STATISTICAL METHODS FOR THE SOCIAL SCIENCES

C. A. HESSE, BSc, MPhil, PhD,
Senior Lecturer of Statistics
Methodist University College Ghana

J. B. OFOSU, BSc, PhD, FSS
Professor of Statistics
Methodist University College Ghana

Published and Printed by

AKRONG PUBLICATIONS LIMITED
P. O. BOX M. 31
ACCRA, GHANA

(0244 648 757, 0264 648 757)

ISBN: 978–9988–2–6060–6

Published, 2017

akrongh@yahoo.com.

PREFACE

The purpose of this book is to acquaint the reader with the increasing number of applications of statistics in engineering and the social sciences. It can be used as a textbook for a first course in statistical methods in Universities and Polytechnics. The book can also be used by decision makers and researchers to either gain basic understanding or to extend their knowledge of some of the most commonly used statistical methods.

Our goal is to introduce the basic theory without getting too involved in mathematical detail, and thus to enable a larger proportion of the book to be devoted to practical applications. Because of this, some results are stated without proof, where this is unlikely to affect the reader's comprehension. However, we have tried to avoid the cook-book approach to statistics by carefully explaining the basic concepts of the subject, such as probability and sampling distributions; these the reader must understand.

The worst abuses of statistics occur when scientists try to analyze their data by substituting measurements into statistical formulae which they do not understand.

The book contains ten Chapters. Chapter 1 deals with overview of statistics. In Chapter 2, we discuss how to describe data, using graphical and summary statistics. Chapter 3 covers probability while Chapters 4 and 5 cover probability distributions. Chapters 6, 7, 8 and 9 present basic tools of statistical inference; point estimation, interval estimation, hypothesis testing and analysis of variance. Chapter 10 presents linear regression and correlation. Our presentation is distinctly applications-oriented.

A prominent feature of the book is the inclusion of many examples. Each example is carefully selected to illustrate the application of a particular statistical technique and or interpretation of results. Another feature is that each chapter has an extensive collection of exercises. Many of these exercises are from published sources, including past examination questions from King Saud University (Saudi Arabia) and Methodist University College Ghana. Answers to all the exercises are given at the end of the book.

We are grateful to Ms. Patience Workpor, secretary to the Dean of Informatics and Mathematical Sciences for her editorial assistance. We are also grateful to Professor Abdullah Al-Shiha of King Saud University (Saudi Arabia) for reading a draft of the book and offering helpful comments and for his permission to publish the statistical tables he used the Minitab software package to prepare. These tables are given in the Appendix. Last, but not least, we thank King Saud University and Methodist University College Ghana, for permission to use their past examination questions in Statistics.

We have gone to great lengths to make this text both pedagogically sound and error free. If you have any suggestions, or find potential errors, please contact us at **jonofosu@hotmail.com or akrongh@yahoo.com**.

J. B. Ofosu

C. A. Hesse

July, 2017

CONTENTS

Chapter One

Overview of Statistics

Chapter Contents

1.1 Population and sample

Consider the following example.

Example 1.1

Suppose we wish to study the body masses of all students of Methodist University. It will take us a long time to measure the body masses of all students of the university and so we may select 20 of the students and measure their body masses. Suppose we obtain the measurements in Table 1.1.

Table 1.1: **Body masses (in kg) of 20 students**

49	56	48	61	59	43	58	52	64	71
57	52	63	58	51	47	57	46	53	59

In this study, we are interested in the body masses of all students of Methodist University. The set of body masses of all students of Methodist University is called the ***population*** of this study. The set of body masses in Table 1.1, $W = \{49, 56, 48, \ldots, 53, 59\}$, is a ***sample*** from this population.

Definition 1.1

A population is the set of all objects we wish to study, for example, all divorced women, or all Methodists. The key word is ***all***.

Definition 1.2

A sample is part of the population we study to learn about the population.

Example 1.2

In a certain study, 900 men were selected from Nsawam. It was found that 25 are smokers.

(a) What is the population in this study? (b) What is the sample size?

Solution

(a) The population is men from Nsawam. (b) The sample size is 900.

1.2 What is statistics?

Statistics is a field of study concerned with:

(a) the collection, organization, and analysis of data, and

(b) the drawing of inferences about a population.

It can be seen that statistics can be classified into two main branches – ***descriptive statistics*** and ***inferential statistics***.

Definition 1.3 (Descriptive statistics)

Descriptive statistics consists of methods dealing with the collection, tabulation, summarization, and presentation of data.

These methods describe the various aspects of a data set. Descriptive statistical methods have their beginning in the inventories kept by early civilizations, such as the Babylonians, Egyptians, and the Chinese. For example, the Old Testament of the Bible refers to the numbering or counting of the people of Israel and to the casting of lots for selection by chance, and the Romans kept careful counts of people, possessions, and wealth in the territories they conquered. Similarly, the Domesday Book of the late eleventh century enumerated the lands and wealth of England. The Middle Ages also saw the growth of governments and religious institutions and their recording of births, deaths, and marriages. These early methods were primarily lists and counts kept for purposes of taxation and military conscription.

Although statistical methods developed throughout history in many different cultures, modern statistical concepts are considered to have developed in Europe in the last seventeenth century with the growth of mathematics and probability theory. Theories of probability have their historical roots in games of chance. By the seventeenth century, interest in gambling and the development of mathematical methods, combined and resulted in early rules of probability. The development of theories of probability led to the inception of ***inferential statistics***, in the beginning of the twentieth century. Statisticians such as Pearson, Fisher, Neyman, Wald, and Tukey, pioneered in the development of the methods of inferential statistics, which are widely applied in so many fields today.

Definition 1.4 (Inferential statistics)

Inferential statistics consist of methods that permit one to reach conclusions and make estimates about populations based upon information from a sample.

Census, parameter and statistic

If every member of a population is evaluated, a ***census*** has been performed, and any summary value of all the individual measurements is called a ***parameter***. If only a subset of a population has been evaluated, any summary value of such measurement is called a ***statistic***. Inferential statistics, therefore, involves using sample statistics to estimate population parameters.

Definition 1.5 (Census)

A ***census*** is an enumeration or evaluation of every member of a population.

Definition 1.6 (Parameter)

A ***parameter*** is any measurement that describes an entire population. Usually, the parameter value is unknown since we rarely can observe the entire population. Parameters are often (but not always) denoted by Greek letters, such as θ, μ and σ.

Definition 1.7 (Statistic)

A ***statistic*** is any measurement computed from a sample of the individual observations made.

It can be seen that a parameter is a numerical summary of a population and a statistic is a numerical summary of a sample data.

One problem with using samples is that, a sample provides only a limited information about the population from which the sample was taken. Although samples are generally representative of their populations, a sample is not expected to give an accurate picture of the whole population. There is usually some discrepancy between a sample statistic and the corresponding population parameter. This discrepancy is called ***sampling error***.

1.3 Why study statistics?

Knowing statistics will make you a better consumer of other people's data. Even if you don't plan to be a professional statistician, you should know enough to handle everyday data problems, to feel confident that others cannot deceive you with spurious arguments, and to know when you've reached the limits of your expertise. Statistical knowledge gives your company a competitive advantage against organizations that cannot understand their internal or external market data. And mastery of basic statistics gives you, the individual manager, a competitive advantage as you work your way through the promotion process, or when you move to a new employer. Here are some reasons why we study statistics.

Communication

The language of statistics is widely used in science, education, health care, engineering, and even the humanities. In all areas of business (accounting, finance, human resources, marketing, information systems, operations management), workers use statistical jargon to facilitate communication. In fact, statistical terminology has reached the highest corporate strategic levels. And in multinational environment, the specialized vocabulary of statistics permeates language barriers to improve problem-solving across national boundaries.

Computer Skills

Whatever your computer skill level, it can be improved. Every time you create a spreadsheet for data analysis, write a report, or make an oral presentation, you bring together skills you already have, and learn new ones. Specialists, with advanced training, design the databases and decision support systems, but you must expert to handle daily data problems without experts. Besides, you can't always find an "expert" and, if you do, the "expert" may not understand your application very well. You need to be able to analyze data, use software with confidence, prepare your own charts, write your own reports, and make electronic presentations on technical topics.

Information Management

Statistics can help you handle either too little or too much information. When insufficient data are available, statistical surveys and samples can be used to obtain the necessary market information. But most large organizations are closer to drowning in data than starving for it. Statistics can help you to summarize large amounts of data and reveal underlying relationships. Have you heard of data mining? Statistics is the pick and shovel that you take to the data mine.

Technical Literacy

Many of the best career opportunities are in growth industries propelled by advanced technology. Marketing staff may work with engineers, scientists, and manufacturing experts as new products and services are developed. Sales representatives must understand and explain technical products like pharmaceutical, medical equipment and industrial tools to potential customers. Purchasing managers must evaluate suppliers' claims about the quality of raw material, components, software, or parts.

Career Advancement

Whenever there are customers to whom services are delivered, statistical literacy can enhance your career mobility. Multi-billion-dollar companies like Barclays Bank, Citibank, Microsoft, and Wal-Mark, use statistics to control cost, achieve efficiency, and improve quality. Without a solid understanding of data and statistical measures, you may be left behind.

Quality improvement

Large manufacturing firms like Coca Cola and General Motors, have formal systems for continuous quality improvement. The same is true of insurance companies and financial service firms like Vanguard, Fidelity, and Barclays Bank. Statistics helps firms oversee their supplies, monitor their internal operations. Quality improvement goes far beyond statistics, but every university college graduate is expected to know enough statistics to understand its role in quality improvement.

Medicine

An experimental drug to treat asthma is given to 75 patients, of whom 24 get better. A placebo is given to a control group of 75 volunteers, of whom 12 get better. Is the new drug better than the placebo, or is the difference within the realm of chance?

Forecasting

A large company carries 50 000 different products. To manage this vast inventory, it needs a weekly order forecasting system that can respond to developing patterns in consumer demand. Is there a way to predict weekly demand and place order from suppliers for every item, without an unreasonable commitment of staff time?

Product warranty

A major automaker wants to know the average dollar cost of engine warranty claim on a new hybrid engine. It has collected warranty cost data on 4 300 warranty claims during the first 6 months after the engines are introduced. Using these warranty claims as an estimate of future costs, what is the margin of error associated with this estimate?

1.4 Opportunities for statisticians

In almost every endeavour of human activity, the scientific method has proven effective for solving problems and improving performance. This approach involves the collection of data pertinent to the particular problem. Statisticians play several important roles in these scientific studies. First, they plan the studies to ensure that the data are collected efficiently and answer the questions relevant to the investigation. Second, they analyze the data to discover what the study has demonstrated and what issues need further investigation.

In industry, statisticians design and analyze experiments to improve the safety, reliability and performance of products of all types. Statisticians are also directly involved with quality control issues in manufacturing to ensure consistent product dependability.

Statisticians work with social scientists to survey attitudes and opinions. In education, statisticians are involved with the assessment of educational aptitude and achievement and with experiments designed to measure the effectiveness of curricular innovations. Statisticians are an important part of research teams which search for better varieties of agricultural crops, and for safer and more effective use of fertilizers.

In major hospitals, medical schools and government agencies, statisticians study the control, prevention, diagnosis and treatment of diseases, injuries and other health abnormalities. They also investigate the efficiency of health delivery systems and practices.

In the pharmaceutical industry, statisticians design experiments to measure the efficacy of drugs in treating illnesses and to assess the likelihood of undesirable side effects.

Statistical methods are also used in business practice, e.g. to forecast demand for goods and services. Actuaries use statistical methods to assess risk levels and set premium rates for insurance and pension industries.

Statisticians also play a vital role in assessing employment levels and needs of the population for health, economic and social services. Without accurate information from agencies like Ghana Statistical Services, Customs Excise and Preventive Services (CEPS), Environmental Protection Agency (EPA), the government cannot effectively allocate its resources.

Research in statistical methods is carried out in universities, government agencies and in private industry. Statisticians employed in these activities, develop new ways to collect and analyze data for the many types of data and experimental settings encountered in practical studies.

1.5 Variables and types of variables

Variables and constants

Variables

Any type of observation which can take different values for different people, or different values at different times, or places, is called a ***variable***. The following are examples of variables:

(a) family size, number of hospital beds, year of birth, number of schools in a country, etc.

(b) height, mass, blood pressure, temperature, blood glucose level, etc.

There are, broadly speaking, two types of variables – ***quantitative*** and ***qualitative variables (or categorical)*****.**

Constants

Constants are characteristics that have values that do not change. Examples of constants are: pi (π), the ratio of the circumference of a circle to its diameter $(\pi = 3.14159...)$ and e, the base of the natural or (Napierian) logarithms $(e = 2.71828)$.

Types of variables

Quantitative variables

A quantitative variable is one that can take numerical values. The variables in (a) and (b), above, are examples of quantitative variables. Quantitative variables may be characterized further as to whether they are ***discrete*** or ***continuous***.

Discrete variables

The variables in (a), above, can be counted. These are examples of discrete variables. A discrete variable is characterized by gaps or interruptions in the values that it can assume. Any variable phrased as “the number of …”, is discrete, because it is possible to list its possible values {0,1, …}. Any variable with a finite number of possible values is discrete. The following example illustrates the point. The number of daily admissions to a hospital is a discrete variable since it can be represented by a whole number, such as 0, 1, 2 or 3. The number of daily admissions on a given day cannot be a number such as 1.8, 3.96 or 5.33.

Continuous variables

The variables in (b), above, can be measured. These are examples of continuous variables. A continuous variable does not possess the gaps or interruptions characteristic of a discrete variable. A continuous variable can assume any value within a specific relevant interval of values assumed by the variable. Notice that age is continuous since an individual does not age in discrete jumps.

Categorical variables

A variable is called categorical when the measurement scale is a set of categories. For example, marital status, with categories (single, married, widowed), is categorical. For Ghanaians, the region of residence, is categorical, with categories Greater Accra, Eastern, and so on. Other categorical variables are whether employed (yes, no), religious affiliation (Protestant, Catholic, Jewish, Muslim, others, none), political party preference and favorite type of music (classical, country, folk, jazz, rock), place of birth, nationality, colour, colour of hair, gender, blood group, smoking habit, surname, rank in military. Categorical variables are often called ***qualitative***. It can be seen that categorical variables can neither be measured nor counted.

1.6 Levels of measurement and measurement scales

Variables can further be classified according to the following four levels of measurement: nominal, ordinal, interval and ratio. A detailed discussion of this can be found in Stevens (1946), and Ofosu and Hesse (2011).

Nominal scale

This scale of measure applies to qualitative variables only. On the nominal scale, no order is required. For example, gender is nominal, blood group is nominal, and marital status is also nominal. On the nominal scale, categories are mutually exclusive. Thus an item must belong to exactly one category. Notice that, we cannot perform arithmetic operations on data measured on the nominal scale.

Ordinal scale

This scale also applies to qualitative data. On the ordinal scale, order is necessary. This means that one category is lower than the next one or vice versa. For example, in the Army, the rank of private is lower than the rank of captain, which is lower than the rank of major, and so on. Thus, the rank of an army officer is measured on the ordinal scale. In universities, the rank of an academic staff is measured on the ordinal scale. Grades are also ordinal, as excellent is higher than very good, which in turn is higher than good, and so on.

It should be noted that, in the ordinal scale, differences between category values have no meaning. For example, although Professor is higher than Lecturer, the difference between these two ranks does not exist numerically. Similarly, if 4 denotes “excellent”, 3 denotes “very good”, 2 denotes “good” and 1 denotes “fair”, it does not mean that a candidate who is rated “excellent” is twice as competent as a candidate who is rated “good”, just because “excellent” is denoted by 4 and “good” is denoted by 2.

Interval scale

This scale of measurement applies to quantitative data only. In this scale, the zero point does not indicate a total absence of the quantity being measured. An example of such a scale is temperature on the Celsius or Fahrenheit scale. Suppose the minimum temperatures of 3 cities, *A*, *B* and *C*, on a particular day were 0 °C, 20 °C and 10 °C, respectively. It is clear that we can find the differences between these temperatures. For example, city *B* is 20 °C hotter than city *A*. However, we cannot say that city *A* has no temperature. Note that city *A* has a temperature equivalent to 32 °F. Moreover, we cannot say that city *B* is twice as hot as city *C*, just because city *B* is 20 °C and city *C* is 10 °C. The reason is that, in the interval scale, the ratio between two numbers is not meaningful.

Ratio scale

This scale of measurement also applies to quantitative data only and has all the properties of the interval scale. In addition to these properties, the ratio scale has a meaningful zero starting point and a meaningful ratio between 2 numbers.

An example of variables measured on the ratio scale, is weight. A weighing scale that reads 0 kg gives an indication that there is absolutely no weight on it. So the zero starting point is meaningful. If Yaw weighs 40 kg and Akosua weighs 20 kg, then Yaw weighs twice as Akosua. Another example of a variable measured on the ratio scale is temperature measured on the ***Kelvin scale***. This has a true zero point.

Summary of types of variables

Fig. 1.1 shows a chart, summarizing the relationships between the various types of variables and measurement scales.

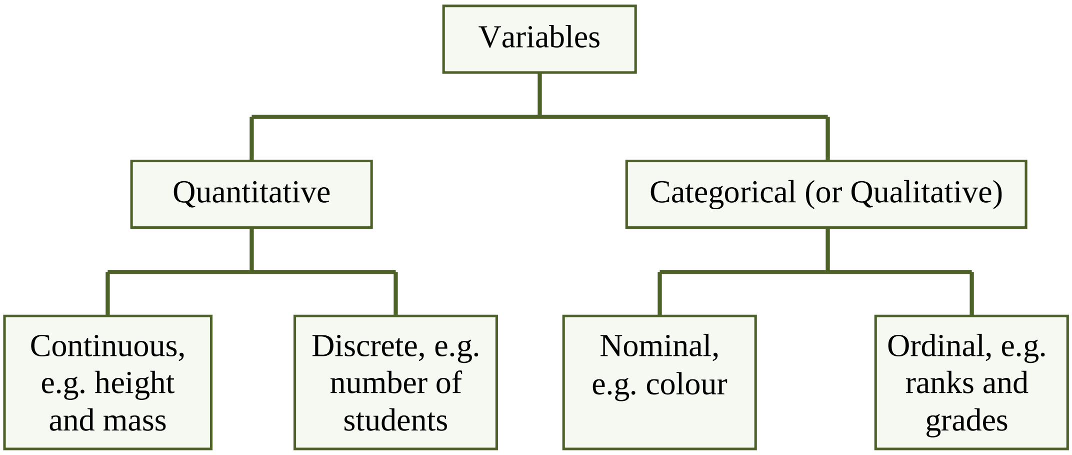

Variables
Quantitative
Categorical (or Qualitative)
Continuous, e.g. height and mass
Discrete, e.g. number of students
Nominal, e.g. colour
Ordinal, e.g. ranks and grades

***Fig. 1.1*: Types of variables**

Exercise 1(a)

1. For each of the following variables, state whether it is quantitative or qualitative and specify the measurement scale that is employed when taking measurements on each.

   (a) gender of babies born in a hospital, (b) marital status,
   (c) temperature measured on the Kelvin scale, (d) nationality,
   (e) masses of babies in kg, (f) temperature in °C,
   (g) prices of items in a shop, (h) position in an exam.
   (i) the rank of an academic staff in a University.

2. For each of the following situations, answer questions (a) through (d):

   (a) What is the variable in the study? (b) What is the population?
   (c) What is the sample size? (d) What measurement scale was used?

   A. A study of 150 students from St. Ann School, showed that 10% of the students had blood group A.

   B. A study of 100 patients admitted to St. Paul's Hospital, showed that 25 patients lived 8 km from the hospital.

   C. A study of 50 teachers in Town A showed that 5% of the teachers earn GH¢800.00 per month.

3. A team of ornithologist is doing field research by using a mist net to capture migrating birds. They collect the following information:

(a) Species, (b) Weight
(c) Wing span (d) Condition, either poor, fair, good, or excellent,
(e) Band ID number, (f) Approximate age.

Indicate whether each of these is an attribute measure or a variables measure.

4. Explain what is meant by inferential statistics.

5. Define the following terms:
(a) population, (b) qualitative variable,
(c) discrete variable, (d) sample,
(e) continuous variable, (f) quantitative variable.

6. For each of the following, indicate whether it is a discrete or a continuous variable.
(a) The number of minutes it takes to read a page in this text.
(b) The number of chapters in the text.
(c) The weight of the text.
(d) The number of problems in the text.
(e) The number of times the letter *e* appears on a page.
(f) The length of a page in inches.

7. Suppose that the following information is obtained from Ms Ofosu on her application for a home mortgage following response, indicate whether it is a continuous variable and which type of measurement scale it represents.
(a) Place of residence: in Accra.
(b) Type of residence: Single family home.
(c) Date of birth: August 13, 1966.
(d) Projected monthly payments: GH¢2 479.
(e) Occupation: Director of Food and Drug Board.
(f) Employer: Methodist University.
(g) Number of years at Job: 10.
(h) Annual income: GH¢140 000.
(i) Amount of mortgage requested: GH¢220 000.

8. Which scale of measurement (norminal, ordinal, or interval) is most appropriate for
(a) Attitude toward legalization of marijuana (favour, neutral, oppose).
(b) Gender (male, female).
(c) Number of children in a family (0, 1, 2, …).
(d) Political party affiliation (NPP, NDC, CPP).
(e) Religious affiliation (Catholic, Jewish, Protestant, Muslim, Others).

(f) Political philosophy (very liberal, somewhat liberal, moderate, somewhat conservative, very conservative).
(g) Years of school completed (0, 1, 2, 3, …).
(h) Highest degree attained (none, high school, bachelor's, master's, doctorate).
(i) Employment status (employed, full time, employed part time, unemployed).

1.7 Methods of data collection

1.7.1 Introduction

Most research techniques and many statistical process-control techniques involve the use of sampling. A sample is selected, evaluated and studied in an effort to gain information about the larger population from which the sample was drawn. In Section 1.1, we learned that a sample is defined as a subset or part of a population. Although, by definition, samples will be smaller than the population from which they are drawn, samples can be very small or very large. A single student can be considered a sample of students from a given university, a very large sample consisting of millions of households can be selected to respond to a lengthy questionnaire that is part of a census.

A sample represents a population, and information obtained from a sample is generalized to be true for the entire population from which it was drawn. The validity or accuracy of generalizations from samples to populations depends on how well a sample represents its population. A well-selected sample can provide information comparable to that obtained by a census.

Advantages of sampling

Studying a sample instead of a population, can have the following advantages.

1. ***Cost*** – Samples can be studied at much lower cost. The smaller number of units or individuals involved in a sample requires less time and money to evaluate. Samples can provide affordable, accurate, and useful information in cases where a census would cost more than the value of the information obtained.

2. ***Time*** – Samples can be evaluated more quickly than a population. If a decision had to wait for the results of a census, a critical advantage might be missed, or the information might be made obsolete by events or changes that took place while the data were being collected and analyzed.

3. ***Accuracy*** – Any time data are collected, there is a chance for errors to occur. Errors of measurement, incorrect recording of data, transposition of digits, recording of information in the wrong area of a form, and errors in entering data into a computer can all influence the accuracy of results. In general, the larger the data set, the more opportunity there is for

errors to occur. A sample can provide a data set that is small enough to monitor carefully and can permit careful training and supervision of data gatherer and handlers.

4. ***Feasibility*** – In some research situations, the population of interest is not available for study. A substantial portion of the population might not yet exist or might no longer be available for evaluation. In other cases, evaluation of an item requires its destruction. For example, a manufacturer interested in how much pressure could be applied to a part before it cracked, could not perform a census without destroying the entire production run.

5. ***Scope of information*** – In a sample survey, there are greater varieties of information that can be considered which may be impracticable in a complete census due to constraints such as limited number of trained personnel and equipment. When evaluating a smaller group, it is sometimes possible to gather more extensive information on each unit evaluated.

1.7.2 Sample designs

There are two categories of sample designs, namely, ***probability*** **(*or random*)** ***sampling*** and ***non-probability sampling***.

1. Probability Sampling

In this sub-section, we introduce important sampling methods which incorporate ***randomization***, which means that the selection is not consciously influenced by human choice. The major principle of these designs is to avoid bias in the selection procedure and to achieve the maximum precision for a given outlay of resources. The main types of probability sampling designs are: simple random sampling, systematic sampling, stratified sampling, cluster sampling and multi-stage sampling.

(i) Simple random sample

Subjects of a population to be sampled could be families, schools, cities, hospitals, records of reported crimes, and so on. ***Simple random sampling*** is a method of sampling for which every possible sample has equal chance of selection. Let n denote the number of subjects in the sample. This number is called the ***sample size***.

Definition 1.8 (Simple random sample)

> A ***simple random sample*** of n subjects from a population is one in which each possible sample of that size has the same probability (chance) of being selected.

A simple random sample is often just called a ***random sample***. The ***simple*** objective is used to distinguish this type of sampling from more complex sampling schemes.

Why is it a good idea to use random sampling? Because everyone has the same chance of inclusion in the sample, so it provides fairness. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population. Most inferential statistical methods assume randomization of the sort provided by random sampling.

How to select a simple random sample

One way of obtaining a simple random sample is to use the 'lottery system'.

The lottery system

The lottery system consists of writing the name of each item in the sample frame on a slip of paper or a card and then drawing them from a container one after the other. To ensure a bias free selection, shuffle the cards or the slips of paper before each draw.

Advantages of the lottery system

- It is independent of the properties of the population.
- It is a very reliable method of selecting random samples.
- It eliminates selection bias.

Disadvantages of the lottery system

- It is time-consuming and cumbersome when the population is large.
- Cannot be used when the population is infinite.

A discussion of methods of data collection can be found from Levy and Lemeshow (1999) and Rao (2000).

Tables of random numbers

Another method for selecting a random sample is to use a table of random numbers. A table of random numbers has the property that, no matter how we select our digits (up, down, diagonally, etc.) each digit, 0 through 9, is equally likely to be selected. Table 1.2, on the next page, shows 48 random digits arranged in 8 columns and 6 rows of five-digit blocks. The random numbers were generated by using the MINITAB software.

Definition 1.9 (Random numbers)

> ***Random numbers*** are numbers that are computer generated according to a scheme whereby each digit is equally likely to be any of the integers 0, 1, 2, …, 9 and does not depend on the other digits generated.

The numbers fluctuate according to no set pattern. Any particular digit has the same chance of being a 0, 1, 2, …, or 9. The numbers are chosen independently, so any digit chosen has no influence on any other selection. If the first digit in a row of the table is 9, for instance, the next digit is still just as likely to be a 9 as a 0 or a 1 or any other number. Random numbers are available in published tables and can be generated with a software and many statistical calculators.

Table 1.2: A table of random numbers

Column Row	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	10480	15011	01536	02011	81647	91646	69179	14194
2	22368	46573	25595	85393	30995	89198	27982	53402
3	24130	48360	22527	97265	76393	64809	15179	24830
4	42167	93093	06243	61680	07856	16376	39440	53537
5	37570	39975	81837	16656	06121	91782	60468	81308
6	77621	06907	11008	42751	27756	53498	18602	70659

Example 1.3

Suppose you want to select a simple random sample of 10 students from a class of 20 students. The sampling frame is a directory of these students. You can select the students by using two-digit random numbers to identify them, as follows:

(1) Assign the numbers 01 to 20 to the students in the directory, using 01 for the first student in the list, 02 for the second student, and so on.

(2) Starting at any point in Table 1.2, choose successive two-digit numbers until you obtain 10 distinct numbers between 01 and 20.

(3) Include in the sample the students with the assigned numbers equal to the random numbers selected.

For example, using the first row of Table 1.2, the first 5 two-digit random numbers are 10, 15, 01, 02 and 14. Notice that we skipped the numbers which are greater than 20 since no student in the directory has an assigned number greater than these numbers.

After using the first row of Table 1.2, move to the next row of numbers and continue. The column (or row) from which you begin selecting the number does not matter, since the numbers have no set pattern. Most statistical software can do this all for you.

(ii) Systematic random sample

Another method of random sampling is to choose every k^{th} item from the list, starting from a randomly chosen entry among the first k items on the list. This is called ***systematic sampling***. The number k is called the ***skip*** number. Fig. 1.2 shows how to sample every fourth item, starting from item 2, resulting in a sample of size $n = 20$ items from a list of $N = 78$ items.

A systematic sample of n items from a population of N items requires that the skip number be approximately N/n. In sampling from a sampling frame, it is simpler to select a systematic random sample than a simple random sample because it uses only one random number.

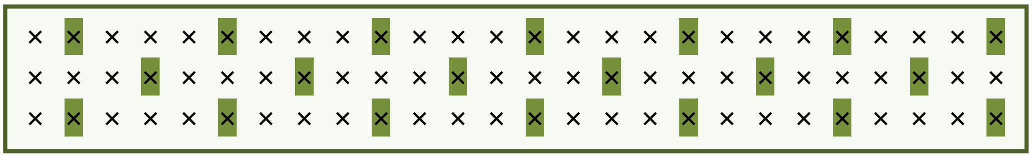


Fig. 1.2: Systematic sampling

An attraction of systematic sampling is that it can be used with unlistable or infinite population, such as production processes (e.g. testing every 5 000th light bulb) or political polling (e.g., surveying every tenth voter who emerges from the polling place). Systematic sampling is also well-suited to linearly organized physical population (e.g., pulling every tenth patient folder from alphabetized filing drawers in a veterinary clinic).

Example 1.4

Suppose we want a systematic random sample of 100 students from a population of $30\,000$ students listed in a campus directory. Here, $n = 100$ and $N = 30\ 000$, and so $k = 30\ 000/100 = 300$. The population size is 300 times the sample size. Therefore we have to select one of every 300 students. We select one student at random using every 300^{th} student after the one selected randomly. This produces a sample of size 100. The first three digits in Table 1.2 are 104, which falls between 001 and 300, so we first select the student numbered 104. The numbers of the other students selected are 104 + 300 = 404, 404 + 300 = 704, 704 + 300 = 1004, 1004 +300 = 1304, and so on. The 100^{th} student selected is listed in the last 300 names in the directory.

(iii) Stratified random sample

Another probability sampling method, useful in social science research for studies comparing groups, is stratified random sampling.

Definition 1.10

A ***stratified random sample*** divides the population into subgroups called ***strata***, and then selects a simple random sample from each stratum.

Stratified random sampling is called ***proportional*** if the sampled strata proportions are the same as those in the entire population. For example, if 90% of the population of interest are men and 10% are women, then the sampling is proportional if the sample size for men is nine times the sample size for women.

Stratified random sampling is called ***disproportional*** if the sampled strata proportion differs from the population proportions. This is useful when the population size for a stratum is relatively small. A group that comprises a small part of the population may not have enough representation in a simple random sample to allow precise inferences.

Example 1.5

Suppose we want to estimate smallpox vaccination rate among employees in a university, and we know that our target population (those individuals we are trying to study) is 55% male and 45% female. Suppose our budget only allows a sample of size 200. To ensure the correct gender balance, we could sample 110 males and 90 females.

(iv) Cluster random sampling

Simple, systematic, and stratified random sampling are often difficult to implement, because they require a complete sampling frame. Such lists are easy to obtain when sampling cities or hospitals for example, but more difficult to obtain when sampling individuals or families. *Cluster samples* are essentially strata consisting of geographical regions. We divide a region (say a city) into sub-regions (say, blocks, sub-divisions, or schools). In a one-stage cluster sampling, our sample consists of all elements in each of k randomly chosen sub-regions (or clusters). In a two-stage cluster sampling, we first randomly select k sub-regions (clusters) and then choose a random sample of elements within each cluster. Fig. 1.3 illustrates how four elements could be sampled from each of five randomly chosen clusters, using a two-stage cluster sampling.

Cluster sampling is useful when:

- Population frame and stratum characteristic are not readily available.
- It is too expensive to obtain a simple or stratified sample.
- The cost of obtaining data increases sharply with distance.
- Some loss of reliability is acceptable.

Although cluster sampling is cheap and quick, it is often reasonably accurate because people in the same neighbourhood tend to be similar in income, ethnicity, educational background, and so on. Cluster sampling is useful in political polling, surveys of gasoline pump prices, studies of crime victimization surveys, or lead contamination in soil. A hospital may contain clusters (floors) of similar patients. A warehouse may have cluster (pallets) of inventory parts. Forest sections may be viewed as clusters to be sampled for disease or timber growth rates.

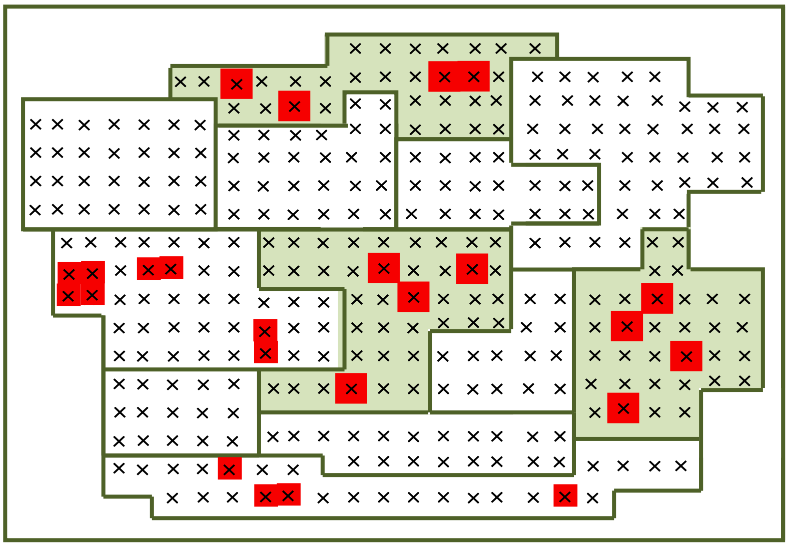


Fig. 1.3: Two-stage cluster sampling

Example 1.6

A study might plan to sample about 1% of the families in a city, using city block as clusters. Using a map to identify city blocks, it could select a simple random sample of 1% of the blocks and then sample every family on each block. A study of patient care in mental hospitals in Ghana could first randomly sample mental hospitals (the clusters) and then collect data for patients within these hospitals.

Example 1.7

What is the difference between a stratified sample and a cluster sample?

Solution

A stratified sample uses every stratum. The strata are usually groups we want to compare. By contrast, a cluster sample uses a sample of the clusters, rather than all of them. In cluster sampling, clusters are merely ways of easily identifying groups of subjects. The goal is not to compare the clusters but to use them to obtain a sample. Most clusters are not represented in the eventual sample.

(v) Multi-stage Sampling

A random sample of a population of interest often incurs considerable expense in collecting the data from a wide area. A cheaper solution is to use multi-stage sampling which starts by dividing the country into a number of regions. Some of these are selected at random and subdivided further, e.g. into rural, suburban and inner city areas. Again, some of these are selected at random and subdivided again, e.g. into parliamentary wards and a further random selection made. The process can be repeated until individual households or companies or units of interest are identified.

The Family Expenditure Survey makes use of multi-stage sampling. The Survey uses the Small Users File of Postcode Address File and the primary sampling unit is postal sectors. The benefit of this approach is that the resulting samples are concentrated in relatively few geographical areas which reduces the cost of data collection.

2. Non-probability sampling

Non-probability sampling designs select samples with features not embodying randomness. The selection of the elements in the sample lies solely on personal judgement. The chance of selecting an element cannot be determined. For this reason, there is no means of measuring the risk of making erroneous conclusion desired from non-probability samples. Thus the reliability of results (i.e. sampling errors) cannot be assessed and also used to make valid conclusions about the population. The main methods of non-probability sampling are Convenience, Judgemental and Quota Sampling

(i) Convenience sample

The sole virtue of ***convenience sampling*** is that it is quick. The idea is to grab whatever sample is handy. The convenience sample is simply one that happens to come your way. An accounting professor who wants to know how many MBA students would take a summer elective in international accounting can just survey the class she is currently teaching. The students polled may not be representative of all MBA students, but an answer (although imperfect) will be available immediately.

A newspaper reporter doing a story on perceived airport security might interview co-workers who travel frequently. An executive might ask department heads if they think non-business Web surfing is widespread.

You might think that convenience sampling is rarely used or, when it is, that the results are used with caution. However, this does not appear to be the case. Since convenience samples often sound the first alarm on timely issue, their results have a way of attracting attention and have probably influenced quite a few business decisions. The mathematical properties of convenience samples are unknowable, but they do serve a purpose and their influence cannot be ignored.

(ii) Judgment sample

Judgment sampling is a non-probability sampling method that relies on the expertise of the sampler to choose items that are representative of the population. The sample obtained by this method is based on personal judgment and some pre-knowledge of the population. For example, to estimate the corporate spending on research and development (R&D) in the medical equipment industry, we might ask an industry expert to select several "typical" firms. Unfortunately, subconscious biases can affect expert, too. In this context, "bias" does not mean prejudice, but rather non-randomness in the choice. Judgment samples may be the best alternative in some cases, but we can't be sure whether the sample was random.

(iii) Quota Sampling

Quota sampling is a special kind of judgment sampling, in which the interviewer chooses a certain number of people in each category (e.g., men/women). Quota sampling involves first classification of the population into non-overlapping sub populations, called strata. The sample is then obtained by selecting the individual elements from each stratum based on a specified quota. In quota sampling the selection of the sample is made by the interviewer, who has been given quotas to fill from specified sub-groups of the population. For example, an interviewer may be told to sample 50 females between the ages of 45 and 60.

Since the selection of the sample is non–random, the enumerator is allowed to use his/her own judgement to meet the various quotas. This introduces a large degree of biasness. The lack of randomness is, however, compensated for by less cost and administrative convenience.

1.7.3 Sampling with or without replacement

Consider the lottery system on page 14. If an item selected is put in the box before taking another item, we are ***sampling with replacement***. Using the box analogy, if we throw each item back in the bowl and stir the contents before the next draw, an item can be chosen again. Duplicates are unlikely when the sample size n is much smaller than the population size N. People instinctively prefer sampling ***without replacement*** because drawing the same item more than once seems to add nothing to our knowledge. However, using the same sample item more than once does not introduce any bias (i.e. no systematic tendency to over or underestimate whatever parameter we are trying to measure).

1.8 Computers and statistical analysis

The recent widespread use of computers has had a tremendous impact on statistical analysis. Computers can perform more calculations faster and far more accurately than can human technicians. The use of computers makes it possible for investigators to devote more time to the improvement of the quality of raw data and the interpretation of the results.

The current prevalence of microcomputers and the abundance of statistical software packages have further revolutionized statistical computing. The researcher in search of a statistical software package will find the book by Woodward et al. (1987) extremely helpful. This book describes approximately 140 packages. Among the most prominent ones are: Statistical Package for the Social Sciences (SPSS), S-plus, MINITAB, SAS and GENSTAT. The spreadsheet, Excel, also has facilities for statistical analysis.

Exercise 1(b)

1. Give two reasons why it is sometimes necessary to take a sample from a population.
2. State two ways of obtaining primary data.
3. State two sources of secondary data.
4. State two advantages and two disadvantages of the lottery system for taking a simple random sample from a population.
5. State two disadvantages and one advantage of telephone interview, as a means of collecting data.
6. Briefly describe the difference between descriptive statistics and inferential statistics.
7. A doctor examined a patient to determine the cause of a disease. He took a drop of blood and used it to determine the state of health of the patient. What aspect of statistics is the doctor employing in order to form a judgement?
8. In your own words, explain and give an example of each of the following statistical terms: (a) population, (b) sample.
9. Mrs. Akrong wants to check whether the pot of soup she is cooking has the right taste and quantity of salt. She did this by tasting a small portion of the soup scooped in a ladle. What aspect of statistics is she employing in order to form a judgement? Briefly explain why she decided to use this particular method?

10. Explain the difference between qualitative and quantitative data. Give examples of qualitative and quantitative data.

11. List the four levels of measurement and give examples.

12. Explain the difference between:
    (a) nominal and ordinal data, (b) a census and a sample survey,

13. Clusters versus strata
    (a) With a cluster random sample, do you take a sample of (i) the clusters? (ii) the subjects within every cluster?
    (b) With a stratified random sample, do you take a sample of (i) the strata? (ii) the subjects within every stratum?
    (c) Summarize the main differences between cluster sampling and stratified sampling in terms of whether you sample the group or sample from within the group the form the clusters or strata.

14. A class has 50 students. Use the column of the first two digits in the random number table (Table 1.2) to select a simple random sample of three students. If the students are numbered 01 to 50, what are the numbers of the three students selected?

15. In cluster random sample with equal-sized clusters, every subject has the same chance of selection. However, the sample is not a simple random sample. Explain why not.

1.9 Chapter summary

Statistical methods analyze data on ***variables***, which are characteristics that vary among subjects. Statistical methods depend on the type of variable.

- Numerically measured variables, such as family income and number of children in a family, are ***quantitative***. They are measured on an ***interval scale***.
- Variables taking values in a set of categories are ***categorical or qualitative***. Those measured with unordered categories, such as religious affiliation and blood group, have a ***nominal scale***. Those measured with ordered categories, such as social class and political ideology, have an ***ordinal scale*** of measurement.
- Variables are also classified as ***discrete***, having possible values that are a set of separate numbers (such as 0, 1, 2, …), or ***continuous***, having a continuous, infinite set of possible values. Categorical variables, whether nominal or ordinal, are discrete. Quantitative variables can be of either type, but in practice are treated as continuous if they can take a large number of values.

Much social science research uses ***observational studies***, which use available subjects to observe variables of interest. One should be cautious in attempting to conduct inferential analyses with data from such studies. Inferential statistical methods require ***probability samples***, which incorporate randomization in some way. Random sampling allows control over the amount of ***sampling error***, which describes how results can vary from sample to sample. Random samples are much more likely to be representative of the population than are non-probability sample such as volunteer samples.

- For a simple random sample, every possible sample of size n has the same chance of selection.
- There are other examples of probability sampling: ***Stratified*** random sampling takes every k^{th} subject in the sampling frame list. ***Stratified*** random sampling divides the population into groups (strata) and takes a random sample from each stratum. ***Cluster*** random sampling takes a random sample of clusters of subjects (such as city blocks) and uses subjects in those clusters as the sample. ***Multistage*** sampling uses combinations of these methods.

References

Levy, P. S. and Lemeshow, S. (1999). *Sampling of populations, Methods and Appications*. John Wiley and Sons Inc., New York.

Rao, P. S. (2000). *Sampling Methodologies with applications*. Chapman and Hall, London.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 677 – 680.

Ofosu, J. B. and Hesse, C. A. (2011). Elementary Statistical Methods (2nd Edition). EPP books services, Accra.

Woodward, W. A., Elliott, A. C. and Gray, H. L. (1987). Directory of Statistical Microcomputer Software. *Marcel Dekker, New York*.

Chapter Two

Descriptive Statistics

Chapter Contents

We have seen that statistical methods are ***descriptive*** or ***inferential***. The purpose of descriptive statistics is to summarize data to make it easier to assimilate the information. In this chapter, we present basic methods of descriptive statistics.

2.1 Frequency distribution

Table 2.1 gives the number of children per family for 54 families selected from Obo, a town in Ghana. The data, presented in this form in which it was collected, is called ***raw data*****.**

Table 2.1: Number of children per family

0	1	4	4	3	2	2	3	1	2	4	3	0	2	1	1	2	2
1	1	3	2	2	4	0	0	4	2	2	3	1	1	2	3	2	2
2	0	3	4	2	1	3	2	2	3	4	4	1	0	3	2	1	1

From Table 2.1, it can be seen that, the minimum and the maximum numbers of children per family are 0 and 4, respectively. Apart from these numbers, it is impossible, without further careful study, to extract any exact information from the data. By breaking down the data into the form of Table 2.2, however, certain features of the data become apparent. For instance, from Table 2.2, it can easily be seen that, most of the 54 families selected have two children. This information cannot easily be obtained from the raw data in Table 2.1.

Table 2.2: **Frequency distribution of the data in Table 2.1**

Number of children	Tally	Frequency
0	~~////~~ /	6
1	~~////~~ ~~////~~ //	12
2	~~////~~ ~~////~~ ~~////~~ ///	18
3	~~////~~ ~~////~~	10
4	~~////~~ ///	8
		Total = 54

Table 2.2 is called a ***frequency table*** or a ***frequency distribution***. It is so called because it gives the frequency or number of times each observation occurs. Thus, by finding the frequency of each observation, a more intelligible picture is obtained.

The steps for constructing a frequency distribution may be summarized as follows:

(i) List all values of the variable in ascending order of magnitude.

(ii) Form a tally column, that is, for each value in the data, record a stroke in the tally column next to that value. In the tally, each fifth stroke is made across the first four. This makes it easy to count the entries and enter the frequency of each observation. (Note: Values with frequency zero are omitted.)

(iii) Check that the frequencies sum to the total number of observations.

2.1.1 Grouped frequency distribution

Table 2.3 gives the body masses of 22 patients, measured to the nearest kilogram.

Table 2.3: **Body masses (in kilograms) of 22 patients**

60	45	72	55	42	65	54	68	74	50	78
70	58	48	67	64	68	52	60	58	75	83

It can be seen that the minimum and the maximum body masses are 42 kg and 83 kg, respectively. A frequency distribution giving every body mass between 42 kg and 83 kg would be very long and would not be very informative. The problem is overcome by grouping the data into classes. If we choose the classes 41 – 49, 50 – 58, 59 – 67, 68 – 76 and 77 – 85, we obtain the frequency distribution given in Table 2.4.

Table 2.4: Grouped frequency distribution of the data in Table 2.3

Mass (kg)	Tally	Frequency
41 – 49	///	3
50 – 58	~~////~~ /	6
59 – 67	~~////~~	5
68 – 76	~~////~~ /	6
77 – 85	//	2
		Total = 22

These are, of course, not the only classes which could be chosen. Table 2.4 gives the frequency of each group or class; it is therefore called a ***grouped frequency table*** or a ***grouped frequency distribution***. Using this grouped frequency distribution, it is easier to obtain information about the data than using the raw data in Table 2.3. For instance, it can be seen from Table 2.4, that 17 of the 22 patients have body masses between 50 kg and 76 kg (both inclusive). This information cannot easily be obtained from the raw data in Table 2.3.

It should be noted that, even though Table 2.4 is concise, some information is lost. For example, the grouped frequency distribution does not give us the exact body masses of the patients. Thus the individual body masses of the patients are lost in our effort to obtain an overall picture. However, Table 2.4 is far more comprehensible and its contents are easier to grasp than Table 2.3.

We now define the terms that are used in grouped frequency tables.

(i) Class limits

The intervals into which the observations are put are called ***class intervals***. The end points of the class intervals are called ***class limits***. For example, the class interval 41 – 49, has lower class limit 41 and upper class limit 49.

(ii) Class boundaries

The raw data in Table 2.3 were recorded to the nearest kilogram. Thus, a body mass of 49.5 kg would have been recorded as 50 kg, a body mass of 58.4 kg would have been recorded as 58 kg, while a body mass of 58.5 kg would have been recorded as 59 kg. It can therefore be seen that, the class interval 50 – 58, consists of measurements greater than or equal to 49.5 kg and less than 58.5 kg. The numbers 49.5 and 58.5 are called the ***lower and upper boundaries*** of the class interval 50 – 58. The class boundaries of the other class intervals are given in Table 2.5.

Table 2.5: **Body masses of 22 patients (to the nearest kg)**

Class interval	Class boundaries	Class mark	Frequency
41 – 49	40.5 – 49.5	45	3
50 – 58	49.5 – 58.5	54	6
59 – 67	58.5 – 67.5	63	5
68 – 76	67.5 – 76.5	72	6
77 – 85	76.5 – 85.5	81	2

$[\mathbf{S}_1]$ Notice that the lower class boundary of the i^{th} class interval is the mean of the lower class limit of the class interval and the upper class limit of the $(i-1)^{th}$ class interval (i = 2, 3, 4, …). For example, in Table 2.5, the lower class boundaries of the second and the fourth class intervals are $\frac{1}{2}(50 + 49) = 49.5$ and $\frac{1}{2}(68 + 67) = 67.5$, respectively.

$[\mathbf{S}_2]$ It can also be seen that the upper class boundary of the i^{th} class interval is the mean of the upper class limit of the class interval and the lower class limit of the $(i+1)^{\text{th}}$ class interval (i = 1, 2, 3, …). Thus, in Table 2.5, the upper class boundary of the fourth class interval $(68-76)$ is $\frac{1}{2}(76 + 77) = 76.5$.

(iii) Class mark

The mid-point of a class interval is called the ***class mark*** or ***class mid-point*** of the class interval. It is the average of the upper and lower class limits of the class interval. It is also the average of the upper and lower class boundaries of the class interval. For example, in Table 2.5, the class mark of the third class interval was found as follows:

$$\text{class mark } = \tfrac{1}{2}(59 + 67) = \tfrac{1}{2}(58.5 + 67.5) = 63.$$

(iv) Class width

The difference between the upper and lower class boundaries of a class interval is called the ***class width*** of the class interval. ***Class widths of class intervals can also be found by subtracting two consecutive lower class limits, or by subtracting two consecutive upper class limits. In particular***:

$[\mathrm{S}_3]$ The width of the i^{th} class interval is the numerical difference between the upper class limits of the i^{th} and the $(i-1)^{\text{th}}$ class intervals (i = 2, 3, …). It is also the numerical difference between the lower class limits of the i^{th} and the $(i+1)^{\text{th}}$ class intervals (i = 1, 2, …).

In Table 2.5, the width of the first class interval is $|41-50|$ = 9. This is the numerical difference between the lower class limits of the first and the second class intervals. The width of the second class interval is $|50-59|$ = 9. This is the numerical difference between the lower class limits of the second and the third class intervals. It is also equal to $|58-49|$, the numerical difference between the upper class limits of the first and the second class intervals.

Example 2.1

Table 2.6 gives the distribution of the lengths of 30 iron rods, measured to the nearest ten centimetres. Find the class widths and the class boundaries of the class intervals.

Table 2.6: Lengths of iron rods (to the nearest 10 cm)

Length (cm)	60 – 90	100 – 150	160 – 200	210 – 250	260 – 310
Frequency	3	6	10	7	4

Solution

Table 2.7, on the next page, shows the required class boundaries and the class widths. The class widths were found as follows:

Class width of the first class interval = 100 – 60 = 40 (**see** $[\mathrm{S}_3]$).

Class width of the second class interval = 150 – 90 = 60, etc.

The class boundaries were found as follows:

By $[\mathrm{S}_2]$, the upper class boundary of the first class interval is $\frac{1}{2}(90+100)=95.$ Since the class width of this class interval is 40, the lower class boundary of the class interval is 95 – 40 = 55.

By $[\mathrm{S}_1]$, the lower class boundary of the last class interval is $\frac{1}{2}(250+260)=255.$ The class width of this class interval is 310 – 250 = 60 (see $[\mathrm{S}_3]$). Therefore, the upper class boundary of the last class interval is 255 + 60 = 315.

Table 2.7: Lengths of iron rods (to the nearest 10 cm)

Length (cm)	Class boundaries	Class width
60 – 90	55 – 95	100 – 60 = 40
100 – 150	95 – 155	150 – 90 = 60
160 – 200	155 – 205	200 – 150 = 50
210 – 250	205 – 255	250 – 200 = 50
260 – 310	255 – 315	310 – 250 = 60

Notice that, since the lengths of the iron rods are recorded to the nearest 10 cm, the lengths 45 cm, 46 cm, …, 54 cm will be recorded as 50 cm, while the lengths 55 cm, 56 cm, …, 64 cm will be recorded as 60 cm. Furthermore, the lengths 85 cm, 86 cm, …, 94 cm will be recorded as 90 cm while the length 95 cm will be recorded as 100 cm. It can therefore be seen that the observations in the first class interval are greater than or equal to 55 cm and less than 95 cm.

2.1.2 Guidelines for choosing class intervals

(1) Given a set of raw data (that is, data which have not been organized numerically), before we construct a grouped frequency distribution, we have to decide on the number of class intervals to use in order to give the best indication of the trends in the data. If too few class intervals are used, important features of the distribution may be overlooked and if too many class intervals are used, then the purpose of the table, the reduction of the data to a manageable size, may be defeated. Experience has shown that the best number of class intervals to choose is between 5 and 20, depending upon such factors as the range and the number of observations. Those who wish to have more specific guidance in the matter of deciding how many class intervals are needed may refer to Sturges (1926).

(2) Class intervals must be uniquely defined, i.e., class intervals must be chosen such that no value in the data can be included in two different classes. Consider, for example, Table 2.9, on the next page, which shows two groupings of the data in Table 2.8. It can be seen that, in Grouping 1, the class intervals are not uniquely defined. For example, the number 58 can be included in the first two class intervals while the number 64 can be included in the second and the third class intervals. Grouping 2, however, defines the class intervals uniquely and it is therefore preferred to Grouping 1.

Table 2.8: Marks obtained by 20 students in an examination

58	74	64	72	60	80	58	78	66	80
76	57	63	70	62	53	66	65	74	67

Table 2.9: Frequency distribution of the data in Table 2.8

Grouping 1	Grouping 2
Class boundaries	Class boundaries
52 – 58	52.5 – 58.5
58 – 64	58.5 – 64.5
64 – 70	64.5 – 70.5
70 – 76	70.5 – 76.5
76 – 85	76.5 – 82.5

(3) As previously pointed out, all values within a class interval are assumed to be concentrated at the class mark of that class interval. Class intervals should therefore be chosen such that the class marks coincide with actually observed data. The advantage of this method is that it tends to reduce errors brought about by grouping. Furthermore, calculations will be performed on the grouped data and the class marks will be used in these calculations. It is therefore convenient to choose class marks that will make these future calculations as simple as possible. Consider, for example, Table 2.11, which shows two groupings of the data given in Table 2.10. The class marks for Grouping 1 coincide with some of the observed data, whereas those for Grouping 2 do not. Moreover, it will be easier to use the class marks for Grouping 1 for further calculations than those for Grouping 2. Hence Grouping 1 would be preferred to Grouping 2.

Table 2.10: Masses of 18 eggs (to the nearest gramme)

47	72	46	68	57	62	62	58	69
50	64	52	49	67	47	71	72	57

Table 2.11: Frequency distribution of the data in Table 2.10

Grouping 1		Grouping 2	
Class limits	Class marks	Class limits	Class marks
45 – 49	47	44 – 49	46.5
50 – 54	52	50 – 55	52.5
55 – 59	57	56 – 61	58.5
60 – 64	62	62 – 67	64.5
65 – 69	67	68 – 73	70.5
70 – 74	72		

(4) Usually, it is convenient to make all class intervals of equal size, but there are occasions when class intervals of varying sizes can be used more effectively than those of equal size. For example, in Table 2.12, we see that the frequencies of the last three class intervals are small in comparison with those of the other class intervals. In cases such as this, we combine the last three class intervals. This gives Table 2.13.

Table 2.12: **Ages of women blood donors**

Age (in years)	Frequency
20.5 – 30.5	687
30.5 – 40.5	705
40.5 – 50.5	998
50.5 – 60.5	453
60.5 – 70.5	142
70.5 – 80.5	93
80.5 – 90.5	10

Table 2.13: **Ages of women blood donors**

Age (in years)	Frequency
20.5 – 30.5	687
30.5 – 40.5	705
40.5 – 50.5	998
50.5 – 60.5	453
60.5 – 90.5	245

2.1.3 Relative frequency

It is sometimes useful to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the frequency of the particular class interval by the total number of observations. We refer to the proportion of values falling within a class interval as the ***relative frequency*** of the class interval. In Table 2.13, the relative frequency of the first class interval is

$$\frac{687}{3088} = 0.2225,$$

since the class frequency is 687 and the sum of the frequencies is 3 088. Note that relative frequencies must add up to 1, allowing for rounding errors.

2.1.4 Cumulative frequency

In many situations, we are not interested in the number of observations in a given class interval, but in the number of observations which are less than (or greater than) a specified value. For example, in Table 2.5, on page 26, it can be seen that 3 patients have body masses less than 49.5 kg and 9 patients (i.e. 3 + 6) have body masses less than 58.5 kg. These frequencies are called ***cumulative frequencies***. A table of such cumulative frequencies is called a ***cumulative frequency table*** or ***cumulative frequency distribution***.

Table 2.14 shows the data in Table 2.5 along with the cumulative frequencies and the relative frequencies. Notice that the last cumulative frequency is equal to the sum of all the frequencies.

Table 2.14: **Frequency, cumulative frequency, and relative frequency distributions of the data in Table 2.5**

Mass (kg)	Frequency	Cumulative frequency	Relative frequency
40.5 – 49.5	3	3	0.1364
49.5 – 58.5	6	9	0.2727
58.5 – 67.5	5	14	0.2273
67.5 – 76.5	6	20	0.2727
76.5 – 85.5	2	22	0.0909
	Total = 22		Total = 1.0000

Example 2.2

Table 2.15 gives the ages of a sample of patients who attended Hope Medical Hospital.

(a) Find the sample size. (b) Complete the blank cells.

Table 2.15: **Ages of patients**

Ages (years)	Freqency	Relative frequency	Cumulative frequency
10 – 14	–	–	–
15 – 19	8	0.16	12
20 – 24	15	–	–
25 – 29	–	–	37
30 – 34	–	–	–

Solution

(a) If the sample size is *n*, then the relative frequency of the second class interval is 8 ÷ *n*. Hence, *n* is a root of the equation

$$\frac{8}{n} = 0.16 \quad \Rightarrow \quad n = \frac{8}{0.16} = 50\,.$$

The sample size is 50.

(b) Table 2.16 gives the completed blank cells.

***Table 2.16:* Ages of patients**

Ages (years)	Freqency	Relative frequency	Cumulative frequency
10 – 14	4	0.08	4
15 – 19	8	0.16	12
20 – 24	15	0.30	27
25 – 29	10	0.20	37
30 – 34	13	0.26	50
	Total = 50	Total = 1.00	

Notice again that:

(i) the last cumulative frequency is equal to the sum of all the frequencies;

(ii) relative frequencies must add up to 1, allowing for rounding errors.

Exercise 2(a)

1. The following are the blood groups of a sample of patients who attend Peace Hospital.

A	B	O	AB	B	A	O	O	AB	B
B	B	A	O	O	AB	O	B	A	B
AB	O	A	B	A	O	A	A	B	A

(a) What is the population in this study? (b) What is the variable in this study?

(c) Construct a frequency table for the data.

2. The following table shows the number of hours 45 hospital patients slept following the administration of a certain anesthetic.

7	10	12	4	8	7	3	8	5
12	11	3	8	1	1	13	10	4
4	5	5	8	7	7	3	2	3
8	13	1	7	17	3	4	5	5
3	1	17	10	4	7	7	11	8

(a) Construct a frequency distribution for the data using the class intervals 0 – 2, 3 – 5, 6 – 8, …, 15 – 17.

(b) Find the relative frequencies and the cumulative frequencies of the frequency distribution in part (a).

3. Find the relative frequencies and the cumulative frequencies of the frequency distribution in Table 2.6 on page 27.

4. In a certain study, blood glucose levels (in mg/dl) of a sample of students of St. Andrew High School were measured.

103	125	120	118	117	109	114	118	131	118
116	119	117	119	110	117	124	117	124	113
127	127	114	129	120	105	121	112	115	126
101	114	128	125	109	122	123	130	115	123

(a) State the population and the variable in the study.

(b) Make a frequency table of the data using the class intervals 100 – 104, 105 – 109, 110 – 114, …, 130 – 134.

(c) Obtain the class boundaries, class mid-points and class widths of the frequency distribution in part (a).

(d) Find the relative frequencies and the cumulative frequencies of the frequency distribution in part (a).

5. Find the class boundaries, class mid-points and class widths of the class intervals of the following grouped frequency distributions.

(a)

Class interval	3 – 7	8 – 10	11 – 15	16 – 18	19 – 25
Frequency	5	15	25	12	7

(b)

Class Interval	**Frequency**
1500 – 1540	4
1550 – 1600	6
1610 – 1690	10
1700 – 1800	7
1810 – 1870	3

(c)

Class Interval	**Frequency**
50 – 90	6
100 – 240	10
250 – 340	14
350 – 440	5

6. The following table gives the distribution of the ages of a sample of patients who attend Hope Hospital.

Age (years)	Frequency	Relative Frequency
5 – 14	6	0.08
15 – 24	9	–
25 – 34	–	0.24
35 – 44	24	–
45 – 54	15	–
55 – 64	–	–

(a) What is the population in this study? (b) What is the variable in this study?
(c) What is the sample size? (d) Complete the blank cells in the table.

7. The following table gives the distribution of the ages of a sample of 25 students from St. Luke's School. Complete the blank cells in the table.

Age (years)	Frequency	Cumulative Frequency	Relative Frequency
10 – 15	2	–	–
16 – 21	4	–	–
22 – 32	–	–	0.32
33 – 43	–	–	–
44 – 60	5	–	–

8. The following are the number of babies born during a year in 60 community hospitals.

30	55	27	45	56	48	45	49	32	57
47	56	37	55	52	34	54	42	32	59
35	46	24	57	32	26	40	28	53	54
29	42	42	54	53	59	39	56	59	58
49	53	30	53	21	34	28	50	52	57
43	46	54	31	22	31	24	24	57	29

From these data:

(a) Construct a frequency distribution using the class intervals 20 – 24, 25 – 29, 30 – 34, …

(b) Find the relative frequencies and the cumulative frequencies of the frequency distribution in part (a).

9. The following are the lengths of 22 iron rods, measured to the nearest centimetre.

83	75	58	60	52	68	64	67	72	58	60
78	50	74	68	54	65	42	55	48	45	70

(a) Construct a frequency distribution using the class intervals 41 – 49, 50 – 58, …, 77 – 85.

(b) Find the relative frequencies and cumulative frequencies of the frequency distribution in part (a).

2.2 Graphical representation of data

In the last section, we found that information given in a frequency distribution is easier to interpret than raw data. Information given in a frequency distribution in a tabular form is easier to grasp if presented graphically. Many types of diagrams are used in statistics, depending on the nature of the data and the purpose for which the diagram is intended. In this section, we discuss how statistical data can be presented by histograms and cumulative frequency curves.

2.2.1 Histogram

A histogram consists of rectangles with:

(i) bases on a horizontal axis, centres at the class marks, and lengths equal to the class widths,

(ii) areas proportional to class frequencies.

If the class intervals are of equal size, then the heights of the rectangles are proportional to the class frequencies and it is then customary to take the heights of the retangles numerically equal to the class frequencies.

If the class intervals are of different widths, then the heights of the rectangles are proportional to $\frac{\text{class frequency}}{\text{class width}}$. This ratio is called ***frequency density***.

Example 2.3

Table 2.17 shows the distribution of the heights of 40 students selected from St. Paul High School. Draw a histogram to represent the data.

***Table 2.17:* Heights of students**

Height (cm)	150 – 154	155 – 159	160 – 169	170 – 174	175 – 184	185 – 189
Frequency	3	4	16	10	6	1

Solution

Since the class intervals have different sizes, the heights of the rectangles of the histogram are proportional to the frequency densities of the class intervals. The calculations of the heights of the rectangles can be set up as shown in Table 2.18. If d_i is the frequency density of class interval i, then the height of the rectangle representing this class interval is cd_i, where c is any positive number (see Table 2.18, column 5). Fig. 2.1 shows a histogram for the data. It was drawn by taking c = 5. Notice that the centres of the bases of the rectangles of the histogram are at the class marks. If preferred, a histogram may be drawn showing class boundaries instead of class marks.

***Table 2.18*: Work table for computing the heights of rectangles of a histogram**

Height (cm)	Class width (*a*)	Frequency (*b*)	Frequency density d_i = (*b*)/(*a*)	Height of rectangle
150 – 154	5	3	0.6	0.6*c*
155 – 159	5	4	0.8	0.8*c*
160 – 169	10	16	1.6	1.6*c*
170 – 174	5	10	2.0	2.0*c*
175 – 184	10	6	0.6	0.6*c*
185 – 189	5	1	0.2	0.2*c*

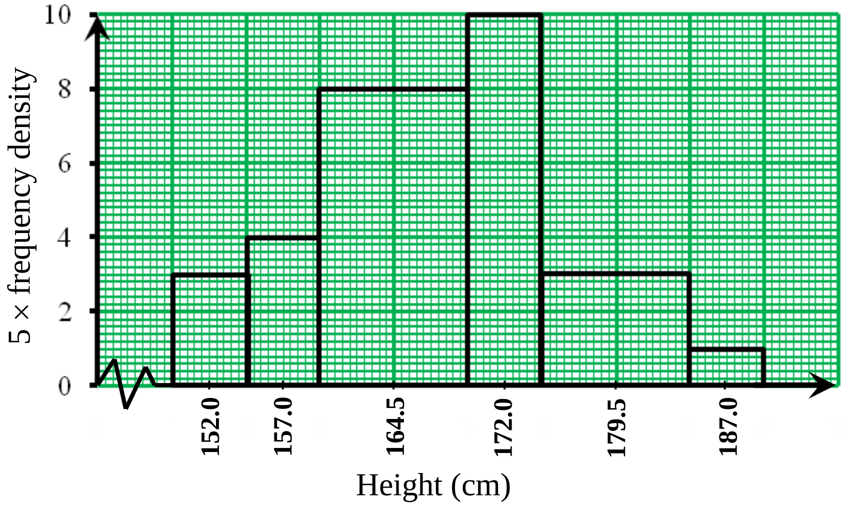


Fig. 2.1: ***Histogram of the data in Table 2.17***

Example 2.4

Table 2.19 shows the distribution of ages of 168 diabetic patients selected from Progress Hospital. A histogram is drawn to represent the data. If the height of the rectangle representing the second class interval is 3 cm, find the height of the rectangle which represents the third class interval.

Table 2.19: **Ages of diabetic patients**

Age (years)	5 – 9	10 – 24	25 – 34	35 – 44
Frequency	20	36	48	64

Solution

The calculations of the heights of the rectangles of the histogram can be set up as shown in Table 2.20. Notice that the heights of the rectangles are proportional to the frequency densities of the class intervals (see Table 2.20, column 5).

Table 2.20: **Work table for computing the heights of rectangles of a histogram**

Age (years)	Class width (a)	Frequency (b)	Frequency density $d_i = (b) \div (a)$	Height of rectangle
5 – 9	5	20	4.0	$4.0c$
10 – 24	15	36	2.4	$2.4c$
25 – 34	10	48	4.8	$4.8c$
35 – 44	10	64	6.4	$6.4c$

If the height of the rectangle representing the second class interval is 3 cm, then

$$2.4c = 3 \quad \Leftrightarrow \quad c = \frac{3}{2.4} = 1.25.$$

The height of the rectangle which represents the third class interval is

$$4.8c \text{ cm} = 4.8 \times 1.25 \text{ cm} = 6 \text{ cm}.$$

Drawing a histogram

1. When drawing a histogram, suitable scales must be chosen for both the vertical and horizontal axes. Scales like “2 cm to 5 units” or “2 cm to 10 units” are the best. Avoid using scales like “2 cm to 3 units” or “2 cm to 7 units”.
2. Label the axes.
3. Give your graph a title.

2.2.2 Cumulative frequency curve

A graph obtained by plotting a cumulative frequency against the upper class boundary and joining the points by a smooth curve, is called a ***cumulative frequency curve***. The following example illustrates an application of a cumulative frequency curve.

Example 2.5

Table 2.21 shows the frequency distribution of the body masses of 50 AIDS patients.

(a) Construct a cumulative frequency curve to represent the data.

(b) Use your cumulative frequency curve to estimate the number of patients whose body masses are: (i) less than 65 kg, (ii) at least 75 kg.

Table 2.21: **Body masses of 50 AIDS patients**

Mass (kg)	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
Frequency	3	6	17	13	8	3

Solution

(a) Table 2.22 gives the cumulative frequency distribution of the data in Table 2.21.

Table 2.22: **Cumulative frequency distribution of the data in Table 2.21**

Mass (kg) less than	Cumulative frequency
29.5	0
39.5	0 + 3 = 3
49.5	3 + 6 = 9
59.5	9 + 17 = 26
69.5	26 + 13 = 39
79.5	39 + 8 = 47
89.5	47 + 3 = 50

Notice that a class with frequency zero is added before the first class. It can be seen that the last cumulative frequency is equal to the total number of observations, a check on the accuracy of our calculation. The corresponding cumulative frequency curve is shown in Fig. 2.2 on page 40. The curve is obtained by marking the upper class boundary on the horizontal axis and the cumulative frequencies on the vertical axis. All the points are joined by a smooth curve.

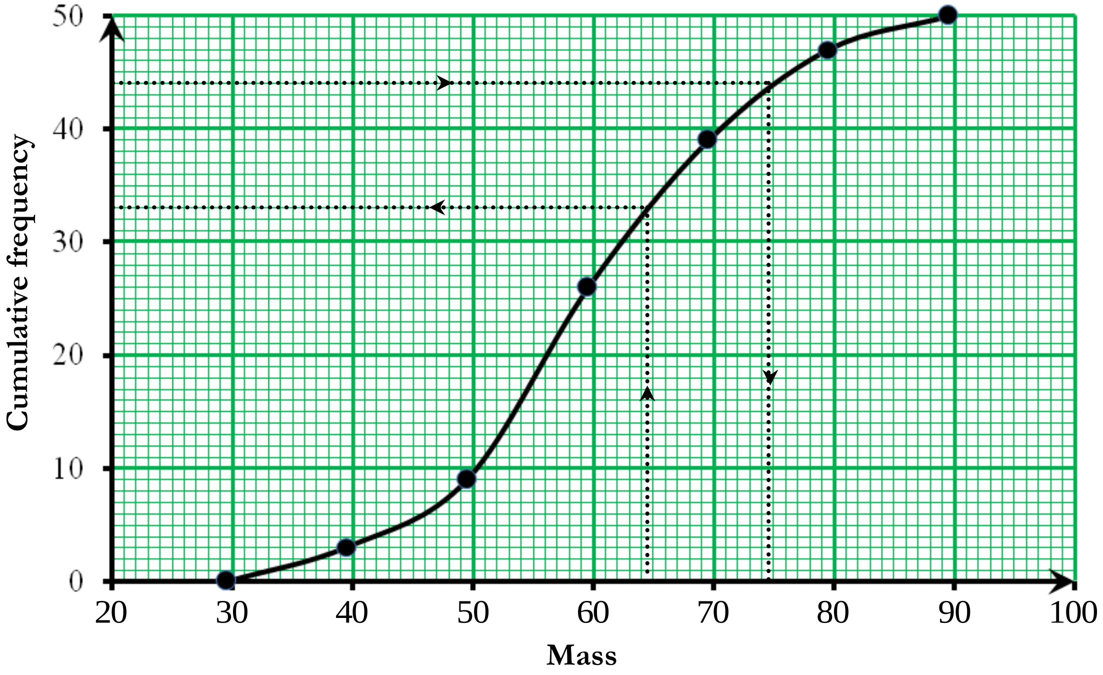


Fig. 2.2: *Cumulative frequency curve of the data in Table 2.21*

(b) (i) Since the body masses of the patients are recorded to the nearest integer, body masses less than 65 kg consist of all body masses less than 64.5 kg. Therefore, to estimate the number of patients whose body masses are less than 65 kg, we obtain the cumulative frequency which corresponds to the point 64.5 kg on the horizontal axis. From Fig. 2.2, we find that 33 patients have body masses less than 65 kg.

(ii) To estimate the number of patients whose body masses are at least 75 kg, we first estimate the number of patients whose body masses are less than 75 kg. Now, the upper boundary of the interval "less than 75" is 74.5. From Fig. 2.2, the cumulative frequency which corresponds to the point 74.5 kg on the horizontal axis, is 44. It follows that 44 patients have body masses less than 75 kg. Thus, the number of patients whose body masses are at least 75 kg is (50 – 44) = 6.

2.2.3 Frequency polygon

A grouped frequency table can also be represented by a frequency polygon, which is a special kind of line graph. To construct a frequency polygon, we plot a graph of class frequencies against the corresponding class mid-points and join successive points with straight lines. Fig. 2.3, on the next page, shows the frequency polygon for the data in Table 2.16, on page 33.

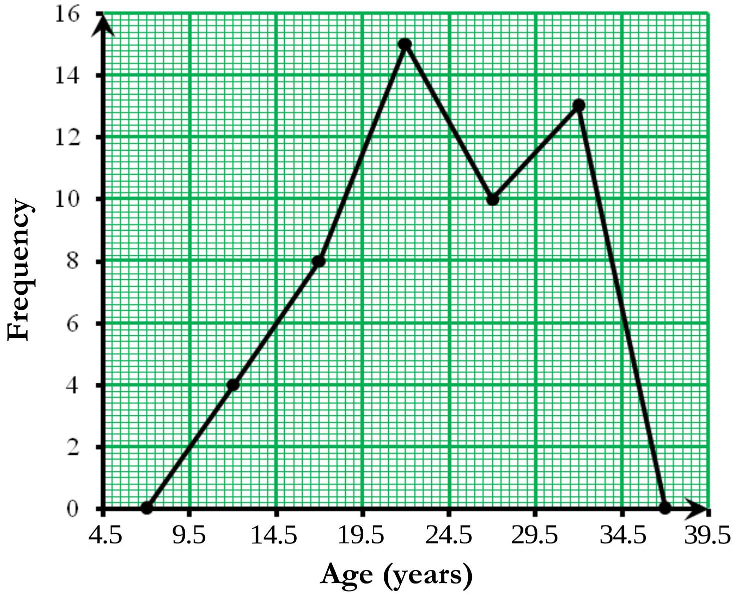


Fig. 2.3: ***Frequency polygon of the data in Table 2.16***

Notice that the polygon is brought down to the horizontal axis at the ends of points that would be the mid-points if there were additional class intervals at each end of the corresponding histogram. This makes the area under a frequency polygon equal to the area under the corresponding histogram.

Fig. 2.4 shows the frequency polygon of Fig. 2.3 superimposed on the corresponding histogram. This figure allows us to see, for the same set of data, the relationship between the two graphic forms.

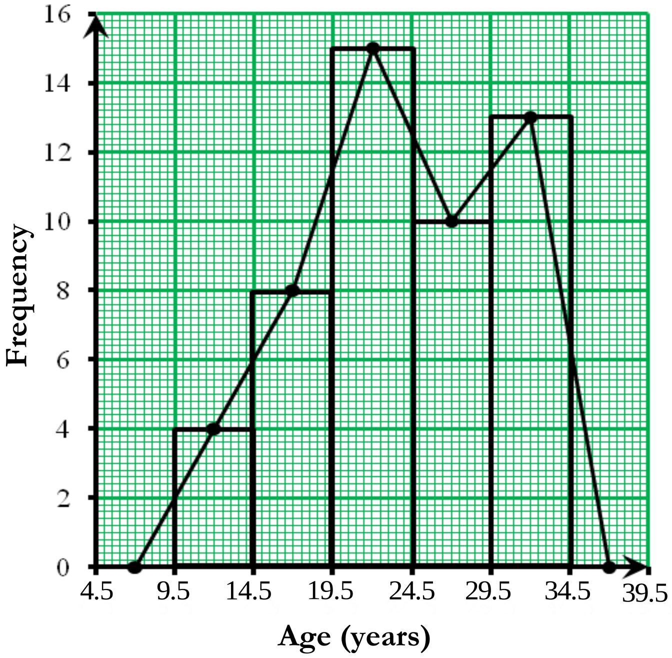


Fig. 2.4: ***Histogram and frequency polygon of the data in Table 2.16***

2.2.4 Stem-and-leaf plot

A stem-and-leaf plot is a graphical device that is useful for representing a relatively small set of data which takes numerical values. To construct a stem-and-leaf plot, we partition each measurement into two parts. The first part is called the ***stem***, and the second part is called the ***leaf***. The stem of a measurement consists of one or more of the remaining digits. The stems form an ordered column with the smallest stem at the top and the largest at the bottom. The stems are separated from their leaves by a vertical line. We include in the stem column all stems within the range of the data even when a measurement with that stem is not in the data set. The rows of a stem-and-leaf plot contain the leaves, ordered and listed to the right of their respective stems. When leaves consist of more than one digit, all digits after the first may be omitted. Decimals, when present in the original data, are omitted in a stem-and-leaf plot.

A stem-and-leaf plot conveys similar information as a histogram. Turned on its side, it has the same shape as the histogram. In fact, since the stem-and-leaf plot shows each observation, it displays information that is lost in a histogram. A properly constructed stem-and-leaf plot, like a histogram, provides information regarding the range of the data set, shows the location of the highest concentration of measurements, and reveals the presence or absence of symmetry. An advantage of a stem-and-leaf plot over a histogram, is the fact that it preserves the information contained in the individual measurements. Such information is lost when we construct a grouped frequency table. Another advantage of a stem-and-leave plot is that it can be constructed during the tallying process, so the intermediate step of preparing an ordered array is eliminated.

Stem-and-leaf plots are useful for quick portrayal of a small data set. As the sample size increases, you can accommodate the increase in leaves by splitting the stems. For instance, you can list each stem twice, putting leaves of 0 to 4 on one line and leaves of 5 to 9 on another. When a number has several digits, it is simplest for graphical portrayal to drop the last digit or two. For instance, for a stem-and-leaf plot of annual income in thousands of dollars, a value of GH¢27.1 thousand has a stem of 2 and a leave of 7 and a value of GH¢106.4 thousand has a stem of 10 and a leaf of 6.

Example 2.6

The following are the marks scored by 30 candidates in an English test. Construct a stem-and-leaf plot for the data.

56	71	62	81	52	61	73	80	84	93
53	75	78	78	56	64	65	76	78	78
85	88	94	96	96	67	89	78	79	68

Solution

Since all the measurements are two-digit numbers, we will have one-digit stems and one-digit leaves. For example, the mark 85 has a stem of 8 and a leaf of 5. Fig. 2.5 is the required stem-and-leaf plot. The four numbers in the first row represent 52, 53, 56 and 56.

Stem	Leaf
5	2 3 6 6
6	1 2 4 5 7 8
7	1 3 5 6 8 8 8 8 8 9
8	0 1 4 5 8 9
9	3 4 6 6

Fig. 2.5: ***Stem-and-leaf plot of the data in Example 2.6***

2.2.5 Bar chart

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

(i) Simple bar chart

In a simple bar chart, the height (or length) of each bar is equal to the frequency it represents.

Example 2.7

Table 2.23 gives the production of timber in five districts of Ghana in a certain year. Draw a bar chart to illustrate the data. The bars are separated to emphasize that the variable is quantitative rather than quantitative.

Table 2.23: **Production of timber in 5 districts in Ghana**

Districts	Production of timber (tonnes)
Bibiani	600
Nkawkaw	900
Wiawso	1800
Ahafo	1500
Agona	2400

Solution

Fig. 2.6, on the next page, represents the required bar chart. Notice that the bars are of equal width and the distances between them are equal. Since district is a nominal variable, there is no particular natural order for the bars.

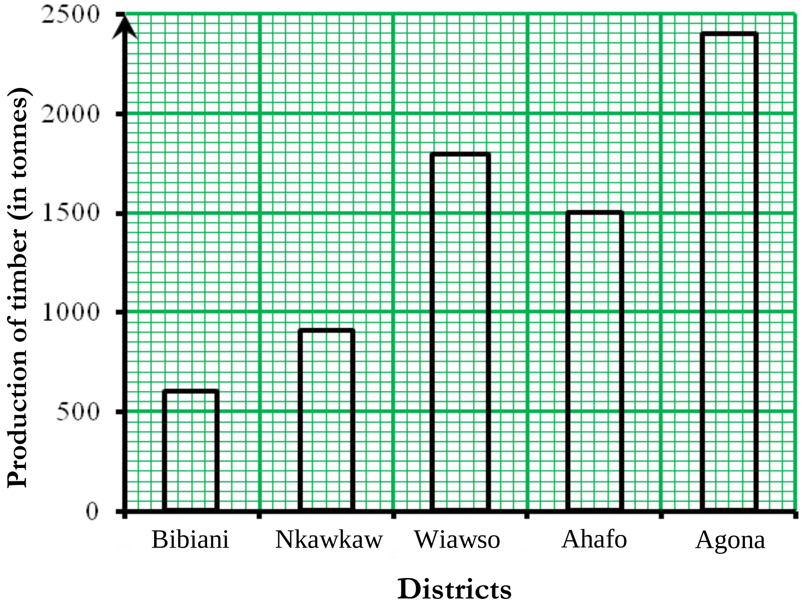


Fig. 2.6: ***A simple bar chart for the data in Table 2.23***

(ii) Component bar chart

In a component bar chart, the bar for each category is subdivided into component parts; hence its name. Component bar charts are therefore used to show the division of items into components. This is illustrated in the following example.

Example 2.8

Table 2.24 shows the distribution of sales of agricultural produce from Asiedu Farm in 1995, 1996 and 1997. Illustrate the information with a component bar chart.

Table 2.24: **Sales of agricultural produce from Asiedu Farm**

		Sales (million dollars)		
		1995	**1996**	**1997**
Agricultural produce	**Coffee**	90	120	180
	Cocoa	180	140	220
	Palm oil	30	30	20

Solution

Fig. 2.7, on the next page, shows a component bar chart for the data. The sales of agricultural produce consist of three components: the sales of coffee, cocoa, and palm oil. The component bar chart shows the changes of each component over the years as well as the comparison of the total sales between different years.

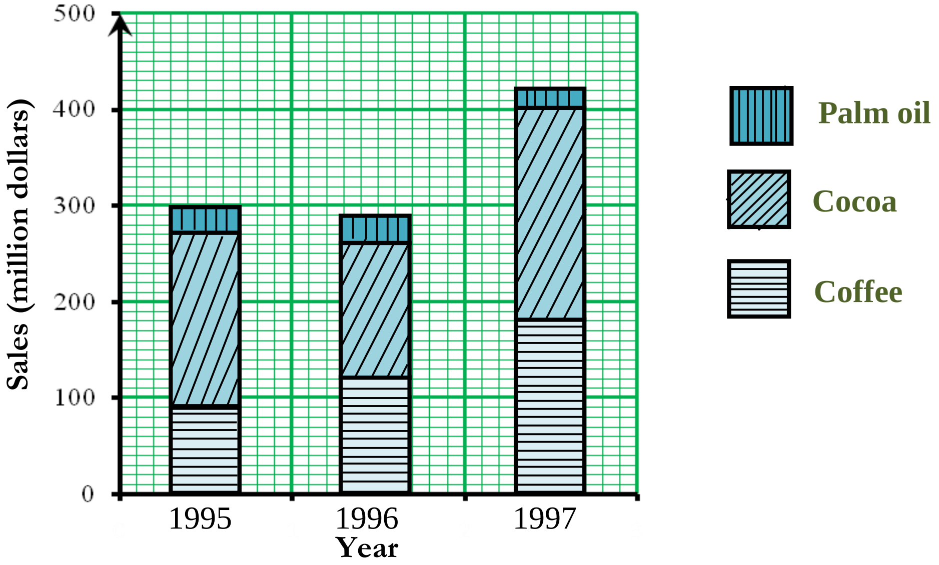


Fig. 2.7: A component bar chart of the data in Table 2.24

(iii) Grouped bar chart

For a grouped bar chart, the components are grouped together and drawn side by side. We illustrate this with the following example.

Example 2.9

Illustrate the data in Table 2.24 with a grouped bar chart.

Solution

Fig. 2.8 shows the required grouped bar chart.

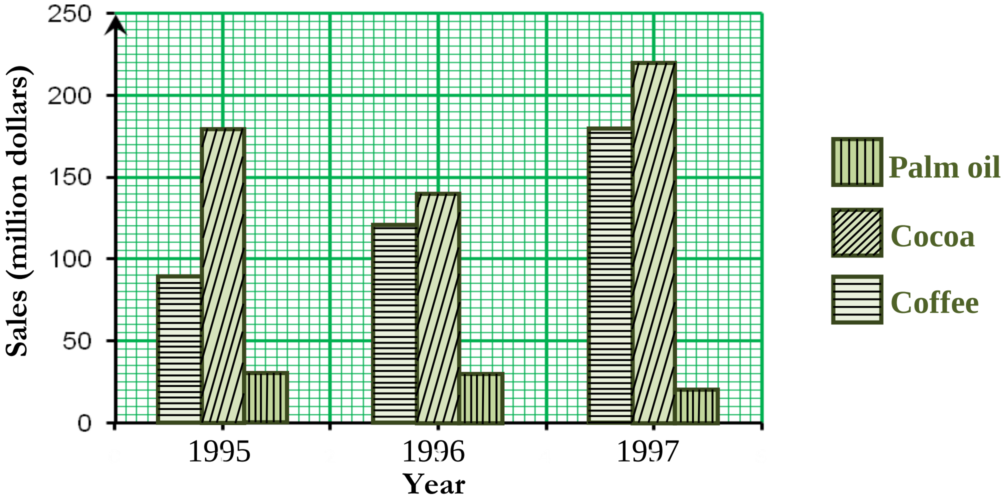


Fig. 2.8: A grouped bar chart of the data in Table 2.24

2.2.6 Pie Charts

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

The following are the steps for constructing a pie chart

(1) Find the sum of the category values.

(2) Calculate the angle of the sector for each category, using the following result:

$$\text{angle of the sector for category } A = \frac{\text{value of category } A}{\text{sum of category values}} \times 360^{\circ}.$$

(3) Construct a circle and mark the centre.

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

(5) Label each sector clearly.

Example 2.10

A housewife spent the following sums of money on buying ingredients for a family Christmas cake in 2007.

Flour	GH¢24
Margarine	GH¢96
Sugar....................................	GH¢18
Eggs	GH¢60
Baking powder......................	GH¢12
Miscellaneous	GH¢30

Represent the above information on a pie chart.

Solution

The angles of the sectors are calculated as shown in Table 2.25. Fig. 2.9, on the next page, shows the required pie chart.

Table 2.25: Work table for computing the angles of the sectors of a pie chart

Item	Amount used (GH¢)	Angle of sector
Flour	24	$\frac{24}{240} \times 360^{o} = 36^{o}$
Margarine	96	$\frac{96}{240} \times 360^{o} = 144^{o}$
Sugar	18	$\frac{18}{240} \times 360^{o} = 27^{o}$
Eggs	60	$\frac{60}{240} \times 360^{o} = 90^{o}$
Baking powder	12	$\frac{12}{240} \times 360^{o} = 18^{o}$
Miscellaneous	30	$\frac{30}{240} \times 360^{o} = 45^{o}$
Total	**240**	**360º**

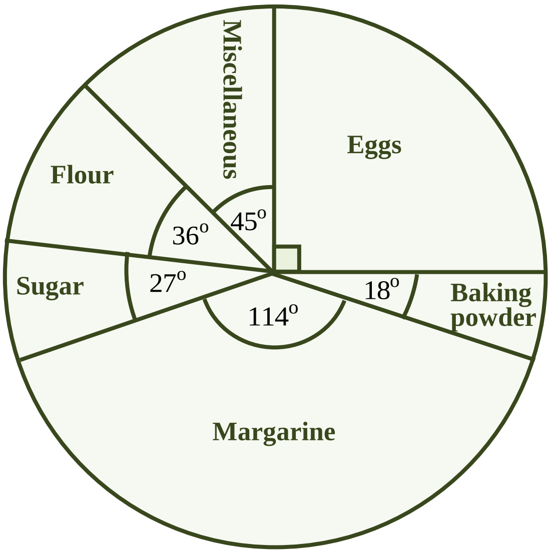


Fig. 2.9: *A pie chart of the data in Table 2.24*

Exercise 2(b)

1. Refer to Exercise 2(a), Question 2. Construct a histogram and a frequency polygon using the frequency distribution in part (a).
2. Refer to Exercise 2(a), Question 8. Use the frequency distribution in part (a) to construct a histogram and a frequency polygon to represent the data.
3. The following are the ages of 30 patients seen in the emergency room of a hospital on a Monday night. Construct a stem-and-leaf plot for the data.

32	21	35	43	39	60	36	12	54	45
37	53	45	23	64	10	34	22	36	45
55	44	55	46	22	38	35	56	45	57

4. The following table gives the ages (in years) of 60 cancer patients.

Age (years)	5 – 14	15 – 19	20 – 24	25 – 29	30 – 44
Frequency	8	16	18	12	6

A histogram is drawn to represent this data. If the height of the rectangle representing the fifth class interval is 2 cm, find the heights of the rectangles representing the first, second and the third class intervals. Construct a histogram to represent the data.

5. The following table gives the distribution of the heights of 100 children, to the nearest centimetre.

Height (cm)	120–129	130–139	140–149	150–159	160–169	170–179
Frequency	6	15	31	37	9	2

Draw a cumulative frequency curve for the data and use it to estimate:
(a) the number of children whose heights are between 142 cm and 152 cm (inclusive),
(b) the number of children whose heights are greater than 156 cm.

6. The following table gives the distribution of the marks scored by 40 students in an examination.

Mark (%)	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54	55 – 59	60 – 64	65 – 69
Frequency	2	4	7	10	8	5	3	1

Draw a cumulative frequency curve for the data and use it to estimate:
(a) the number of students who scored between 42% and 62% (inclusive),
(b) the least mark a student must score if he/she is to be placed in the top 25% of the class.

7. The following table gives the enrolments in primary and secondary schools in Kenya.

		Number of students (thousands)	
		Primary School	Secondary School
Year	1971	350	154
	1975	351	176
	1979	353	177
	1983	354	180

Illustrate the information with
(a) a component bar chart, (b) a grouped bar chart.

8. The heights, in centimeters, of 30 boys are as follows:

168	149	156	161	172	171	156	141	151	142
155	172	155	162	164	165	166	173	174	176
177	177	168	168	157	143	158	157	145	146

Construct a stem- and- leaf plot of the data (for a boy whose height is 162 cm, record this as a stem of 16 and a leaf of 2.)

9. The following table shows the amount of rainfall in Asarekrom during the first five months of 2006. Construct a bar chart to illustrate the data.

Month	January	February	March	April	May
Rainfall (cm)	5.3	4.9	6.0	5.2	3.4

10. The following table gives the frequency distribution of the results of an examination taken by students from two schools *M* and *N*. Construct a grouped bar chart to represent this information.

Grade	*A*	*B*	*C*	*D*	*E*	*F*
School *M*	5	8	38	14	25	10
School *N*	18	22	20	10	25	5

11. The following information gives the proportion in which Yaro spends his annual salary.

Food	30%	Income Tax	20%
Rent	15%	Savings	5%
Transport	7.5%	Miscellaneous	22.5%

(a) Construct a pie chart to illustrate the above information.

(b) If Yaro’s annual salary is GH¢1 800.00, calculate the amount he spends on food.

12. The level of water in a reservoir is checked every morning at 9 o’clock. One Monday, the level was 61 mm above the zero mark. On each of the next three days, the level fell by 12 mm per day. However, because of a storm in the night, it was found that the level on Friday was 38 mm higher than on Thursday. For the next three days, the level fell by 18 mm per day.

(a) Construct a table of the water level during this week.

(b) Construct a bar chart to illustrate this information.

13. The following table shows the distribution of academic staff by faculty and rank in a certain university. Illustrate the information with

(a) a component bar chart (b) a grouped bar chart

Faculty	**Professors**	**Senior Lecturers**	**Lecturers**
Business Administration	2	3	12
Social Studies	6	2	13
General Studies	3	1	6

14. In an election, the number of votes won by political parties *A*, *B*, *C*, *D* and *E* in a village are as follows:

Party	*A*	*B*	*C*	*D*	*E*
No. of votes	140	110	190	520	240

(a) Construct a pie chart to illustrate this information.
(b) What percentage of the total votes did the winner obtain?

15. The table below shows the number of cars sold by a company from January to June, 1990:

January	February	March	April	May	June
7 100	7 668	10 366	9 940	8 236	7 810

(a) Construct a pie chart to illustrate this information.
(b) What is the percentage of cars sold in February?

2.3 Measures of central tendency

In the above sections, you have learnt how data can be summarised in the form of tables and presented in the form of graphs so that important features can be illustrated easily and more effectively. In this section, we consider statistical measures which can be used to describe the characteristics of a set of data. We are interested in a single value that serves as a representative value of the overall data. Three of such measures are the ***mean***, the ***mode***, and the ***median***. These three measures reflect numerical values in the centre of a set of data and are therefore called measures of ***central tendency***.

2.3.1 The mean

The mean of a set of numbers $x_1, ..., x_n$ is denoted by $\overline{x}$, and is defined by the equation

$$[\mathbf{S}_4] \quad \overline{x} = \frac{1}{n}(x_1 + x_2 + x_3 + ... + x_n) = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}} x_{\mathrm{i}}.$$

It can be seen from $[\mathbf{S}_4]$, that:

$$[\mathbf{S}_5] \quad n\overline{x} = x_1 + x_2 + x_3 + ... + x_n.$$

Example 2.11

Find the mean of the numbers 2, 4 , 7, 8, 11, 12.

Solution

The mean $= \dfrac{2+4+7+8+11+12}{6} = \dfrac{44}{6} = 7\tfrac{1}{3}$.

Example 2.12

The set of numbers $x^2,\ 3,\ 3x-4, 7, 9,$ where x is a positive integer, has a mean of 5. Find the value of x.

Solution

The value of x is given by the equation

$$\tfrac{1}{5}(x^2+3+3x-4+7+9)=5$$

$$\Leftrightarrow \quad x^2+3x+15=25$$

$$\Leftrightarrow \quad x^2+3x-10=0$$

$$\Leftrightarrow \quad (x+5)(x-2)=0$$

$$\Leftrightarrow \quad x=-5 \text{ or } x=2.$$

Since x is a positive integer, we reject the negative root. Hence $x = 2$.

Example 2.13

The maximum load that a lift can take is 1 000 kg. If 5 men with a mean weight of 61 kg and 12 women with a mean weight of 52 kg take the lift, will their total weight exceed the maximum load?

Solution

The total weight of the 5 men and the 12 women is (see $[\mathbf{S}_5]$)

5 × 61 kg + 12 × 52 kg = 929 kg.

This total weight is less than 1 000 kg and so does not exceed the maximum load the lift can take.

The mean of a frequency distribution

If the numbers $x_1, x_2, x_3, ..., x_k$ occur with frequencies $f_1, f_2, f_3, \ldots, f_k$, respectively, then their mean is given by

$$\bar{x} = \frac{f_1x_1+f_2x_2+f_3x_3+...+f_kx_k}{f_1+f_2+f_3+...+f_k} = \frac{\Sigma f_i x_i}{\Sigma f_i} \quad \text{...(2.3.1)}$$

Example 2.14

Table 2.26 shows the body masses of 50 men. Find the mean body mass.

Table 2.26: Body masses of 50 men

Mass (kg)	59	60	61	62	63
Frequency	3	9	23	11	4

Solution

The calculation can be arranged as shown in Table 2.27.

Table 2.27: Work table for calculating the mean

Mass (x)	Frequency (f)	fx
59	3	177
60	9	540
61	23	1 403
62	11	682
63	4	252
	$\sum f = 50$	$\sum fx = 3\,054$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{3\,054}{50} = 61.08.$$

The mean body mass is 61.08 kg.

Assumed mean method

The amount of computation involved in using Equation (2.3.1) can be reduced by using the following result:

If M is any guessed mean or assumed mean (which may be any number) and if $d_{\mathrm{i}} = x_{\mathrm{i}} - M$ ($i = 1, 2, ..., k$), then Equation (2.3.1) becomes

$$\bar{x} = M + \frac{\sum f_{\mathrm{i}} d_{\mathrm{i}}}{\sum f_{\mathrm{i}}} \quad \text{..(2.3.2)}$$

This method is called "***finding the mean by the assumed mean method***"**.** The following example illustrates an application of the assumed mean method.

Example 2.15

Solve Example 2.14 by using a suitable assumed mean.

Solution

We choose 61 (the number with the highest frequency) as the assumed mean. The solution can be arranged as shown in Table 2.28.

Table 2.28: **Calculations for Example 2.15**

Mass (x)	Frequency (f)	$d = x - 61$	fd
59	3	–2	–6
60	9	–1	–9
61	23	0	0
62	11	1	11
63	4	2	8
	$\sum f = 50$		$\sum fd = 19 - 15 = 4$

$$\bar{x} = \left(61 + \tfrac{1}{50}\sum fd\right) = \left(61 + \tfrac{4}{50}\right) = 61.08\text{ kg, as before.}$$

It should be noted that the assumed mean can be any real number. However, the amount of computation can be reduced further if we take the number with the highest frequency as the assumed mean.

The mean of a grouped frequency distribution

Equations (2.3.1) and (2.3.2) are valid for grouped frequency distributions if we interpret x_i as the class mark of a class interval and f_i the corresponding class frequency. In the following example, we apply Equation (2.3.1) to find the mean of a grouped frequency distribution.

Example 2.16

Table 2.29 shows the distribution of the marks scored by 60 students in a Physics examination. Find the mean mark.

Table 2.29: **Marks scored in a Physics examination**

Mark (%)	60 – 64	65 – 69	70 – 74	75 – 79	80 – 84
Number of students	2	15	25	14	4

Solution

The solution can be arranged as shown in Table 2.30, on the next page.

Table 2.30: Calculations for Example 2.16

Marks	Class mark (x)	Frequency (f)	fx
60 – 64	62	2	124
65 – 69	67	15	1 005
70 – 74	72	25	1 800
75 – 79	77	14	1 078
80 – 84	82	4	328
		$\sum f = 60$	$\sum fx = 4\,335$

$$\overline{x} = \frac{\sum fx}{\sum f} = \frac{4335}{60} = 72.25.$$

The mean mark is 72.25 %.

If all the class intervals of a grouped frequency distribution have equal size c, then, Equation (2.3.2) takes the form

$$\overline{x} = M + c\frac{\sum f_i u_i}{\sum f_i} \quad \text{..}(2.3.3)$$

where $u_i = \dfrac{x_i - M}{c}$, $\quad (i = 1, 2, \ldots, k)$.

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

Example 2.17

Use the coding method to solve Example 2.16.

Solution

The width of each class interval is 5 and so c = 5. We take M = 72, the class mark with the highest frequency. The calculations can be arranged as shown in Table 2.31, on the next page.

$$\overline{x} = 72 + 5 \times \frac{3}{60} = 72 + \frac{1}{4} = 72.25.$$

Table 2.31: Calculations for Example 2.17

Mark	Class mark (x)	Frequency (f)	$u = \frac{x-72}{5}$	fu
60 – 64	62	2	–2	– 4
65 – 69	67	15	–1	–15
70 – 74	72	25	0	0
75 – 79	77	14	1	14
80 – 84	82	4	2	8
		60		$\sum fx = 22 - 19 = 3$

Remarks

In calculating the mean of a grouped frequency distribution, we assume that all values within a class interval are coincident with the class mark of that class interval. The fact that this is not usually the case, means that the mean calculated from a grouped data is likely to differ from the mean of the original (ungrouped) data. As already pointed out on page 29, this error, brought about by grouping, can be minimized by choosing class intervals such that class marks coincide with actually observed data.

2.3.2 The median

The median of a set of data is defined as the ***middle value when the data is arranged in order of magnitude***. If there are no ties, half of the observations will be smaller than the median, and half of the observations will be larger than the median. For a set of N observations $x_1, x_2, \dots, x_N$ arranged in order of magnitude, there are two cases:

$[\mathbf{S_6}]$ If N is odd, then the median is given by

$$\text{median} = \text{the } \tfrac{\mathbf{1}}{\mathbf{2}}(N + \mathbf{1})^{\text{th}} \text{ ordered observation.}$$

$[\mathbf{S_7}]$ If N is even, then the median is given by

$$\text{median} = \tfrac{1}{2}\Big\{ \text{the } \left(\tfrac{1}{2}N\right)^{\text{th}} \text{ ordered observation} + \text{the } \left(\tfrac{1}{2}N + 1\right)^{\text{th}} \text{ ordered observation}\Big\}.$$

Example 2.18

Find the median of each of the following sets of numbers.

(a) 12, 15, 22, 17, 20, 26, 22, 26, 12 (b) 4, 7, 9, 10, 5, 1, 3, 4, 12, 10

Solution

(a) Arranging the data in an increasing order of magnitude, we obtain

12, 12, 15, 17, 20, 22, 22, 26, 26.

Here, N (= 9) is odd, and so (see $[\mathbf{S_6}]$),

median = the $\frac{1}{2}(9+1)^{th}$ ordered observation = the 5^{th} ordered observation = 20.

Notice that ***if a number is repeated, we still count it the number of times it appears when we calculate the median***.

(b) Arranging the data in an increasing order of magnitude, we obtain

1, 3, 4, 4, 5, 7, 9, 10, 10, 12.

Here, N(=10) is an even number and so (see $[\mathbf{S_7}]$)

$$\text{median} = \frac{1}{2}\{\text{the } 5^{th} \text{ ordered observation + the } 6^{th} \text{ ordered observation}\} = \frac{1}{2}(5+7)=6.$$

Notice that, in each case, the median divides the distribution into two equal parts, with 50% of the observations greater than it and the other 50% less than it.

Example 2.19

The following are the ages (in years) of 30 children at a birthday party. Find the median age of the 30 children:

4, 3, 5, 8, 4, 6, 7, 8, 6, 4, 5, 6, 7, 5, 7,
6, 6, 5, 4, 4, 4, 3, 5, 6, 8, 7, 3, 6, 5, 8.

Solution

In order to find the median of the data, we first prepare a frequency table for the data. This method is recommended when we have a large number of observations. Table 2.32 gives a frequency table of the data.

Table 2.32: **Ages, in years, of children at a birthday party**

Age (x)	Tally	Frequency (f)
3	///	3
4	~~////~~ /	6
5	~~////~~ /	6
6	~~////~~ //	7
7	////	4
8	////	4

The total number of observations is 30, an even number, so the median is given by (see $[\mathbf{S}_7]$ on page 54)

$$\text{median} = \tfrac{1}{2}\text{(the } 15^{th} \text{ ordered observation + the } 16^{th} \text{ ordered observation).}$$

Now, the sum of the first 3 frequencies is 15, while the sum of the first four frequencies is 22. Hence, the 15^{th} and 16^{th} ordered observations are 5 and 6, respectively. Therefore,

$$\text{median} = \tfrac{1}{2}(5+6) = 5.5.$$

The median age is 5.5 years.

Example 2.20

The monthly salaries of five employees of a certain firm are given as: \$252.00, \$396.00, \$328.00, \$924.00, \$375.00.

Find (a) the mean monthly salary, (b) the median monthly salary. Which of these two measures is more typical of the salaries of the five employees? Give reasons.

Solution

(a) The mean monthly salary is $\$\left\{\frac{1}{5}(252+396+328+924+375)\right\}$ = \$455.00.

(b) We first arrange the salaries in order of magnitude. This gives:

\$252.00, \$328.00, \$375.00, \$396.00, \$924.00.

Since there is an odd number of observations, the middle value, \$375.00, is the median monthly salary.

In this example, the mean salary gives a false picture since it is greater than the salaries of 4 of the 5 employees. The median salary is, however, "close" to the salaries of most of the employees. It is therefore more representative of the data than the mean salary. (Notice that the median salary is not affected by the ***extreme value***, \$924.00, while the mean salary is affected by it.)

The median of a grouped frequency distribution

The exact value of the median of a grouped data cannot be obtained because the actual values of a grouped data are not known. For a grouped frequency distribution, the median is in the class interval which contains the $(\frac{1}{2}N)^{th}$ ordered observation, where N is the total number of observations. This class interval is called the ***median class***. The median of a grouped frequency distribution can be estimated by either of the following two methods:

(i) Linear interpolation method for estimating the median

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

$$[\mathbf{S_8}] \qquad \text{Median} = L + \left(\frac{\frac{1}{2}N - F}{f_m}\right)c,$$

where N = total number of observations,
L = lower class boundary of the median class,
F = sum of all frequencies below L,
f_m = frequency of the median class,
c = class width of the median class.

An application of this formula is given in Example 2.21.

(ii) Estimation of the median from a cumulative frequency curve

The median of a grouped frequency distribution can be estimated from a cumulative frequency curve. A horizontal line is drawn from the point $\frac{1}{2}N$ on the vertical axis to meet the cumulative frequency curve. From the point of intersection, a vertical line is dropped to the horizontal axis. The value on the horizontal axis is equal to the median, as shown is Fig. 2.10. An application of this method is given in Example 2.21.

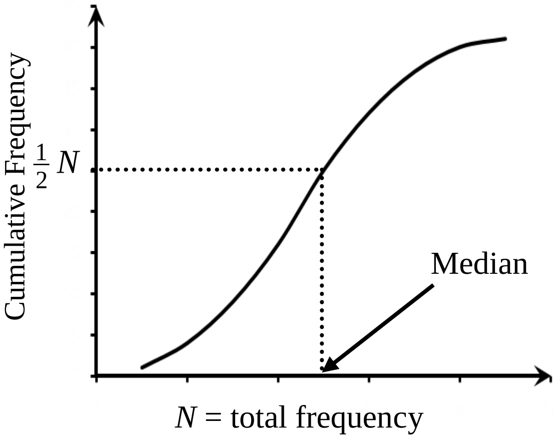


Fig. 2.10: ***Estimation of the median***

It should be noted that in determining the median of a grouped frequency distribution by these two methods, we assume that the original data, ungrouped, are evenly spread in the median class. The fact that this is not usually the case, means that the value obtained is likely to differ from that obtained by using the original data.

Example 2.21

Table 2.33, on the next page, gives the distribution of the heights of 60 students in a Senior High school. Find the median height of the students and explain the significance of the result.

Table 2.33: **Heights of students**

Height (cm)	145 – 149	150 – 154	155 – 159	160 – 164	165 – 169	170 – 174
Number of students	3	9	16	18	10	4

Solution

We give two methods for solving the problem.

First method

Here, we estimate the median by linear interpolation. We first determine the median class. Now, $N = \sum f = 60.$ Therefore the median is the $\left(\frac{1}{2}\times 60 = 30\right)^{th}$ ordered observation. The sum of the first three class frequencies is 28 while the sum of the first four class frequencies is 46. The median class is therefore the fourth class interval with class boundaries 159.5 and 164.5. Thus, (see $\left[\mathbf{S_8}\right]$) $L = 159.5,\ c = 164.5 - 159.5 = 5,\ \ f_m = 18$, and $F = 28$. The median height is therefore equal to

$$159.5\ +\ \left(\frac{\frac{1}{2}\times 60 - 28}{18}\right)\times 5\ \text{cm}\ =\ \left(159.5\ +\ \frac{2}{18}\ \times\ 5\right)\text{cm}\ =\ 160.1\,\text{cm}.$$

This means that 50% of the students are less than 160.1 cm tall and the other 50% are more than 160.1 cm tall.

Second method

Here, we estimate the median from a cumulative frequency curve. We first prepare the cumulative frequency table for the data in Table 2.33. This is given in Table 2.34.

Table 2.34: **Cumulative frequency table of the data in Table 2.33**

Height (cm) less than	Cumulative frequency
144.5	0
149.5	3
154.5	12
159.5	28
164.5	46
169.5	56
174.5	60

Fig. 2.11 shows the cumulative frequency curve for the data. To estimate the median height, we draw a horizontal line from the point $\left(\frac{1}{2}\times 60 = 30\right)$ on the vertical axis to meet the cumulative frequency curve. From the point of intersection, a vertical line is dropped to the horizontal axis, meeting it at the point 160 cm. The median height of the students is therefore 160 cm.

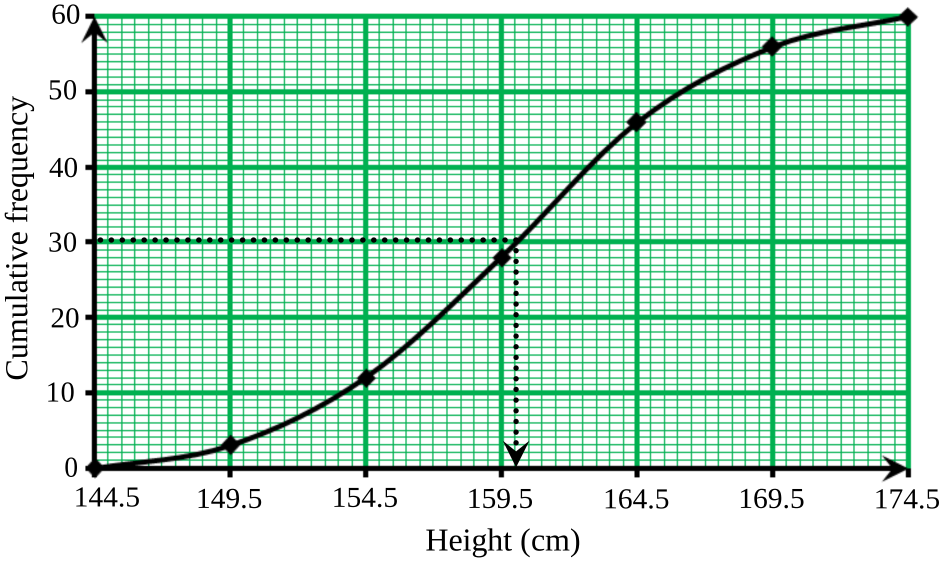


Fig. 2.11: ***Cumulative frequency curve of the data in Table 2.33***

2.3.3 The mode

The mode of a set of data is the value which occurs with the greatest frequency. ***The mode is therefore the most common value***.

Example 2.22

(a) The mode of 1, 2, 2, 2, 3 is 2.

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

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

It can be seen that the mode of a distribution may not exist, and even if it exists, it may not be unique.

Example 2.23

20 patients selected at random had their blood groups determined. The results are given in Table 2.35.

Table 2.35: **Blood groups of 20 patients**

Blood group	*A*	*B*	*AB*	*O*
Number of patients	2	4	6	8

The blood group with the highest frequency is *O*. The mode of the data is therefore blood group *O*. We can say that most of the patients selected have blood group *O*.

Notice that the mean and the median cannot be applied to the data in Example 2.23. This is because the variable “blood group” cannot take numerical values. However, it can be seen from Examples 2.22 and 2.23, that ***the mode can be used to describe both quantitative and qualitative data.***

The mode of a grouped frequency distribution

For a grouped frequency distribution, the class interval with the highest frequency is called the ***modal class***.

Fig. 2.12 shows a histogram for a grouped frequency distribution. The modal class is the class interval which corresponds to rectangle *ABCD*. An estimate of the mode of the distribution is the abscissa of the point of intersection of the line segments $\overline{AE}$ and $\overline{BF}$ in Fig. 2.12.

The following example illustrates how to estimate the mode of a distribution from a histogram.

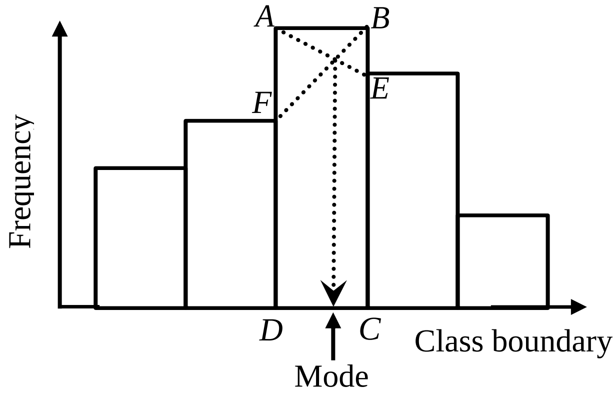


Fig. 2.12: ***A histogram, showing how to estimate the mode***

Example 2.24

Table 2.36 gives the distribution of the marks scored by 20 students in a Mathematics quiz.

Table 2.36: **Marks scored by students**

Mark	1 – 5	6 – 10	11 – 15	16 – 20	21 – 25
Frequency	3	4	7	2	4

Construct a histogram for the data and use it to estimate the mode of the data.

Solution

Fig. 2.13 shows a histogram for the data. To estimate the mode of the data from Fig. 2.13, we determine the abscissa of the point of intersection of the line segments $\overline{AC}$ and $\overline{BD}$. This gives the estimated modal mark as 12.5.

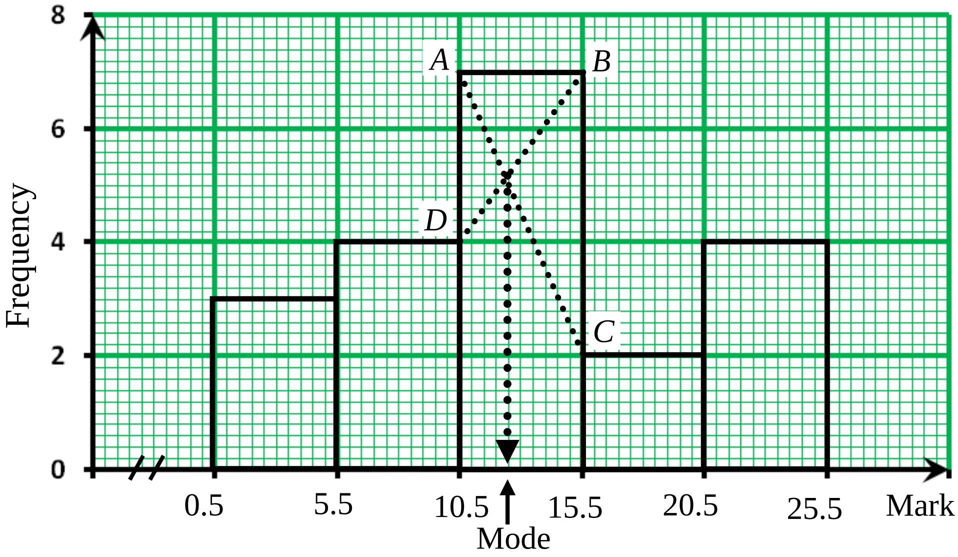


Fig. 2.13: *Histogram of the data in Table 2.36*

2.3.4 The midrange

Another measure of central tendency that is commonly used to report daily atmospheric temperature, is the ***midrange***. The midrange is the average of the smallest and largest observations. Thus,

$$\text{midrange} = \tfrac{1}{2}\left(X_{\text{smallest}} + X_{\text{largest}}\right).$$

Despite its simplicity, the midrange must be used cautiously. Because it involves only the smallest and largest observations in a data set, it becomes distorted as a measure of central tendency if an outlier or extreme value is present.

2.3.5 Relative merits of the mean, the median and the mode

We have looked at three different measures of central tendency and we now consider them in the light of the various information they give about sets of data. By examining the advantages and limitations of each of the three measures, we may know what information they give.

The mean

(i) The mean is unique for any set of quantitative data. That is, there is one and only one mean for a given set of quantitative data.

(ii) The formula for calculating the mean, uses numerical values for the observations. So the mean is appropriate only for quantitative data. It is not sensible to compute the mean of observations on a nominal scale. For instance, categorical variable such as religion, measured with categories such as (Protestant, Catholic, Jewish, Other), the mean religion does not make sense, even though these levels may sometimes be coded by numbers for convenience. Similarly, we cannot find the mean of observations on an ordinal rating, such as excellent, good, fair, and poor, unless we assign numbers such as 4, 3, 2, 1 to the ordered levels, treating it as quantitative.

(iii) Its main characteristic and virtue is that in its calculation, every value in the data is used. To this extent, the mean may be regarded as more representative than the other two.

(iv) Since it is the result of arithmetic processes, it can be used for further calculation. For example, knowing the mean and the total frequency of a set of data, their product gives the sum of all the observations in the data.

(v) Its main defect is that it is affected by an observation that falls well above or well below in the bulk of the data, called an ***outlier*** (see Example 2.20 on page 57).

The median

(i) It is unique; that is, like the mean, there is one and only one median for a given set of data.

(ii) The median cannot be found for nominal data.

(iii) The median, like the mean, is appropriate for quantitative data. Since it requires only ordered observations to compute it, it is also valid for ordinal-scale data.

(iv) Because of its definition, the median is especially useful in describing data that naturally fall into rank order, such as grades, and salaries.

(v) It is preferred to the mean as a measure of central tendency if the distribution is skewed.

(vi) Its main defect is that, in its calculation, every value of the data is not used.

The mode

(i) The mode is not unique. That is, there can be more than one mode for a given set of data. Distributions with a single mode are referred to as ***unimodal***. Distributions with two modes are referred to as ***bimodal***. Distributions may have several modes, in which case they are referred to as ***multimodal***.

(ii) The mode of a set of data may not exist (see Example 2.22 on page 60).

(iii) It is not affected by outliers.

(iv) It is mostly used by manufacturers since it gives a better idea of what particular size of a product to manufacture in excess of the others. For instance, a shoemaker is more interested in the modal size of the shoes he manufactures than the mean or the median size.

Exercise 2(c)

1. Find the mean of the following data.

x	0	1	2	3	4	5
f	2	3	5	8	7	3

2. Using a suitable assumed mean, find the mean of the following data.

x	152	156	160	164	168	172	176	180
f	3	6	10	20	30	20	8	3

3. The following table gives the distribution of the lengths of 40 iron rods. Using a suitable assumed mean, find the mean length of the iron rods.

Length (cm)	190	195	200	205	210	215	220
Frequency	3	5	7	10	6	7	2

4. The following table gives the distribution of the marks scored by 40 students in a Physics examination. Calculate, using the assumed mean method, the mean mark.

Mark (%)	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
Frequency	1	3	10	12	7	4	3

5. The following table gives the distribution of the ages of 40 patients who attended a clinic on a certain day. Calculate, using the assumed mean method, the mean age of the patients.

Age (years)	21 – 25	26 – 30	31 – 35	36 – 40	41 – 45	46 – 50
Frequency	2	5	10	12	8	3

6. The following table gives the distribution of the lengths of a sample of leaves from a tree. Find the mean length of the leaves.

Length (cm)	4 – 5	6 – 7	8 – 9	10 – 11	12 – 13	14 – 15
Frequency	2	6	14	31	30	7

7. The following table gives the distribution of the number of eggs laid by a chicken each

day in 15 days. Find the median number of eggs.

Number of eggs (x)	0	1	2	3	4
Number of days (f)	1	2	4	5	3

8. The following are the marks obtained be 40 students in a quiz that contains 10 questions. Find the median mark. (Hint: See Example 2.19 on page 56).

   4, 3, 5, 6, 8, 7, 3, 6, 5, 8, 4, 4, 7, 6, 6, 5, 6, 7, 4, 6,
   3, 4, 6, 4, 8, 7, 8, 3, 4, 5, 5, 6, 3, 4, 8, 3, 4, 8, 6, 7.

9. The following table gives the ages of 40 patients who attended a certain clinic on a given day. Calculate the median age.

Age (years)	10 –14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44
Frequency	1	2	6	14	13	3	1

10. The following are the ages (in years) of 5 patients selected from Peace Hospital: 30, 35, 33, 29, 83.
    (a) Calculate the mean and the median ages of the patients.
    (b) Which of the two measures is more representative of the data? Give your reasons.

11. The following table gives the distribution of the masses of 60 eggs.
    (a) Calculate the median mass.
    (b) Estimate the median mass from a cumulative frequency curve.

Mass (g)	41 – 46	47 – 49	50 – 52	53 – 55	56 – 61
Number of eggs	12	14	16	10	8

12. The following table gives the distribution of the ages of 120 nurses of a certain hospital.
    (a) What is the upper class boundary of the modal class?
    (b) Estimate the modal age from a histogram.

Age (years)	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49
Frequency	4	14	32	38	24	8

13. Find the mean or median, whichever you consider more suitable in the following data. Monthly salaries of five nurses:
    GH¢480.00, GH¢220.00, GH¢200.00, GH¢208.00, GH¢224.00.

14. With reference to the data in Question 6: (a) find the modal class, (b) estimate the modal mark from a histogram.

15. With reference to the data in Question 6: (a) find the modal class, (b) estimate the modal length of the leaves from a histogram.

16. With reference to the data in Question 11: (a) construct a histogram to represent the data, (b) find the modal class, (c) estimate the modal mass of the eggs.

17. Give examples to show when:
    (a) the mode would be a better average than the mean,
    (b) the mean and the median would all be equally satisfactory,
    (c) the median would be a better average to use than the mean.

2.4 Quartiles and Percentiles

2.4.1. Quartiles

The median divides a set of data into two equal parts. We can also divide a set of data into more than two parts. When an ordered set of data is divided into four equal parts, the division points are called ***quartiles***.

The ***first*** or ***lower quartile***, Q_1, is a value that has one fourth, or 25% of the observations below its value.

The ***second quartile***, Q_2, has one-half, or 50% of the observations below its value. The second quartile is equal to the median.

The ***third*** or ***upper quartile***, Q_3, is a value that has three-fourths, or 75% of the observations below it.

Example 2.25

Find the quartiles of the following data: 11, 14, 2, 6, 5, 18, 9, 6, 11, 18, 15, 10.

Solution

Arranging the data in ascending order of magnitude, we obtain

2, 5, 6, 6, 9, 10, 11, 11, 14, 15, 18, 18.

We first find Q_2, the median of the data. The total frequency is 12, an even number.

It follows that (see $[\mathbf{S}_7]$, on page 55).

$$Q_2 = \text{median} = \frac{1}{2}(\text{ the } 6^{th} \text{ ordered observation } + \text{the } 7^{th} \text{ ordered observation })$$

$$= \frac{1}{2}(10+11) = 10.5.$$

Notice that Q_2 divides the data into two equal parts with six observations less than it and six observations greater than it. The first quartile, Q_1, is the median of the 6 observations less than Q_2. It follows that

$$Q_1 = \tfrac{1}{2}(6+6) = 6.$$

The third quartile, Q_3, is the median of the six observations greater than Q_2. Hence,

$$Q_3 = \tfrac{1}{2}(14+15) = 14.5.$$

Quartiles from a grouped frequency distribution

We give two methods for estimating quartiles of a grouped frequency distribution. Quartiles of a grouped frequency distribution can be estimated by linear interpolation. Assuming that the data are evenly distributed in the class interval in which Q_k lies, we obtain, by linear interpolation,

$$[\mathbf{S_9}] \quad Q_k = L + \left\{\frac{\tfrac{k}{4}N - F}{f_{Q_k}}\right\}c, \qquad (k = 1, 2, 3),$$

where $N = \sum f$

L = lower class boundary of the class interval in which Q_k lies,

c = size of the class interval in which Q_k lies,

f_{Q_k} = frequency of the class interval in which Q_k lies,

and F = sum of all frequencies below L.

Example 2.26

Table 2.37 shows the distribution of the lengths of 100 iron rods. Find the lower and the upper quartiles of the distribution.

Table 2.37: Lengths of iron rods

Length (cm)	40 – 44	45 – 49	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74
Frequency	6	12	22	30	15	10	5

Solution

Here, $N = \sum f = 100.$ The lower quartile corresponds to the $\left(\tfrac{1}{4}\times 100 = 25\right)^{\text{th}}$ ordered observation. The sum of the first two frequencies is 18 while the sum of the first three frequencies is 40. Q_1 therefore, lies in the third class interval with class boundaries 49.5 and 54.5. Therefore (see $[\mathbf{S_9}]$), $L = 49.5$, F = 18, $f_{Q_1} = 22$, c = 54.5 – 49.5 = 5.0 and

$$Q_1 = L + \left\{\frac{\frac{1}{4}N - F}{f_{Q_1}}\right\}c = \left\{49.5 + \left(\frac{25-18}{22}\right)\times 5\right\}\text{cm}$$
$$= (49.5 + 1.59)\text{ cm} = 51.09\text{ cm}.$$

The upper quartile corresponds to the $\left(\frac{3}{4}\times 100 = 75\right)^{th}$ ordered observation. The sum of the first four frequencies is 70 while the sum of the first five frequencies is 85. Q_3 therefore lies in the fifth class interval with class boundaries 59.5 and 64.5. Hence, using $[\mathbf{S_9}]$, we obtain L = 59.5, $F = 70$, $f_{Q_3} = 15$, c = 64.5 – 59.5 = 5.0 and

$$Q_3 = L + \left\{\frac{\frac{3}{4}N - F}{f_{Q_3}}\right\}c = \left\{59.5 + \left(\frac{75-70}{15}\right)\times 5\right\}\text{ cm} = 61.17\text{ cm}.$$

2.4.2 The Box-and-Whisker plot

Another graphical display of a set of data is the ***box-and-whisker plot*** (or simply, the ***box plot***). The following are the steps for drawing a box-and-whisker plot.

(1) Represent the variable of interest on a horizontal line (or sometimes on a vertical line).
(2) Draw a box in the space above the horizontal axis in such a way that the left end of the box aligns with the first quartile (Q_1) and the right end of the box aligns with the third quartile (Q_3).
(3) Divide the box into two parts by a vertical line that aligns with the median, Q_2.
(4) Draw a horizontal line, called the ***lower whisker***, from the left end of the box to a point that aligns with the smallest measurement in the data set.
(5) Draw another horizontal line from the right end of the box to a point that aligns with the largest measurement in the data set. This line is called the ***upper whisker***.

It can be seen that the whiskers of a box plot extend to the maximum and minimum observations, except for ***outliers***, which are marked separately. Fig. 2.14, on the next page, shows a box plot representing data, whose minimum value is 2, lower quartile is 4, median is 5, upper quartile is 7, and the maximum value is 10.

Notice that the upper whisker and the upper half of the central box are larger than the lower ones. Therefore, Fig. 2.14 shows that the right tail of the distribution, is longer than the left tail. The plot also reflects the skewness of the data to the right.

It can be seen that, a box plot of a set of data, gives a visual summary of ***five key numbers that are associated with the data***. These are: the minimum value, the lower quartile, the median, the upper quartile and the maximum value.

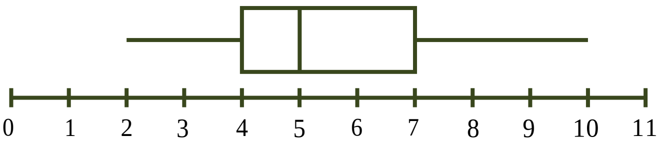


Fig. 2.14: ***A box-and-whisker plot***

Side-by-side box plots are useful for comparing two distributions (see Example 2.33 on page 75).

Example 2.27

Figure 2.15 shows the box plot of a set of data. Write down:

(a) the quartiles of the data,

(b) the maximum value of the data.

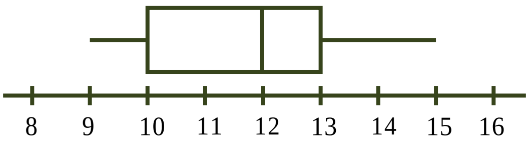


Fig. 2.15: ***A box plot***

Solution

(a) $Q_1 = 10$, $Q_2 = 12$, $Q_3 = 13$. (b) the maximum value of the data is 15.

Example 2.28

Draw a box plot to represent the data in Exmaple 2.25.

Solution

From Example 2.25, $Q_1 = 6$, $Q_2 = 10.5$, $Q_3 = 14.5$, the minmum value of the data is 2, and the maximum value of the deta is 18. Fig. 2.16 shows the required box plot.

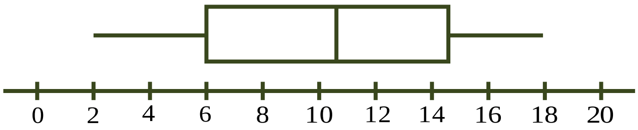


Fig. 2.16: ***Box plot of the data in Example 2.25***

Exploratory data analysis (EDA)

Box-and-whisker plots and stem-and-leaf plots are examples of what are known as ***exploratory data analysis techniques***. These techniques allow the investigator to examine data in ways that reveal trends and relationships, identify unique features of data sets, and facilitate their description and summarization. Books by Tukey (1977) and Du Toit et al. (1986) provide an overview of most of the well known methods of analyzing and portraying data graphically with emphasis on exploratory techniques.

2.4.3 Central tendency using quartiles

Quartiles can be used to define additional measures of central tendency and dispersion. They have the advantage of not being influenced by outliers. This makes them useful for analyzing changes over time. For example, in stock portfolios.

The midhinge: A measure of central tendency based on the quartiles

Quartiles are useful in the development of a measure of location that is called the ***midhinge***. The midhinge is the mean of the first and third quartiles in a set of data. Thus,

$$\text{midhinge} = \tfrac{1}{2}\left(Q_1 + Q_2\right).$$

The midhinge is composed only of the first and third quartiles, so it is not affected by extreme values in the data as the mean. The hinges were introduced by Tukey (1977).

2.4.4 Percentiles

When an ordered set of data is divided into 100 equal parts, the division points are called ***percentiles***. More generally, the (100*k*)th percentile P_k, is a value such that 100*k*% of the observations are below this value and 100(1 – *k*)% of the observations are above the value.

It can be seen that $P_{0.25}$, the 25th percentile, has 25% of the observations below it, $P_{0.50}$, the 50th percentile, has 50% of the observations below it and $P_{0.75}$, the 75th percentile, has 75% of the observations below it. Thus the quartiles are the 25th, 50th, and 75th percentiles.

Calculating the (100*p*)th percentile

The following rule simplifies the calculation of percentiles.

(1) Order the *n* observations from smallest to largest.

(2) Determine the product *np*.

(a) If *np* is not an integer, round it up to the next integer and find the corresponding ordered value.

(b) If *np* is an integer, say *k*, calculate the mean of the *k*th and (*k* + 1)th ordered observations.

The following example illustrates an application of the above rule.

Example 2.29

Twenty observations on the time to failure, in hours, of electrical insulation materials (adapted from Nelson's ***Applied Life Data Analysis***, 1982) are given below (in order). Obtain the quartiles and the 84th percentile.

204	228	252	300	324	444	620	720	816	912
1 176	1 296	1 392	1 488	1 512	2 520	2 856	3 192	3 528	3 710

Solution

The first quartile corresponds to $P_{0.25,}$ the 25th percentile. To find $P_{0.25,}$ we determine $0.25n = 0.25 \times 20 = 5$. This is an integer so we take the mean of the 5th and 6th ordered observations. Hence,

$$Q_1 = \tfrac{1}{2}(324+444) = 384.$$

The second quartile corresponds to $P_{0.50}$, the 50th percentile. Since $np = 20(0.50) = 10$ is an integer, we take the mean of the 10th and 11th ordered observations. Thus,

$$Q_2 = \tfrac{1}{2}(912+1\,176) = 1\,044.$$

Q_3 corresponds to $P_{0.75,}$ the 75th percentile. Since $0.75n = 0.75 \times 20 = 15$ is an integer, we find the mean of the 15th and 16th ordered observations. Thus,

$$Q_3 = \tfrac{1}{2}(1\,512 + 2\,520) = 2\,016.$$

To find the 84th percentile, we calculate $0.84n = 0.84 \times 20 = 16.8$, which we round up to 17. Hence,

$$P_{0.84} = \text{the 17th ordered observation} = 2\,856.$$

Exercise 2(d)

1. Find Q_1, Q_2 and Q_3 for each of the following sets of data.
   (a) 16, 20, 9, 15, 8, 21, 22 (b) 2, 6, 8, 3, 10, 5, 11, 13, 16, 19, 14
   (c) 21, 6, 2, 6, 4, 10, 12, 3, 1, 2, 5, 4.

2. Draw a box plot for each set of data in Question 1.

3. The following are the number of minutes that a person had to wait for a bus to work on 15 working days:

10	1	13	9	5	9	2	10	3	8	6	17	2	10	15

   (a) Find the quartiles. (b) Draw a box plot.

4. The following are determinations of a river's annual maximum flow in cubic metres per second:

405,	355,	419,	267,	370,	391,	612,	383,
434,	462,	288,	317,	540,	295,	508	

   (a) Construct a stem-and-leaf plot with two-digit leaves.
   (b) Use the stem-and-leaf plot to calculate the quartiles of the data.

5. The following are figures on an oil well's daily production in barrels;

214,	204,	226,	198,	243,	225,	207,	203,
209,	200,	217,	202,	208,	212,	205,	220.

(a) Construct a stem-and-leaf plot with stem labels 19, 20, …, 24.

(b) Use the stem-and-leaf plot to find the quartiles of the data.

6. The following table gives the distribution of the body masses of 180 cancer patients who attend King Fahd Hospital.

Mass (kg)	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74	75 – 79
Frequency	12	18	40	56	30	24

Calculate the first and third quartiles of the distribution.

7. The following table gives the distances travelled by 70 workers to their offices.

Distance (km)	0 – 4	5 – 9	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34
Frequency	4	6	14	26	10	8	2

Calculate the first and third quartiles of the distribution.

8. In Applied Life Data Analysis (Wiley, 1982), Wayne Nelson presents the breakdown time of an insulating fluid between electrodes at 34 kV. The times, in minutes, are as follows:

0.19, 0.75, 0.96, 1.31, 2.78, 3.16, 4.15,
4.67, 4.85, 6.50, 7.35, 8.10, 8.27. 12.06, 31.75.

(a) Find the lower and upper quartiles of breakdown time.

(b) Find the 30th and 85th percentiles of breakdown time.

9. The following are the masses of 30 eggs (to the nearest gramme).

47	72	46	68	57	62	62	58	69	51
50	64	52	49	67	47	71	72	57	61
53	53	44	62	53	53	61	68	58	48

(a) Construct a stem-and-leaf plot to represent the data.

(b) Find the modal mass of the data.

(c) Calculate the 25th, 50th and 75th percentiles of the data.

(d) Calculate the 95th and 64th percentiles.

10. Find the 10th and 88th percentiles of the life data in Example 2.29.

2.5 Measures of dispersion

Table 2.38: Data sets *A* and *B*

A	1	2	3	3	4	5
B	–1	0	3	3	5	8

In Section 2.3, we discussed how a set of data can be summarized by a single representative value which describes the central value of the data. Consider the two sets of data in Table 2.38.

The mean, the mode and the median of each set of data is equal to 3. Fig. 2.17 shows the dot diagrams of data sets *A* and *B*. It can be seen that, while values of data set *A* are grouped close to their mean, values of data set *B* are more spread out. We say that values of data set *B* are ***more dispersed*** (or ***scattered***) than those of data set *A*.

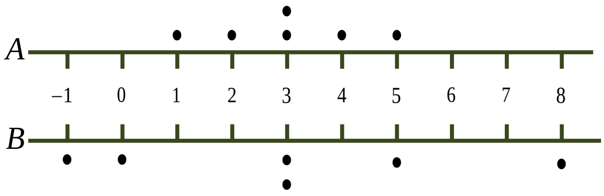


Fig. 2.17: Dot diagrams of data sets *A* and *B*

This example shows that the mean, the mode and the median, are not enough in describing a set of data. In addition to using these measures, we need a numerical measure of ***dispersion*** (or ***variation***) of a set of data. The most important measures of dispersion are the ***range***, the ***interquartile range*** and the ***standard deviation***. These measures are discussed in this section.

It should be noted that:

(1) ***If all values in a set of data are equal, then there is no dispersion***. For example, the data 5, 5, 5 has no dispersion.

(2) ***If values of a set of data are not equal, but very close to each other, then there is a small dispersion***.

2.5.1 The range

The range of a set of data is defined as the difference between the largest observation and the smallest observation in the set of data. Thus,

Range = largest observation – smallest observation.

Clearly, the larger the range, the greater the variability in the data. Thus, if the range of data set *A* is greater than that of data set *B*, then data set *A* is more dispersed than data set *B*.

Example 2.30

Consider again, the sets of data in Table 2.38. The range of data set *A* is $(5-1) = 4$, while the range of data set *B* is $8-(-1) = 9.$ The range of data set *B* is greater than that of data set *A*. This confirms that data set *B* is more dispersed than data set *A*.

Example 2.31

The marks obtained by 8 students in Mathematics and Physics examinations are as follows:

Mathematics: 35, 60, 70, 40, 85, 96, 55, 65.
Physics: 50, 55, 70, 65, 89, 68, 72, 80.

Find the ranges of the two sets of data. Are the Physics marks more dispersed than the Mathematics marks?

Solution

For Mathematics,

highest mark = 96, lowest mark $= 35$, range $= 96-35 = 61$.

For Physics,

highest mark = 89, lowest mark $= 50$, range $= 89-50 = 39$.

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

Example 2.32

Examine the following sets of data:

A: 3, 4, 5, 6, 8, 9, 10, 12, 15. *B*: 3, 8, 8, 9, 9, 9, 10, 10, 15.

(a) Which of the two sets of data is more dispersed?

(b) Calculate the range of each set of data. What can you say about the range of a set of data as a measure of dispersion of the set of data?

Solution

(a) Fig. 2.18 shows the dot diagrams of the two sets of data. It can be seen that values of data set *A* are more scattered than those of data set *B*. There is therefore more variation or dispersion in data set *A* than in data set *B*.

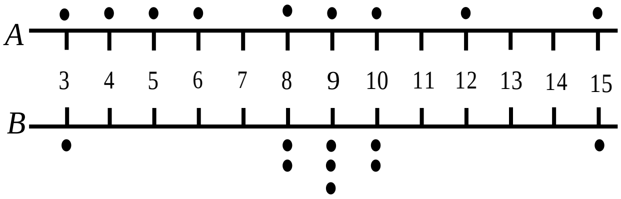


Fig. 2.18: *Dot diagrams of data sets A and B in Example 2.32*

(b) In both cases, the range is

largest value $-$ smallest value $= 15-3 = 12$.

In part (a), we found that data set *A* is more dispersed than data set *B*. Since the range indicates that the two sets of data have the same dispersion, the range is not a good measure of dispersion in this case.

Remarks

The range has the advantage that it is quick and easy to calculate. However, since it depends only on the maximum and the minimum values of a set of data, it does not show how the whole data is distributed between these two values. ***The range is therefore not a good measure of dispersion if one or both of these two values differ greatly from other values of the data***. To overcome this problem, we sometimes use the interquartile range which we now discuss.

2.5.2 Dispersion using quartiles

The interquartile range (IQR) or midspread

A robust measure of dispersion is the interquartile range. The interquartile range of a set of data is the difference between the upper and lower quartiles of the data. Thus,

Interquartile range $= Q_3 - Q_1$.

Notice that, since 25% of a set of data is less than or equal to Q_1 and 75% of the data is less than or equal to Q_3, the central 50% of a set of data lies within the interquartile range of the data. The interquartile range of a set of data is therefore not affected by values of the data outside this range. The interquartile range is sometimes used as a measure of dispersion.

Consider the two sets of data in Example 2.32. For data set *A*, $Q_1 = 5$, $Q_3 = 10$, and so the interquartile range of data set *A* is $(10 - 5) = 5$.

For data set *B*, $Q_1 = 8$, $Q_3 = 10$, and also the interquartile range for data set *B* is $(10 - 8) = 2$.

Since the interquartile range of data set *A* is greater than that of data set *B*, these results confirm that data set *A* is more dispersed than data set *B*. Recall that the range gave us a misleading result.

Example 2.33

The box plots in Fig. 2.19 give the results of a study of the ages, in years, of students from two schools, *A* and *B*.

(a) What is the median age of students from School *A*?

(b) Determine whether the ages of students from School *A* are more variable than those of students from School *B*.

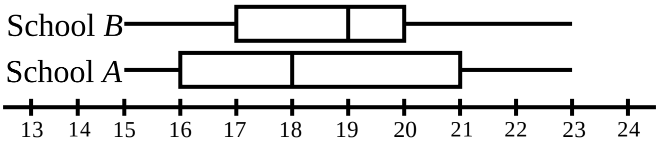


Fig. 2.19: Box plots of ages of students in Schools A and B

Solution

(a) The median age of students from School A is 18 years.

(b) We first calculate the interquartile range of the data from each school.

School *A*: Q_1 = 16 years, Q_3 = 21 years,
Interquartile range = 21 years – 16 years = 5 years.

School *B*: Q_1 = 17 years, Q_3 = 20 years,

Interquartile range = 20 years – 17 years = 3 years.

The interquartile range of the data from School *A* is greater than that of the data from School *B*, and so the ages of students from School *A* are more variable than those from School *A*.

Notice that the two sets of data have the same range but different dispersions. This example shows again, that the range is a poor measure of dispersion (see Example 2.32).

2.5.3 The variance and standard deviation

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

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2.5.1)$$

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

The reason for dividing by $n-1$ rather than *n*, as we might have expected, is that s^2 is based on $(n-1)$ ***degrees of freedom***. The term degrees of freedom results from the fact that the *n* deviations $x_1 - \bar{x}$, $x_2 - \bar{x}$, …, $x_n - \bar{x}$ always sum to zero, and so specifying the values of any $(n-1)$ of these automatically determines the value of the remaining one. Thus, only $(n-1)$ of the *n* deviations, $x_i - \bar{x}$, are freely determined. From a practical point of view, dividing the squared differences by $(n-1)$ rather than *n* is necessary in order to use the sample variance in the inference procedures discussed in Chapters 6 and 7. Students interested in pursuing the matter further should refer to the article by Walker (1040).

If s_A, the standard deviation of data set *A*, is greater than s_B, the standard deviation of data set *B*, then data set *A* is more dispersed than data set *B*. It should be noted that the standard deviation of a set of data is a non-negative number. It follows that

$$s_A > s_B \Leftrightarrow s_A^2 > s_B^2 .$$

Example 2.34

Calculate the variances and standard deviations of the sets of data, *A* and *B*, in Table 2.38 on page 72.

Solution

The calculations can be arranged as shown in Tables 2.39 and 2.40, on the next page.

Table 2.39: Data set *A*

x	$x-\bar{x}=x-3$	$(x-\bar{x})^2$
1	–2	4
2	–1	1
3	0	0
3	0	0
4	1	1
5	2	4
18	0	10

Table 2.40: Data set *B*

x	$x-\bar{x}=x-3$	$(x-\bar{x})^2$
–1	–4	16
0	–3	9
3	0	0
3	0	0
5	2	4
8	5	25
18	0	54

$$\bar{x}_A=\tfrac{1}{6}\sum_{i=1}^{6}x_i=\tfrac{18}{6}=3$$

$$s_A^2=\tfrac{1}{5}\sum_{i=1}^{6}(x_i-\bar{x})^2=\tfrac{10}{5}=2\,.$$

$$s_A=\sqrt{2}=1.41\,.$$

$$\bar{x}_B=\tfrac{1}{6}\sum_{i=1}^{6}x_i=\tfrac{18}{6}=3$$

$$s_B^2=\tfrac{1}{5}\sum_{i=1}^{6}(x_i-\bar{x})^2=\tfrac{54}{5}=10.8\,.$$

$$s_B=\sqrt{10.8}=3.29\,.$$

It can be seen that $s_B > s_A$, confirming that data set *B* is more dispersed than data set *A* (see the dot diagrams in Fig. 2.17 on page 72).

The unit of measurement of the sample variance is the square of the unit of measurement of the data. Thus, if *x* is measured in centimetres (cm), then the unit of measurement of the sample variance is cm^2. The standard deviation has the desirable property of measuring variability in the same unit as the data.

An alternative formula for computing the variance

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

$$s^2=\frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right\}. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.5.2)$$

The following two examples illustrate applications of Equation (2.5.2) in calculating s^2.

Example 2.35

Calculate the variances and the standard deviations of the sets of data, *A* and *B*, in Example 2.32, on page 73. Which of the two sets of data is more dispersed?

Solution

The calculations can be arranged as shown in Table 2.41.

$$s_A^2 = \tfrac{1}{8}\left[700 - \tfrac{1}{9}(72)^2\right] \;=\; 15.5, \text{ and so } s_A = 3.94$$

$$s_B^2 = \tfrac{1}{8}\left[805 - \tfrac{1}{9}(81)^2\right] \;=\; 9.5, \text{ and so } s_B = 3.08.$$

It can be seen that $s_A > s_B$, confirming that data set *A* is more dispersed than data set *B* (see Example 2.32 on page 73).

Table 2.41: **Calculations for Example 2.35**

	Data set *A*		Data set *B*	
i	x_i	x_i^2	x_i	x_i^2
1	3	9	3	9
2	4	16	8	64
3	5	25	8	64
4	6	36	9	81
5	8	64	9	81
6	9	81	9	81
7	10	100	10	100
8	12	144	10	100
9	15	225	15	225
	72	**700**	**81**	**805**

Example 2.36

The following are the haemoglobin levels (g/dl) of 10 patients selected from St. Paul Hospital.

$$\sum_{i=1}^{10} x_i = 160, \qquad \sum_{i=1}^{10} x_i^2 = 2\,596.$$

Find the mean and the standard deviation of the data.

Solution

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10} \times 160 = 16 \text{ g/dl.}$$

$$s^2 = \frac{1}{9}\left[\sum_{i=1}^{10} x_i^2 - \frac{1}{10}\left(\sum_{i=1}^{10} x_i\right)^2\right]$$

$$= \frac{1}{9}\left[2\ 596 - \frac{1}{10}(160)^2\right] = 4 \text{ (g/dl)}^2.$$

Thus, s = 2 g/dl.

The sample variance of an ungrouped frequency distribution

If, in a frequency distribution, x_i, …, x_k occur with frequencies f_1, …, f_k, respectively, then Equation (2.5.2) becomes

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{k} f_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} f_i x_i\right)^2\right\} \text{..(2.5.3)}$$

where $n = \sum_{i=1}^{k} f_i$.

Assumed mean method

The amount of computation involved in using Equation (2.5.3) can be reduced by using the assumed mean method we introduced in Section 2.3. If M is any assumed mean, and if $d_\text{i} = x_\text{i} - M$, $(i = 1, 2, ..., k)$, then Equation (2.5.3) becomes

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{k} f_i d_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} f_i d_i\right)^2\right\}, \text{...(2.5.4)}$$

where n is the sum of the frequencies.

The following two examples illustrate an application of Equation (2.5.4).

Example 2.37

Table 2.42, on the next page, gives the lengths (to the nearest mm) of 100 leaves from a given plant. Using a suitable assumed mean, calculate the mean length and the standard deviation.

Table 2.42: **Lengths of 100 leaves**

Length (mm)	57	58	59	60	61	62
Number of leaves	9	8	29	22	18	14

Solution

We take 59 (the number with the highest frequency) as the assumed mean. This gives Table 2.43.

$$\bar{x} = 59 + \frac{1}{100}\sum fd = 59 + \frac{74}{100} = 59.74 \text{ mm.}$$

$$s^2 = \tfrac{1}{99}\left\{\sum fd^2 - \tfrac{1}{100}\left(\sum fd\right)^2\right\} = \tfrac{1}{99}\left\{264 - \tfrac{1}{100}(74)^2\right\}$$

$$= 2.1135 \text{ mm}^2$$

So, $s = \sqrt{2.1135 \text{ mm}^2} = 1.45 \text{ mm.}$

The mean length of the leaves is 59.74 mm and the standard deviation is 1.45 mm.

Table 2.43: **Calculations for Example 2.37**

Length (x)	Frequency (f)	$d = x - 59$	fd	fd^2
57	9	–2	–18	36
58	8	–1	–8	8
59	29	0	0	0
60	22	1	22	22
61	18	2	36	72
62	14	3	42	126
	$\sum f = 100$		$\sum fd = 100 - 26$ $= 74$	$\sum fd^2 = 264$

Example 2.38

In order to choose between two measuring instruments, *A* and *B*, each instrument was used to measure the diameters of 18 coins. Instrument *A* gave the following measurements in centimetres:

3.53, 3.54, 3.55, 3.52, 3.54, 3.51, 3.54, 3.56, 3.53,
3.54, 3.53, 3.55, 3.52, 3.53, 3.55, 3.54, 3.55, 3.53.

(a) Using an assumed mean of 3.54 cm, calculate the mean and the standard deviation of these measurements.

(b) Instrument *B* gave the same mean as *A* but its standard deviation was 0.0107 cm.

Which of the two instruments is better? Give reasons for your answer.

Solution

(a) The solution is best arranged as in Table 2.44.

Table 2.44: Calculations for Example 2.38

x	$d = x - 3.54$	Frequency (f)	fd	fd^2
3.51	–0.03	1	–0.03	0.0009
3.52	–0.02	2	–0.04	0.0008
3.53	–0.01	5	–0.05	0.0005
3.54	0	5	0	0.0000
3.55	0.01	4	0.04	0.0004
3.56	0.02	1	0.02	0.0004
		$\sum f = 18$	$\sum fd = -0.06$	$\sum fd^2 = 0.0030$

From Table 2.44,

$$\bar{x} = 3.54 + \frac{1}{18}\sum fd = 3.54 - \frac{0.06}{18} = 3.537\,\text{cm}.$$

$$s_A^2 = \frac{1}{17}\left\{\sum fd^2 - \frac{1}{18}\left(\sum fd\right)^2\right\} = \frac{1}{17}\left\{0.003 - \frac{1}{18}(-0.06)^2\right\} = 0.000\,1647\,\text{cm}^2.$$

So, $s_A = 0.0128$ cm.

(b) Since the standard deviation of the measurements obtained by using instrument A is greater than that obtained by using insrument B, there will be a smaller variation in the measurements obtained by using instrument B than that obtained by using instrument A. Instrument B is better than instrument A since it will give more consistent measurements.

The sample variance and sample standard deviation of a grouped frequency distribution

Equation (2.5.3) is valid for grouped frequency distributions if we interpret x_i as the class mark of a class interval and f_i the corresponding class frequency. In the following example, we apply Equation (2.5.3) to find the standard deviation of a grouped frequency distribution.

Example 2.39

Table 2.45, on the next page, shows the ages (in years) of 50 cancer patients. Find the standard deviation of the distribution.

Table 2.45: Ages of cancer patients

Age (years)	35 – 39	40 – 44	45 – 49	50 – 54	55 – 59	60 – 64
Frequency	1	4	12	23	7	3

Solution

The work can be arranged as shown in Table 2.46. We take the assumed mean M = 52.

$$s^2 = \tfrac{1}{49}\left\{ \sum fd^2 - \tfrac{1}{50}\left(\sum fd\right)^2 \right\} = \tfrac{1}{49}\left\{ 1\,400 - \tfrac{1}{50}(-50)^2 \right\} = 27.551 \ \text{(years)}^2$$

So, s = 5.249 years.
The standard deviation of the distribution is 5.25 years.

Table 2.46: Calculations for Example 2.39

Age (years)	Class mark (x)	$d = x - 52$	f	fd	fd^2
35 – 39	37	–15	1	–15	225
40 – 44	42	–10	4	–40	400
45 – 49	47	–5	12	–60	300
50 – 54	52	0	23	0	0
55 – 59	57	5	7	35	175
60 – 64	62	10	3	30	300
			$\sum f = 50$	$\sum fd = 65 - 115$ $= -50$	$\sum fd^2 = 1\,400$

2.5.4 The coefficient of variation

The coefficient of variation of a set of data is defined by the equation

$$\text{CV} = \tfrac{s}{\bar{x}}(100)\% \quad \ldots\ldots(2.5.5)$$

It can be seen that, since $\bar{x}$ and s are expressed in the same unit of measurement, the unit of measurement cancels out in computing the coefficient of variation. Coefficient of variation is therefore a measure which is independent of the unit of measurement.

Applications

(a) Coefficient of variation is used to compare the variability of sets of data measured in different units. For example, we may wish to know, for a certain population, whether body masses, measured in kilograms, are more variable than heights, measured in centimetres.

(b) Coefficient of variation is also used to compare the variability of sets of data measured in the same unit, but whose means are quite different. For example, if we compare the standard deviation of heights of primary school students with the standard deviation of heights of university students, we may find that the latter standard deviation is numerically larger than the former, because the heights themselves are larger, not because the dispersion is greater.

Example 2.40

Measurements made with a micrometer of the diameters of ball bearings have a mean of 4.52 mm and a standard deviation of 0.0142 mm, whereas measurements made with a second micrometer of the lengths of screws have a mean of 1.64 inches and a standard deviation of 0.0075inch. Which of the two micrometers gives more precise measurements?

Solution

Since the two sets of data are measured in different units, we use coefficient of variation to compare variability. The coefficients of variation are:

$$\mathrm{CV}_{\text{ball bearing}} = \left(\frac{0.0142 \text{ mm}}{4.52 \text{ mm}}\right)\times 100\% = 0.314\%,$$

$$\mathrm{CV}_{\text{screw}} = \left(\frac{0.0075 \text{ in}}{1.64 \text{ in}}\right)\times 100\% = 0.457\%,$$

respectively. Since $\mathrm{CV}_{\text{ball bearing}} < \mathrm{CV}_{\text{screw}}$, the measurements made with the first micrometer exhibit relatively less variability than those made with the second micrometer. The first micrometer therefore gives more precise measurements than the second micrometer.

Example 2.41

The following table gives the results of a survey to study the body masses of primary and high school students in a certain country.

	Mean mass (kg)	Standard deviation (kg)
Primary School	24	8
High School	45	12

We wish to know which is more variable, the body masses of the High School students or the body masses of the Primary School students.

Solution

Since the means of the two sets of data are very different, we use coefficient of variation to compare variability. The coefficients of variation are.

$$\text{CV}_{\text{Primary school}} = \left(\frac{8\text{ kg}}{24\text{ kg}}\right)\times 100\% = 33.3\%,$$

$$\text{CV}_{\text{High school}} = \left(\frac{12\text{ kg}}{45\text{ kg}}\right)\times 100\% = 26.6\%,$$

respectively. It can be seen that the body masses of the Primary School students have a greater relative variation than those of the High School students.

Exercise 2(e)

1. Find the range and the interquartile range of each of the following sets of data.
   (a) 8, 6, 11, 21, 14, 9, 3
   (b) 28, 61, 26, 44, 39, 27, 45, 24, 32, 47
   (c) 33, 24, 29, 27, 21, 12, 16, 23, 9, 18, 14

2. Find: (a) the range, (b) the interquartile range, of each of the sets of data in Exercise 2(d), Question 1 (see page 70).

3. Calculate the standard deviation of each of the following sets of data.
   (a) 5, 7, 11, 2, 6, 3, 15 (b) 1, 4, 6, 8, 9, 11
   (c) $3x, 8x, 9x, 11x, 14x$ (d) 4, 5, 7, 9, 10, 14

4. Calculate the standard deviation of the following set of data.

x	5	6	7	8	9
f	43	39	14	3	1

5. Calculate the standard deviation of each of the following sets of data, using a suitable assumed mean.

(a)

x	37	42	47	52	57	62
f	1	4	12	23	7	3

(b)

x	152	157	162	167	172	177	182
f	4	14	26	32	21	10	3

(c)

x	53	55	57	61
f	3	4	8	5

6. The masses, to the nearest kilogram, of 100 students, were as given below. Calculate the standard deviation of the data, using a suitable assumed mean.

Mass (kg)	55	56	57	58	59	60	61	62	63	64
Frequency	2	1	6	8	22	29	18	10	2	2

7. The following table gives the ages of 50 blood donors. Calculate the standard deviation of the data, using a suitable assumed mean.

Age (years)	35 – 39	40 – 44	45 – 49	50 – 54	55 – 59	60 – 64
Frequency	1	4	12	23	7	3

8. The following table gives the distribution of the lengths of 40 leaves from a certain plant. Find the mean and the standard deviation, using a suitable assumed mean.

Length (mm)	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39
Number of leaves	1	5	16	14	3	1

9. The body masses of students selected at random from two schools, *P* and *Q*, were measured. The results are given in the following table.

	Mean (kg)	Standard deviation (kg)
School *P*	15 kg	3 kg
School *Q*	64 kg	8 kg

Determine whether the body masses of students from School *P* are more variable than those of students from School *Q*.

10 (a) The following are the body masses, measured to the nearest kilogram, of 5 students from St. Andrew School: 35, 40, 45, 45, 50.

Find the sample mean and the sample standard deviation.

(b) The following are the body masses, measured to the nearest kg, of 10 students from St. Paul College: $\sum_{i=1}^{10} x_i = 450$, $\sum_{i=1}^{10} x_i^2 = 20\,439$.

Determine whether the body masses of this group are more variable than those of the 5 students from St. Andrew School.

11. The following table gives the results of a survey to study the haemoglobin and blood glucose levels of patients selected from King James Hospital.

	Mean	Standard Deviation
Haemoglobin level (g/dl)	10	6
Blood glucose level (mmol/l)	5	4

Determine whether haemoglobin levels of the patients are more variable than their blood glucose levels.

12. The following are the days seven (7) patients stayed in a hospital:

    5, 5, 7, 10, 5, 20, 102.

    Find the mean and the median of the data. Which of the two measures is more representative of the data? Give reasons for your answer.

13. The following table gives the shoe sizes of 20 women selected from Sapele, a town in Nigeria. Find the mean and the modal shoe sizes. Which of these measures is more useful to the manufacturer of shoes?

    36 38 39 37 38 36 40 39 37 38
    39 38 40 39 39 40 38 42 38 38

14. The following are the blood glucose levels of six (6) patients in mg/dl:

    78, 80, 540, 77, 80, 78.

    (a) Find the mean and the median blood glucose levels.
    (b) Which of the two measures is more representative of the data? Give reasons for your answer.

2.6 Shapes of distributions

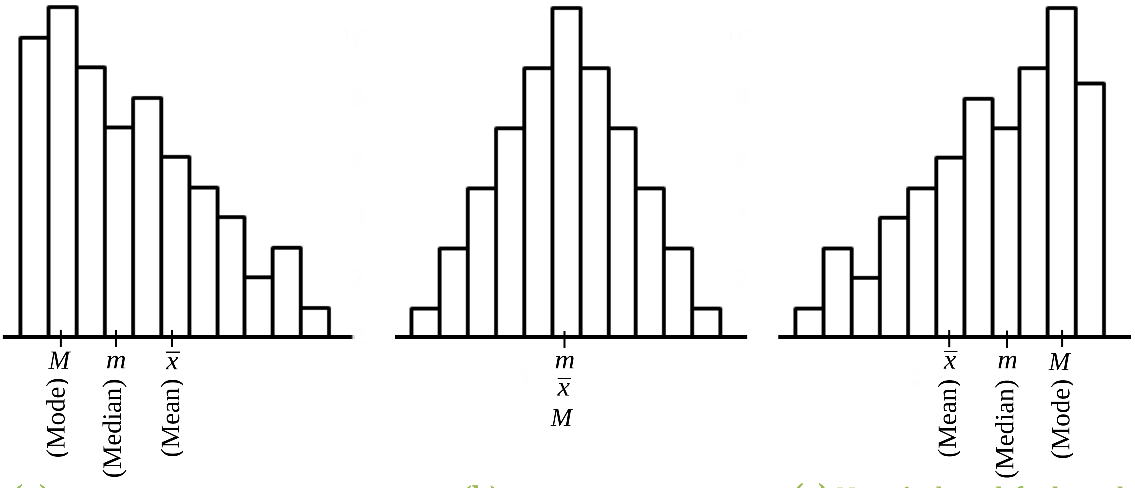
Another important property of a set of data is the shape of its distribution. We can evaluate the shape of a distribution by considering two characteristics of data sets: ***symmetry*** and ***kurtosis***. Both symmetry and kurtosis, evaluate the manner in which the data are distributed around their mean.

Symmetry

A symmetrical distribution can be defined as one in which the upper half is a mirror image of the lower half of the distribution. If a vertical line is drawn through the mean of a distribution depicted by a histogram, the lower half could be "folded over" and would coincide with the upper half of the distribution.

One way to identify symmetry involves a comparison of the mean and the median. If these two measures are equal, we may generally consider the distribution to be symmetrical. If the mean is greater than the median, the distribution may be described as ***positively or right-skewed*** (that is, has a long tail to the right). If the mean is less than the median, the distribution is considered to be ***negatively or left-skewed*** (that is, has a long tail to the left).

It can be seen that, for skewed distributions, the mean lies towards the direction of skew (the longer tail) relative to the median. Fig. 2.20 illustrates the relationship between the mean, the mode and the median in these three types of distributions. In summary: mean > median ⇒ right-skewed, mean = median ⇒ symmetrical, mean < median ⇒ left-skewed.



(a) Positively or right-skewed (b) Symmetrical (c) Negatively or left-skewed

Fig. 2.20: Comparison of the mean $\overline{x}$, the median (m), and the mode (M), when a histogram is skewed to the right (a), symmetrical (b), skewed to the left (c)

In a general way, skewness may be judged by looking at the sample histogram, or by comparing the mean and the median. However, this comparison is imprecise and does not take the sample size into account. When more precision is needed, we use the sample coefficient of skewness given by Microsoft Excel (see Redmond, 1999).

$$\text{Skewness} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3 .$$

This unit-free statistic can be used to compare two samples measured in different units (say, dollars and cedis) or to compare one sample with known reference distribution such as the normal (bell-shaped) distribution. The skewness coefficient is obtained from Excel's Tools > Data Analysis > Descriptive Statistics or by the function = **SKEW(Data)**.

2.7 Determining symmetry by using the five key numbers of a statistic

The symmetry of a set of data can be determined by using the five keys numbers of the data. A set of data is symmetrical, if the following are true:

(i) Q_2 is equidistant from Q_1 and Q_3.

(ii) The distance from the minimum value to Q_1 is equal to the distance from the maximum value to Q_3.

Fig. 2.21 shows how the shape of a distribution affects the box-and-whisker plot of the distribution. In Fig. 2.21 (a), it can be seen that;

(i) Q_2 is equidistant from Q_1 and Q_3,

(ii) the distance from the minimum value to Q_1 is equal to the distance from the maximum value to Q_3.

The distribution is therefore symmetrical.

In Fig. 2.21(b), a right-skewed distribution, we note that there is a long tail on the right side of the distribution. The distance between Q_1 and the median is less than the distance between the median and Q_3, while the distance between Q_1 and the minimum value is less than the distance between Q_3 and the largest value.

In Fig. 2.21(c), a left-skewed distribution, we note that there is a long tail on the left side of the distribution. The distance between Q_1 and the median is greater than the distance between the median and Q_3, while the distance between the smallest value and Q_1 is greater than the distance between Q_3 and the largest value.

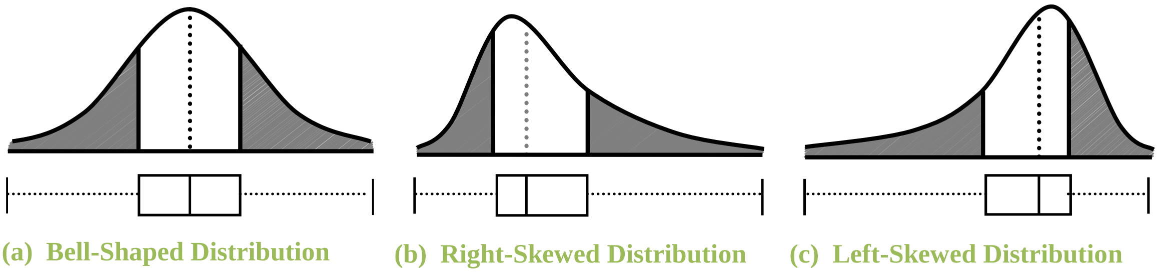

(a) Bell-Shaped Distribution (b) Right-Skewed Distribution (c) Left-Skewed Distribution

Fig. 2.21: ***Box-and-whisker plots for three hypothetical distributions***

Kurtosis

Kurtosis concerns the relative concentration of values in the centre of the distribution as compared to the tails. In terms of this property, we can define three types of distributions: ***leptokurtic, mesokurtic***, and ***platykurtic***. A leptokurtic distribution is characterized by a prominent peak. The prefix **lepto** means thin and refers to the taller, thinner peak of the distribution (see Fig. 2.22, on the next page). A mesokurtic distribution is one in which the values are predominantly located in the centre of the distribution, with relatively few values falling in the tails. The normal distribution (see Section 4.8) is an example of a mesokurtic distribution. A platykurtic distribution is one in which the values are relatively spread out through the range of the distribution, so that the peak is relatively flat and very few values appear in the tails. The prefix ***platy*** means flat and refers to the relatively flattened peak of the distribution (see Fig. 2.22, on the next page).

A histogram is an unreliable guide to kurtosis because its scale and axis proportions may vary, so a numerical statistic is needed. Microsoft Excel uses the statistic

$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

This sample kurtosis coefficient is obtained from Excel's function **= KURT(Data)**.

Kurtosis is not the same thing as dispersion, although the two are sometimes confused. For a study of measures of kurtosis, see Ramsey and Ramsey (1990).

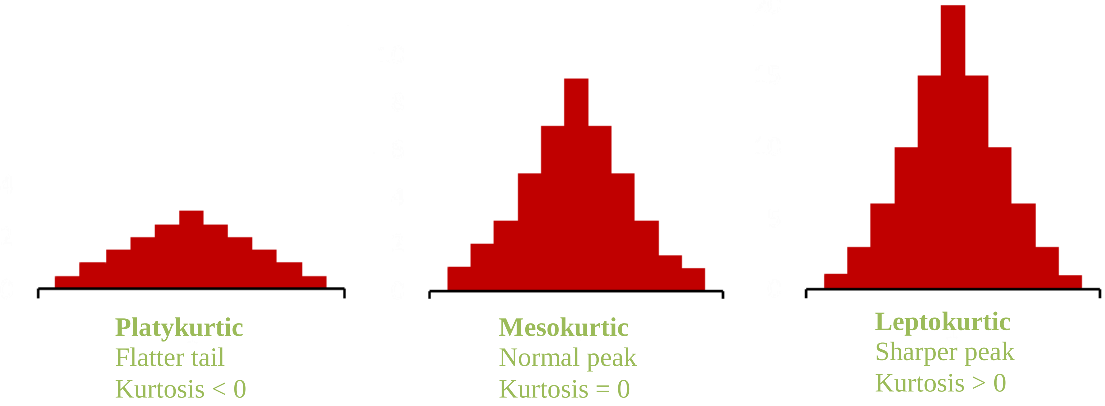


Fig. 2.22

Notation for parameters

> μ (Greek mu) and σ (Greek lowercase sigma) denote the mean and standard deviation of a variable for the population.

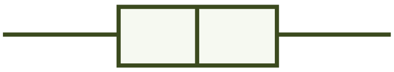
We call μ and σ the ***population mean*** and ***population standard deviation***, respectively. The population mean is ***the average of all the observations for the entire population***. The population standard deviation describes the variability of those observations about the population mean.

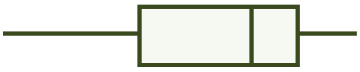
Whereas the statistics $\bar{X}$ and S are variables, with values $\bar{x}$ and s depending on the sample chosen, the parameters μ and σ are constants. This is because μ and σ refer to just one particular group of observations, namely, the observations in the entire population. The parameter values are usually unknown, which is the reason for sampling and calculating sample statistics to estimate their values. Much of the rest of this text deals with ways of making inferences about unknown parameters (such as μ) using sample statistics (such as $\bar{X}$). Before studying these inferential methods, though, you need to learn some basic ideas of ***probability***, which serves as the foundation for methods. Probability is the subject of the next chapter.

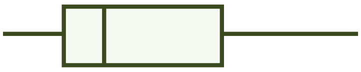
Exercise 2(f)

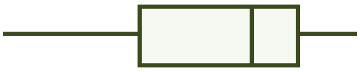
1. For each of the following sets of data, find the mean, the mode and the median. State whether the data are skewed. If so, how?
   (a) 40, 50, 50, 60, 60, 50, 50, 70, 100, 50, 50, 40, 90, 80, 60.
   (b) 40, 60, 50, 60, 70, 60, 80, 80, 70, 70, 80, 70, 90, 70, 100.
   (c) 90, 90, 40, 50, 90, 90, 80, 80, 80, 60, 90, 100, 90, 100, 70.

2. State whether each of the following box-and-whisker plots, represents a symmetric, a right-skewed or a left-skewed distribution.

(a) 

(b) 

(c) 

(d) 

3. An examination was conducted for 100 students and the results were given in the table below.

Marks scored	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74
No. of students	5	15	20	28	12
Marks scored	75 – 79	80 – 84	85 – 89	90 – 94	95 – 99
No. of students	9	5	3	2	1

   (a) Draw a histogram to represent the data.
   (b) From your histogram, estimate, correct to one decimal place, the modal score.
   (c) Using an assumed mean of 67, calculate, correct to one decimal place,
   (i) the mean, (ii) the standard deviation.

4. (a) The following are haemoglobin level (g/dl) of 5 children who are receiving treatment from St. Thomas Hospital:

8, 11, 14, 12, 10.

   Find the mean and the standard deviation of the data.

   (b) The following are the haemoglobin levels (g/dl) of a sample of 10 healthy children:

$$\sum_{i=1}^{10} x_i = 160, \qquad \sum_{i=1}^{10} x_i^2 = 2\,596.$$

   Determine whether haemoglobin levels of this group are more variable than those of the children who are receiving treatment from St. Thomas Hospital.

5. The following table shows the distribution of the marks scored by 100 students in a mathematics examination.

Marks scored	Under 20	20 – 29	30 – 39	40 – 49
Number of students	5	13	30	22
Marks scored	50 – 59	60 – 69	70 – 79	80 – 89
Number of students	16	8	5	1

Draw a suitable diagram and use it to find (a) the median, (b) the upper quartile, (c) the pass mark if 60% of the students pass the examination.

6. The table below shows the distribution of the marks of 60 students in a test.

Marks	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
Frequency	2	5	20	23	7	3

Calculate, correct to two decimal places, (a) the mean and (b) the standard deviation.

7. Forty small-scale industries in a certain country are classified according to their size (i.e. the number of people employed). The table below shows the classification.

Number of employees	1– 20	21– 40	41– 60	61– 80	81 – 100	101 – 120
Number of industries	3	8	13	10	5	1

(a) State
(i) the modal class of the distribution, (ii) the median class of the distribution.
(b) Draw a cumulative frequency curve and use it to estimate:
(i) the median of the distribution,
(ii) the number of industries with more than 45 but less than 85 employees.

8. The data below are the masses (in kg) at birth of 32 babies born at a maternity home in a week.

3.2	3.0	3.8	4.2	2.8	3.3	3.3	2.7
3.6	2.8	3.5	4.6	4.3	3.7	4.4	3.1
3.9	2.5	3.7	4.3	3.7	4.0	2.8	3.6
3.1	3.0	3.4	2.8	3.7	3.3	3.0	3.2

(a) Construct a frequency table using the class intervals 2.25 – 2.75, 2.75 – 3.25, 3.25 – 3.75, 3.75 – 4.25 and 4.25 – 4.75.
(b) Draw a histogram to illustrate the data, and use it to estimate:
(i) the median of the distribution, (ii) the mode of the distribution.
(c) Calculate the mean of the distribution.

9. The table below gives the distribution of marks obtained by 500 candidates in an examination.

Mark	1 – 10	11 – 20	21 – 30	31 – 40	41 – 50
Frequency	5	15	40	105	150
Mark	51 – 60	61 – 70	71 – 80	81 – 90	91 – 100
Frequency	120	50	10	5	0

Draw a cumulative frequency curve for the distribution

(a) Use your graph to estimate:

(i) the median mark, (ii) the interquartile range.

(b) If only 5 % of the candidates attained the distinction level, estimate the lowest mark for this level.

10. The table below gives the distribution of the marks of 200 pupils in a certain test.

Mark	11 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 – 70	71 – 80
Frequency	18	36	34	44	42	16	10

Using an assumed mean of 45.5 marks, calculate, correct to two decimal places

(a) the mean mark, (b) the standard deviation of the distribution.

11. The table below shows the distribution of marks (x) obtained in a Chemistry examination.

Mark (x)	45 – 49	50 – 54	55 – 59	60 – 64	65 – 69
Number of candidates	0	2	21	36	56
Mark (x)	70 – 74	75 – 79	80 – 84	85 – 89	
Number of candidates	70	42	31	28	

(a) Draw:

(i) a histogram, (ii) a cumulative frequency curve for the distribution.

(b) Use your diagrams for the distribution to estimate:

(i) the mode, (ii) the median.

12. The marks obtained in a test by 40 pupils are as follows:

78	60	76	66	33	81	67	84	72	60
54	42	27	33	24	66	27	63	18	30
39	44	30	30	33	45	39	33	30	45
27	36	42	18	42	36	60	72	72	63

(a) Construct a frequency table, using class intervals 10–19, 20–29, 30–39, etc.

(b) Draw a histogram for the data.

(c) (i) Use your histogram to estimate the mode.

(ii) Calculate the mean of the distribution.

2.8 Chapter Summary

This chapter introduced ***descriptive statistics*** – ways of ***describing*** data to summarize key characteristics of data.

2.8.1 Overview of tables and graphs

- A ***frequency distribution*** summarizes the counts for possible values or intervals of values. A ***relative frequency*** distribution reports this information using percentages or proportions.
- A ***bar graph*** uses bar over possible values to portray a frequency distribution for a categorical variable. For a quantitative variable, a similar graphic is called a ***histogram***. It shows whether the distribution is approximately bell shaped, U shaped, skewed to the right (longer tail pointing to the right), or whatever.
- The ***stem-and-leaf plot*** is an alternative portrayal of data for a quantitative variable. It groups together observations having the same leading digit (stem), and shows also their final digit (leaf). For small samples, it displays the individual observations.
- The ***box plot*** portrays the quartiles, the extreme values, and outliers. A box plot and a stem-and-leaf plot can also provide back-to-back comparisons of two groups of data.

Stem-and-leaf plots and box plots, simple as they are, are relatively recent innovations in statistics. They were introduced by the great statistician John Tukey (see Tukey, 1977). See Tufte (2001) for other innovative ways to present data graphically.

2.8.2 Overview of measures of central tendency

Measures of central tendency describe the centre of the data, in terms of a typical observation.

- The ***mean*** is the sum of the observations divided by the sample size. It is the centre of gravity of the data.
- The ***median*** divides the ordered data set into two parts of equal numbers of observations. Half below and half above that point.
- The lower quarter of the observations fall below the ***lower quartile***, and the upper quarter fall above the ***upper quartile***. These are the 25th and 75th percentiles. The median is the 50th percentile. The quartiles and median split the data into four equal parts. They are less affected than the mean by outliers or extreme skew.
- The ***mode*** is the most commonly occurring value. It is valid for any type of data, though usually used with categorical data or discrete variables taking relatively few values.

2.8.3 Overview of measures of variability

- The ***range*** is the difference between the largest and smallest observations. The ***interquartile range*** is the range of the middle half of the data between the upper and lower quartiles. It is less affected by outliers.
- The ***variance*** averages the squared deviations about the mean. Its square root, the ***standard deviation***, is easier to interpret, describing a typical distance from the mean.
- The ***empirical rule*** states that for a bell-shaped distribution, about 68% of the observations fall within one standard deviation of the mean, about 95% fall within two standard deviations, and nearly all, if not all, fall within three standard deviations.

Table 2.47 summarizes the measure of central tendency and variability. A ***statistic*** summarizes a sample. A ***parameter*** summarizes a population. ***Statistical inference*** uses statistics to make predictions about parameters.

Table 2.47: Summary of measures of central tendency and variability

Measure	Definition	Interpretation
Central tendency		
Mean	$\bar{x} = \sum_{i=1}^{n} x_i \Big/ n$	Centre of gravity
Median	Middle observation of ordered sample	50th percentile, splits sample into two equal parts
Mode	Most frequently occurring value	Most likely outcome, valid for all types of data
Variability		
Standard deviation	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$	Empirical Rule: If bell shaped, 65%, 95% within s, 2s of $\bar{x}$
Range	Difference between largest and smallest observation	Greater with more variability.
Interquartile range	Difference between upper quartile (75th percentile) and lower quartile (25th percentile)	Encompasses middle half of data.

References

Du Toit, S. H. C., Steyn, A. G. W. and Stumpf, R. H. (1986). Graphical Exploratory Data Analysis. *Springer-Verlag, New York*.

Nelson, W. (1982). Applied Life Data Analysis. *John Wiley & Sons*, *New York*.

Ramsey, P. P. and Ramsey, P. H. (1990). Simple tests of normality in small samples. *Journal of Quality Technology*, **22**, 299 – 309.

Redmond, W. A. (1999). Microsoft Excel 2000, Microsoft Press.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, **21**, 65 – 66.

Tufte, E. R. (2001). The Visual Display of Quantitative Information, 2nd ed. Grapic Press.

Turkey, J. W. (1977). Exploratory Data Analysis, *Addison-Wesley*, *Reading*, *Mass*.

Walker, H. W. (1040). Degrees of freedom. *Journal of Educational Psychology*, **31**, 253 – 269.