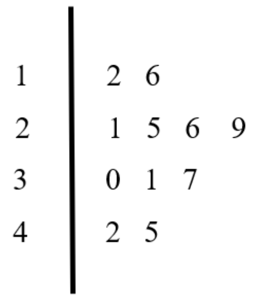


Figure 3.6: Stem and Leaf plot

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 3.7

Table 3.5: hypothetical data on coconut production

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

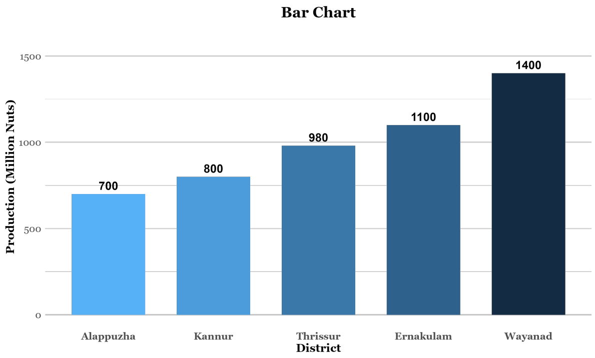
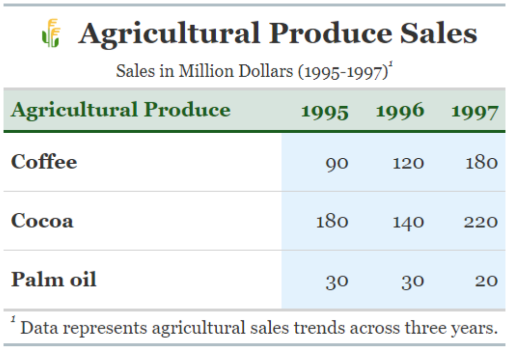


Figure 3.7: Barchart

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. This is illustrated in the following example.

Figure 3.8 shows the distribution of sales of agricultural produce from a Farm in 1995, 1996 and 1997 and its corresponding component barchart in Figure 3.9.



🌾 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 3.8: Sales data of agricultural produce

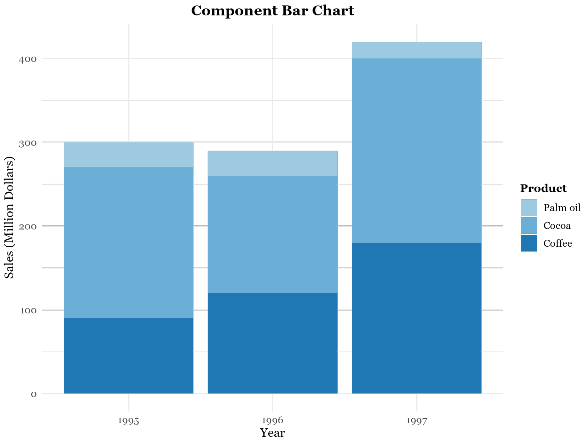


Figure 3.9: Component Barchart

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 3.8 can also be represented using a grouped bar chart shown in Figure 3.10. 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 both within and across groups.

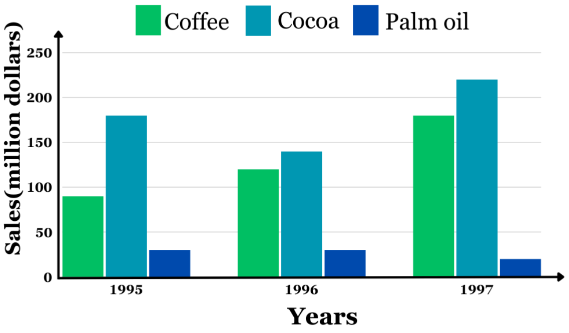


Figure 3.10: Grouped bar chart

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 barchart

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 charts

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 base on this data is shown in Figure 3.11

Table 3.7: Hypothetical data on the production of different commodities

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

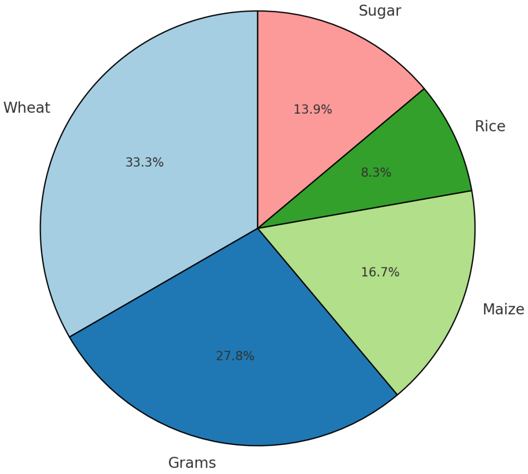


Figure 3.11: Piechart

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 3.12 below shows the parts of a box plot

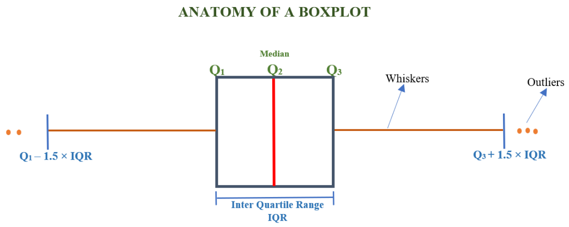


Figure 3.12: Anatomy of box plot

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 3.12 to 3.23 you can see a few popular and advanced graph types widely used in modern data analysis.

Quotes to Inspire

"Statistics is the grammer of science"
- Karl Pearson

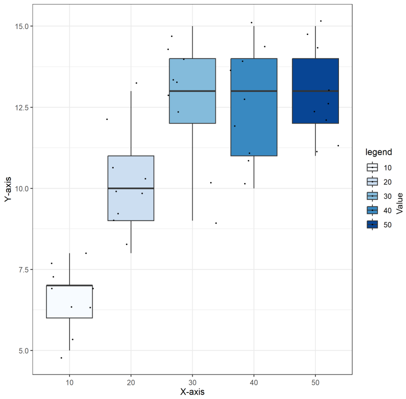


Figure 3.13: Box Plot

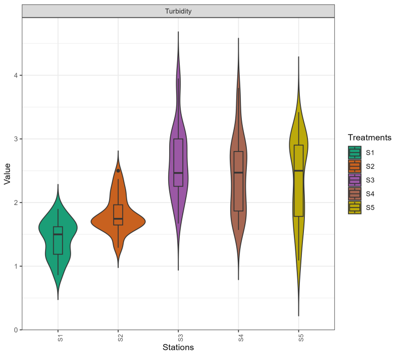


Figure 3.14: Violin Plot

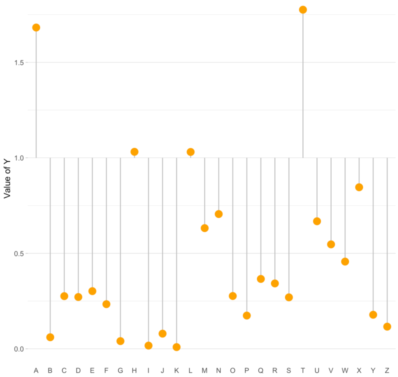


Figure 3.15: Lollipop Plot

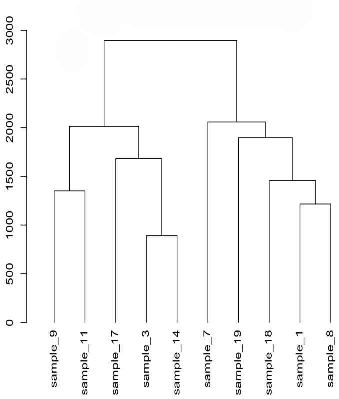


Figure 3.16: Dendrogram

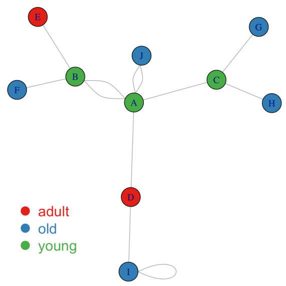


Figure 3.17: Network Graph

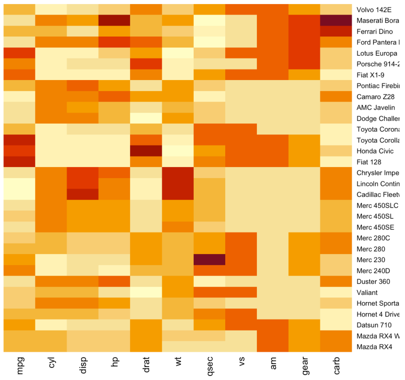


Figure 3.18: Heat Map

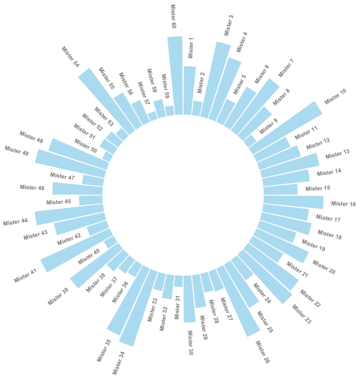


Figure 3.19: Circular Bar Plot

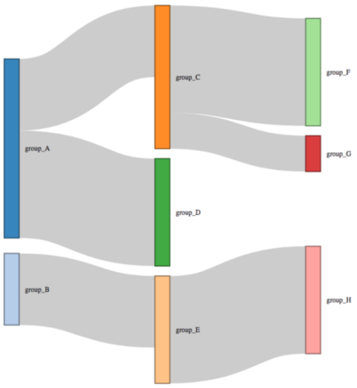


Figure 3.20: Sankey Diagram

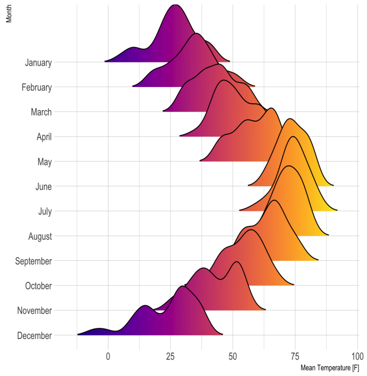


Figure 3.21: Ridgeline Plot

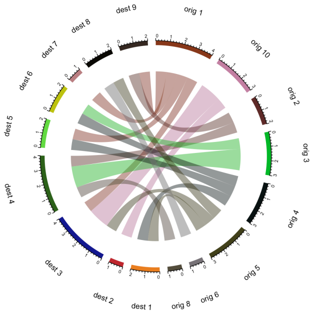


Figure 3.22: Chord Diagram

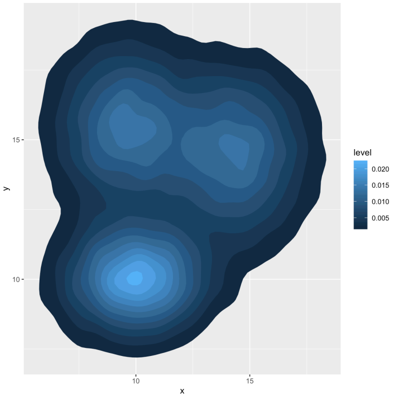


Figure 3.23: Density Plot

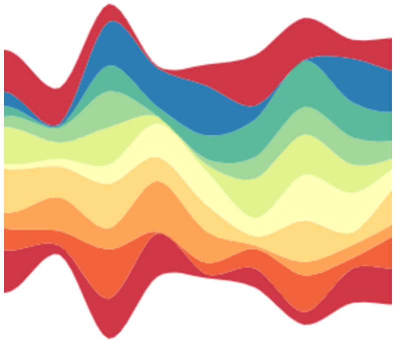


Figure 3.24: Stream Graph

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**. These measures provide insights into the central value of the data, making them fundamental tools for understanding and interpreting data distributions. We will discuss these measures in two sections, here we will discuss on simple averages.

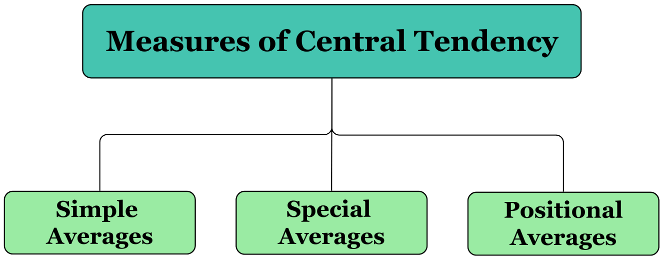


Figure 4.1: Measures of central tendency

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

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

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

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

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

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

Example 4.2 Table 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

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

<u>Assumed mean method (Indirect method)</u>

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

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

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 Example 4.2

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

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

$\overline{x} = 161 + \frac{4}{50} = 161.08$ 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:

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

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 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

Marks	Class mark($\mathbf{x_i}$)	Frequency($\mathbf{f_i}$)	$\mathbf{f_i x_i}$
60-65	62.5	2	125
65-70	67.5	15	1012.5
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.

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

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 Example 4.3, see Table 4.4

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

Marks	Class mark($\mathbf{x_i}$)	Frequency($\mathbf{f_i}$)	$\mathbf{u_i = \frac{x_i - 72.5}{5}}$	$\mathbf{f_i u_i}$
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
		$\sum_{i=1}^{k} f_i = 60$		$\sum_{i=1}^{k} f_i u_i = 3$

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

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.

4.2 The Median

The median of a set of data is defined as the middle value when the data is arranged in order of magnitude. If there are no ties, half of the observations will be smaller than the median, and half of the observations will be larger than the median. The median can be the middle most item that divides the group into two equal parts, one part comprising all values greater, and the other, all values less than that item. It is a positional measure.

4.3 Median of ungrouped or raw data

Arrange the given n observations x_1 ,$x_2, \dots , x_n$ 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 4.4 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 4.4

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^{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 is 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 4.5: The following is a frequency table 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 4.5:

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. To find out the 12^{th} position, we use less than cumulative frequencies as shown:

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

The 12th position is after the 7th position but before the 14th position. 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 4.6: The table 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 4.6:

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 25th position and 26th position, we add up the frequencies as shown:

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 25th position is 2 and the mark at the 26th position is 3. The median is the average of the scores at 25th and 26th 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:

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

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 $\frac{\mathrm{N}}{2}$ 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.

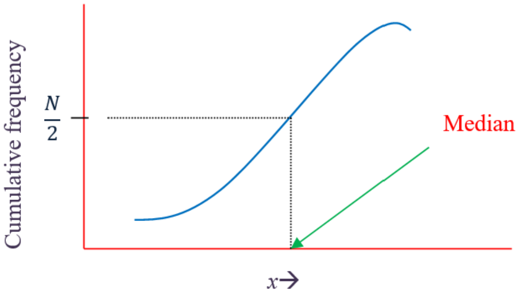


Figure 4.2: median from a cumulative frequency curve

Example 4.7 Table 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 4.7

(i) Linear interpolation method

$N = 60$

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

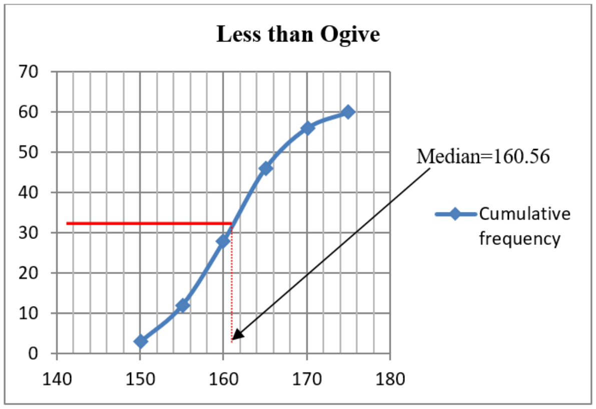
F = sum of all frequencies below 160(cumulative frequency) = 16+9+3= 28

f_m = frequency of the median class=18

C = class width of the median class=5

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

(ii) From a cumulative frequency curve



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 common value. 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 4.8

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 4.9 20 patients selected at random had their blood groups determined. The results are given in the table below

Table 4.12: Blood group of 20 patients

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

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

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

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 4.10 For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the mode.

Table 4.13: frequency distribution of weights of sorghum ear heads

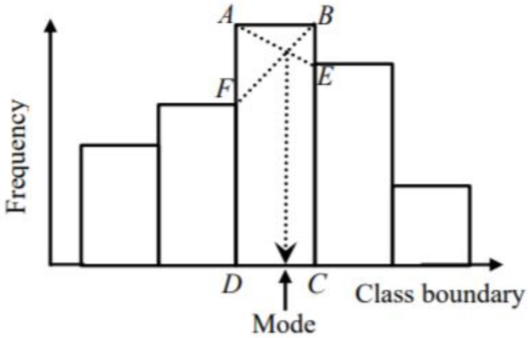
Weights of ear heads (g)	No of ear heads (f)
60-80	22
80-100	38
100-120	45
120-140	35
140-160	20

Modal class is **100-120**

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

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.



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. That is, 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.

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** 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$^{\text{th}}$ 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 $x_1, x_2, \ldots,\ x_n$are observations then

$$\text{Geometric mean, } \mathbf{GM} = \sqrt[\mathbf{n}]{\mathbf{x_1 x_2 \ldots x_n}}$$

$$= \left(\mathbf{x_1 x_2 \ldots x_n}\right)^{\frac{\mathbf{1}}{\mathbf{n}}}$$

$$\text{logGM}{=}\frac{\mathbf{1}}{\mathbf{n}}\log\left(\mathbf{x_1 x_2 \ldots x_n}\right)$$

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

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

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

5.1.1 Geometric mean for grouped frequency table data

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

where x_i is the mid-value, f_i is the frequency , 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?

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

Geometric mean=

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

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

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

note: here Antilog $(x) = 10^x$ *i.e.*

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

Example 5.2: Geometric mean of a Frequency Distribution

Weight of ear head (x)	Frequency(f)	log(x)	flog(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

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

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

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

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

Example 5.3: Geometric mean of a Grouped Frequency Distribution

Class	Mid value (x)	Frequency(f)	log(x)	flog(x)
60-80	70	5	1.845	9.225
80-100	90	4	1.954	7.817
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 5.4:
Here n =32

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

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

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

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

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.

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

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

In case of Frequency distribution

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

where x_i is the mid-value, f_i is the frequency , k is the number of classes

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 $n = 5$

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

$$\text{H.M} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{5}{0.433824} = 11.525$$

Example 5.5: Number of tomatoes per plant are given below. Calculate the harmonic mean.

No. of Tomato per plants	20	21	22	23	24	25
No. of Plants	4	2	7	1	3	1

Solution 5.5:

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

$$\text{H.M} = \frac{n}{\sum_{i=1}^{n} f_i \frac{1}{x_i}} = \frac{18}{0.821898} = 21.90$$

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 Relation between AM, GM and HM

If AM stands for Arithmetic Mean, GM stands for Geometric Mean and HM stands for Harmonic Mean; then

$$\text{AM}\times\text{HM}=\text{GM}^{\mathbf{2}}$$

also

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

5.4 AM, GM or HM ?

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—it's a direct measurement.

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.

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.

5.5 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.

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

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

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

Example 5.6: Compute quartiles for the data 25, 18, 30, 8, 15, 5, 10, 35, 40, 45

Solution 5.6:

First arrange the data in ascending order

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

here $n = 10$

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

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

$Q_1 = 2.75^{\text{th}}$ item = 2^{nd} item + $0.75(3^{\text{rd}}$ item – 2^{nd} item)

So from the given data Q_1= 8+0.75(10– 8) = **9.5**

$\mathbf{Q_2}$**= median**

here Q_2 =(18+25)/2 = **21.5**

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

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

5.6.1 Quartiles of a discrete frequency data

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

Example 5.7: Compute quartiles for the data given bellow

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

Solution 5.7:

x	f	cf
5	4	4
8	3	7
12	2	9
15	4	13
19	5	18
24	2	20
30	4	24

Here n =24

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

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

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

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

5.6.2 Quartiles of a continuous frequency data

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

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

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

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 3^{rd} quartile class

f_3= frequency of the 3^{rd} quartile class

c_3= width of the 3^{rd} quartile class

m_3 = cumulative frequency preceding the 3^{rd} quartile class

Example 5.8: Find the quartiles for the grouped frequency data given

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

Class	frequency	cumulative frequency
80-90	12	194
90-100	10	204

Solution 5.8:

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

The cumulative frequency value just greater than 51 is 54 so the class 20-30 is the 1$^{\text{st}}$ quartile class

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

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

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

The cumulative frequency value just greater than 153 is 167 so the class 60-70 is the 3$^{\text{rd}}$ quartile class

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

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

5.7 Percentiles

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

For example:
- The **50th percentile** is equivalent to the **median**, representing the middle value of the dataset.
- The **25th percentile** corresponds to the first quartile (Q_1), which marks the lower 25% of the data.
- The **75th 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

$\mathbf{P_x}= \left(\frac{\mathbf{x(n+1)}}{\mathbf{100}}\right)^{\text{th}}$**item**

For a frequency distribution the x^{th} percentile is given by following steps

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

$$\mathbf{P_x}= \mathbf{l}+\frac{\left(\frac{\mathbf{x\times n}}{\mathbf{100}}\right)-\mathbf{cf}}{\mathbf{f}}\times\mathbf{c}$$

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 5.9: Compute $\mathbf{P_{25}}$**and** $\mathbf{P_{75}}$ for the data 25, 18, 30, 8, 15, 5, 10, 35, 40, 45

Solution 5.9:

First arrange the data in ascending order

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

Here n =10

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

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

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

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

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

! Note

Data in Example 5.9 is same as Example 4.6; 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 **5.7** & **5.8**; verify that $P_{50} = Q_2$, $P_{25} = Q_1$ & $P_{75} = Q_3$

5.8 Deciles

Deciles are similar to quartiles, but while quartiles consist of three points that divide an ordered dataset into four equal parts, deciles consist of 9 points that divide the dataset into ten equal parts. The $\mathbf{x}^{\text{th}}$ decile is denoted as d_x. It is important to note that the **median** is the **5th decile**.

$\mathbf{d_x} = \left(\frac{\mathbf{x(n+1)}}{\mathbf{10}}\right)^{\text{th}}$ **item**

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

1. Find cumulative frequencies
2. Find $\left(\frac{x.n}{10}\right)$
3. See in the cumulative frequencies, the value just greater than$\left(\frac{x.n}{10}\right)$and then the corresponding class interval is called **decile class**.
4. Use the following formula

$$\mathbf{d_x} = \mathbf{l} + \frac{\left(\frac{\mathbf{x \times n}}{\mathbf{10}}\right) - \mathbf{cf}}{\mathbf{f}} \times \mathbf{c}$$

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 **5.6**, **5.7** & **5.8**; verify that $\mathrm{D}_5 = \mathrm{Q}_2 = \mathrm{P}_{50} = \textit{median}$

Quotes to Inspire

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

6 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

7 References

Ball, Philip. 2004. *Critical Mass.* Farrar, Straus; Giroux.

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.