NATIONAL OPEN UNIVERSITY OF NIGERIA

STA111

DESCRIPTIVE STATISTICS

COMPILED COURSE MATERIAL

Six uploaded course files | Original learning pages preserved

Compiled for convenient student access by NOUN Update

NOUN Update is an independent educational support platform.

COURSE CONTENTS

SOURCE LESSONS AND START PAGES

Week One: Statistical Data

19 source pages • begins on PDF page 3

Module Two: Presentation of Data

23 source pages • begins on PDF page 22

Week Three: Measures of Central Tendency

20 source pages • begins on PDF page 45

Week Four: Geometric and Harmonic Mean

7 source pages • begins on PDF page 65

Week Five: Measures of Partition

18 source pages • begins on PDF page 72

Week Six + Module Five: Dispersion, Correlation and Regression

41 source pages • begins on PDF page 90

PUBLICATION NOTE

The original lesson pages are reproduced in sequence without editorial rewriting.
Any errors or inconsistencies in the source are not changed in this compiled copy.
For clarified worked examples and exam practice, use the separate NOUN Update CE&CS.

Week one: Statistical Data

- Unit 1: Definition, Scope and Limitation of Statistics
- Unit 2: Source of Data
- Unit 3: Method of Data Collection

Unit 1 Definition, Scope and Limitation of Statistics
1.1 Introduction
1.2 Learning Outcomes
1.3 Definition, Scope and Limitation of Statistics
1.3.1 Definition of Statistics
1.3.2 Scope of Statistics
1.3.3 Limitation of Statistics
1.4 Self-Assessment Exercises
1.5 Summary
1.6 References, Further Readings and Web Sources
1.7 Possible Answers to Self-Assessment Exercises



1.1 introduction

In this unit, you will be introduced to the basic concepts of statistics and how they are applied in real-world situations. You will learn how data is collected, organized, analyzed, and interpreted to draw meaningful conclusions. Statistics helps us to understand what is happening around us by giving us tools to handle data carefully and accurately.

This unit will guide you through the definition of statistics, its scope, and its limitations. These foundational ideas are important because they will help you understand what statistics can do, where it is useful, and what you should be careful about when using statistical methods.



1.2 Learning Outcomes

Before end of this unit students should be able to:

- Explain the meaning of statistics.
- Explain the Scope and limitations of Statistics
- Identify the Importance of Statistics



1.3 Definition, Scope and Limitation of Statistics

1.3.1 Definition of Statistics

Statistics is a scientific method that focuses on the collection, organization, analysis, interpretation, and presentation of data to aid in decision-making and problem solving. It provides methodologies for understanding patterns, making predictions, and drawing meaningful conclusions from numerical and categorical data.

Key Aspects of Statistics

1. Data Collection
    - Involves gathering raw information from various sources, such as surveys, experiments, or observations.
    - Methods include sampling techniques (random, stratified, systematic) and data acquisition from populations or subsets.
2. Organization and Classification
    - Raw data is structured into meaningful formats such as tables, frequency distributions, and grouped data for easier analysis.
    - Helps in identifying trends, patterns, and anomalies.
3. Data Analysis
    - Statistical techniques, including descriptive measures (mean, median, mode, variance, standard deviation), help summarize and understand data.
    - Advanced methods include probability theory, hypothesis testing, and regression analysis to infer deeper insights.
4. Interpretation

- Involves making sense of statistical findings and drawing conclusions based on evidence.
- Helps in decision-making processes, such as predicting future trends, assessing risks, or validating theories.

5. Presentation

- Data is communicated effectively through visual tools like charts, graphs, histograms, and reports to enhance understanding.
- Statistical reports are often used in business, economics, healthcare, engineering, and social sciences.

1.3.2 Scope of Statistics

Statistics is more than just a tool for collecting numerical data; it is also a means of developing sound techniques for handling, analyzing, and drawing valid conclusions from it. Statistics is used in every aspect of human activity, both social and physical, such as biology, commerce, education, planning, business management, information technology, and so on. It is difficult to think of a single area of human activity where statistics cannot be used.

Types of statistics

Statistics can be broadly divided into two, viz.: Descriptive Statistics and Inferential Statistics.

Descriptive Statistics:

Its primary concern is the description of large amounts of data. It includes data classification and diagrammatic representation of data such as drawing histograms, bar charts, pictograms, and so on. It also includes the computation of other descriptive statistics such as mean, median, mode, range, and variance, among others.

Inferential or Inductive Statistics

This aspect allows us to draw conclusions about the population based on samples drawn from the population. It is generalizing about the population based on some specific samples. In inferential statistics, we use samples to draw conclusions about the population from which the samples were drawn. Nature of Statistical Data

Quantitative Variables:

Statistics deals with numerical data. Business, economic, social, and scientific data can all be measured or counted directly. For example, data on daily sales, the value of a commodity's export or import, the number of completed buildings, the number of registered contractors, the quantity of commodity produced, wages, and so on. These are known as quantitative variables.

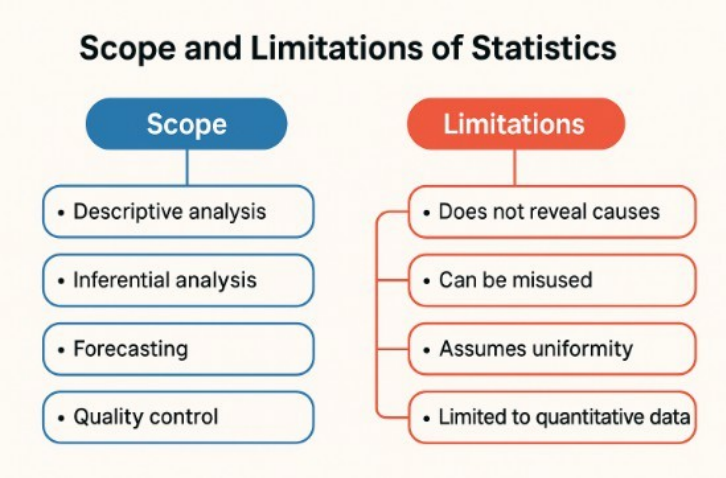
Qualitative Variables:

Information that is not numerical in nature, such as skin, eye, or hair color, level of education, sex, marital status, or building type. Although qualitative variables cannot be numerically as- signed, the process of assigning numerical values to qualitative variables is known as coding. Similarly, qualitative data can be arranged in descending order of importance and assigned val- ues in that order. This is known as ranking.

1.3.3 Limitation of Statistics

Statistics, despite its widespread application in all aspects of human activity, has its own set of limitations. Some of them are listed below.

1. Statistics is unsuitable for studying qualitative phenomena.
2. Statistics does not study individuals.
3. Statistical laws are not exact.
4. The statistics table could be misused.
5. Statistics is only one method of researching a problem.





1.4 Self-Assesement Exercises

1. What is the primary function of statistics?

A) Making laws B) Summarizing and analyzing data

C) Replacing experiments D) Predicting the future with certainty

2. Which of the following best defines statistics?

A) The study of human behavior B) The science of counting money

C) The science of collecting, organizing, analyzing, and interpreting data

D) A branch of mathematics dealing with algebra

3. One key feature of statistics is:
A) Guessing without data B) Decision-making under uncertainty
C) Proving hypotheses with 100% accuracy D) Avoiding numerical data

4. The scope of statistics does NOT include:
A) Economics B) Biology C) Music taste analysis D) Astrology

5. Which of the following is part of the statistical process?
A) Voting B) Data interpretation C) Cooking D) Painting

6. In which field is statistics commonly applied?
A) Mathematics only B) Sociology only
C) All fields involving data D) None of the above

7. Which of the following is a *limitation* of statistics?
A) It works well with emotions B) It relies on numerical data
C) It interprets moral values D) It excludes estimation

8. Why must one be cautious while interpreting statistical results?
A) Statistics are usually wrong B) Results are always emotional
C) Misleading conclusions may occur without context D) It is impossible to understand graphs

9. A major scope of statistics includes:
A) Arithmetic operations B) Logical fallacies
C) Collection and classification of data D) Writing fiction

10. Which of the following is NOT a limitation of statistics?
A) Requires specialized knowledge B) Cannot work without numbers
C) Used for all data types including qualitative D) Misuse can lead to wrong conclusions

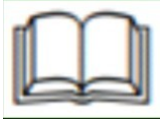


1.5 Summary

In this unit, we explored the foundational concepts of Statistics.We began by understanding that statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data. It helps us make informed decisions based on numerical evidence.

We also examined the scope of statistics, which spans across many fields such as business, economics, education, medicine, agriculture, engineering, and government. Whether used for research, policy-making, forecasting, statistics provides valuable tools to solve real-world prob- lems.

However, we must also recognize the limitations of statistics. While statistical methods are powerful, they depend heavily on the quality of data, the appropriateness of the methods used, and the honesty of interpretation. Misuse or misrepresentation of statistics can lead to incorrect conclusions.



1.6 References, Further Readings, Web Sources

1. Tmt. V.Varalakshmi, Thiru G.Gnana Sundaram, Tmt. B.Indrani, Tmt. N.Suseela and Tmt.S.Ezhilarasi (2004). Statistics for Higher Secondary – First Year.
2. Obasohan P.E and Abd-elfatai A.M (2012). An Introduction to Statistical Method.
3. Allan G. Bluman (2012).Elementary of Statistics *Eighth edition*



1.7 Possible Answers to SAEs

1.B, 2.C, 3.B, 4.D, 5.B, 6.C, 7.B, 8.C, 9.C, 10.C

Unit 2 Source of Data

2.1 Introduction
2.2 Learning Outcomes
2.3 Source of Data
2.3.1 Primary Source
2.3.2 Secondary Source
2.4 Self-Assessment Exercises
2.5 Summary
2.6 References, Further Readings, and Web Sources
2.7 Possible Answers to Self-Assessment Exercises



2.1 Introduction

In the field of statistics and research methodology, data is the foundation of all analysis and inference.Without data, statistical methods cannot be applied, and conclusions cannot be drawn. Therefore, understanding the source of data is fundamental to any study, as it determines the accuracy, reliability, and relevance of the results obtained. Data can be defined as a collection of facts, figures, observations, or measurements used for reference or analysis. However, for data to be meaningful, it must be collected from a dependable source and in a manner that aligns with the objectives of the study.



2.2 Learning Outcomes

Before end of this unit students should be able to:

- Defining and differentiating between primary and secondary sources of data.
- Recognizing the ethical considerations in data collection, especially when using primary sources.
- Identifying the advantages and disadvantages of each type of data source.



2.3 Source of Data

2.3.1 Primary Source

Primary data is information gathered by the investigator for the purpose of conducting an investigation or conducting a study. Such data is unique in nature and is derived from surveys conducted by individuals, research institutions, or any organization. This can be accomplished

through personal investigation, investigation teams, and questionnaires. The investigator is the original collector of the data in primary sources, so he is aware of any flaws in the data.Although primary data are important, they are only used to supplement data obtained outside the organi- zation, particularly from government agencies.

Advantages of primary data:

i. Collecting primary data by creating schedules or distributing and collecting questionnaires by mail is less expensive and takes less time.

ii. Information gathered from primary data is more reliable than information gathered from secondary data.

iii. They usually contain more specific information than secondary data. This is due to the researcher's knowledge of where to gather information.

iv. They frequently define the terms and units that are used.

Disadvantages of primary data:

i. Data collection by personal survey is only possible if the area covered by the investigator is small. The cost of collecting data by sending the enumerator is unavoidable. The enumerator should double-check that the information provided by the informants is correct.

ii. If the questions are embarrassing or complicated, or if they probe into individual personnel matters, the schedules may not be filled with accurate and correct information, making this method unsuitable.

iii. One significant disadvantage of primary data is that it is limited in its application. It cannot be used for long-term policy decisions or future planning unless it looks outside the organiza- tion for relevant data.

iv. The process of generating primary data takes a significant amount of time and money.

2.3.2 **Secondary**

As the name implies, these are pre-existing data collected by other people for another pur- pose. Secondary data are primarily the result of administrative activities. Secondary data must be used with caution. They may not provide the exact information requested or may not be in the appropriate format. They may be insufficient, outdated, or do not cover the required area of interest. Before secondary data can be used reliably, the source, definition of terms, and method of collection and compilation must be determined.

They are published statistics and historic records for secondary sources, and the users of the statistical information are not the original data collectors.

The following are the secondary data sources: Statistical abstract, bulletins and reports issued by government departments especially their statistical units.

i. Miscellaneous reports issued by government and non-governmental agencies.

ii. Research report and Learned Journals.

iii. Daily newspaper, magazines and miscellaneous periodicals.

iv. Central Bank of Nigeria statistical bulletin, annual report and statement of accounts.

v. Manpower Board of Surveys.

Precautions in the use of Secondary Data

The following are some considerations for using secondary data.

1. How the data was collected and processed

2. The data's accuracy

3. How far the data has been summarized

4. How comparable the data is to other tabulations.

5. How to interpret the data, especially when figures collected for one purpose are used for another.

In general, when dealing with secondary data, people must make a choice between what they want and what they can find.

Advantage of Secondary Data

i. Secondary data is inexpensive to obtain. Many government publications are inexpensive, and libraries hold large amounts of secondary data produced by the government, businesses, and other organizations.

ii. The internet provides access to a large amount of secondary data.

iii. Because much of the secondary data available has been collected over many years, it can be used to plot trends.

iv. Secondary data is useful to:

- the government because it aids in decision making and policy planning.
- business and industry - in areas such as marketing and sales to gain an understanding of general economic and social conditions and to provide information on competitors.
- research organizations - by disseminating social, economic, and industrial data.

Disadvantages of Secondary data

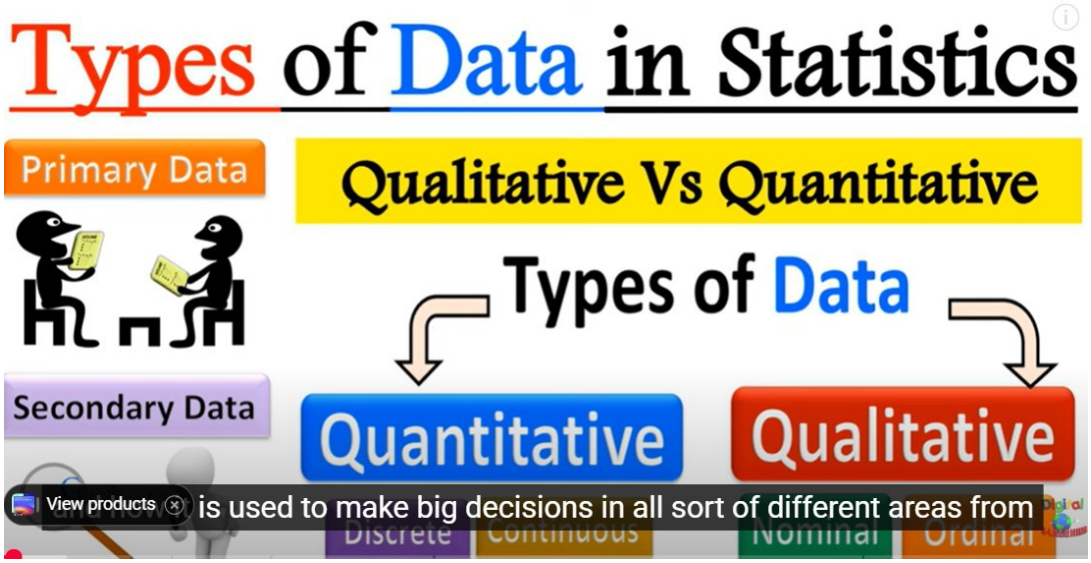
i. Quality and Accuracy Issues:- The reliability of secondary data depends on the original source, which may contain errors, biases, or outdated information.

ii. Limited Control Over Data Collection:- Since the researcher did not collect the data first-hand, they have no control over how it was gathered, making it difficult to assess the method- ology used.

iii. Outdated Information:- Some secondary data may no longer be applicable due to changes in trends, policies, or societal factors.

iv. Lack of Specificity:- The data might be too general or aggregated, preventing a detailed analysis of specific variables.

v. Inconsistencies in Data Sources:- Different sources may present conflicting information, making it difficult to determine the most accurate version.



You can click here to watch the video.https://youtu.be/DRlriVQf3Sw?si=6XIlzmEwIHDkZkeo

2.4 Self-Assesement Exercises

1. Which of the following is a primary source of data?
   A) A published textbook B) Data from a field survey
   C) News articles D) Encyclopedia entries

2. Secondary data is best described as:
   A) Data collected for the first time B) Data from interviews
   C) Previously collected and published data D) Data from observations

3. A census conducted by a national agency is an example of:
   A) Primary data B) Secondary data
   C) Simulated data D) Informal data

4. Which of the following methods is used to collect primary data?
   A) Library research B) Conducting questionnaires
   C) Reading published papers D) Citing Wikipedia

5. Which is NOT an advantage of secondary data?
   A) Saves time and cost B) Easy to access
   C) Always accurate D) Useful for background research

6. The source of secondary data could be:
A) Field work B) Government records
C) Lab experiment D) Direct measurement

7. Data collected directly by the researcher is known as:
A) Secondary data B) Experimental data
C) Primary data D) Processed data

8. Which of the following is an example of quantitative primary data?
A) Opinions from a blog B) Temperature readings from a thermometer
C) Notes from a textbook D) Data copied from a journal

9. One disadvantage of primary data is:
A) It is always outdated B) It lacks detail
C) It is time-consuming to collect D) It is not reliable

10. Which source provides the most recent and raw data?
A) Secondary source B) Textbooks
C) Primary source D) Statistical reviews



2.5. Summary

What you have learned in this unit concerns the types and sources of data used in statistics.You now understand the differences between primary and secondary data, and how each is collected and used.The unit also introduced methods such as surveys, interviews, and document analysis, all of which play a key role in data gathering.



2.6 References, Further Readings, Web Sources

1. Tmt. V.Varalakshmi, Thiru G.Gnana Sundaram, Tmt. B.Indrani, Tmt. N.Suseela and Tmt.S.Ezhilarasi (2004). Statistics for Higher Secondary – First Year.
2. Obasohan P.E and Abd-elfatai A.M (2012). An Introduction to Statistical Method.
3. Allan G. Bluman (2012).Elementary of Statistics *Eighth edition*



2.7 Possible Answers to SAEs

1.B, 2.C, 3.A, 4.B, 5.C, 6.B, 7.C, 8.B, 9.C, 10.C

Unit 3 Method of Data Collection

3.1 Introduction
3.2 Learning Outcomes
3.3 Method of Data Collection
3.3.1. Direct Observation
3.3.2. Personal Interview
3.3.3 Questionnaires
3.3.4 Telephone Method
3.3.5 Survey
3.3.6 Census
3.4 Self-Assessment Exercises
3.5 Summary
3.6 References, Further Readings, and Web Sources
3.7 Possible Answers to SAEs



3.1 introduction

In research and statistical studies, data collection refers to the systematic process of gathering relevant information to analyze, interpret, and draw valid conclusions. The accuracy and credibility of any study largely depend on how well the data is collected. Several practical methods exist for collecting firsthand or original data directly from the respondents. This section focuses specifically on six key methods: Direct Observation, Personal Interview, Questionnaires, Telephone Method, Survey, and Census. Each method is applied based on the nature of the study, target population, and available resources



3.2 Learning Outcomes

Before end of this unit students should be able to:

- Define and explain the concept of data collection.
- Describe six recognized methods of primary data collection.
- Understand the conditions under which each method is most effective.



3.3 Method of Data Collection

3.3.1 Direct Observation

This is the most accurate method of data collection, but it is not suitable for all situations. It entails the investigator personally observing what happens at the scene of the incident. For example, an investigator may want to know how many vehicles travel a specific route in order to locate a gas station. The investigator or his agents must be present on the scene to observe the vehicular movement and record his findings.

Advantages

i. The accuracy of the data collected is guaranteed.
ii. The investigator can assess the data's quality.

Disadvantages

i. It is inappropriate in many situations.
ii. It takes time because you must remain in that location while observing the unit or item of interest. It is time-consuming.
iii. When the observed realizes you are watching him, he may be influenced.

3.3.2 Personal Interview

a. Direct Personal Interview

Direct personal interviews: Informants are people who provide information. To gather the nec- essary information, the investigator personally meets with them and asks questions. It is best suited for intensive field surveys rather than extensive ones. It is best suited for intensive study of a narrow field.

Advantages

i. It is a method that is fairly comprehensive.
ii. Reliable responses are assured.
iii. Any question that has been misunderstood can be clarified.

Disadvantages

i. It is both costly and time-consuming.
ii. When there are a large number of people to interview and they are spread out over a large area, it is extremely difficult.
iii. This method reinforces personal prejudice and bias.

b. Indirect Personal Interview

This method is appropriate for use by committees of public inquiries whose responses are con-

sidered complex. The interviewer conducts an interview with a third party (a witness) who has information about the case at hand. For example, in an accident scene where the people involved are either dead or unable to give account, an eye witness to the accident can be asked to give account of the incident.

Advantages

i. Complex responses are readily accepted.
ii. Even if information has been completely lost, a witness can still provide it.

Disadvantages

i. Witness responses may contradict one another.
ii. The witness could purposefully misrepresent the information given.

3.3.3 Questionnaire

These are questions designed to elicit responses from respondents that will lead to the investigator's objectives being met. The success of any questionnaire is heavily influenced by three main factors: the subject matter, the target population, and the survey's sponsor.

a. Mailed Questionnaire

Using this method, a list of questions is prepared and mailed to all informants. Technically, the list of questions is known as a questionnaire. A covering letter accompanying the questionnaire explains the purpose of the investigation and the importance of correct information, and asks informants to fill in the blank spaces provided and return the form within a certain time frame. This method is appropriate when the informants are literate and dispersed over a large area.

Advantages

i. It is relatively cheap.
ii. It is preferable when the informants are dispersed across a large geographical area.

Disadvantages

i. A low response rate is possible.
ii. There could be a high proportion of incorrect responses.
iii. The respondent may interpret the question incorrectly.

b. Questionnaires by Enumerators

This method places the job of administering the questionnaires to the respondents in the hands of enumerators (agents) who are usually trained in skills of administering the questionnaires. The enumerators take the questionnaires to the respondent who fill in and return to the enu- merators or the enumerators fill in the questionnaires by asking the respondents the various questions- the latter is called "questionnaire to be filled by the enumerators".

Advantages

i. It is appropriate for a large investigation.
ii. It is a comprehensive data collection method.
iii. This method produces relatively accurate data.

Disadvantages

i. Incompetent enumerators may be chosen.
ii. Inadequate training of enumerators may result in misrepresentation.

3.3.4 Telephone Method

People's information is frequently gathered through communication networks such as the tele- phone. This is only applicable where the telephone system is efficient. Due to inefficient communication systems, this method of data collection is not widely used in Nigeria. Many villages and towns do not have telephone networks.

Advantages

i. It provides a quick and efficient data collection method.
ii. There is enough time to answer the questions.
iii. The interviewer's bias is minimized or completely absent.
iv. No clustering in the sample.

Disadvantages

i. The qualities of the respondents cannot be observed.
ii. There is little or no control over the respondents' identities.
iii. The logical sequence of thoughts, in terms of the order of the questions, cannot be followed.
iv. No complex questions can be posed.
v. The immediate reaction may not be reliable or provided due to insufficient information.

3.3.5 Survey

Survey is the systematic method of collecting data from a define area. It involves collection of necessary information that will present a good picture of the situation. For instance, we have business surveys which cover all the variables that normally affect the conduct of a business. If you want to go into the production of wheat on a large scale, you may wish to know the business environment that will guarantee the successful production of wheat.

In this case, you might need to conduct a survey in wheat production which will among other things help you to determine the profitable share of wheat market. Hence you may have to look at variables like the existing price, the level of wheat production already in existence, the con- sumption level, the import level and opportunities for financial markets. All these factors are necessary for serious consideration before embarking on expansion or establishment of wheat production.

Advantages

i. Relatively easy to administer.
ii. Can be developed in less time.
iii. Cost-effective, but cost depends on survey mode.

Disadvantages

i. Respondents may not feel encouraged to provide accurate answer.
ii. You might discover that some questions don't get answer.
iii. There can be difference in how people understand the survey questions.
iv. Some answers can be challenge to clarify.

3.3.6 Census

A census is a study of every unit, everyone or everything, in a population. It is known as a com- plete enumeration, which means a complete count. The National Census, apart from knowing the population of a country is most useful as basis for planning and ensuring an equitable dis- tribution of the national resources. Therefore, surveys and census are important procedures for generating primary data and the methods outlined above can be adopted in surveys or census in an attempt to generate the required information or data for a particular use.

Advantage

i. Accuracy and Reliability.
ii. This method is effective if the universe is small.
iii. Census method completely examines each unit and gathers important data for intensive study.
iv. Census method is applicable to examine heterogeneous units.

Disadvantages

i. It is very costly method of data collection.
ii. It consumes more time and labour to complete data collecting task.
iii. This method is not applicable or suitable if the population is large. It is suitable only for small population.
iv. There is a high chance of statistical errors in this method.



3.4 Self-Assesement Exercises

1. Which of the following is **not** a method of primary data collection?
   (a) Observation

(b) Questionnaire
(c) Census report
(d) Interview

2. Which method is best for collecting data from illiterate respondents?
   (a) Self-administered questionnaire
   (b) Online survey
   (c) Personal interview
   (d) Email survey

3. A secondary source of data is:
   (a) Information from direct observation
   (b) Data obtained from textbooks or journals
   (c) Results of laboratory experiments
   (d) Data collected using structured interviews

4. A major disadvantage of using questionnaires is:
   (a) It is cheap to administer
   (b) It can be sent to many people at once
   (c) Respondents may not return them
   (d) Responses can be easily coded



3.5. Summary

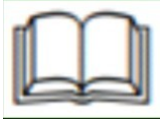
Data collection is the process of gathering information needed for research, decision-making, or analysis. It is an essential step in any study because the quality of the data collected directly affects the accuracy and usefulness of the results.

There are two main types of data:

Primary data is collected directly by the researcher through methods like observation, interviews, questionnaires, surveys, and telephone calls.

Secondary data is collected by someone else and made available through sources like govern- ment publications, books, journals, websites, and organizational records.

Each method of data collection has its own strengths and weaknesses. Primary data is more specific and reliable but may be time-consuming and costly. Secondary data is easy to access and cheaper, but it may not always meet the specific needs of the study



3.6 References, Further Readings, Web Sources

1. Tmt. V.Varalakshmi, Thiru G.Gnana Sundaram, Tmt. B.Indrani, Tmt. N.Suseela and Tmt.S.Ezhilarasi (2004). Statistics for Higher Secondary – First Year.

2. Obasohan P.E and Abd-elfatai A.M (2012). An Introduction to Statistical Method.
3. Allan G. Bluman (2012).Elementary of Statistics *Eighth edition*



3.7 Possible Answers to SAEs

1. Correct Answer: (c) Census report
Reason: A census report is typically a secondary source because it's already compiled data.

2. Correct Answer: (c) Personal interview
Reason: Illiterate respondents may not be able to read or write, so oral interviews are more appropriate.

3. Correct Answer: (b) Data obtained from textbooks or journals Reason: These are already published data, hence secondary.

4. Correct Answer: (c) Respondents may not return them Reason: Non-response is a common problem with questionnaires.

Module Two: Presentation of Data

- Unit 1: Presentation of Data in Table
- Unit 2: Presentation of Data in Chart
- Unit 3: Presentation of Data in Graph

Unit 1 Presentation of Data in Table

1.1 Introduction
1.2 Learning Outcomes
1.3 Presentation of Data in Table
  1.3.1 Frequency Distribution
  1.3.2 Types of Frequency
  1.3.3 Nature of Class
  1.3.4 Preparation of Frequency Table
1.4 Self-Assessment
1.5 Summary
1.6 References, Further Readings, and Web Sources
1.7 Possible Answers to Self-Assessment Exercises



1.1 introduction

When data is collected, especially in large quantities, it can be overwhelming and difficult to interpret in its raw form. To make sense of such data, it is important to organize and summarize it. One of the most useful ways of organizing data is through a frequency distribution.

A frequency distribution is a method of tabulating data by showing how often each value (or range of values) occurs in a dataset. This helps in understanding the pattern, spread, and central tendency of the data



1.2 Learning Outcomes

Before end of this unit students should be able to:

- Define and explain the concept of frequency distribution.
- Differentiate between grouped and ungrouped frequency distributions.

- Construct a frequency distribution table from raw data.
- Identify and interpret key components like class interval, frequency, and class width.



1.3 Presentation of Data in Table

1.3.1 Frequency Distribution

A frequency distribution is a table or arrangement that displays the number of times (frequency) each data value or class interval appears. This approach simplifies large data sets, making them easier to understand and analyze. By organizing data into a frequency distribution, we can quickly identify common patterns, such as the most or least frequent values, and understand how the data is spread out across the range

1.3.2 Types of Frequency Distribution

There are two main types of frequency distributions based on how the data is organized:

1. **Ungroup frequency distribution**

An ungrouped frequency distribution is used when the dataset is small and the values are distinct (i.e., not continuous or wide-ranging). Each individual value from the dataset is listed separately with its corresponding frequency.

It's mostly used when:

- Data values are not too many
- Values are repeated a few times
- There's no need to group into intervals

Example: In a demographic study involving 40 households within a village, data was collected on the number of children per household, yielding the following results.

Table 2.1: Number of Children per Family (Raw Data)

1	0	3	2	1	5	6	2
2	1	0	3	4	2	1	6
3	2	1	5	3	3	2	4
2	2	3	0	2	1	4	5
3	3	4	4	1	2	4	5

Represent the data in the form of a ungroup frequency distribution. Frequency distribution of the number of children is given below:

Number of Children	Tally marks	Frequency
0	|||	3
1	𝍸 ||	7
2	𝍸 𝍸	10
3	𝍸 |||	8
4	𝍸 |	6
5	||||	4
6	||	2
	Total	**40**

2. **Group frequency distribution**

A grouped frequency distribution is used when the dataset is large or when the data values have a wide range. In this case, the data is divided into class intervals (e.g., 10–19, 20–29), and the frequency for each interval is recorded.The advantage of a group frequency distribution is that it enables a very large array of data to be reduced to a smaller manageable size.

It's mostly used when:

- Data values are too many or continuous
- Values range widely
- Data needs to be summarized

Example: A study was conducted among 60 students in the Department of Mathematics, National Open University of Nigeria, to determine the number of hours they spend studying each week. The frequency distribution of their responses is shown below:

Raw Data (Hours per Week):

5, 6, 7, 8, 5, 6, 7, 9, 10, 8, 6, 7, 10, 11, 12,
9, 10, 13, 14, 12, 10, 9, 13, 11, 12, 14, 15, 16,
15, 14, 13, 17, 16, 15, 14, 13, 12, 17, 18, 17, 16,
18, 19, 20, 18, 19, 17, 16, 20, 19, 18, 21, 20, 22,
23, 21, 20, 22, 24, 25

Based on this data, the following grouped frequency distribution was formed:

1.3.3 Nature of Class

When creating a continuous or group frequency distribution, there are several key terms you need to understand. These concepts help us describe how data is organized into intervals (classes).

Table 2.2: Frequency Distribution of Study Hours per Week

Class Interval (C.I.)	Frequency
5 – 8	10
9 – 12	14
13 – 16	14
17 – 20	16
21 – 24	6
25 – 28	1

1. Class Limits

Each class is defined by two boundary values the smallest and the largest values that can fall within that class. The lower boundary is called the **lower class limit**, and the upper boundary is the **upper class limit**. For instance, in the class 50–100, the number 50 is the lower limit (denoted by L), while 100 is the upper limit (denoted by U).

2. Class Interval A class interval represents a specific range of values in which data points fall. For example, intervals like 50–75, 75–100, 100–125, etc., are used to divide the dataset into meaningful segments. Each interval starts from a lower limit and ends just before the next one begins.

3. Class Width or Size

The width of a class interval shows how broad each group is. It is obtained by subtracting the lower limit from the upper limit of that class. For exclusive class intervals, the formula is:

$$C = U - L$$

where C represents the class width.

4. Range

The range of a dataset tells us the total spread of the values. It is the difference between the highest and the lowest observations:

$$R = Largest\ Value - Smallest\ Value$$

It helps in determining the size of intervals or the scale of grouping.

5. Mid-value or Midpoint The central value of a class interval is called the **midpoint**. It provides a representative value for each class and is calculated by taking the average of the class limits:

$$Midpoint = \frac{L + U}{2}$$

For example, in the class $20 - 30$,the midpoint is $\frac{20+30}{2} = 25$.

6. Number of Class Intervals Deciding how many class intervals to use is important. Too many intervals can complicate the table, while too few can hide useful patterns. Generally, a distribution may have between 5 and 15 classes. To find an appropriate number, we can use **Sturges' Rule**:

$$K = 1 + 3.322 \log_{10}(N)$$

where:

- K = number of class intervals,
- N = total number of observations.
- *log*= logarithm of the number

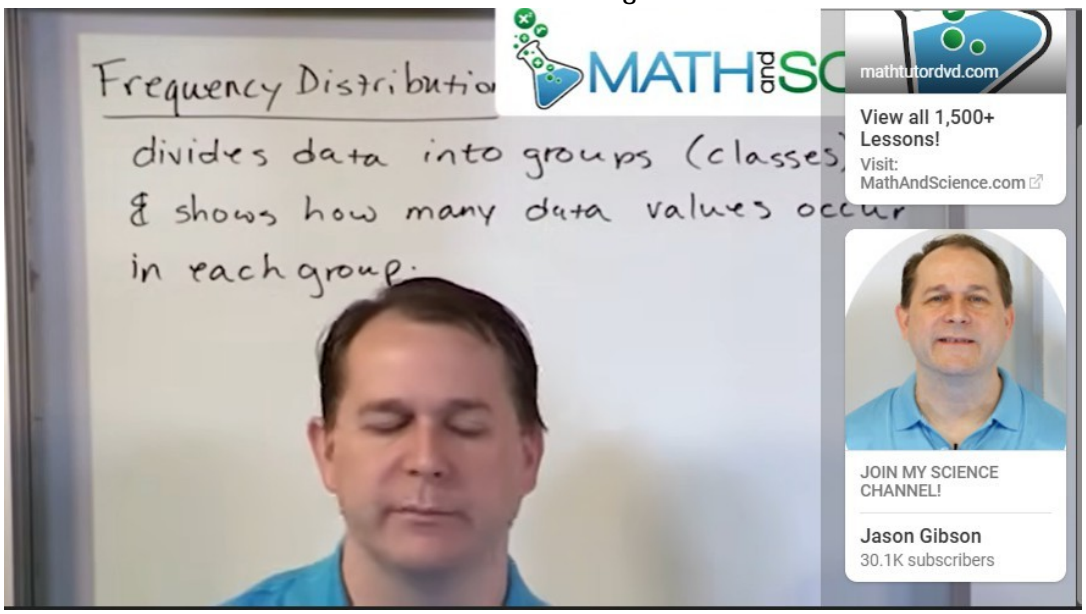
Example: If $N = 10$, then $K = 1 + 3.322 \log_{10}(10) = 1 + 3.322 = 4.322 \approx 4$

7. Size of the Class Interval

The approximate size of each class interval can also be found by dividing the range by the number of classes:

$$C = \frac{R}{K} \quad \frac{Range}{1 + 3.322 \log_{10}(N)}$$

This ensures that each class has a consistent and logical width.



You can click here to access the video.https://youtu.be/xe4Ok1LEQe8?t=7

1.4 Self-Assesement Exercises

1. What does a frequency distribution table show?

(a) The relationship between two variables

(b) How values are coded alphabetically

(c) The number of times each value or range appears

(d) The average of a dataset

2. What is the term for the smallest value that can be included in a class?

(a) Mid-value

(b) Lower class limit

(c) Frequency

(d) Class width

3. The class interval 30–39 has a width of:

(a) 9

(b) 8

(c) 10

(d) 69

4. What do we call the difference between the highest and lowest values in a dataset?

(a) Class width

(b) Frequency

(c) Range

(d) Mid-value

5. The midpoint of a class interval is calculated as:

(a) Upper limit – Lower limit

(b) Lower limit × Upper limit

(c) (Lower limit + Upper limit) ÷ 2

(d) Frequency ÷ Class width

6. According to Sturges' Rule, the number of class intervals K is estimated using:

(a) $K = N \times \log_{10} 2$

(b) $K = 1 + 3.322 \log_{10} N$

(c) $K = \log_2 N$

(d) $K = N \div 2$

7. The column in a frequency table that contains class ranges is called:

   (a) Frequency column

   (b) Data column

   (c) Class interval column

   (d) Observation column

8. If a class interval is 40–49, what is its mid-value?

   (a) 44.5

   (b) 89

   (c) 45

   (d) 40

9. What is the purpose of classifying data into classes?

   (a) To make the data less accurate

   (b) To increase data volume

   (c) To simplify data analysis

   (d) To hide raw data

10. What happens to the class width if the number of class intervals increases?

    (a) It increases

    (b) It decreases

    (c) It remains the same

    (d) It doubles



1.5. Summary Frequency distribution is a method used in statistics to organize raw data into classes or categories and show how often each class occurs. It helps to simplify large sets of data by grouping values into intervals and counting how many times each group appears in the data set. The resulting table is called a frequency table, and it often includes columns for the class intervals, the corresponding frequencies, and sometimes additional columns like cumulative and relative frequencies. This form of presentation makes it easier to identify patterns, trends, and the general distribution of values within the data. Frequency distribution is especially useful when dealing with large data sets, as it condenses the information into a more readable and interpretable format.



1.7 References, Further Readings, Web Sources

1. Newbold, P., Carlson, W. L., and Thorne, B. (2021). Statistics for business and economics (9th ed.). Pearson Education.

2. Sullivan, M. (2022). Fundamentals of statistics (5th ed.). Pearson.

3. Bluman, A. G. (2021). Elementary statistics: A step-by-step approach (11th ed.). McGraw- Hill Education.



1.8 Possible Answers to SAEs

1.C,2.B 3.A 4.C 5.C 6.B 7.C 8.C 9.C 10.B

Unit 2 Presentation of Data in Chart

2.1 Introduction
2.2 Learning Outcomes
2.3 Presentation of Data in Chart
2.3.1 Histogram
2.3.2 Bar Chart
2.3.2 Pie Chart
2.4 Summary
2.5 References, Further Readings, and Web Sources
2.6 Possible Answers to Self-Assessment Exercises



2.1 introduction

Presenting data visually helps to simplify large amounts of information, making it easier to analyze, compare, and communicate. Charts are effective tools in descriptive statistics to rep- resent different types of data distributions. This note covers common types of charts used to display data, including their explanations, numerical examples, and corresponding visual rep- resentations.



2.2 Learning Outcomes

Before end of this unit students should be able to:

- Define and distinguish between a histogram, bar chart, and pie chart.
- Identify the appropriate graphical representation for different types of data.
- Construct a histogram for grouped numerical data.
- Create a bar chart to display categorical variables.



2.3 Presentation of Data in Chart

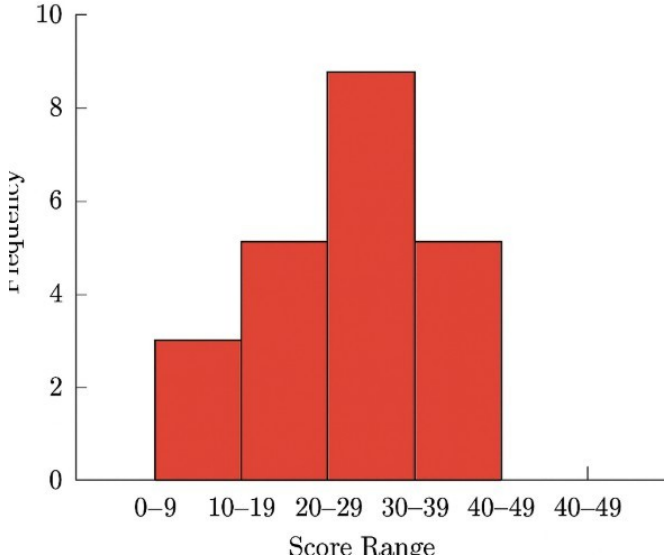
2.3.1 Histogram

A histogram is a bar graph representing the frequency distribution of continuous (interval or ratio) data. The bars touch each other to reflect the continuity of the data.

Example: In a class test, the scores of 30 students were grouped into intervals. The frequency distribution is shown below: Draw the histogram

Table 2.3: Students' Scores Distribution

Score Range	Frequency
0–9	2
10–19	5
20–29	8
30–39	10
40–49	5



2.3.2 Bar Chart

A **bar chart** (or bar graph) is a graphical method of presenting categorical data using rectangular bars. Each bar represents a category of data, and the height (or length) of the bar is proportional to the value or frequency of the category. Bar charts are widely used in statistics, business, economics, and education to visually compare different sets of data.

Types of Bar Charts

Bar charts come in different types depending on how the bars are presented:

- **Vertical Bar Chart:** Bars stand vertically, with height representing the frequency or value.
- **Horizontal Bar Chart:** Bars are drawn horizontally; length shows the value.
- **Grouped Bar Chart:** Used to show multiple data sets grouped side by side for comparison.
- **Stacked Bar Chart:** Combines multiple data sets stacked on top of each other in one bar.

Uses of Bar Charts

- Comparing different categories of data.
- Presenting survey results or frequencies.
- Showing changes over time (when data is discrete).
- Making data easier to understand at a glance.

Advantages of Bar Charts

- Simple and easy to interpret.
- Visually effective in comparing categories.
- Can display both positive and negative values.
- Useful for large sets of categorical data.

Limitations

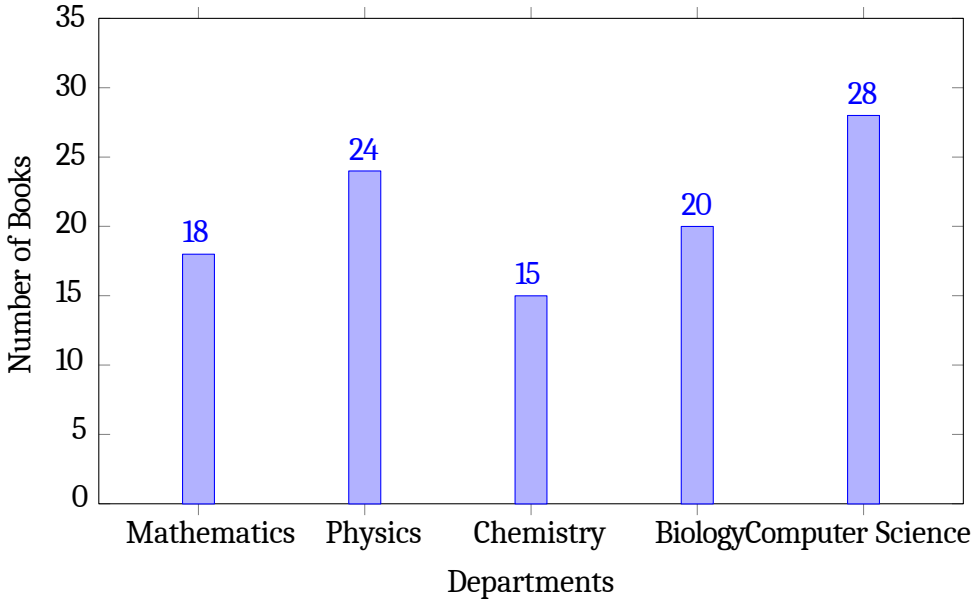
- Not suitable for continuous data.
- Too many bars can make the chart cluttered.
- Does not show exact data values like a table.

Numerical Example

The table below shows the number of books read by students in different Departments National Open University of Nigeria in one month:

Department	Number of Books Read
Mathematics	18
Physics	24
Chemistry	15
Biology	20
Computer Science	28

Bar Chart Representation



2.3.3 Pie Chart

A **pie chart** is a circular statistical graphic that is divided into slices to illustrate numerical proportions. Each slice (sector) of the pie represents a category, and the size of each slice is proportional to the quantity it represents. Pie charts are commonly used to show parts of a whole.

Uses of Pie Chart

- Representing categorical data as parts of a whole.
- Comparing proportions among different categories.
- Displaying percentage distribution.

Advantages of Pie Chart

- Visually appealing and easy to understand.
- Shows the relationship of parts to the whole.
- Useful for quick comparisons.

Limitations

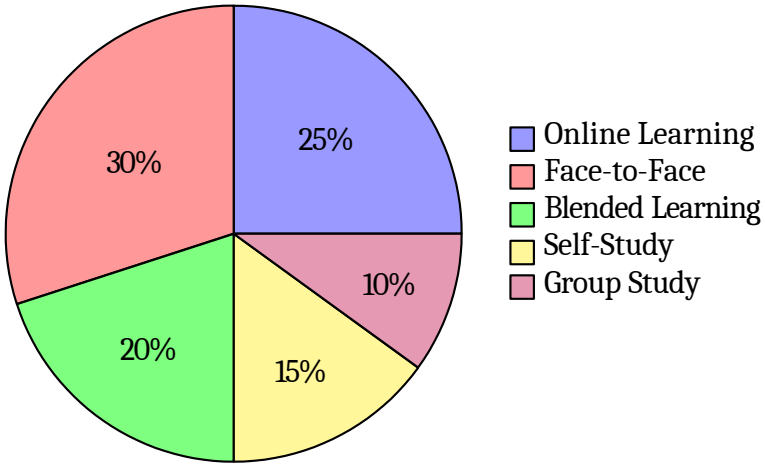
- Not suitable for large data sets with many categories.
- Difficult to interpret precisely.
- Cannot show trends over time.

Numerical Example

The table below shows how 100 students from different departments prefer learning methods:

Learning Method	Number of Students
Online Learning	25
Face-to-Face	30
Blended Learning	20
Self-Study	15
Group Study	10

Pie Chart Representation



2.4 Self-Assesement Exercises

1. Which of the following charts is best used to represent continuous numerical data?
   (a) Pie chart (b) Bar chart (c) Histogram (d) Line graph

2. In a histogram, the bars are:
   (a) Separated by gaps (b) Drawn horizontally (c) Placed side-by-side with no space
   (d) Colored in different patterns

3. A pie chart is most suitable for:
   (a) Continuous data (b) Categorical proportions (c) Grouped frequency data (d) Time series data

4. Bar charts represent:
   (a) Continuous intervals (b) Cumulative frequency (c) Categorical data (d) Interquartile range

5. The main difference between a histogram and a bar chart is:
   (a) One uses colors (b) One is 3D (c) Histogram has no gaps; bar chart has gaps
   (d) They are always the same

6. Which chart is best suited for comparing parts of a whole?
   (a) Histogram (b) Pie chart (c) Bar chart (d) Line graph

7. In a pie chart, if a category represents 25% of the data, its angle in degrees is:
   (a) 45° (b) 60° (c) 90° (d) 180°

8. Which of these charts uses **class boundaries**?
   (a) Pie chart (b) Histogram (c) Bar chart (d) Scatter plot

9. The bars in a bar chart can be:
   (a) Only vertical (b) Only horizontal (c) Both vertical or horizontal (d) Only curved

10. In a histogram, the height of each bar represents:
    (a) Midpoint (b) Frequency (c) Angle (d) Pie size



2.5 Summary

Histogram, bar chart, and pie chart are essential graphical tools used in statistics for data presentation and interpretation. A histogram is typically used to display continuous numerical data organized in intervals, with adjacent bars indicating frequency distribution. A bar chart, on the other hand, is designed for categorical data, where separate bars represent different categories or groups, making comparison easy. Pie charts are circular representations of data, best used for showing proportional relationships or percentages within a whole. Each chart serves a unique purpose in summarizing data, making it easier to understand patterns, trends, and comparisons.



2.6 References, Further Readings, Web Sources

1. Uduak, E. A., & Ogbonna, M. I. (2021). *Fundamentals of statistical methods for social and management sciences*. Ibadan: Anchor Publishers.

2. Gupta, S. C. (2022). *Fundamentals of statistics*. Himalaya Publishing House.

3. Triola, M. F. (2023). *Elementary statistics* (14th ed.). Pearson Education.



2.7 Possible Answers to SAEs

1.C, 2.C, 3.B, 4.C, 5.C, 6.B, 7.C, 8.B, 9.C, 10.B

Unit 3 Data Presentation in Graph

3.1 Introduction
3.2 Learning Outcomes
3.3 Data Presentation in Graph
3.3.1 Frequency Polygon
3.3.2 Cumulative Frequency Distribution
3.4 Self-Assessment Exercises
3.5 Summary
3.6 References, Further Readings, and Web Sources
3.7 Possible Answers to Self-Assessment Exercises



3.1 introduction

In statistics, data presentation in graph refers to the visual representation of data using charts, diagrams, and plots to communicate information clearly and effectively. Graphical methods provide a quick way to identify patterns, trends, and anomalies in datasets, making them essen- tial tools for data analysis and interpretation



3.2 Learning Outcomes

Before end of this unit students should be able to:

- Understand the purpose and importance of using graphs in statistical data presentation and interpretation.
- Identify different types of statistical graphs, including frequency polygon and ogive, and explain their characteristics.
- Construct accurate frequency polygons and ogives using grouped data and cumulative frequencies.



3.3 Data Presentation in Graph

3.3.1 Frequency Polygon

A frequency polygon is a graphical representation of data that shows the shape of a distribu- tion. It is constructed by plotting points that represent the midpoints of each class interval against their corresponding frequencies, and then connecting these points using straight lines. To complete the shape, extra points with zero frequency are added before the first and after the last midpoint, connecting the polygon to the horizontal axis. This gives the graph its distinct polygonal shape. The construction of a frequency polygon starts with organizing data in a frequency table, identifying the class intervals, and calculating the midpoints. These midpoints are then used as the x-values, and the corresponding class frequencies are the y-values. Once plotted, the points are connected in sequence using straight lines. The result is a line graph that visually presents how data is distributed across intervals.

Frequency polygons are very useful in statistical analysis. They provide a clear and simple way to understand the distribution, shape, and spread of data. One of their key advantages is that they are ideal for comparing two or more distributions on the same graph, which is difficult to achieve with histograms. This makes frequency polygons valuable in education, business, and research for comparing exam scores, sales performance, or survey results across different groups or time periods.

Compared to other graphical tools like bar charts or histograms, frequency polygons are smoother and more compact, especially when dealing with continuous data. They help in identifying pat- terns such as skewness, central tendency, and dispersion more clearly.

Steps to Construct a Frequency Polygon:

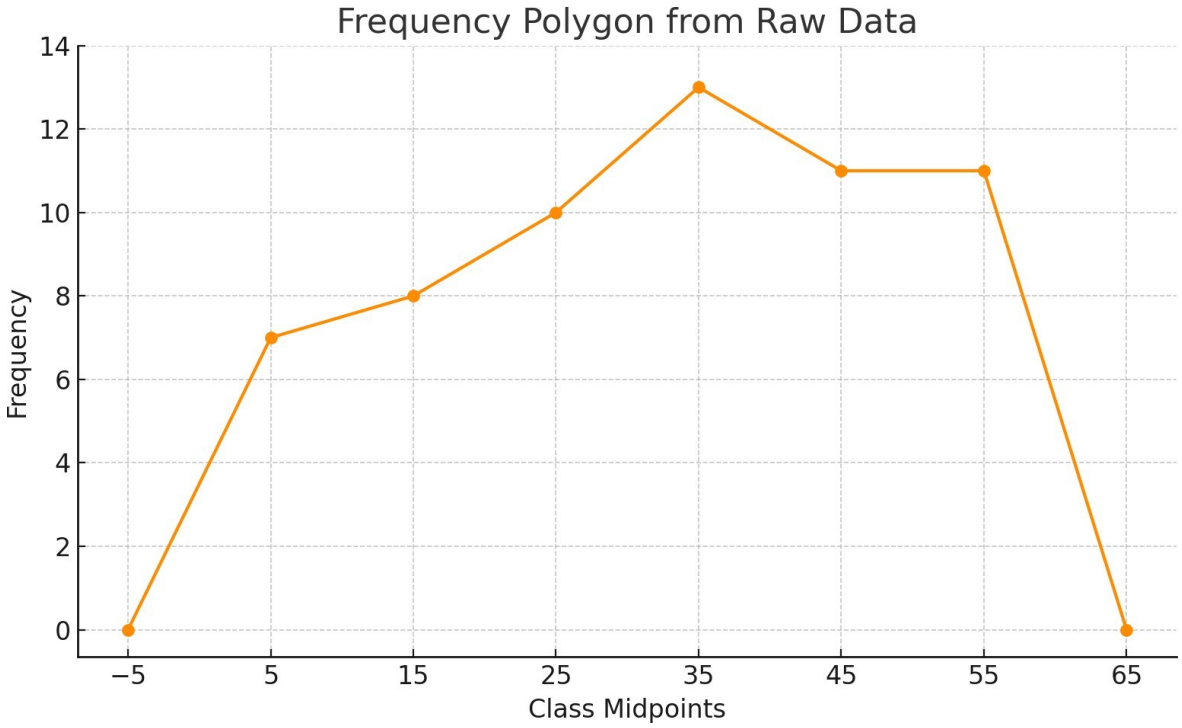
1. Find the midpoints of each class interval.
2. Plot the midpoints on the x-axis and the frequencies on the y-axis.
3. Join the points with straight lines.
4. Add a class before the first and after the last with zero frequency to close the polygon.

This is an example of a frequency polygon constructed from the scores of students in the Department of Mathematics. The data was grouped into class intervals, and the corresponding frequencies were used to create a frequency distribution table as shown below:

Table: Frequency Distribution and Midpoints

Class Interval	Frequency	Midpoint
0 – 10	7	5
10 – 20	8	15
20 – 30	10	25
30 – 40	13	35
40 – 50	11	45
50 – 60	11	55

Below is the graph of the example



Advantages of Frequency Polygon

- Frequency polygons provide a clear visual impression of the shape or distribution of a dataset.
- They are ideal for comparing two or more distributions on the same graph.
- They can be drawn easily from a frequency table by using class midpoints.
- Unlike histograms, frequency polygons use lines, making them suitable for presentations where trends need to be highlighted.
- They are less cluttered than histograms, especially when dealing with overlapping datasets.

Disadvantages of Frequency Polygon

- Frequency polygons are not as effective with small datasets.

- They may be less intuitive to interpret than histograms, especially for beginners.
- Accuracy depends on consistent and correct determination of class midpoints.
- They do not show actual frequencies of specific classes as clearly as histograms do.

Limitations

- Frequency polygons assume uniform distribution within each class interval, which may not always be accurate.
- They cannot represent categorical data.
- Additional endpoints (with zero frequency) must be added to close the polygon, which may affect perception if not properly explained.

3.3.2 Cumulative Frequency Distribution

A cumulative frequency distribution is a method of summarizing data that shows the total num- ber of observations that fall below or above a particular value in a dataset. Unlike a simple frequency distribution that shows how many times each class or group occurs, the cumulative frequency shows how the data accumulates as you move from one class to the next. This helps to understand the data's structure more clearly, especially when dealing with large amounts of information.

There are two main types of cumulative frequency: less than cumulative frequency and greater than cumulative frequency. In a less than cumulative frequency distribution, each class shows the total number of observations that are less than or equal to the upper boundary of that class. In contrast, a "greater than" cumulative frequency distribution shows the number of observa- tions that are greater than or equal to the lower boundary of each class. Both methods are useful for identifying percentiles, medians, and other statistical measures.

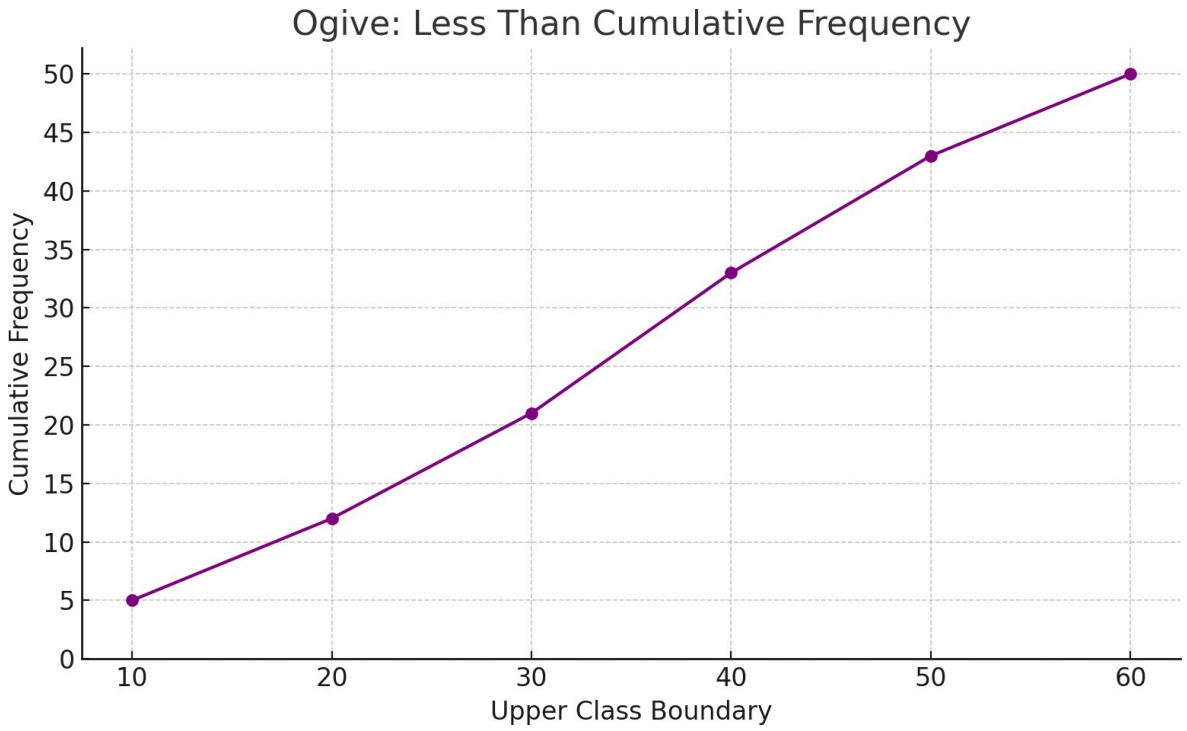
To construct a cumulative frequency distribution, begin with a grouped frequency table, then progressively add (for "less than") or subtract (for "greater than") the frequencies in each class. For example, if the first class has a frequency of 5, the second has 8, and the third has 12, then the "less than" cumulative frequencies would be 5, 13(5 + 8), and 25(5 + 8 + 12). This running total helps to track how values build up across the dataset.

Cumulative frequency is particularly useful when creating a type of graph called an ogive. An ogive is a curve drawn by plotting cumulative frequencies against class boundaries, and it helps in finding the median, quartiles, and percentiles visually. It also gives a general idea of how values are distributed over the range. This makes cumulative frequency distribution a helpful tool not just in statistics classes, but in fields like economics, business, and health sciences where data trends need to be analyzed.

Example: The table below shows a cumulative frequency distribution based on class intervals and their corresponding frequencies.

Cumulative Frequency Distribution table

Class Interval	Frequency	Upper Class Boundary	Cumulative Frequency (Less Than)
0 – 10	5	10	5
10 – 20	7	20	12
20 – 30	9	30	21
30 – 40	12	40	33
40 – 50	10	50	43
50 – 60	7	60	50



Advantages of Cumulative Frequency Distribution

- It helps in identifying key statistical measures such as the median, quartiles, percentiles, and interquartile range.
- It provides a quick summary of how data values are distributed across intervals.
- Cumulative frequency curves (ogives) make it easier to visually analyze data trends and distributions.
- It is useful for comparing different datasets on the same scale.
- Helps in decision making when analyzing performance, growth, or trends.

Disadvantages of Cumulative Frequency Distribution

- It may hide the actual frequency of individual class intervals.
- Interpreting cumulative data can be more complex than simple frequency tables.
- Constructing ogives manually can be time-consuming and error-prone.

Limitations

- Cumulative frequency distribution is not suitable for very small datasets.
- It assumes uniform distribution within class intervals, which may not reflect the actual data distribution.
- It provides limited insight into measures of central tendency like mean or mode.



3.4 Self-Assesement Exercises

1. Which of the following is an advantage of using a cumulative frequency distribution?
    A. It clearly shows individual class frequencies
    B. It helps in estimating percentiles and quartiles
    C. It avoids use of class boundaries
    D. It displays circular comparison
2. One disadvantage of cumulative frequency distribution is that it:
    A. Is difficult to graph
    B. Requires a pie chart
    C. Can hide exact class frequencies
    D. Always needs class midpoints
3. Which is a limitation of cumulative frequency distribution?
    A. It shows all class details
    B. It's suitable for small datasets
    C. It assumes uniform distribution in classes
    D. It's best for nominal data
4. What is a major advantage of a frequency polygon?

A. It uses bars to show cumulative totals

B. It compares multiple datasets clearly

C. It is limited to one dataset

D. It cannot show class trends

5. Frequency polygons are less ideal for small datasets because:

A. They become too complex

B. There are not enough midpoints to form a line

C. They use too many colors

D. They need pie slices

6. What is one disadvantage of a frequency polygon compared to a histogram?

A. It is easier to read

B. It shows clearer bar heights

C. It can be used for categorical data

D. It doesn't need midpoints

7. Which of the following is a limitation of using a frequency polygon?

A. It cannot show trends

B. It requires class frequencies only

C. It assumes even distribution across intervals

D. It uses colors excessively

8. Why must zero-frequency endpoints be added to a frequency polygon?

A. To balance color contrast

B. To help close the graph visually

C. To reduce frequencies

D. To remove class boundaries

9. In cumulative frequency distributions, the ogive curve is especially useful for:

A. Finding quartiles and medians

B. Creating pie charts

C. Determining mode directly

D. Analyzing nominal data

10. A key limitation of both CFD and frequency polygon is that:

    A. They are easy to use

    B. They require no table

    C. They assume equal class width and uniform data spread

    D. They apply to pie charts



3.5 Summary

Cumulative Frequency Distribution is a statistical method used to show how the total number of observations builds up across class intervals. It helps in determining how many data values fall below (or above) a particular point. The less than cumulative frequency adds values progressively from the lowest class, while the greater than type works downward from the highest class. The cumulative frequency data is usually represented using a graph called an ogive, which is useful in estimating statistical measures such as the median, quartiles, and percentiles. This method is especially helpful when analyzing large datasets to visualize the accumulation pattern of values over a range.

A frequency polygon is a type of line graph used to display the shape of a data distribution. It is constructed by plotting the midpoints of class intervals against their respective frequencies and connecting the points with straight lines. To complete the polygon, zero-frequency points are added at both ends. Frequency polygons are especially useful for comparing two or more distributions on the same graph and for identifying trends in data. Unlike histograms, they use lines instead of bars, making them cleaner and more effective when visual clarity or overlaying multiple data series is required.



3.6 References, Further Readings, Web Sources

1. Uduak, E. A., & Ogbonna, M. I. (2021). *Fundamentals of statistical methods for social and management sciences*. Ibadan: Anchor Publishers.

2. Gupta, S. C. (2022). *Fundamentals of statistics*. Himalaya Publishing House.

3. Triola, M. F. (2023). *Elementary statistics* (14th ed.). Pearson Education.



3.7 Possible Answers to SAEs

1.B, 2.C 3.C, 4.B, 5.B, 6.B, 7.C, 8.B 9.A 10.C

Week Three: Measure of Central Tendency

- Unit 1: Arithmetic Mean
- Unit 2: Median
- Unit 3: Mode
- Unit 4: Geometric mean
- Uint 5: Harmonic mean

Unit 1 Measure of Central Tendency

1.1 Introduction
1.2 Learning Outcomes
1.3 Arithmetic Mean
1.4 Self-Assessment Exercises
1.5 Summary
1.6 References, Further Readings, and Web Sources
1.7 Possible Answers to Self-Assessment Exercises



1.1 introduction

The Arithmetic Mean is the most commonly used measure of central tendency in statistics. It is often referred to as the ”average” and represents the sum of all data values divided by the number of values. It provides a single representative value that describes the center point of a data set. The arithmetic mean is widely used in various fields such as economics, education, health, engineering, and social sciences to summarize and compare data.



1.2 Learning Outcomes

Before end of this unit students should be able to:

- Define the arithmetic mean as a fundamental measure of central tendency in statistics.
- Identify situations where the arithmetic mean is an appropriate summary measure of a dataset.

- Calculate the arithmetic mean for both ungrouped and grouped data using the correct formulas.



1.3 Arithmetic Mean

The **Arithmetic Mean**, commonly referred to as the average, is one of the most widely used measures of central tendency. It represents the central or typical value in a data set and is obtained by summing all the values and dividing by the number of observations. The arithmetic mean gives a sense of the overall level of the data, especially when values are evenly distributed. **Ungrouped data**, which consists of raw data values listed individually, the arithmetic mean is computed using the formula:

$$\bar{x} = \frac{\sum x}{n}$$

where $\bar{x}$ is the mean, $\sum x$ is the sum of all data values, and n is the total number of values. For example, for the numbers 5, 7, 8, and 10, the mean is:

$$\bar{x} = \frac{5 + 7 + 8 + 10}{4} = \frac{30}{4} = 7.5$$

Ungrouped data is easy to work with when the dataset is small, and each observation is known. However, in situations where the data is large or presented in frequency format (i.e., how many times each value occurs), it becomes necessary to use a more structured approach, such as grouping the data into class intervals.

Grouped data, which is presented in a frequency distribution table (with class intervals and frequencies), the arithmetic mean is calculated using the formula:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

where f is the frequency of each class and x is the midpoint of the class interval. The midpoint is calculated by:

$$x = \frac{\text{Lower Limit} + \text{Upper Limit}}{2}$$

This method provides a good estimate of the average, especially when exact values are not available and data is grouped.

The arithmetic mean for grouped data is practical and efficient when handling large volumes of data. Since the individual values are not known, we assume each value in a class is equal to the midpoint, which introduces a small approximation. Nevertheless, this method is highly effective in summarizing and analyzing statistical data.

The arithmetic mean is particularly useful in situations where all observations carry equal importance and are of the same kind. For example, in finding the average score of students in a class or average monthly rainfall, the arithmetic mean provides a straightforward and reliable

answer.

Despite its usefulness, the arithmetic mean has some limitations. It is sensitive to extreme values or outliers, which can distort the result and give a misleading impression. For instance, a single very large or very small value can significantly affect the mean, especially in small datasets. In such cases, other measures like the median may be more appropriate.



Example 1: The table below shows the distribution of marks:

Class Interval	Frequency (f)	Midpoint (x)
0–9	4	4.5
10–19	6	14.5
20–29	8	24.5
30–39	7	34.5
40–49	5	44.5

Now calculate fx for each class:

Class Interval	f	x	fx
0–9	4	4.5	18.0
10–19	6	14.5	87.0
20–29	8	24.5	196.0
30–39	7	34.5	241.5
40–49	5	44.5	222.5
Total	30		765.0

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{765}{30} = 25.5$$

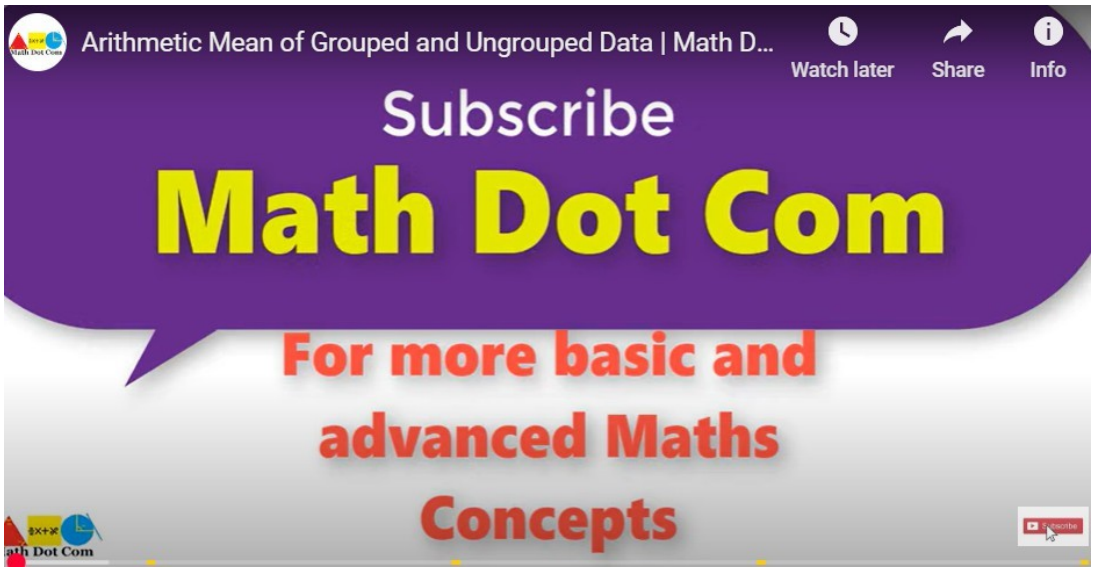
Example 2: The table below shows the ages of participants in a training program.

Age Interval (Years)	Frequency (f)	Midpoint (x)
15–19	3	17
20–24	7	22
25–29	10	27
30–34	5	32
35–39	2	37

Now calculate fx for each class:

Class Interval	f	x	fx
15–19	3	17	51
20–24	7	22	154
25–29	10	27	270
30–34	5	32	160
35–39	2	37	74
Total	27		709

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{709}{27} \approx 26.26$$



You can click here to access the video.https://youtu.be/I33mWfH8Noc

1.6 Self-Assesement Exercises

Choose the most appropriate option (A, B, C, or D) for each question.

1. What is the arithmetic mean of the following numbers: 10, 12, 14, 16, 18?
   A) 12 B) 14 C) 15 D) 13

2. Which of the following best defines the arithmetic mean?
   A) The difference between the highest and lowest values
   B) The value that appears most often
   C) The average obtained by dividing the total of all values by the number of values
   D) The middle value in an ordered data set

3. What is the formula for the arithmetic mean of ungrouped data?
   A) $\bar{x} = \frac{f \cdot x}{N}$ B) $\bar{x} = \frac{\Sigma x}{n}$ C) $\bar{x} = \frac{\Sigma f x}{n}$ D) $\bar{x} = \frac{\Sigma f}{x}$

4. If the sum of 8 numbers is 64, what is the arithmetic mean?
   A) 6 B) 7 C) 8 D) 10

5. What is the midpoint of the class interval 20–24?
   A) 20 B) 21 C) 22 D) 23

6. In grouped data, how is the arithmetic mean calculated?
   A) By using the exact data values
   B) By using class width only
   C) By using the midpoints and frequencies of the classes
   D) By using the median values

7. What does the arithmetic mean represent in a dataset?
   A) The most extreme value
   B) The central location or average of the data
   C) The mode of the data
   D) The value with highest frequency

8. Why might the arithmetic mean be misleading in some cases?
   A) It is always greater than the median
   B) It cannot be used with grouped data
   C) It is affected by extreme values or outliers
   D) It is equal to zero

9. What is the arithmetic mean of the following frequencies and midpoints:

$$f = [2, 3, 5], \quad x = [10, 20, 30]$$

   A) 20 B) 22 C) 23 D) 24

10. If a student scores 80, 85, 90, and 95 in four subjects, what is the mean score?
    A) 87.5 B) 90 C) 85 D) 88



1.5 Summary

The Arithmetic Mean is one of the most commonly used measures of central tendency. It is calculated by adding all the values in a dataset and dividing the sum by the number of observations. For ungrouped data, it provides a single representative value that reflects the overall distribution. In grouped data, the arithmetic mean is determined using class midpoints and frequencies, offering a way to estimate the central value even when exact data points are unavailable.

The arithmetic mean is especially useful in quantitative analysis because it considers all data values in its calculation. However, it can be sensitive to extreme values (outliers), which may distort its representativeness in skewed distributions. Despite this, it remains a fundamental tool in statistics, economics, education, and research for summarizing and comparing datasets



1.6 References, Further Readings, Web Sources

Gupta, S. C., & Kapoor, V. K. (2020). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.

Spiegel, M. R., & Stephens, L. J. (2017). *Statistics* (4th ed.). McGraw-Hill Education.



1.7 Possible Answers to SAEs

1.B, 2.C, 3.B, 4.C, 5.C, 6.C, 7.B, 8.C, 9.A, 10.A

Unit 2 Median

2.1 Introduction
2.2 Learning Outcomes
2.3 Median for Ungroup and Group Data
2.3.1 Median for Ungroup Data
2.3.2 Median for Group Data
2.4 Self-Assessment Exercises.
2.5 Summary
2.6 References, Further Readings, and Web Sources
2.7 Possible Answers to SAEs



2.1 introduction

The Median is a measure of central tendency that represents the middle value in a dataset when the data are arranged in ascending or descending order. It divides the data into two equal halves half of the observations lie below the median and half lie above it. Unlike the arithmetic mean, the median is not affected by extreme values or outliers, making it especially useful for ana- lyzing skewed distributions.In ungrouped data, finding the median simply involves sorting the values and identifying the middle one (or the average of the two middle values if the dataset has an even number of elements). For grouped data, the median is calculated using a formula that involves the cumulative frequency, class boundaries, and class width of the median class.



2.2 Learning Outcomes

Before end of this unit students should be able to:

- Understand the concept and significance of the median as a measure of central tendency.
- Correctly identify and arrange data in order to compute the median for ungrouped datasets.
- Apply the formula for finding the median in grouped frequency distributions.



2.3 Median for Ungroup and Group Data

2.3.1 Median for Ungroup Data

The median is a statistical measure used to find the central value of a data set that has been arranged in order. It divides the data into two equal parts: 50% of the values fall below the

median, and 50% fall above it. The median is a positional average, which means its value depends on the position of data points rather than their actual magnitude. This makes the median different from the arithmetic mean, which considers the numerical values directly in its computation.

To calculate the median of ungrouped data, the first step is to sort the data in either ascending or descending order. Once sorted, if the number of observations n is odd, the median is the value at the middle position, specifically the $\frac{n+1}{2}$ th value. For instance, if there are 9 values, the median is the 5th value in the ordered list. If n is even, the median is found by taking the average of the two middle values, located at the $\frac{n}{2}$ th and $\frac{n}{2}+1$ th positions.

The median is particularly effective when dealing with data that includes extreme values or outliers. For example, if you are examining the monthly incomes of a group where one individual earns significantly more than others, the arithmetic mean might be misleadingly high. In such cases, the median provides a more accurate representation of a typical income, since it is not influenced by such extreme values.

Another advantage of the median is that it can be used with ordinal data, where values have a logical order but may not be measured on an interval scale. For example, rankings such as "strongly agree," "agree," "neutral," "disagree," and "strongly disagree" can be ordered, and a median can be determined, even though calculating a mean would not be appropriate. This makes the median a flexible and valuable measure in both qualitative and quantitative research.

Example: Find the median of the following dataset:

$$14,\ 9,\ 12,\ 18,\ 20,\ 11,\ 10,\ 16,\ 13,\ 15$$

Step 1: Arrange the data in ascending order:

$$9,\ 10,\ 11,\ 12,\ 13,\ 14,\ 15,\ 16,\ 18,\ 20$$

Step 2: Count the number of observations. There are 10 values.

Step 3: Since the number of values is even, take the average of the 5th and 6th values:

$$\text{Median} = \frac{13+14}{2} = \frac{27}{2} = 13.5$$

2.3.2 **Median for Group Data**

In grouped data, individual values are not listed; instead, data is organized into class intervals with corresponding frequencies. As a result, we cannot directly observe the middle value like we do in ungrouped data. To find the median in grouped data, we first need to determine the class interval in which the median lies this is known as the *median class*. The median class is identified by locating the class whose cumulative frequency includes the middle observation (i.e., the $\frac{n}{2}$ th item, where n is the total frequency).

Once the median class is identified, we apply a specific formula to estimate the median

value. The formula is:

$$\text{Median} = L + \left(\frac{\frac{n}{2} - F}{f} \right) \times h$$

where:

L = lower boundary of the median class,

n = total frequency,

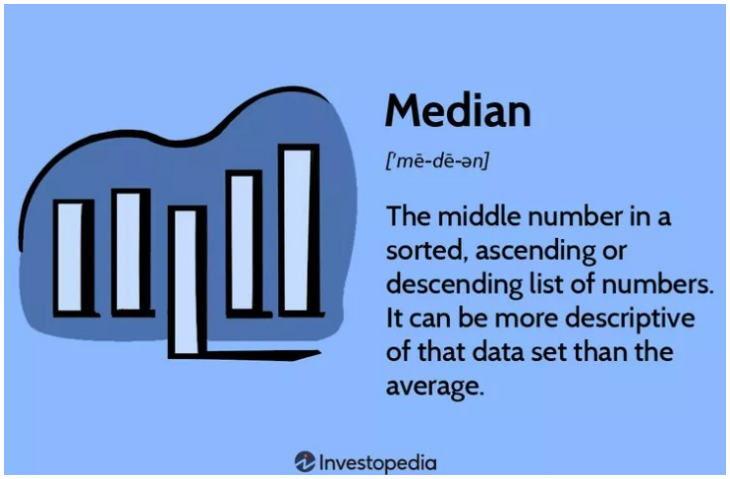
F = cumulative frequency before the median class,

f = frequency of the median class,

h = class width.

This formula provides an estimate of the median by assuming a uniform distribution of values within the median class interval.

The median of grouped data is particularly useful when dealing with large datasets that have been summarized into class intervals, such as age groups, income brackets, or exam scores. It is less affected by extreme values and provides a reliable measure of central tendency when the data is skewed or unevenly distributed. Though the median is only an estimate in grouped data, it is widely used in statistical analysis due to its simplicity and robustness.



Example 2: The following table shows the distribution of scores for a class of students:

Class Interval	Frequency (f)
0–10	3
10–20	5
20–30	8
30–40	12
40–50	7
50–60	5

Step 1: Compute cumulative frequency:

Class Interval	f	Cumulative Frequency (CF)
0–10	3	3
10–20	5	8
20–30	8	16
30–40	12	28
40–50	7	35
50–60	5	40

Step 2: Total frequency: $n = 40 \Rightarrow \frac{n}{2} = 20$

Step 3: Median class is 30–40

Step 4: Apply the formula:

$$\text{Median} = L + \left(\frac{\frac{n}{2} - F}{f}\right) \times h$$

$$= 30 + \frac{20 - 16}{12} \times 10 = 30 + \frac{40}{12} = 30 + 3.33$$

$$\text{Median} = 33.33$$



You can click here to access the video.https://youtu.be/quwOqNuOZMU

2.4 Self-Assesement Exercises

1. What is the median of the ungrouped data set: 6, 4, 9, 3, 7?
   A. 6 B. 4 C. 7 D. 5

2. When the number of observations is even in ungrouped data, the median is:
A. The highest value B. The lowest value C. The mean of the two middle values D. The total sum divided by frequency

3. In ungrouped data, what is the first step to find the median?
A. Draw a graph B. Sort the data C. Add frequencies D. Multiply values

4. In grouped data, the median is found using which of the following components?
A. Only the class mark B. Frequency only C. Class boundaries and cumulative frequency D. Midpoints

5. What is the formula for finding the median in grouped data?
A. Median = $\frac{L+U}{2}$ B. Median = $\frac{n}{f}$ C. Median = $L + \left(\frac{\frac{n}{2} - F}{f}\right) \times h$
D. Median = $\frac{CF}{n}$

6. What does L represent in the grouped data median formula?
A. Lower frequency B. Lower boundary of the median class C. Limit of class width D. Label of frequency

7. The median class is identified by:
A. Smallest frequency B. Middle value of frequencies C. The class whose cumulative frequency contains $\frac{n}{2}$ D. Last class in the table

8. In grouped data, the class width h is calculated by:
A. Subtracting upper from lower limit B. Adding lower and upper limit C. Subtracting lower limit of one class from next D. Counting class boundaries

9. Which of the following statements is TRUE about the median?
A. It is always equal to the mean B. It is greatly affected by outliers C. It always requires class marks D. It is a positional measure of central tendency

10. Why is the median often used in skewed data distributions?
A. It's easier to calculate B. It avoids the influence of extreme values C. It gives a larger average D. It requires no formulas



2.5. Summary

The median is the middle value in a data set when the values are arranged in ascending or descending order. It divides the data into two equal halves. For ungrouped data, if the number of observations is odd, the median is the middle value; if even, it is the average of the two middle values. The median is not affected by extreme values, making it a good measure of central tendency when dealing with skewed data.

In grouped data, the median is determined by identifying the median class, which is the class interval where the cumulative frequency reaches or exceeds half the total number of observations. A formula is then applied to estimate the precise median value within that class. This helps to summarize large datasets in a meaningful way even when individual data points are not visible.



2.6 References, Further Readings, Web Sources

Gupta, S. C., & Kapoor, V. K. (2020). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.

Spiegel, M. R., & Stephens, L. J. (2017). *Statistics* (4th ed.). McGraw-Hill Education.



2.7 Possible Answers to SAEs

1.A, 2.C, 3.B, 4.C, 5.C, 6.B, 7.C, 8.C, 9.D, 10.B

Unit 3 Mode

3.1 Introduction
3.2 Learning Outcomes
3.3 Mode for Ungroup and Group Data
3.3.1 Mode for Ungroup
3.3.2 Mode for Group
3.4 Self-Assessment Exercises
3.5 Summary
3.6 References, Further Readings, and Web Sources
3.7 Possible Answers to Self-Assessment Exercises



1.1 introduction

The mode is a measure of central tendency that identifies the value or values that occur most frequently in a data set. It is the only average that can be used with nominal (categorical) data, such as colors, names, or brands. Unlike the mean and median, which are calculated using arithmetic operations, the mode simply reflects the most common observation(s).

In ungrouped data, the mode is the value that appears the most. In grouped data, the mode is estimated using a formula that identifies the modal class (the class interval with the highest frequency). The mode is especially useful when analyzing large datasets with repeated values or when identifying the most popular or common item is important, such as in market research, education, or social sciences



1.2 Learning Outcomes

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

- Identify the mode from a given set of ungrouped or grouped data.
- Differentiate mode from other measures of central tendency such as mean and median.
- Apply the appropriate formula to calculate the mode for grouped data using the modal class.



1.3 Mode for Ungroup and Group Data

1.3.1 Mode for Ungroup Data

The mode is the value that appears most frequently in a data set. Unlike the mean and median, which involve arithmetic operations, the mode is determined simply by identifying which num- ber or category occurs most often. It is one of the three common measures of central tendency and is particularly useful in situations where we are interested in knowing the most common or popular item in a set of values.

In ungrouped data, the mode is found by examining each data value and counting how many times each one appears. The value with the highest frequency is the mode. For example, in the dataset: 5, 7, 3, 7, 2, 7, 6, the mode is 7 because it appears more times than any other number. If no number repeats, we say that there is no mode. If two or more values occur with the same highest frequency, then the data is said to be bimodal or multimodal, respectively.

One of the key advantages of the mode is that it can be used with categorical data, unlike the mean or median which require numerical values. For example, if we survey the favorite colors of a group of people and find that ”blue” is mentioned most often, then ”blue” is the mode of that dataset. This makes the mode highly applicable in fields like market research, social sciences, and education, where qualitative data is common.

However, the mode also has limitations. It may not accurately represent the center of the data, especially if the dataset is small or has multiple modes. In addition, it can sometimes be misleading if the value that occurs most frequently is an outlier or does not reflect the overall trend of the data. Unlike the mean, the mode does not take into account every value in the dataset.

Despite its limitations, the mode remains a valuable and simple statistical tool. It is especially helpful when dealing with non-numeric or discrete data where other averages are not meaningful.

Example: Consider the following data set representing the number of books read by 10 students:

$$3,\ 5,\ 2,\ 4,\ 5,\ 6,\ 5,\ 3,\ 4,\ 2$$

Step 1: Count the frequency of each value:

- 2 appears 2 times
- 3 appears 2 times
- 4 appears 2 times
- 5 appears 3 times
- 6 appears 1 time

Step 2: Identify the number that occurs most frequently.

Mode = 5 (because it appears 3 times, more than any other value)

1.3.2 **Mode for Group Data**

In grouped data, the mode refers to the value that appears most frequently, but since individual data points are grouped into class intervals, we cannot determine the mode directly from the raw data. Instead, we estimate the mode using the modal class, which is the class interval with the highest frequency in the distribution table. The assumption is that the most frequently occurring value lies somewhere within this modal class.

To calculate the mode for grouped data, we use the formula:

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where:

L = lower boundary of the modal class

f_1 = frequency of the modal class

f_0 = frequency of the class before the modal class

f_2 = frequency of the class after the modal class

h = class width

This formula uses interpolation to estimate the mode based on the frequencies surrounding the modal class. The logic is that if a class has more frequency than its neighboring classes, the true modal value lies closer to its lower boundary, and the formula adjusts for how much higher the frequency is compared to the classes before and after.

The mode in grouped data is especially useful in identifying peaks in data distribution, such as in histograms, where it shows the concentration of values. It is ideal for large datasets where individual values are not visible, and is often used in fields like demography, economics, and education to understand the most common category within a dataset.

Example: Below is a frequency distribution showing the scores of students in a test:

Class Interval	Frequency
10 – 20	5
20 – 30	8
30 – 40	12
40 – 50	20
50 – 60	10
60 – 70	5

Step 1: Identify the modal class.

The class with the highest frequency is 40 – 50 (frequency = 20), so it is the modal class.

Step 2: Assign values for the formula

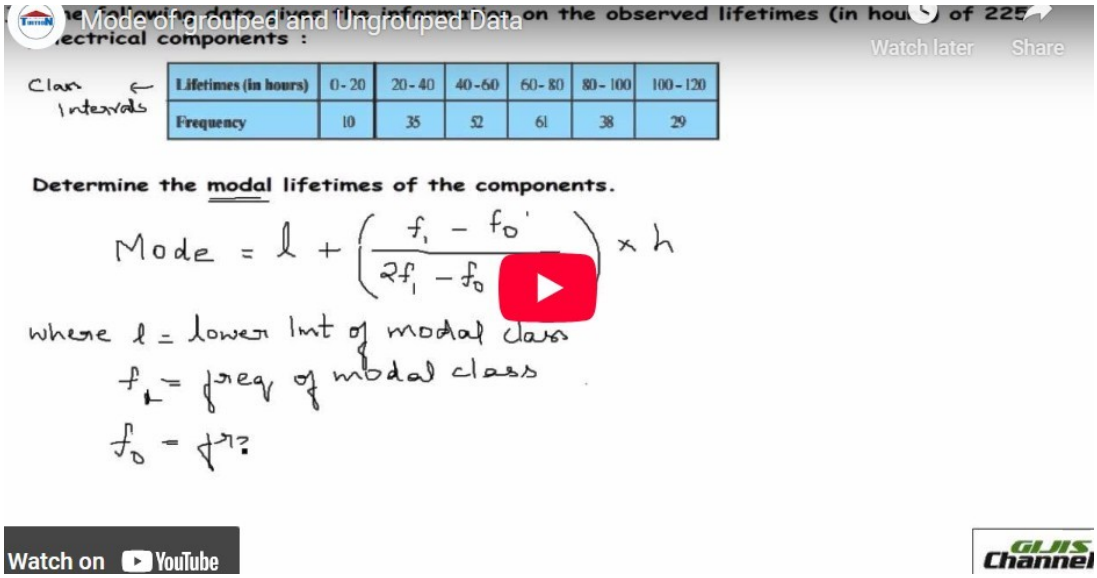
- L = 40 (lower boundary of the modal class)
- f_1 = 20 (frequency of modal class)
- f_0 = 12 (frequency of the class before: 30 – 40)
- f_2 = 10 (frequency of the class after: 50 – 60)
- h = 10 (class width)

Step 3: Apply the formula

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

$$= 40 + \frac{20 - 12}{2(20) - 12 - 10} \times 10 = 40 + \frac{8}{18} \times 10$$

$$= 40 + 4.44 = 44.44$$



You can click here to access the video.https://youtu.be/quwOqNuOZMU



1.4 Self-Assesement Exercises

1. What is the mode in a dataset?
   A. The average of values
   B. The value that appears most frequently
   C. The difference between maximum and minimum
   D. The middle value

2. Which of the following data sets is bimodal?
   A. 2, 3, 4, 5, 6
   B. 1, 2, 2, 3, 3
   C. 5, 6, 7, 8, 9
   D. 2, 2, 2, 2, 2

3. In grouped data, the class with the highest frequency is called:
   A. Median class
   B. Modal class
   C. Mean class
   D. Class midpoint

4. The mode of the ungrouped data: 4, 5, 4, 6, 7, 4, 3 is:
   A. 3
   B. 4
   C. 5
   D. 6

5. Which of the following is a limitation of mode?
   A. It considers all values
   B. It can be calculated for grouped data only
   C. It is not affected by outliers
   D. It may not exist or may not be unique

6. What does the symbol f_1 represent in the grouped mode formula?
   A. Frequency before modal class
   B. Frequency of modal class
   C. Frequency after modal class
   D. Total frequency

7. In the formula for grouped mode, what is h?
   A. Height of bar

B. Class mark
C. Class width
D. Cumulative frequency

8. What happens when two classes have the same highest frequency?
   A. Mode is undefined
   B. Mode is the average of class midpoints
   C. The distribution is bimodal
   D. Mode becomes the mean

9. Which measure of central tendency is most suitable for categorical data?
   A. Mean
   B. Median
   C. Mode
   D. Standard deviation

10. In a grouped frequency table, how is the modal class identified?
    A. The class with the smallest lower limit
    B. The class with the highest frequency
    C. The class with the widest range
    D. The class at the middle of the table



3.5. Summary

The mode is the value that occurs most frequently in a data set. It is one of the three main measures of central tendency, alongside the mean and the median. For ungrouped data, the mode is simply the number that appears most often in the data set. A data set can be unimodal (one mode), bimodal (two modes), multimodal (more than two modes), or have no mode if all values occur with equal frequency.

In the case of grouped data, where data is organized into class intervals, the mode is estimated using the modal class the class interval with the highest frequency. A formula is then applied to calculate the mode within that class, providing a more accurate representation of the most common value in the distribution.



3.6 References, Further Readings, Web Sources

Gupta, S. C., & Kapoor, V. K. (2020). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons.

Spiegel, M. R., & Stephens, L. J. (2017). *Statistics* (4th ed.). McGraw-Hill Education.



3.7 Possible Answers to SAEs

1.B, 2.B, 3.B, 4.B, 5.D, 6.B, 7.C, 8.C, 9.C, 10.B

Week Four

Geometry Mean and Harmonic Mean

4.1 Introduction
4.2 Learning Outcomes
4.3 Geometry Mean and Harmonic Mean
4.4 Self-Assessment Exercises 4.5 Summary
4.6 References, Further Readings, and Web Sources
4.7 Possible Answers to Self-Assessment Exercises



4.1 introduction

The geometric mean is a measure of central tendency that is used to find the average of a set of positive numbers by multiplying them together and then taking the n^{th} root (where n is the total number of values). It is especially useful when comparing ratios, percentages, or growth rates across time. Unlike the arithmetic mean, the geometric mean dampens the effect of very high or very low values, making it ideal in financial and economic data analysis.

The harmonic mean is another type of average that is calculated by dividing the number of observations by the sum of the reciprocals of the data values. It is most appropriate in situations involving rates or ratios, such as average speed or price per unit. The harmonic mean gives more weight to smaller values and is typically used when the data contains values expressed as ”per unit” or ”per time”.



4.2 Learning Outcomes

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

- Understand the meaning and appropriate applications of geometric and harmonic means.
- Differentiate between arithmetic, geometric, and harmonic means.
- Calculate the geometric and harmonic means of ungrouped data.



4.3 Geometry Mean and Harmonic Mean Geometric Mean (GM)

The Geometric Mean (GM) is one of the fundamental measures of central tendency used in statistics, particularly suitable when dealing with data that involves multiplication or exponential growth. Unlike the arithmetic mean, which is based on addition, the geometric mean focuses on the product of data values. It is especially appropriate in analyzing quantities that

are multiplied together or when values exhibit proportional change over time.

The geometric mean is defined as the nth root of the product of n positive values. Mathemati- cally, it is expressed as:

$$\text{Geometric Mean} = \sqrt[n]{x_1 \cdot x_2 \cdots\cdots\cdots x_n}$$

where $x_1, x_2, \ldots, x_n$ are all positive numbers. This formula requires that all values must be greater than zero, since the presence of a zero or a negative value would make the geometric mean undefined or invalid.

A key advantage of using the geometric mean is that it reduces the impact of extremely large or small values that might distort the arithmetic mean. This makes it particularly useful in fields like finance, where investment returns are compounded, or in environmental science, where measurements such as pollutant levels are multiplicative and often span several orders of mag- nitude.

Another important application of the geometric mean lies in comparing data across different units or scales, especially when values are given as ratios, percentages, or index numbers. For example, when analyzing growth factors over multiple years, the geometric mean provides a more accurate average growth rate than the arithmetic mean, which assumes linear progression rather than compounding.

Despite its usefulness, the geometric mean has some limitations. It cannot be applied to datasets containing zero or negative values, as the computation would either return zero or become mathematically undefined. Furthermore, for some students or practitioners, the geometric mean may be less intuitive than other averages due to its reliance on roots and multiplicative calcula- tions. However, understanding its strengths and limitations is essential for selecting the appro- priate average in real-world contexts.

Example: Find the geometric mean of the following ten numbers:

$$2, 4, 6, 3, 5, 7, 9, 8, 10, 6$$

Solution: Step 1: Multiply all the numbers together:

$$2 \times 4 \times 6 \times 3 \times 5 \times 7 \times 9 \times 8 \times 10 \times 6 = 21772800$$

Step 2: Take the 10th root of the product:

$$\text{Geometric Mean} = \sqrt[10]{21772800 \text{ Step}}$$

3: Calculate the 10th root (approximate value):

$$\text{Geometric Mean} \approx 5.89$$

Unit 5 Harmonic Mean

The harmonic mean is a type of average that is calculated as the reciprocal of the arithmetic mean of the reciprocals of a data set. It is most appropriate for situations involving rates, ratios, or quantities expressed in relation to time, distance, or per unit. Unlike the arithmetic mean which adds values directly, the harmonic mean gives greater weight to smaller numbers and is especially useful when averaging values like speed, density, or cost per unit.

Mathematically, for a set of n non-zero values $x_1, x_2, x_3, \ldots, x_n$, the harmonic mean (HM) is defined as:

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

This formula shows that each value in the dataset contributes to the mean through its recip- rocal. This makes the harmonic mean ideal for situations where we want to average rates. For example, if a car travels at different speeds over equal distances, the harmonic mean gives the correct average speed over the entire journey.

One of the key features of the harmonic mean is its sensitivity to small values. Even a single low number in the dataset can significantly lower the harmonic mean. This makes it different from the arithmetic and geometric means, which are less affected by smaller values. Because of this property, the harmonic mean is best suited for datasets that do not include zeros or neg- ative values, as division by zero or negative reciprocals would make the mean undefined or misleading.

In practical applications, the harmonic mean is commonly used in fields such as physics (for average speed), economics (average cost), and finance (price-to-earnings ratios). It helps pro- fessionals draw more accurate conclusions where other means might distort the result due to high outliers. However, it's important to use the harmonic mean only when appropriate, espe- cially when the values being averaged represent the same quantity type and unit.

Example: A company records the cost per unit (in Naira) of producing five different items as follows: 4, 5, 6, 10, and 12. What is the harmonic mean of the cost per unit?

Solution:

Let the data values be:

$$x_1 = 4, \quad x_2 = 5, \quad x_3 = 6, \quad x_4 = 10, \quad x_5 = 12$$

Apply the harmonic mean formula:

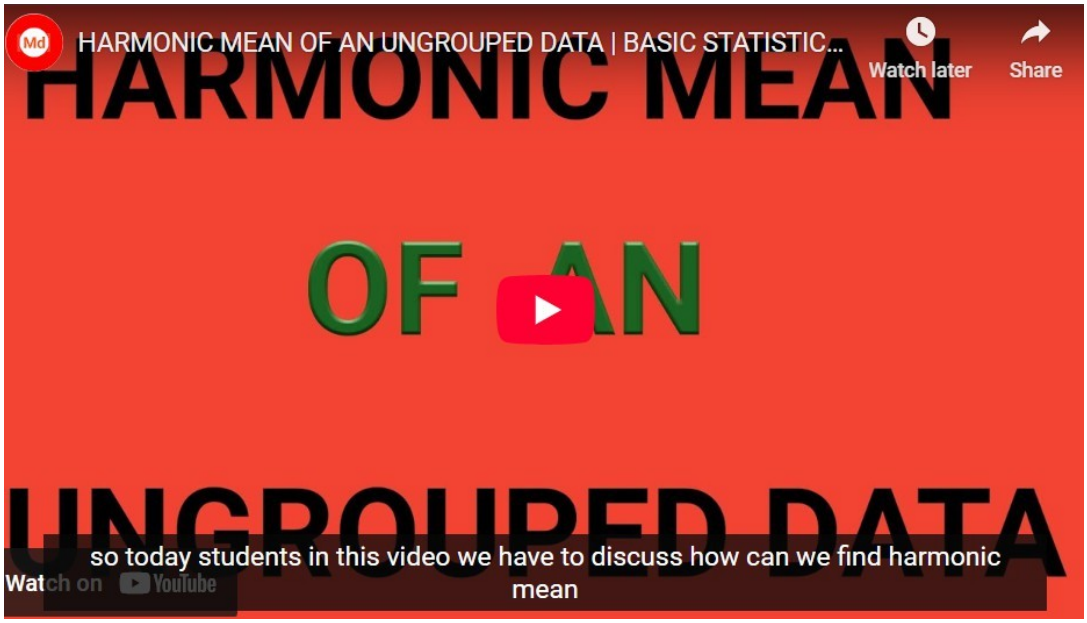
$$HM = \frac{5}{\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{10} + \frac{1}{12}}$$

Compute the reciprocals:

$$\frac{1}{4} = 0.25, \quad \frac{1}{5} = 0.20, \quad \frac{1}{6} \approx 0.1667, \quad \frac{1}{10} = 0.10, \quad \frac{1}{12} \approx 0.0833$$

$$HM = \frac{5}{0.25 + 0.20 + 0.1667 + 0.10 + 0.0833} = \frac{5}{0.80}$$

$$HM = 6.25$$



You can click here to access the video.https://youtu.be/r0r0okX76pk



4.4 Self-Assesement Exercises

1. The geometric mean is most appropriate for data involving:
   A. Simple averages B. Percentages and growth rates
   C. Frequency distribution D. Class boundaries

2. The harmonic mean is useful when dealing with:
   A. Mode and median B. Cumulative frequencies
   C. Rates and ratios D. Tallied observations

3. The geometric mean of 4 and 16 is:
   A. 10 B. 6 C. 8 D. 12

4. The harmonic mean of 4 and 6 is:
   A. 5 B. 4.8 C. 4.5 D. 4.9

5. Which of the following is not a property of the harmonic mean?
   A. Sensitive to smaller values B. Can be used for negative numbers
   C. Appropriate for rates D. Is always less than or equal to the arithmetic mean

6. The geometric mean of $x_1, x_2, \ldots, x_n$ is calculated as:
   A. $\frac{x_1 + x_2 + \cdots + x_n}{n}$
   B. $\sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$
   C. $\frac{1}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$
   D. $\frac{x_1 - x_2 + \cdots + x_n}{n}$

7. If the values are 2, 4, and 8, the geometric mean is:
   A. 5 B. 4 C. 8 D. 6

8. Harmonic mean is defined only for:
   A. Any real numbers B. Negative values only
   C. Non-zero positive numbers D. Integers only

9. The geometric mean is always:
   A. Greater than the arithmetic mean
   B. Equal to the harmonic mean
   C. Less than or equal to the arithmetic mean
   D. More sensitive to large values

10. The harmonic mean is most often used to calculate:
    A. Average marks B. Average income
    C. Average speed D. Average height



4.5. Summary

The geometric mean is a measure of central tendency used to find the average of a set of positive numbers by multiplying all the values together and then taking the nth root (where n is the total number of values). It is especially useful in situations involving growth rates, ratios, and percentages such as population growth, financial returns, or compounded interest. Unlike the arithmetic mean, the geometric mean reduces the impact of extremely high or low values, making it more representative in multiplicative processes.The harmonic mean is most applicable when averaging quantities expressed as rates or ratios, such as speed, efficiency, or cost per unit. It is calculated as the reciprocal of the average of the reciprocals of the data

values. The harmonic mean is especially useful when all data points carry equal weight in a given context, and it tends to be lower than both the arithmetic and geometric means, making it suitable for minimizing the impact of large outliers.



4.6 References, Further Readings, Web Sources

Kamat, K. (2023). *Understanding Mean, Harmonic Mean and Geometric Mean: A Comprehensive Guide*. Medium. Retrieved from https://medium.com/@karan.kamat1406/understanding-mean-harmonic-mean-and-geometric-mean-a-comprehensive-guide-f61e1310feaa

Hackaday Staff. (2024). *Intuitive Explanation Of Arithmetic, Geometric & Harmonic Mean*. Hackaday. Retrieved from https://hackaday.com/2024/08/24/intuitive-explanation-of-arithmetic-geometric-harmonic-mean



4.7 Possible Answers to SAEs

1.B, 2.C, 3.C, 4.C, 5.B, 6.B, 7.B, 8.C, 9.C, 10.C

Week Five: Measure of Partition

Unit 1 Quartiles
1.1 Introduction
1.2 Learning Outcomes
1.3 Quartiles
1.4 Self-Assessment
1.5 Summary
1.6 References, Further Readings, and Web Sources
1.7 Possible Answers to Self-Assessment Exercises



1.1 introduction

In statistics, quartiles are a type of quantile that divide a dataset into four equal parts after the data has been arranged in ascending order. Each quartile represents a portion of the dataset, helping to describe the spread and center of the data. They are essential in descriptive statistics because they provide a deeper understanding of data distribution, especially in relation to me- dians and percentiles.

Quartiles are especially useful in constructing box plots and detecting outliers in a dataset. They are widely applied in fields such as economics, medicine, education, and business for data analysis and decision-making.



1.2 Learning Outcomes

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

- Explaining the concept of quartiles as measures of position within a data distribution
- Interpret the Interquartile Range (IQR) as a measure of data variability and its use in identifying data spread.
- Explain the Semi-Interquartile Range (SIQR) and its value in describing consistency around the median.
- Analyze the robustness of IQR and SIQR in the presence of extreme values or outliers.



1.3 Quartiles

Quartiles are values that divide a dataset into four equal parts, making them a key component of descriptive statistics. When a set of data is arranged in ascending order, the quartiles help identify the distribution and spread of the data by marking three important cut-off points. These are the first quartile (Q1), the second quartile (Q2, also called the median), and the third quartile (Q3). Each quartile represents a specific percentile in the data: Q1 at the 25th percentile, Q2 at the 50th percentile, and Q3 at the 75th percentile.

The first quartile (Q1) is the value below which 25% of the data falls. It provides a measure of the lower spread of the data and can help identify lower outliers. The second quartile (Q2) is the median of the dataset, dividing it into two equal halves. This is the midpoint and is useful for understanding the center of the data. The third quartile (Q3) is the point below which 75% of the data lies, offering insight into the upper portion of the dataset.

Quartiles are especially helpful for detecting skewness in a distribution. If the quartiles are not evenly spaced, it suggests the data may be skewed to the left or right. They are also fundamen- tal in constructing boxplots, which visually summarize the central tendency, dispersion, and presence of outliers in data. These plots rely on Q1, Q2, and Q3, along with the minimum and maximum values, to represent the range and shape of the distribution.

In practice, quartiles can be calculated for both ungrouped (raw) and grouped data. For un- grouped data, the values are found by locating specific positions in the sorted data. For grouped data, quartiles are estimated using interpolation within class intervals based on cumulative fre- quencies. This makes quartiles applicable to small data sets as well as large, summarized datasets.

Formulas for Calculating Quartiles, Interquartile and Semi-Interquartile

For Ungrouped Data (Raw Data)

Let n be the total number of ordered data points:

$$Q_1 = \frac{1}{4}(n+1)^{th} \text{ value}$$

$$Q_2 = \frac{1}{2}(n+1)^{th} \text{ value}$$

$$Q_3 = \frac{3}{4}(n+1)^{th} \text{ value}$$

If the result is not a whole number, interpolate between values.

Example 1: Test scores of 8 students in Economics

Data: 18, 12, 24, 10, 14, 16, 20, 22

Sorted: 10, 12, 14, 16, 18, 20, 22, 24

$n = 8$

- $Q_1 = 2.25^{th}$ value $= 12 + 0.25(14 - 12) = 12.5$
- $Q_2 = 4.5^{th}$ value $= \frac{16 + 18}{2} = 17$
- $Q_3 = 6.75^{th}$ value $= 20 + 0.75(22 - 20) = 21.5$

For Grouped Data (Frequency Distribution
Use the formula:

$$Q_k = L + \left(\frac{kN}{4} - F\right) \times \frac{h}{f}, \quad \text{where } k = 1, 2, 3$$

- L: lower boundary of quartile class
- N: total frequency
- F: cumulative frequency before quartile class
- f: frequency of the quartile class
- h: class width

Example 2: Distribution of scores from 30 students in a Mathematics test

Class Interval	Frequency	Cumulative Frequency
0–10	4	4
10–20	6	10
20–30	10	20
30–40	8	28
40–50	2	30

Total Frequency: $N = 30$

First Quartile (Q_1)

$$\frac{N}{4} = \frac{30}{4} = 7.5 \Rightarrow Q_1 \text{ is in } 10\text{–}20$$

$$Q_1 = 10 + \frac{7.5 - 4}{6} \cdot 10 = 10 + 5.83 = 15.83$$

Second Quartile (Q_2)

$$\frac{N}{2} = 15 \Rightarrow Q_2 \text{ is in } 20\text{–}30$$

$$Q_2 = 20 + \frac{15 - 10}{10} \times 10 = 20 + 5 = 25$$

Third Quartile (Q_3)

$$\frac{3N}{4} = 22.5 \Rightarrow Q_3 \text{ is in } 30\text{–}40$$

$$Q_3 = 30 + \frac{22.5 - 8}{20} \times 10 = 30 + 3.13 = 33.13$$

1.3.1 Interquartile Range (IQR)

The interquartile range (IQR) is a statistical measure used to describe the spread of the middle 50% of a dataset. It is calculated by subtracting the first quartile (Q_1) from the third quartile (Q_3), i.e., IQR = $Q_3 - Q_1$. This range focuses on the central part of the data, leaving out the lowest 25% and the highest 25%, making it very useful in analyzing data that may contain extreme values or outliers.

IQR is especially valuable when trying to understand the variability within a dataset while min- imizing the effect of skewed data or anomalies. Unlike the total range, which includes the maximum and minimum values (and can be heavily influenced by a single outlier), the IQR gives a better sense of where most of the data lies. This makes it a robust measure of disper- sion.

In practice, the IQR is used in boxplots (box-and-whisker plots) to visually show the spread of data. In such plots, the box spans from Q_1 to Q_3, and the line inside the box shows the median (Q_2). Any points lying beyond 1.5 times the IQR above Q_3 or below Q_1 are considered potential outliers.

Therefore, the IQR is not only a numerical measure of spread but also a tool for identifying outliers and understanding the overall structure of a dataset. It is commonly used in fields such as finance, healthcare, and education where data consistency and reliability are crucial.

1.3.2 Semi-Interquartile Range (SIQR)

The semi-interquartile range (SIQR) is half of the interquartile range and provides a measure of the average distance between the first and third quartiles from the median. It is calculated as SIQR = $\frac{Q_3 - Q_1}{2}$. This measure gives a concise summary of the spread of the middle half of a dataset, showing how tightly the data is grouped around the center.

SIQR is particularly useful in situations where we want to communicate the variability of the data in a simpler form. By halving the IQR, it reflects how far typical values deviate from the median, providing a cleaner picture of central concentration, especially in data that are not symmetrically distributed.

Because it ignores both the highest and lowest 25% of the data, the SIQR is considered a resis- tant measure of spread, meaning it is not influenced by outliers or extreme values. This makes it ideal for summarizing the variability in skewed distributions or in data with irregular patterns. In essence, the SIQR complements the IQR by offering a focused view of the consistency of the middle values. It is frequently used in reports and academic studies where a more compact summary of variation is preferred, especially when comparing datasets or reporting variability in a simple and understandable way.



1.4 Self-Assesement Exercises

1. What does the first quartile (Q1) represent?

   (A) 10% of the data

   (B) 25% of the data below it

   (C) 50% of the data below it

   (D) The average of the data

2. Which of the following best describes the interquartile range (IQR)?

   (A) The difference between the highest and lowest values

   (B) The range of all data points

   (C) The difference between Q1 and Q2

   (D) The difference between Q3 and Q1

3. What does Q2 usually represent?

   (A) Mode

   (B) Median

   (C) Mean

   (D) Minimum

4. Which range contains the middle 50% of a dataset?

   (A) Q1 to Q3

   (B) Minimum to Maximum

   (C) Q2 to Q3

   (D) Mean to Median

5. What is the Semi-Interquartile Range (SIQR)?

   (A) Q1 + Q3

   (B) Half of the IQR

   (C) Double the IQR

   (D) Mean - Q2

6. Which measure is most resistant to outliers?

   (A) Range

   (B) Mean

   (C) Interquartile Range

   (D) Standard deviation

7. If Q1 = 40 and Q3 = 60, what is the SIQR?

   (A) 10

   (B) 20

   (C) 100

   (D) 40

8. What type of data is appropriate for calculating quartiles?

   (A) Categorical data

   (B) Nominal data

   (C) Ordinal or numerical data

   (D) Binary data

9. What does a large IQR indicate?

   (A) Low variability

   (B) High concentration around the mean

   (C) High variability

   (D) All data are identical

10. If IQR = 80, what is the SIQR?

    (A) 20

    (B) 80

    (C) 40

    (D) 160



1.5. Summary

Quartiles are values that divide a dataset into four equal parts. The first quartile (Q1) marks the 25th percentile, the second quartile (Q2 or median) marks the 50th percentile, and the third

quartile (Q3) marks the 75th percentile. They help describe the spread and center of data, especially in non-parametric statistics.

The Interquartile Range (IQR) is the range between the third and first quartiles, given by:

$$\text{IQR} = Q_3 - Q_1$$

It measures the middle 50% of the data and is useful for identifying outliers and understanding data variability.

The Semi-Interquartile Range (SIQR) is half of the IQR:

$$\text{SIQR} = \frac{Q_3 - Q_1}{2}$$

It provides a measure of spread around the median and is less affected by extreme values, mak- ing it a robust measure of dispersion.



1.6 References, Further Readings, Web Sources

- Diez, D. M., Barr, C. D., & C¸ etinkaya-Rundel, M. (2024). *OpenIntro Statistics* (2nd ed.). Open Intro, Inc.
- Ross, S. M. (2023). *Introductory Statistics* (5th ed.). Academic Press.



1.7 Possible Answers to SAEs

1. B 2. D 3. B 4. A 5. B 6. C 7. A 8. C 9. C 10. C

Unit 2 Deciles

2.1 Introduction
2.2 Learning Outcomes
2.3 Deciles
2.4 Self-Assessment
2.5 Summary
2.6 References, Further Readings, and Web Sources
2.7 Possible Answers to Self-Assessment Exercises



2.1 introduction

Deciles are a type of quantile that divide a data set into ten equal parts. Each decile represents 10% of the ordered data, making it a useful measure to understand the distribution and spread of a dataset.

- The first decile (D_1) is the value below which 10% of the data fall.
- The fifth decile (D_5) is the median, splitting the data into two equal halves.
- The ninth decile (D_9) marks the point below which 90% of the data lie.

Deciles help us observe how data is concentrated and can be especially helpful when comparing relative performance, income levels, or other ranked information in fields such as economics, education, and public health.



2.2 Learning Outcomes

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

- Explain what deciles are and how they divide a dataset into ten equal parts.
- Describe the real-life relevance of deciles in interpreting distributions such as income or performance.
- Explain the usefulness of deciles in understanding inequality and data spread.



2.3 Deciles

Deciles are statistical measures used to divide a dataset into ten equal parts after arranging the data in ascending order. Each decile represents 10% of the data, making them useful for ana- lyzing the spread, distribution, and concentration of values. They are part of a broader family of measures called quantiles, which include quartiles, percentiles, and medians.

There are nine decile points in any dataset, labeled D_1 to D_9. These divide the data into ten segments of equal frequency. For example, D_1 marks the point below which 10% of the data lie, D_2 marks the point below which 20% lie, and so on. The fifth decile D_5 corresponds to the median, and D_9 indicates the value below which 90% of the data lie.

To determine deciles, the data must first be ordered from smallest to largest. Each decile is then located using its respective position formula depending on the total number of observations. Although there are several formulas for computing decile positions, the most basic concept in- volves locating the value that separates each 10% block of the data. Interpolation may be used if the position falls between two values.

Deciles are commonly used in fields such as economics, education, and social sciences to assess inequality, rank individuals or groups, or evaluate performance. For instance, income deciles are used to classify populations from the poorest 10% to the richest 10%. In education, deciles can be used to identify students in the top or bottom 10% of exam performance.

Interpreting deciles involves understanding where a particular observation falls within the dis- tribution. For example, if a student scores in the 8th decile, it means they performed better than 70%–80% of their peers. This kind of information can be crucial for making decisions about interventions, scholarships, or policy changes.

While deciles are similar to quartiles (which divide data into four parts) and percentiles (which divide data into 100 parts), they strike a balance between detail and simplicity. Deciles provide a more detailed view than quartiles but are less granular than percentiles.

1. Deciles for Ungrouped Data

To find the k^{th} decile (D_k) in ungrouped data:

$$D_k = \frac{k(n+1)}{10}^{th} \text{ item}, \quad \text{for } k = 1, 2, \ldots, 9$$

where:

- n is the number of observations,
- k indicates the decile position (e.g., $k = 1$ for D_1, $k = 5$ for D_5).

2. Deciles for Grouped Data

To calculate the k^{th} decile (D_k) in grouped data:

$$D_k = L + \left(\frac{\frac{kN}{10} - F}{f} \right) \times h$$

where:

- L = lower boundary of the decile class,
- N = total frequency,
- F = cumulative frequency before the decile class,
- f = frequency of the decile class,
- h = class width.

Example 1: Given the following dataset for mathematics test scored:

5, 8, 10, 12, 13, 14, 16, 18, 20, 22

Arrange in ascending order (already sorted). There are $n = 10$ observations. To find the third decile (D_3):

$$D_3 = \frac{3(n+1)}{10}^{th} = \frac{3(11)}{10}^{th} = 3.3^{rd} \text{ item}$$

Interpolate between the 3rd and 4th items:

$$D_3 = 10 + 0.3(12 - 10) = 10 + 0.6 = 10.6$$

Example 2: Consider the grouped frequency table below:

Class Interval	Frequency
0–9	5
10–19	8
20–29	12
30–39	7
40–49	3

Total frequency: $N = 5 + 8 + 12 + 7 + 3 = 35$ To find the 5th decile (D_5):

$$\frac{5N}{10} = \frac{5 \times 35}{10} = 17.5$$

Now compute the cumulative frequencies:

- Cumulative frequency before 0–9: 0

- Cumulative frequency before 10–19: 5
- Cumulative frequency before 20–29: 5 + 8 = 13
- Cumulative frequency before 30–39: 13 + 12 = 25

17.5 falls in the 20–29 class, so:

- $L = 20$,
- $F = 13$,
- $f = 12$,
- $h = 10$

Apply the formula:

$$D_5 = 20 + \frac{17.5 - 13}{12} \times 10 = 20 + \frac{4.5}{12} \times 10 = 20 + 3.75 = 23.75$$



2.4 Self-Assesement Exercises

1. What do deciles divide a dataset into?
   A. Four equal parts B. Five equal parts C. Ten equal parts D. One hundred equal parts

2. The fifth decile (D_5) corresponds to which measure?
   A. Mode B. Mean C. Median D. Range

3. What is the formula to calculate the k^{th} decile for ungrouped data?
   A. $\frac{k(n)}{10}$ B. $\frac{k(n+1)}{10}$ C. $\frac{k(n-1)}{10}$ D. $\frac{k(n+5)}{10}$

4. In grouped data, which of the following is **not** part of the decile formula?
   A. Lower boundary B. Cumulative frequency before the class C. Total frequency D. Mode

5. What is the class width in the interval 10–19?
   A. 10 B. 8 C. 9 D. 11

6. If a value falls at the 3.3rd position in a sorted ungrouped dataset, how is the decile calculated?
   A. Take the 4th item directly B. Interpolate between the 3rd and 4th items C. Take the 3rd item directly D. Average the 3rd and 4th items

7. Which field most commonly uses deciles to assess income levels?
   A. Chemistry B. Literature C. Economics D. Music

8. What is the purpose of using deciles in data analysis?
   A. To count the number of items B. To sort data alphabetically C. To understand distribution and concentration of values D. To remove outliers

9. If the total frequency in a grouped dataset is 35, what value is used to find D_5?
   A. $5\frac{\times}{5}\frac{35}{}$ B. $\frac{5 \times 35}{100}$ C. $\frac{5 \times 35}{10}$ D. 35

10. Which of the following is true about deciles?
    A. They are unrelated to percentiles B. They only apply to ungrouped data C. They divide data into 100 parts D. They are quantiles that divide data into 10 equal parts



2.5. Summary

Deciles are statistical measures that divide a dataset into ten equal parts, each representing 10% of the ordered data. The first decile (D_1) is the point below which 10% of the data fall, while the ninth decile (D_9) marks the value below which 90% of the data fall. The fifth decile (D_5) is equivalent to the median. Deciles help to analyze data distribution, concentration, and inequal- ity. They are particularly useful in fields such as economics, education, and social sciences. Deciles can be calculated differently depending on whether the data is ungrouped or grouped.



2.6 References, Further Readings, Web Sources

Diez, D. M., Barr, C. D., & C¸ etinkaya-Rundel, M. (2024). *OpenIntro Statistics* (2nd ed.). Open Intro, Inc.

Ross, S. M. (2023). *Introductory Statistics* (5th ed.). Academic Press.



2.7 Possible Answers to SAEs

1.C 2.C 3.B 4.D 5.A 6.B 7.C 8.C 9.C 10.D

Unit 1 Percentiles

3.1 Introduction
3.2 Learning Outcomes
3.3 Percentiles
3.4 Self-Assessment
3.5 Summary
3.6 References, Further Readings, and Web Sources
3.7 Possible Answers to Self-Assessment Exercises



3.1 introduction

Percentiles are statistical measures that divide a dataset into one hundred equal parts. Each percentile indicates the value below which a given percentage of the data falls. For example, the 25th percentile (P_{25}) represents the point below which 25% of the observations lie, while the 90th percentile (P_{90}) marks the value below which 90% of the data are found. Percentiles are especially useful when comparing individual performance or status against a larger group, such as test scores, income levels, or health indicators like child growth charts.



3.2 Learning Outcomes

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

- Explain what percentiles represent and how they divide a dataset.
- Interpret percentile values in real-world situations like test scores or health statistics.
- Differentiate percentiles from other statistical measures such as quartiles and deciles.



3.3 Percentiles

Percentiles are powerful descriptive statistics that divide a dataset into one hundred equal parts. Each percentile indicates the value below which a given percentage of the data falls. For in- stance, the 60th percentile (P_{60}) means 60% of the data points lie below that value. Percentiles provide insights into the relative standing of a value within a dataset and are widely used in many fields such as education, health, and business analytics.

The 50th percentile, also known as the median, is the middle value of a data set when arranged in order. Other commonly used percentiles include the 25th percentile (first quartile or Q_1) and the 75th percentile (third quartile or Q_3). These points are essential in understanding the spread and skewness of a dataset and are frequently used to construct boxplots and identify outliers.

Percentiles are particularly useful in interpreting standardized test results or medical records. For example, if a student's score is in the 85th percentile, it means that the student scored better than 85% of all test-takers. Similarly, if a child's height is at the 30th percentile for their age, it means 30% of children of the same age are shorter.

There are two main approaches to calculating percentiles: one for ungrouped (raw) data and another for grouped (frequency) data. Each method has its own formula and technique depend- ing on the structure of the data.

For ungrouped data, the percentile is found using the position formula:

$$P_k = \frac{k(n+1)}{100}^{\text{th}} \text{ item}$$

where P_k is the k^{th} percentile, k is the desired percentile (1 to 99), and n is the number of ordered observations. If the result is not a whole number, interpolation between two values is used to obtain the percentile.

Example 1: Suppose the dataset is: 7, 12, 15, 18, 20, 21, 25, 27, 30, 35. To find the 40th percentile (P_{40}), use:

$$P_{40} = \frac{40(10+1)}{100} = 4.4^{\text{th item}}$$

The 4th value is 18 and the 5th is 20, so interpolate:

$$P_{40} = 18 + 0.4(20 - 18) = 18.8$$

Thus, the 40th percentile is 18.8.

For grouped data, percentiles are calculated using the following formula:

$$P_k = L + \left(\frac{\frac{kN}{100} - F}{f} \right) \times h$$

where L is the lower boundary of the percentile class, N is the total frequency, F is the cumulative frequency before the percentile class, f is the frequency of the percentile class, and h is the class width.

Example 2: Consider the grouped frequency distribution below:

Class Interval	Frequency
0–9	5
10–19	8
20–29	12
30–39	10
40–49	5

The total frequency $N = 40$. To find the 30th percentile (P_{30}), compute:

$$\frac{30}{100} \times 40 = 12$$

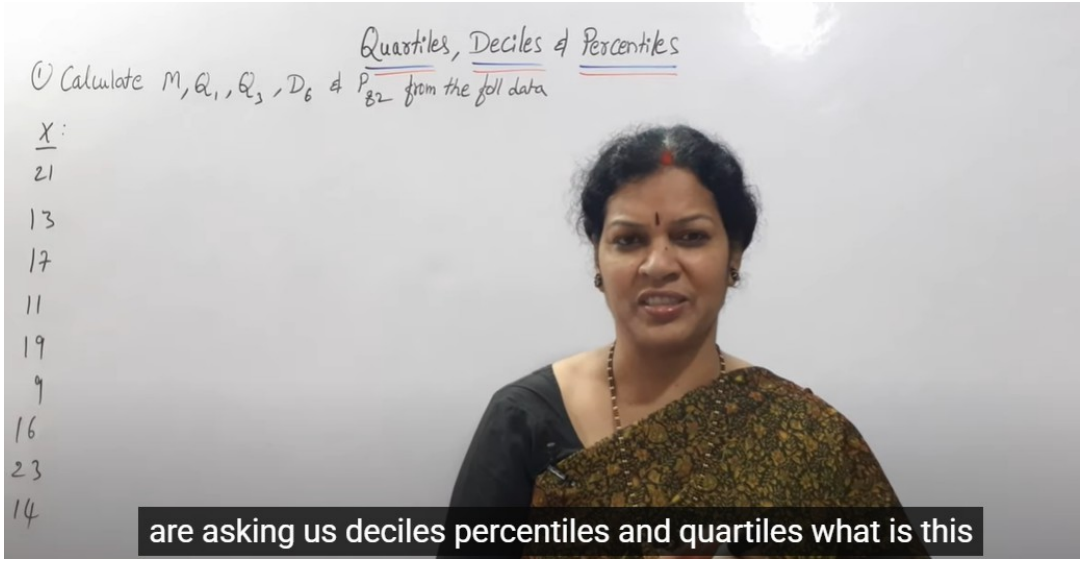
The 12th value falls in the 20–29 class. So:

- $L = 20$ (lower boundary),
- $F = 5 + 8 = 13$ is cumulative frequency before the class,
- $f = 12$,
- $h = 10$

Now apply the formula:

$$P_{30} = 20 + \frac{12 - 13}{12} \times 10 = 20 + \frac{1}{-12} \times 10 = 20 - 0.83 = 19.17$$

So, the 30th percentile is approximately 19.17.



You can click here to watch the video.https://youtu.be/CJMi91OB3PA?si=gRZZmOuEwCfgMcXx



3.4 Self-Assesement Exercises

1. How many equal parts do percentiles divide a dataset into?
   A. 10 B. 4 C. 100 D. 25

2. The 50th percentile corresponds to which statistical measure?
   A. Mode B. Mean C. Quartile D. Median

3. What does the 90th percentile represent in a dataset?
   A. 90 data points B. Top 10 values C. Value below which 90% of data lie D. Maximum value

4. Which of the following is the correct formula for calculating the k^{th} percentile in ungrouped data?
   A. $\frac{k(n)}{100}$ B. $\frac{k(n+1)}{100}$ C. $\frac{k(n-1)}{100}$ D. $\frac{k}{n+1}$

5. What should be done when the position for a percentile in ungrouped data is not a whole number?
   A. Use nearest value B. Ignore the decimal C. Interpolate between two values D. Use maximum value

6. Which of the following values is required to calculate percentiles in grouped data?
   A. Median B. Mean C. Mode D. Class width

7. In the formula for percentiles in grouped data, what does F represent?
   A. Final value B. Cumulative frequency before the percentile class C. Frequency of percentile class D. Full data count

8. What is the percentile class in grouped data?
   A. Class containing the average B. Class with the smallest frequency C. Class containing the $\frac{kN}{100}$ th value D. Class with mode

9. What is the main difference between grouped and ungrouped data in percentile calculation?
   A. Grouped data uses decimal values only B. Grouped data uses class intervals and interpolation C. Ungrouped data uses frequency D. Ungrouped data is always larger

10. In percentile analysis, what does a lower percentile indicate?
    A. Lower position in dataset B. Higher value in dataset C. Maximum score D. More accuracy



3.5 Summary

Percentiles are statistical measures that divide a dataset into 100 equal parts. Each percentile indicates the value below which a specific percentage of data falls. They help us understand the position of a value within a dataset and are widely applied in education, health, and economics. Two main methods are used to calculate percentiles: one for ungrouped (raw) data, using a position-based formula, and the other for grouped (frequency) data, using an interpolation for- mula based on class boundaries, frequencies, and cumulative frequencies. Percentiles are es- sential for identifying distribution patterns, understanding variability, and locating outliers in data.



3.6 References, Further Readings, Web Sources

Barr, C. D., & C¸ etinkaya-Rundel, M. (2024). *OpenIntro Statistics* (2nd ed.). Open Intro, Inc.

Ross, S. M. (2023). *Introductory Statistics* (5th ed.). Academic Press.



3.7 Possible Answers to SAEs

1.C 2.D 3.C 4.B 5.C 6.D 7.B 8.C 9.B 10.A

Week Six

Variance and Standard Deviation of Ungroup and Group Data

3.1 Introduction
3.2 Learning Outcomes
3.3 please of or for Variance and Standard Deviation of Ungroup and Group Data
3.3.1 Variance
3.3.2 Standard deviation
3.4 Seft-Assessment Exercises 3.5 Summary
3.6 References, Further Readings, and Web Sources
3.7 Possible Answers to Self-Assessment Exercises



3.1 introduction

Variance and standard deviation are important measures in statistics that describe how data val- ues are spread out around the mean. While the mean gives us the central value of a data set, it does not show how much the values differ from one another. Variance helps us measure this by calculating the average of the squared differences between each data point and the mean. A higher variance means that the data points are more spread out, while a lower variance shows that they are closer to the mean.

Standard deviation, on the other hand, is the square root of variance and is often preferred because it is expressed in the same units as the original data. This makes it easier to interpret compared to variance. Both variance and standard deviation are widely used in real life applications such as analyzing exam scores, measuring financial risks, and understanding scientific data. They provide a clearer picture of consistency, reliability, and variation within a set of values.



3.2 Learning Outcomes

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

- Define variance and standard deviation in statistics.
- Differentiate between ungrouped and grouped data.
- Calculate variance and standard deviation for both ungrouped and grouped data sets.



3.3 Variance and Standard Deviation of Ungroup and Group Data

3.3.1 Variance Deviation of Ungroup and Group Data

Variance is a measure of how far data values are spread out from the mean. In grouped data, it shows whether the class midpoints are close to the mean or widely scattered. A small variance indicates that most values are near the mean, while a large variance shows that values are spread over a wider range. Squaring the deviations ensures that positive and negative differences do not cancel each other and gives more weight to larger deviations.

In grouped data, the actual data values are summarized into classes with specific frequencies. Each class is represented by its midpoint, and the frequency indicates how many observations fall in that class. Variance for grouped data uses these midpoints and frequencies to calculate a weighted measure of how spread out the data are around the mean.

The formula for variance in ungroup and grouped data are given below:

$$\text{Variance: } \sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

Where:

x_i =individual observation

$\bar{x}$ = mean of the data

N =total number of observations (population)

n =total number of observations (sample)

$$\text{Variance: } \sigma^2 = \frac{\sum f(x - \bar{x})2}{\sum f}$$

Where x is the class midpoint, f is the frequency of the class, and $\bar{x} = \frac{\sum fx}{\sum f}$ is the mean of grouped data. The denominator, $\sum f$, represents the total number of observations.

3.3.2 Standard Deviation of Ungroup and Group Data

Standard deviation is a statistical measure that shows how spread out the values in a data set are around the mean. Unlike variance, which is expressed in squared units, standard deviation is in the same units as the original data. This makes it easier to interpret. A smaller standard deviation indicates that the data values are close to the mean, while a larger standard deviation shows that the values are more spread out.

For ungrouped data, standard deviation is calculated by first finding the mean of the data. Then, the deviation of each observation from the mean is squared, and the sum of these squared devi- ations is divided by the total number of observations (for population) or one less than the total number (for sample). The square root of this result gives the standard deviation.

Formula for Ungrouped Data:

Standard deviation: $\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$

Where x_i is each observation, $\bar{x}$ is the mean, N is the total number of observations in the population, and n is the number of observations in the sample.

For grouped data, the calculation uses class midpoints and frequencies. Each class midpoint represents all values in that class, and its frequency shows how many observations fall into that class. The deviations of midpoints from the mean are squared, multiplied by the class fre- quency, summed, and then divided by the total frequency (for population) or one less than the total frequency (for sample). Taking the square root gives the standard deviation.

Formula for Grouped Data:

$$\text{Standard deviation: } \sigma = \sqrt{\frac{\sum f(x - \bar{x})2}{\sum f}}$$

Where x is the class midpoint, f is the frequency of the class, and $\bar{x} = \frac{\sum fx}{\sum f}$ is the mean of the grouped data. These formulas provide a measure of how spread out the data are in the same units as the original values.

Example 1 Consider the following ungrouped data and compute the variance

Data: 5, 7, 3, 7, 9, 4, 6, 8, 5, 10

Step 1 – Find the mean:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5 + 7 + 3 + 7 + 9 + 4 + 6 + 8 + 5 + 10}{10} = \frac{64}{10} = 6.4$$

Step 2 – Compute squared deviations:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	-1.4	1.96
7	0.6	0.36
3	-3.4	11.56
7	0.6	0.36
9	2.6	6.76
4	-2.4	5.76
6	-0.4	0.16
8	1.6	2.56
5	-1.4	1.96
10	3.6	12.96

$$\sum (x_i - \bar{x})^2 = 44.4$$

Step 3 – Compute Variance :

$$\text{Population variance: } \sigma^2 = \frac{44.4}{10} = 4.44$$

Example 2: Consider the following 10 classes of group data and compute variance and

standard deviation **Data:**

Class	Midpoint (x)	Frequency (f)
0–4	2	2
5–9	7	3
10–14	12	4
15–19	17	3
20–24	22	2
25–29	27	1
30–34	32	1
35–39	37	1
40–44	42	1
45–49	47	1

Step 1 – Compute the mean:

$$\bar{x} = \frac{(2\times 2) + (7\times 3) + (12\times 4) + (17\times 3) + (22\times 2) + (27\times 1) + (32\times 1) + (37\times 1) + (42\times 1) + (47\times 1)}{2+3+4+3+2+1+1+1+1+1} \approx \frac{353}{19} \approx 18.58$$

Step 2 – Compute $f(x-\bar{x})^2$:

x	f	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
2	2	-16.58	274.99	549.98
7	3	-11.58	134.14	402.42
12	4	-6.58	43.29	173.16
17	3	-1.58	2.50	7.50
22	2	3.42	11.69	23.38
27	1	8.42	70.87	70.87
32	1	13.42	180.11	180.11
37	1	18.42	339.34	339.34
42	1	23.42	548.66	548.66
47	1	28.42	807.71	807.71

$$\sum f(x-\bar{x})^2 = 3303.13$$

Step 3 – Compute Variance and Standard Deviation:

$$\text{Population variance: } \sigma^2 = \frac{3303.13}{19} \approx 173.85$$

$$\text{Population SD: } \sigma = \sqrt{173.85} \approx 13.19$$



3.4 Self-Assesement Exercises

1. Which of the following best describes the variance of a data set?

   (a) The average of the data points

   (b) The square of the mean

   (c) The average of the squared deviations from the mean

   (d) The difference between the largest and smallest value

2. The standard deviation is:

   (a) Always smaller than the variance

   (b) The square root of the variance

   (c) Equal to the mean

   (d) The sum of the deviations from the mean

3. If all values in a data set are equal, the standard deviation is:

   (a) Equal to the mean

   (b) Zero

   (c) One

   (d) Cannot be determined

4. A data set has values: 4, 6, 8, 10. What is the variance?

   (a) 5

   (b) 6.6667

   (c) 8

   (d) 4

5. For the same data set (4, 6, 8, 10), what is the standard deviation?

   (a) 2.58

   (b) 2.65

   (c) 2.73

   (d) 2.89

6. Which of the following statements is TRUE?

(a) Variance can never be negative

(b) Standard deviation can be negative

(c) Variance is measured in the same units as the data

(d) Standard deviation is always larger than the range

7. A sample of 5 numbers: 3, 7, 5, 9, 11. What is the sample variance?

(a) 10

(b) 8

(c) 9.5

(d) 12.5

8. Why do we square the deviations in variance?

(a) To make all deviations positive

(b) To give more weight to larger deviations

(c) Both A and B

(d) To make calculation easier

9. Which measure is more affected by extreme values?

(a) Mean

(b) Variance

(c) Standard deviation

(d) Both Variance and Standard deviation

10. A data set has a variance of 25. What is the standard deviation?

(a) 5

(b) 25

(c) 50

(d) 12.5

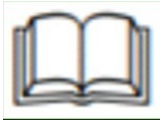


3.5 Summary

Variance and standard deviation are measures of the spread or dispersion of a data set. Variance, denoted as σ^2, represents the average of the squared deviations of each data point from the mean, providing a quantitative measure of how data points deviate from the central value.

Standard deviation, denoted as σ , is the square root of the variance and expresses the dispersion in the same units as the original data. Both measures are always non-negative, and larger values indicate greater variability within the data set.

For ungrouped data, variance and standard deviation are calculated directly from individual data points, while for grouped data, class midpoints and frequencies are used to compute these measures. These statistics are widely applied in fields such as research, finance, and quality control to assess variability, consistency, and risk. A data set in which all values are equal has a variance and standard deviation of zero, reflecting the absence of dispersion among the data points.



3.6 References, Further Readings, Web Sources

(2021). *Elementary Statistics* (14th ed.). Pearson.

Diez, D. M., Barr, C. D., & C¸ etinkaya-Rundel, M. (2021). *OpenIntro Statistics* (3rd ed.). OpenIntro.



3.7 Possible Answers to SAEs

1. C 2. B 3. B 4. B 5. A 6. A 7. C 8. C 9. D 10. A

Module Five: Correlation and Regression Analysis

- Unit 1: Correlation Analysis
- Unit 2: Spearman's Rank Correlation Coefficient
- Unit 3:Simple linear regression
- Unit 4: Coefficient of Determination (R^2)
- Unit 5: Assumptions and Limitations of Correlation and Regression

Unit 1 Correlation Analysis
1.1 Introduction
1.2 Learning Outcomes
1.3 Correlation Analysis
1.3.1 Uses of correlation
1.3.2 Types of correlation
1.3.3 Product Moment Correlation Coefficient
1.4 Seft-Assessment Exercises
1.5 Summary
1.6 References, Further Readings, and Web Sources
1.7 Possible Answers to Self-Assessment Exercises



1.1 introduction

In statistics, understanding how two variables are related is a key step in analyzing data and making informed decisions. Correlation analysis provides a systematic way to measure the strength and direction of the relationship between two variables, allowing us to identify whether changes in one are associated with changes in the other. This concept is widely applied across fields such as business, science, health, and social studies, making it an essential tool for interpreting patterns and drawing meaningful conclusions from data.



1.2 Learning Outcomes

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

- Describe the concept and purpose of correlation in statistical analysis.
- Distinguish between positive, negative, and zero correlation.
- Calculate and interpret correlation coefficients.



1.3 Correlation Analysis

In statistics, correlation is simply about understanding how two things move in relation to each other. It helps us see whether changes in one variable are linked to changes in another. For ex- ample, we might want to know if students who spend more time studying tend to score higher on exams, or if an increase in temperature leads to higher electricity consumption. Correlation gives us a way to measure this connection and express it in numbers, making it easier to see and explain patterns in data.

We use correlation in many areas of life and research. In business, it might help a company see whether advertising spend is connected to sales. In health, it could show if exercise is related to lower blood pressure. In education, it might reveal how class attendance relates to academic performance. Often, correlation is a first step before more detailed analysis, like regression, because it gives us an early idea of whether a relationship exists and how strong it might be.

When carrying out correlation analysis, we usually start by gathering pairs of data for the two variables we're studying and then plot them on a scatter diagram. This simple graph lets us quickly spot patterns whether the points slope upward, slope downward, or show no clear direction at all. Once we have that visual clue, we calculate a correlation value to put a precise number on the relationship, which helps us compare different sets of data or track how a relationship changes over time.

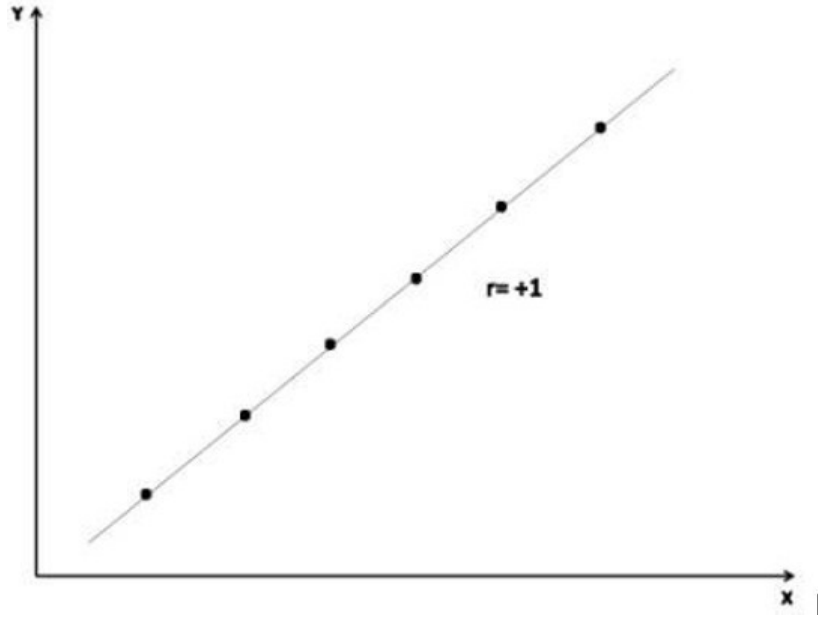
1.3.1 Uses of Correlation

- It helps in predicting the value of one variable from another related variable.
- It measures the degree and direction of association between variables.
- It assists in making informed decisions by revealing significant relationships.
- It identifies factors that may influence outcomes in research and experiments.
- It supports quality control by linking process variables to product performance.

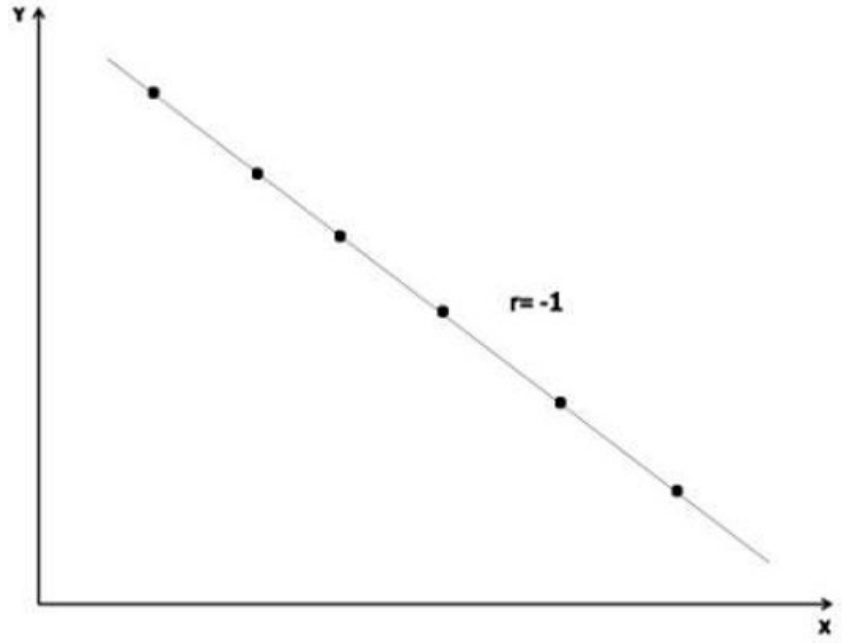
- It aids in planning and policy formulation through evidence-based relationships.

1.3.2 Types of Correlation

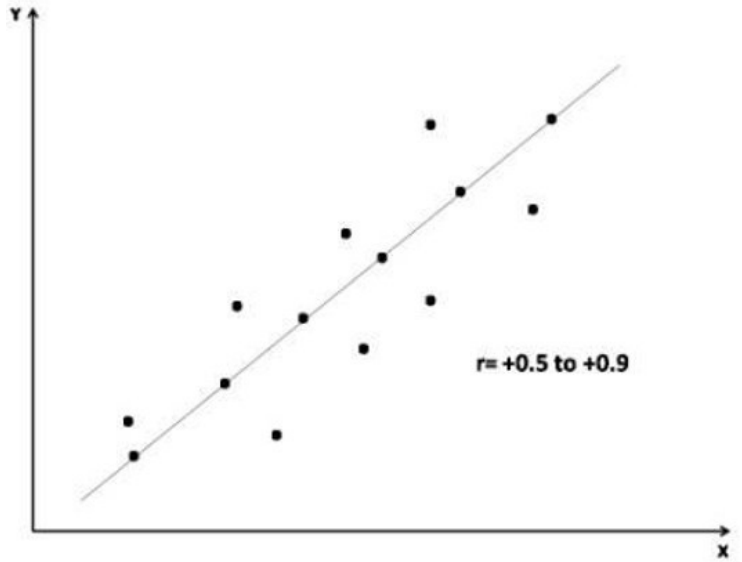
1. Perfect positive correlation: A perfect positive correlation occurs when two variables move in exactly the same direction, with every increase in one variable being matched by a propor- tional increase in the other. In this case, the correlation coefficient takes the maximum value of +1, indicating a complete and consistent linear relationship. For example, if the number of study hours and exam scores were perfectly positively correlated, then any increase in study time would lead to an exact and predictable increase in scores. Such a relationship is rarely ob- served in real-life data, but it serves as an important theoretical benchmark for understanding the strength of associations between variables.



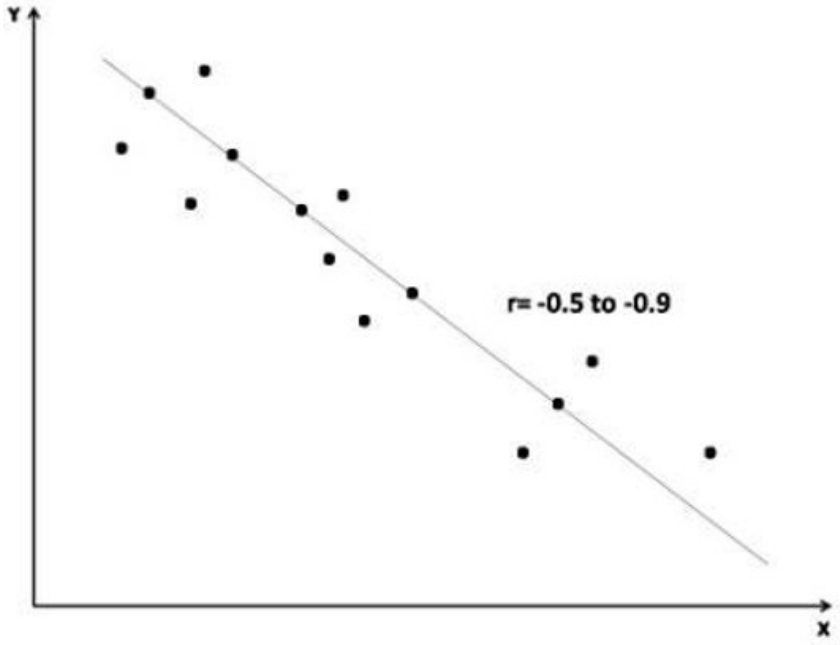
2. Perfect negative correlation: A perfect negative correlation occurs when two variables move in exactly opposite directions with complete consistency. In this case, whenever one variable increases, the other decreases in the same proportion, giving a correlation value of −1. For example, if every additional hour spent on leisure reduces study time by exactly one hour, the two variables would exhibit a perfect negative correlation. While such exact relationships are uncommon in real-life situations, the concept helps illustrate the strongest possible opposite relationship between two factors.



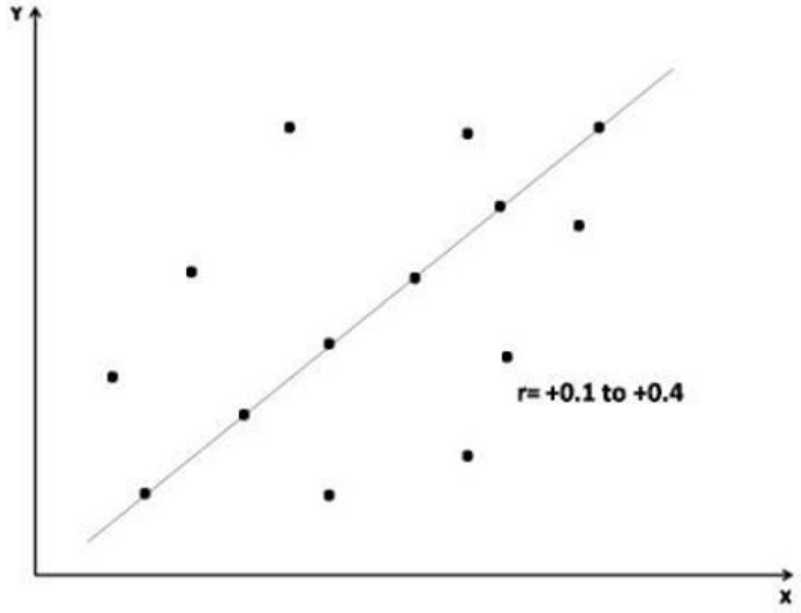
3. Strong Positive Correlation: A strong positive correlation occurs when two variables tend to increase together in a consistent pattern, but not as perfectly as in a perfect positive corre- lation. In this case, as one variable rises, the other also increases significantly, though small deviations from a perfect straight-line relationship may occur. For example, the relationship between hours of study and examination scores often shows a strong positive correlation, as higher study time generally leads to higher scores, even if the increase is not perfectly uniform.



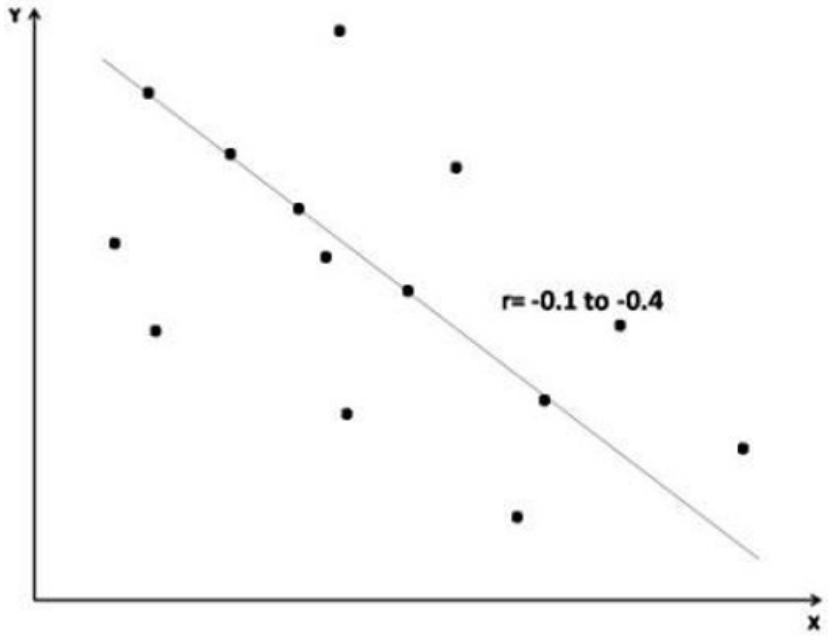
4. Strong Negative Correlation: A strong negative correlation occurs when two variables move in opposite directions in a consistent manner, though not perfectly. In this case, as one variable increases, the other tends to decrease significantly, but with slight variations from a per- fect straight-line relationship. For example, the relationship between the amount of time spent on leisure activities and academic performance for some students may show a strong negative correlation, where more leisure time is generally associated with lower academic scores, even though the relationship is not exact.



5. Weak Positive Correlation: A weak positive correlation exists when two variables show a slight upward trend, but the relationship is not strong or consistent. In this situation, as one vari- able increases, the other also tends to increase, though the points are scattered and the upward pattern is not very clear. For instance, the relationship between the number of hours a student studies and their exam score may show a weak positive correlation if many other factors, such as health or exam difficulty, also influence the outcome.

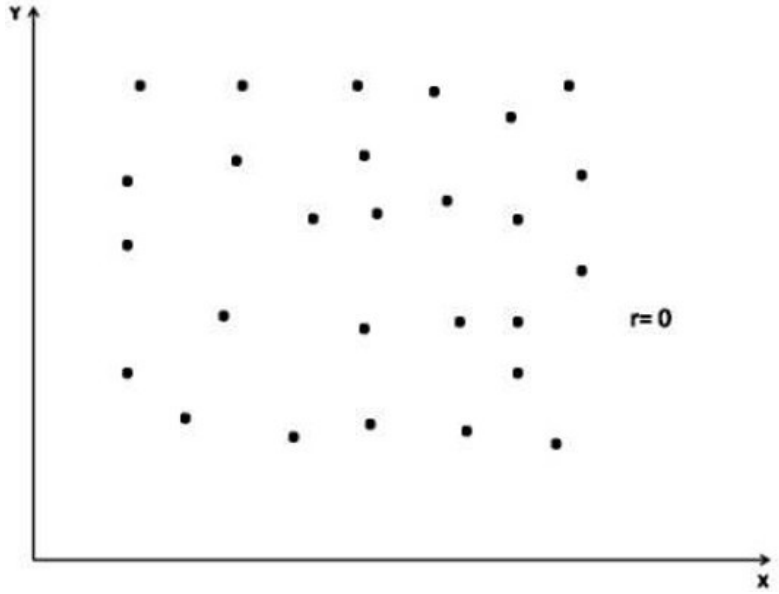


6. Weak Negative Correlation: A weak negative correlation occurs when two variables show a slight tendency to move in opposite directions, but the relationship is not very clear or consistent. In this case, as one variable increases, the other tends to decrease, though the points are widely scattered and do not form a strong downward pattern. For example, the relationship between hours spent watching television and physical fitness might show a weak negative cor- relation, since watching more TV could be linked to lower fitness levels, but many other factors play a role.



7. No Correlation: No correlation exists when there is no observable relationship between the two variables. In this case, a change in one variable does not show any consistent increase or decrease in the other, and the data points appear randomly scattered without forming any clear trend. For example, the number of shoes a person owns and their performance in mathematics

usually show no correlation, since the two variables are unrelated.



1.3.3 **Product Moment correction Coefficient**

The Product Moment Correlation, also known as Pearson's Correlation Coefficient, is one of the most widely used methods of measuring the strength and direction of the linear relation- ship between two continuous variables. It provides a single value, usually denoted as r, that indicates how strongly the variables are associated. The value of r always lies between -1 and $+1$, where $+1$ represents a perfect positive correlation, -1 a perfect negative correlation, and 0 indicates no linear relationship. This measure is particularly useful because it not only shows the strength of the association but also its direction.

The formula for calculating Pearson's Correlation Coefficient is given by:

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x^2 - (\sum x_i)^2 \quad n\sum y^2 - (\sum y_i)^2}}$$

Examp: Suppose we have the following data:

X	Y	x_iy_i	x_i^2	y_i^2
1	2	2	1	4
2	4	8	4	16
3	5	15	9	25
4	4	16	16	16
5	6	30	25	36
6	7	42	36	49
7	8	56	49	64
8	9	72	64	81
9	10	90	81	100
10	12	120	100	144
$\sum x_i = 55$	$\sum y_i = 67$	$\sum x_iy_i = 451$	$\sum x_i^2 = 385$	$\sum y_i^2 = 535$

Step 1: Numerator

$$n\sum x_iy_i - \sum x_i \sum y_i = 10(451) - (55)(67) = 4510 - 3685 = 825$$

Step 2: Denominator

$$\sqrt{n\sum x^2 - (\sum x_i)^2}\ \sqrt{n\sum y^2 - (\sum y_i)^2} = \sqrt{10(385) - 55^2}\ \sqrt{10(535) - 67^2}\ .$$

$$= \sqrt{(3850 - 3025)(5350 - 4489)} = \sqrt{825 \cdot 861} = \sqrt{710{,}325}$$

$$\approx 842.22$$

Step 3: Correlation Coefficient

$$r = \frac{n\sum x_iy_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x^2 - (\sum x_i)^2}\ \sqrt{n\sum y^2 - (\sum y_i)^2}}$$

$$r = \frac{825}{842.22} \approx 0.979$$

Interpretation: The computed correlation coefficient is $r \approx 0.98$, which shows a very strong positive linear relationship between X and Y : as X increases, Y increases almost proportionally.



1.4 Self-Assesement Exercises

1. The product moment correlation coefficient (r) measures:
    a) The average of X and Y
    b) The strength and direction of linear relationship between X and Y
    c) The difference between X and Y
    d) The ratio of variance of X and Y
2. The value of correlation coefficient r always lies between:
    a) 0 and 1
    b) -1 and 1
    c) $-\infty$ and $+\infty$
    d) 0 and ∞
3. If $r = +1$, it indicates:
    a) Weak positive correlation
    b) Perfect positive correlation
    c) Perfect negative correlation
    d) No correlation
4. If $r = -1$, it implies:
    a) No relationship
    b) Weak negative correlation
    c) Perfect negative correlation
    d) Strong positive correlation
5. Which of the following is **not** a type of correlation?
    a) Perfect positive correlation
    b) Strong negative correlation
    c) Zero correlation
    d) Infinite correlation

6. Correlation can be used to:

   a) Predict the value of one variable from another

   b) Measure the strength of association between two variables

   c) Establish cause and effect between two variables

   d) Both (a) and (b)

7. A correlation coefficient close to zero indicates:

   a) Strong positive relationship

   b) Strong negative relationship

   c) No linear relationship

   d) Perfect positive correlation

8. Which of the following statements is true?

   a) Correlation proves causation

   b) Correlation does not imply causation

   c) Correlation always equals regression slope

   d) Correlation is always positive

9. The denominator in the correlation formula involves:

   a) Product of means of X and Y

   b) Square root of product of sums of squares corrected for mean

   c) Only $\sum x_i y_i$

   d) Variance of X minus variance of Y

10. In practice, a correlation of $r \approx 0.98$ means:

    a) X and Y are unrelated

    b) X and Y have a very strong positive relationship

    c) X and Y have a very weak relationship

    d) The relationship between X and Y is negative

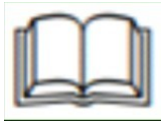


1.5 Summary

The correlation coefficient is one of the most widely used tools in statistics for describing the strength and direction of a relationship between two variables. It provides a single value, usually denoted as *r*, that ranges from −1 to +1. A value close to +1 indicates a strong pos- itive association, where increases in one variable are closely linked to increases in the other. Conversely, a value near −1 reflects a strong negative association, meaning that as one variable increases, the other tends to decrease. When the value is close to zero, it suggests little or no linear relationship between the two variables.

Pearson's correlation coefficient is the most common form and is particularly suited for contin- uous data that follows an approximately linear trend without extreme outliers. Its calculation involves standardizing the covariance between two variables by dividing it by the product of their standard deviations. This adjustment allows the measure to remain scale-free, making it possible to compare relationships across very different types of data. Spearman's rank correla- tion and Kendall's tau are alternative methods that are often applied when data do not meet the assumptions required for Pearson's coefficient, such as in cases of non-linear but monotonic relationships or ordinal data.

In practice, the correlation coefficient is extremely useful for exploring real-world questions. For example, when analyzing the link between hours of study and academic performance, a high positive correlation would suggest that more study time generally improves outcomes. A negative correlation might be observed between exercise and body weight, highlighting an inverse relationship. Ultimately, while the correlation coefficient does not imply causation, it remains a fundamental descriptive statistic that helps researchers, educators, and decision- makers quickly assess patterns and associations in data.



1.6 References, Further Readings, Web Sources

Navarro, D., & Foxcroft, D. (2025). Learning Statistics with jamovi: A Tutorial for Beginners in Statistical Analysis. Open Book Publishers.

Wisniewski, S. J., & Brannan, G. D. (2024). Correlation (Coefficient, Partial, and Spearman Rank) and Regression Analysis. In *StatPearls*. StatPearls Publishing.



1. Possible Answers to SAEs

1.*b*, 2.*b*, 3.*b*, 4.*c*, 5.*d*, 6.*d*, 7.*c*, 8.*b*, 9.*b*, 10.*b*

Unit 2 Spearman's Rank Correlation Coefficient

2.1 Introduction
2.2 Learning Outcomes
2.3 Spearman's Rank Correlation Coefficient
2.3.1 Uses of Spearman's Rank Correlation Coefficient
2.3.2 Advantages of Spearman's Rank Correlation Coefficient
Disadvantage of Spearman's Rank Correlation Coefficient
2.4 Self-Assessment Exercises 2.5 Summary
2.6 References, Further Readings, and Web Sources
2.7 Possible Answers to Self-Assessment Exercises



2.1 introduction

Spearman's Rank Correlation Coefficient is a statistical tool used to measure the strength and direction of association between two variables when the data are ranked rather than measured directly. Unlike Pearson's product-moment correlation, which works with actual numerical values and assumes a linear relationship, Spearman's method focuses on the relative positions (or ranks) of the data points. This makes it particularly useful when dealing with ordinal data, non-linear relationships, or when the assumptions of Pearson's correlation are not satisfied. Es- sentially, Spearman's correlation helps us understand how well one variable tends to increase or decrease as the other does, based purely on their ranked order.



2.2 Learning Outcomes

At the end of this unit, students should be able to:

- Explain the concept of Spearman's Rank Correlation Coefficient and identify when it is more suitable than Pearson's correlation.
- Compute Spearman's correlation coefficient using ranked data.
- Interpret the coefficient in terms of the strength and direction of association.



2.3 Spearman's Rank Correlation Coefficient

Spearman's Rank Correlation Coefficient, denoted as r_s, is a non-parametric measure used to assess the strength and direction of association between two variables. It relies on ranking the data values instead of using their actual magnitudes. This makes it especially useful in situa- tions where data may not be normally distributed or when the relationship between variables is

not linear.

To compute r_s, the first step is to assign ranks to each observation for both variables. The smallest value is given rank 1, the next rank 2, and so on. If two or more values are equal, each tied value receives the average of the ranks they would have occupied. After assigning ranks, the differences between the ranks of the two variables are calculated and denoted by d_i. These differences are then squared to get d_i^2, and the sum $\sum d_i^2$ is used in the formula:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where n is the number of paired observations.

2.3.1 Uses of Spearman's Rank Correlation

- It is commonly used in social sciences, psychology, and education to study relationships where data are ordinal (ranked) rather than continuous.
- It is useful for determining monotonic relationships, where variables move in the same or opposite direction but not necessarily at a constant rate.
- It is applied when data violate the assumptions of Pearson's correlation, such as normality or linearity.

2.3.2 Advantages of Spearman's Rank Correlation

- Simple to compute and interpret.
- Works well with both ordinal and continuous data.
- Less sensitive to extreme values or outliers compared to Pearson's correlation.
- Suitable for small sample sizes.

2.3.3 Disadvantages of Spearman's Rank Correlation

- Less precise than Pearson's correlation when the actual relationship is linear.
- Cannot capture relationships that are not monotonic.
- Handling tied ranks can sometimes reduce accuracy.

Example: Suppose we have the following data on students' scores in Mathematics (X) and English (Y):

Student	X (Math)	Y (English)	Rank X	Rank Y	d_i^2
1	85	78	5	6	1
2	92	95	2	2	0
3	76	84	8	5	9
4	88	86	4	4	0
5	80	70	7	8	1
6	95	99	1	1	0
7	70	65	10	10	0
8	82	74	6	7	1
9	90	88	3	3	0
10	78	72	9	9	0
				$\sum d^2$	12

Using the formula:

$$r_s = 1 - \frac{6(12)}{10(10^2 - 1)} = 1 - \frac{72}{990} = 1 - 0.0727 \approx 0.93$$

Thus, $r_s \approx 0.93$, which indicates a very strong positive correlation between students' performance in Mathematics and English.

2.4 Self-Assesement Exercises

1. Spearman's Rank Correlation is mainly used for:
   a) Ratio data
   b) Ordinal data
   c) Interval data
   d) Nominal data
2. The formula for Spearman's Rank Correlation is: a) $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$
   b) $r_s = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$
   c) $r_s = \sqrt{\frac{N\sum XY - \sum X \sum Y}{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}}$
   d) $r_s = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$

3. In Spearman's Rank Correlation, d represents:
   a) Difference between means
   b) Difference between variances
   c) Difference between ranks of two variables
   d) Difference between correlation values

4. The value of Spearman's Rank Correlation lies between:
   a) -2 and $+2$
   b) -1 and $+1$
   c) 0 and $+1$
   d) None of the above

5. A positive value of Spearman's Rank Correlation suggests:
   a) As one variable increases, the other also tends to increase
   b) As one variable increases, the other tends to decrease
   c) The two variables are unrelated
   d) The data is normally distributed

6. A negative value of Spearman's Rank Correlation suggests: a) Both variables move in the same direction
   b) Both variables are independent
   c) As one variable increases, the other tends to decrease
   d) There is no ranking possible

7. One advantage of Spearman's Rank Correlation over Pearson's is:
   a) It requires interval data
   b) It can handle tied ranks and ordinal data
   c) It always gives higher correlation
   d) It is more accurate in all cases

8. One disadvantage of Spearman's Rank Correlation is:
   a) It is less effective with small data sets
   b) It cannot capture non-monotonic relationships
   c) It does not allow ranking of qualitative data
   d) It always gives negative correlation

9. Which of the following is a common use of Spearman's Rank Correlation?
   a) Predicting future sales trends
   b) Measuring agreement between judges or examiners
   c) Estimating population mean
   d) Testing for normal distribution

10. An advantage of Spearman's Rank Correlation is:
    a) It is only applicable for very large data
    b) It can be applied even when the data is not normally distributed
    c) It is strictly limited to continuous data
    d) It always gives the same value as Pearson's correlation



2.5 Summary

Spearman's Rank Correlation Coefficient is a statistical measure used to determine the strength and direction of a relationship between two ranked variables. Unlike Pearson's correlation, which works with actual numerical values, Spearman's method relies on the relative positions or ranks of the data, making it especially useful when dealing with ordinal data or situations where the assumptions of normality and linearity are not satisfied.

The coefficient, denoted by r_s, ranges between -1 and $+1$. A value close to $+1$ indicates a strong positive association, meaning that as the rank of one variable increases, the rank of the other variable also increases. Conversely, a value close to -1 shows a strong negative associ- ation, where higher ranks in one variable correspond to lower ranks in the other. A value near zero suggests little or no relationship between the two sets of ranks.

Spearman's correlation is particularly useful in social sciences, psychology, education, and other fields where data is often ordinal in nature, or when precise measurement is difficult. While it is simple to compute and interpret, it mainly captures monotonic relationships, mean- ing it may not fully describe more complex patterns.



1.6 References, Further Readings, Web Sources

Navarro, D., & Foxcroft, D. (2025). Learning Statistics with jamovi: A Tutorial for Beginners in Statistical Analysis. Open Book Publishers.

Wisniewski, S. J., & Brannan, G. D. (2024). Correlation (Coefficient, Partial, and Spearman Rank) and Regression Analysis. In *StatPearls*. StatPearls Publishing.



1.7 Possible Answers to SAEs

1.b 2.a 3.c 4.b 5.a 6.c 7.b 8.b 9.b 10.b

Unit 3 Simple Linear Regression

3.1 Introduction
3.2 Learning Outcomes
3.3 Simple Linear Regression
3.3.1 Uses of Simple Linear Regression
3.3.2 Advantages of Simple Linear Regression
3.3.3 Disadvantages of Simple Linear Regression
3.4Self-Assessment Exercises
3.5 Summary 3.6 References, Further Readings, and Web Sources
3.7 Possible Answers to Self-Assessment Exercises



3.1 introduction

Simple linear regression is a basic statistical technique used to study the relationship between two variables. One variable is considered the predictor (independent variable) and the other is the outcome (dependent variable). The method helps us understand whether, and to what extent, changes in the predictor are associated with changes in the outcome. For example, it can be used to examine how the number of hours studied influences a student's exam score, or how advertising expenditure relates to sales revenue.

The main purpose of simple linear regression is not only to measure the strength of the relationship but also to make predictions about the outcome based on the predictor. It provides a clear and straightforward way of summarizing the relationship between two variables using a straight line, making it one of the most widely applied and easy-to-understand tools in statis- tics, economics, and the social sciences.



3.2 Learning Outcomes

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

- Explain the concept of simple linear regression and identify the dependent and independent variables.
- Apply the regression equation to estimate the relationship between two variables.
- Interpret the slope and intercept of a regression.



3.3 Simple Linear regression

Regression analysis is a statistical technique used to study the relationship between a depen- dent variable (also called the response variable) and one or more independent variables (also called predictors or explanatory variables). It helps us understand how the dependent variable changes when the independent variable(s) change. By providing a mathematical model, regres- sion allows researchers to explain patterns, make predictions, and support decision-making in fields such as economics, education, health, engineering, and the social sciences.

The basic form of regression is *simple linear regression*, which examines the relationship be- tween two variables. Its equation is written as:

$$Y = a + bX + \varepsilon$$

where Y is the dependent variable, X is the independent variable, a is the intercept (the value of Y when $X = 0$), b is the slope (which shows how much Y changes when X increases by one unit), and ε is the error term (representing other factors not included in the model). This equation shows that regression is essentially about fitting the "best line" through the data to explain the relationship.

In practice, regression coefficients (a and b) are estimated from data using the method of least squares, which minimizes the sum of squared errors between the observed values and the pre- dicted values from the model. The slope coefficient b tells us the direction and strength of the relationship, while the intercept a anchors the line on the vertical axis. Alongside this, the coefficient of determination (R^2) indicates how well the model explains the variability in the dependent variable.

While regression is powerful, it relies on certain assumptions: the relationship between vari- ables should be linear, errors should be independent and normally distributed, and the variance of errors should remain constant. Violating these assumptions can weaken the validity of the model. Nevertheless, regression analysis remains one of the most widely used methods in statis- tics because it provides both explanation and prediction, helping researchers translate data into meaningful insights.

3.3.1 Uses of Simple Linear Regression

1. Regression analysis is used for prediction and forecasting, for example estimating future sales, demand, or exam scores.
2. It helps identify relationships between variables, such as the effect of education on in- come.
3. It is applied to quantify the effect of one or more factors on an outcome, like measuring the impact of advertising on sales.

4. It supports policy and business decision-making, where models guide strategies in areas such as unemployment or inflation control.

5. It reduces error and controls for confounding variables, ensuring more reliable conclusions about the true effect of predictors.

3.3.1 Advantages of Simple Linear Regression

1. Regression analysis helps in predicting future outcomes based on past data, making it useful for planning and forecasting.

2. It quantifies the strength and direction of the relationship between variables, giving a clearer understanding of their association.

3. The method allows researchers to identify the most important factors influencing an outcome, which aids in better decision-making.

4. Regression can control for multiple variables at once, helping to reduce bias and isolate the effect of key predictors.

5. It provides measurable statistics, such as coefficients and R^2, which make interpretation and comparison more objective.

3.3.1 Disadvantages of Simple Linear Regression

1. Regression assumes a linear relationship between variables, which may not always hold true in real-world data.

2. The results can be highly sensitive to outliers, which may distort the regression line and lead to misleading conclusions.

3. It requires several assumptions (such as normality, independence, and constant variance of errors), and violating these can weaken the validity of the model.

4. Regression analysis shows correlation but does not always prove causation, so interpretations must be made carefully.

5. When many independent variables are included, multicollinearity can occur, making it difficult to determine the individual effect of predictors.

Model and Formular for Simple Linear Regression

Model: $Y = a + bX$

Slope: $b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$

Intercept: $a = \bar{Y} - b\,\bar{X}$

Example: Consider the Data: $X = \{1, 2, 3, 4, 5\}$, $Y = \{2, 4, 5, 4, 5\}$

$$\bar{X} = \frac{1+2+3+4+5}{5} = 3, \qquad \bar{Y} = \frac{2+4+5+4+5}{5} = 4$$

X_i	Y_i	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	2	−2	−2	4	4
2	4	−1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
			Sums:	$\sum(X_i - \bar{X})^2 = 10$	$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 6$

$$b = \frac{6}{10} = 0.6, \qquad a = \bar{Y} - b\,\bar{X} = 4 - 0.6 \times 3 = 2.2$$

$$\hat{Y} = a + bX = 2.2 + 0.6X$$

Interpretation: On average, each 1-unit increase in X is associated with a 0.6 increase in Y ; when $X = 0$, the model predicts $Y \approx 2.2$.



3.4 Self-Assesement Exercises

1. Regression analysis is mainly used to study the relationship between:
   a) Two dependent variables
   b) A dependent and independent variable(s)
   c) Two independent variables
   d) None of the above

2. The equation of a simple linear regression model is:
   a) $Y = aX + b + \varepsilon$
   b) $Y = a + bX + \varepsilon$
   c) $Y = X + a + b$
   d) $Y = abX$

3. In the regression equation $Y = a + bX + \varepsilon$, the slope coefficient b:
   a) Shows the value of Y when $X = 0$
   b) Measures how much Y changes when X increases by one unit
   c) Represents the error term
   d) Is always equal to zero

4. Which method is commonly used to estimate regression coefficients?
   a) Maximum likelihood
   b) Method of moments
   c) Least squares method
   d) Chi-square test

5. One important use of regression analysis is:
   a) Measuring the mean only
   b) Predicting future outcomes
   c) Avoiding data collection
   d) Eliminating sampling errors

6. Which of the following is **not** an advantage of regression analysis?
   a) Helps in forecasting
   b) Identifies important predictors
   c) Proves causation directly
   d) Provides objective measures like coefficients and R^2

7. A major limitation of regression analysis is that:
   a) It can only handle large datasets
   b) It assumes linear relationships between variables
   c) It cannot compute coefficients
   d) It requires no assumptions

8. The coefficient of determination, R^2, is used to:
   a) Test independence of variables
   b) Show how much variation in Y is explained by X
   c) Measure the intercept in regression
   d) Detect multicollinearity

9. Multicollinearity in regression occurs when:
   a) There is no relationship between variables
   b) Independent variables are highly correlated with each other
   c) Dependent variables are constant
   d) Errors are normally distributed

10. Which of the following best explains why regression does not always prove causation?
    a) Because regression only measures relationships, not underlying causes
    b) Because regression requires large samples
    c) Because regression ignores independent variables
    d) Because regression always has an R^2 less than 1



3.5 Summary

egression analysis is a statistical method used to study the relationship between a dependent variable and one or more independent variables. In its simplest form, *simple linear regression* expresses this relationship with the equation:

$$Y = a + bX + \varepsilon$$

where the slope and intercept describe how changes in the independent variable affect the dependent variable. Regression is widely applied in prediction and forecasting, helping researchers and decision-makers estimate future outcomes. It also provides a way to quantify the strength of relationships and identify the most important factors influencing results, making it a valuable tool in fields such as economics, business, education, and the social sciences.



3.6 References, Further Readings, Web Sources

Navarro, D., & Foxcroft, D. (2025). Learning Statistics with jamovi: A Tutorial for Beginners in Statistical Analysis. Open Book Publishers.

Wisniewski, S. J., & Brannan, G. D. (2024). Correlation (Coefficient, Partial, and Spearman Rank) and Regression Analysis. In *StatPearls*. StatPearls Publishing.



3.7 Possible Answers to SAEs

1.b 2.b 3.b 4.c 5.b 6.c 7.b 8.b 9.b 10.a

Unit 4 Coefficient of Determination (R^2)

4.1 Introduction

4.2 Learning Outcomes

4.3 Coefficient of Determination (R^2)

- 4.3.1 Uses of Coefficient of Determination
- 4.3.2 Advantages of Coefficient of Determination
- 4.3.3 Disadvantages of Coefficient of Determination

4.4 Self-Assessment Exercises 4.5 Summary

4.6 References, Further Readings, and Web Sources

4.7 Possible Answers to Self-Assessment Exercises



4.1 introduction

The coefficient of determination, denoted as R^2, is a statistical measure that indicates how well a regression model explains the variation in the dependent variable. It takes values between 0 and 1, where 0 means the model does not explain any variation, and 1 means the model ex- plains all of the variation perfectly. Mathematically, R^2 is defined as the ratio of the explained variation to the total variation in the dependent variable. In simple terms, it represents the pro- portion of changes in the dependent variable (Y) that can be accounted for by the independent variable(s) (X) in the model.



4.2 Learning Outcomes

At the end of this unit, students should be able to:

- Define the coefficient of determination (R^2) and explain its role in regression analysis.
- Interpret the meaning of different values of R^2 in terms of model fit.
- Evaluate the usefulness of R^2 as a measure of how well independent variables explain the variation in the dependent variable.



4.3 Coefficient of Determination (R^2)

The coefficient of determination, commonly denoted as R^2, is a key statistic in regression anal- ysis that tells us how well the independent variable(s) explain the variation in the dependent variable. It is essentially a measure of the "goodness of fit" of the regression model. By pro- viding a single value between 0 and 1, R^2 gives an indication of how much of the total variation in the outcome can be accounted for by the predictors included in the model.

Mathematically, R^2 is defined as the ratio of the explained sum of squares to the total sum of squares. In simpler terms, it compares the variation explained by the regression line with the total variation present in the data. A higher R^2 value means that the regression model is doing a better job at explaining the relationship between variables, while a lower R^2 suggests that much of the variation remains unexplained.

In practice, R^2 helps researchers judge the effectiveness of their model. For instance, an R^2 value close to 1 suggests that most of the variation in the dependent variable can be explained by the independent variable(s), indicating a strong fit. On the other hand, an R^2 value closer to 0 indicates that the independent variables do not provide much explanatory power. This makes

R^2 a valuable summary statistic for model evaluation.

Although R^2 is useful, it should always be interpreted carefully in the context of the research problem. A high R^2 does not automatically mean that the model is appropriate, nor does a low R^2 necessarily mean the model is useless. Factors such as the nature of the data, the number of predictors, and the research objectives all influence how R^2 should be understood. As a result, R^2 is best viewed as one of several tools for assessing model performance rather than as a final judgment on its own.

4.3.1 Uses of Coefficient of Determination

1. R^2 is used to measure the goodness of fit of a regression model.
2. It helps determine how much of the variation in the dependent variable can be explained by the independent variable(s).
3. It allows comparison between different regression models to assess which model provides a better explanation of the data.
4. R^2 assists researchers and decision-makers in judging the predictive power of a model.
5. It provides a simple and intuitive summary statistic for evaluating regression results.

4.3.2 Advantages of Coefficient of Determination

1. It is easy to understand and interpret as a percentage of explained variation.
2. It provides a standardized measure (between 0 and 1) for assessing model performance.
3. R^2 enables quick comparisons between different models.
4. It offers a quantitative way to evaluate how well independent variables collectively explain the outcome.
5. It is widely recognized and accepted in statistical analysis and research.

4.3.3 Disadvantages of Coefficient of Determination

1. A high R^2 does not necessarily mean the model is appropriate or free from bias.
2. It cannot prove causation, only association between variables.
3. Adding more variables to a model can artificially increase R^2, even if those variables are irrelevant.
4. It does not account for overfitting or the quality of individual predictors.

5. In some fields, meaningful models may have low R^2, making interpretation context dependent.



4.4 Self-Assesement Exercises

1. The coefficient of determination (R^2) is mainly used to:
   a) Measure the strength of correlation
   b) Measure the goodness of fit of a regression model
   c) Calculate regression coefficients
   d) Test for independence

2. Which of the following values of R^2 indicates a perfect fit
   a) 0
   b) 0.25
   c) 0.50
   d) 1

3. One use of R^2 is:
   a) To compute the sample mean
   b) To determine how much variation in Y is explained by X
   c) To replace regression coefficients
   d) To remove outliers

4. An advantage of R^2 is that:
   a) It always proves causation
   b) It is easy to interpret as a percentage of explained variation
   c) It prevents multicollinearity
   d) It eliminates measurement error

5. Which of the following is **not** an advantage of R^2
   a) Provides a standardized measure between 0 and 1
   b) Enables comparison of different models
   c) Guarantees the model is appropriate
   d) Widely recognized in research

6. A disadvantage of R^2 is that:
   a) It cannot exceed 1
   b) It increases automatically when more variables are added
   c) It requires complex interpretation
   d) It is not accepted in statistics

7. A high value of R^2:
   a) Always indicates the best model
   b) Means independent variables explain most of the variation in the dependent variable
   c) Proves there is no error term
   d) Implies causation between variables

8. The coefficient of determination is expressed as:
   a) A ratio of explained to total variation
   b) A ratio of residuals to sample size
   c) A difference between slope and intercept
   d) The square root of correlation coefficient

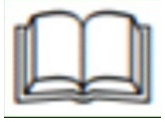
9. One limitation of R^2 is that:
   a) It decreases when more predictors are added
   b) It does not account for overfitting
   c) It cannot be used for regression analysis
   d) It always equals zero in practice

10. In regression analysis, R^2 values close to zero suggest:
    a) The model explains most of the variation
    b) The independent variables have little explanatory power
    c) The regression coefficients are always significant
    d) The model has no intercept



4.5 Summary

The coefficient of determination (R^2) is an important measure in regression analysis that explains how well the independent variable(s) account for the variation in the dependent variable. It takes values between 0 and 1, with higher values indicating a better fit of the model.(R^2) is widely used for assessing the goodness of fit, comparing models, and evaluating predictive power. Its main advantage is simplicity and clear interpretation, though it should always be considered alongside other statistics for a complete understanding of model performance.



4.6 References, Further Readings, Web Sources

Navarro, D., & Foxcroft, D. (2025). Learning Statistics with jamovi: A Tutorial for Beginners in Statistical Analysis. Open Book Publishers.

Wisniewski, S. J., & Brannan, G. D. (2024). Correlation (Coefficient, Partial, and Spearman Rank) and Regression Analysis. In *StatPearls*. StatPearls Publishing.



4.7 Possible Answers to SAEs

1.b 2.d 3.b 4.b 5.c 6.b 7.b 8.a 9.b 10.b

Unit 5 Assumptions and Limitation of Correaltion and Regression Analysis

5.1 Introduction
5.2 Learning Outcomes
5.3 Assumptions and Limitation of Correaltion and Regression Analysis
5.3.1 Assumptions of Correlation
5.3.2 Limitations of Correlation
5.3.3 Assumptions of Regression
5.3.4 Limitations of Regression
5.4 Self-Assessment Exercises
5.5 Summary
5.6 References, Further Readings, and Web Sources
5.7 Possible Answers to Self-Assessment Exercises



1.1 introduction

Correlation and regression analysis are powerful statistical tools used to study relationships be- tween variables and make predictions. However, like all statistical methods, they rely on certain assumptions to produce valid and reliable results. These assumptions include conditions such as linearity, independence, normality, and homoscedasticity, which ensure that the techniques work as intended. At the same time, both correlation and regression come with limitations. For instance, correlation does not imply causation, and regression models may give misleading results if key assumptions are violated or important variables are omitted. Understanding these assumptions and limitations is therefore essential for proper application, accurate interpreta- tion, and avoidance of common errors in data analysis.



4.2 Learning Outcomes

At the end of this unit, students should be able to:

- Describe the assumptions of correlation and regression, such as linearity, independence, normality, and homoscedasticity.
- Recognize the limitations of correlation and regression, including issues like correlation not implying causation, the effect of outliers, and model misspecification.
- Critically assess the validity of correlation and regression results by checking whether assumptions are met and understanding the potential impact of limitations.

5.1 Assumptions and Limitation of Correaltion and Regression Analysis

5.3.1 Assumptions of Correlation

1. The relationship between the two variables is linear.
2. Both variables are measured on interval or ratio scales.
3. Variables are continuous, not categorical.
4. The data are free from significant outliers.
5. The relationship is homoscedastic (constant variance across values).
6. The data come from a random and representative sample.

5.3.2 Assumptions of Correlation

1. Correlation does not imply causation.
2. It only measures linear relationships, not nonlinear ones.
3. Correlation values can be distorted by outliers.
4. A strong correlation may occur due to a third (lurking) variable.
5. Correlation may vary depending on sample size.
6. It cannot capture complex or dynamic relationships between variables.

5.3.3 Assumptions of Regression

1. Linearity: the relationship between independent and dependent variables is linear.
2. Independence: errors (residuals) are independent of each other.
3. Homoscedasticity: constant variance of residuals across all values of the predictor(s).
4. Normality: residuals are normally distributed.
5. No multicollinearity: independent variables are not highly correlated with each other.
6. The model is correctly specified (includes all relevant variables and excludes irrelevant ones).

5.3.4 **Limitations of Regression**

1. Regression does not prove causation, only association.
2. It is sensitive to outliers, which can distort results.
3. Violations of assumptions (linearity, normality, homoscedasticity) lead to unreliable results.
4. Multicollinearity makes it hard to interpret the effect of predictors.
5. Overfitting may occur if too many predictors are included relative to sample size.
6. Omitted variables or poor model specification can bias the estimates.



5.4 Self-Assesement Exercises

1. Which of the following is an assumption of correlation analysis?
   a) Variables are categorical
   b) Relationship between variables is linear
   c) Errors are normally distributed
   d) Variables must always be independent
2. A major assumption of regression analysis is:
   a) Independence of residuals
   b) Causation between variables
   c) Perfect correlation between predictors
   d) Presence of outliers
3. Which of the following is a limitation of correlation?
   a) It explains nonlinear relationships
   b) It cannot imply causation
   c) It ensures regression coefficients are unbiased
   d) It removes the effect of outliers
4. Which assumption is required for regression residuals?
   a) They must be normally distributed
   b) They must all equal zero
   c) They must be highly correlated
   d) They must increase with sample size

5. One limitation of regression analysis is:
   a) It prevents overfitting
   b) It is unaffected by irrelevant variables
   c) It may be biased if the model is misspecified
   d) It always proves causation

6. Correlation is only appropriate when:
   a) The relationship is linear
   b) The dependent variable is categorical
   c) Residuals are heteroscedastic
   d) Predictors are highly collinear

7. Multicollinearity in regression means:
   a) Predictors are not correlated
   b) Predictors are highly correlated with each other
   c) Residuals are normally distributed
   d) Dependent variable is constant

8. A limitation of correlation coefficients is that:
   a) They are unaffected by sample size
   b) They only measure the strength of linear relationships
   c) They guarantee a causal link between variables
   d) They eliminate lurking variables

9. Which assumption ensures that variance of residuals is constant across predictors?
   a) Homoscedasticity
   b) Linearity
   c) Multicollinearity
   d) Independence

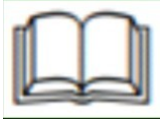
10. When R^2 is high but assumptions are violated, the model:
    a) Is still reliable
    b) May give misleading results
    c) Proves causation
    d) Removes the effect of outliers



5.5 Summary

Correlation and regression are important tools for examining relationships between variables, but they are only reliable when their assumptions are met. Correlation assumes a linear rela- tionship, continuous data, and the absence of extreme outliers, while regression further requires

conditions like independence of residuals, homoscedasticity, and normality. Despite their use- fulness, both methods have limitations: correlation cannot prove causation and only measures linear associations, while regression may give misleading results if assumptions are violated, variables are misspecified, or multicollinearity and overfitting occur. Therefore, careful atten- tion to assumptions and recognition of limitations are essential for accurate interpretation of results.



5.6 References, Further Readings, Web Sources

Navarro, D., & Foxcroft, D. (2025). Learning Statistics with jamovi: A Tutorial for Beginners in Statistical Analysis. Open Book Publishers.

Wisniewski, S. J., & Brannan, G. D. (2024). Correlation (Coefficient, Partial, and Spearman Rank) and Regression Analysis. In *StatPearls*. StatPearls Publishing.



5.7 Possible Answers to SAEs

1.b 2.a 3.b 4.a 5.c 6.a 7.b 8.b 9.a 10.b