Module 6

Junior Secondary Mathematics

Data Handling

THE COMMONWEALTH of LEARNING

Science, Technology and Mathematics Modules

for Upper Primary and Junior Secondary School Teachers of Science, Technology and Mathematics by Distance in the Southern African Development Community (SADC)

Developed by
The Southern African Development Community (SADC)

Ministries of Education in:

- **Botswana**
- **Malawi**
- **Mozambique**
- **Namibia**
- **South Africa**
- **Tanzania**
- **Zambia**
- **Zimbabwe**

In partnership with The Commonwealth of Learning

Published jointly by The Commonwealth of Learning and the SADC Ministries of Education.

Residents of the eight countries listed above may obtain modules from their respective Ministries of Education. The Commonwealth of Learning will consider requests for modules from residents of other countries.

ISBN 1-895369-65-7

SCIENCE, TECHNOLOGY AND MATHEMATICS MODULES

This module is one of a series prepared under the auspices of the participating Southern African Development Community (SADC) and The Commonwealth of Learning as part of the Training of Upper Primary and Junior Secondary Science, Technology and Mathematics Teachers in Africa by Distance. These modules enable teachers to enhance their professional skills through distance and open learning. Many individuals and groups have been involved in writing and producing these modules. We trust that they will benefit not only the teachers who use them, but also, ultimately, their students and the communities and nations in which they live.

The twenty-eight Science, Technology and Mathematics modules are as follows:

Upper Primary Science
Module 1: *My Built Environment*
Module 2: *Materials in my Environment*
Module 3: *My Health*
Module 4: *My Natural Environment*

Junior Secondary Science
Module 1: *Energy and Energy Transfer*
Module 2: *Energy Use in Electronic Communication*
Module 3: *Living Organisms' Environment and Resources*
Module 4: *Scientific Processes*

Upper Primary Technology
Module 1: *Teaching Technology in the Primary School*
Module 2: *Making Things Move*
Module 3: *Structures*
Module 4: *Materials*
Module 5: *Processing*

Junior Secondary Technology
Module 1: *Introduction to Teaching Technology*
Module 2: *Systems and Controls*
Module 3: *Tools and Materials*
Module 4: *Structures*

Upper Primary Mathematics
Module 1: *Number and Numeration*
Module 2: *Fractions*
Module 3: *Measures*
Module 4: *Social Arithmetic*
Module 5: *Geometry*

Junior Secondary Mathematics
Module 1: *Number Systems*
Module 2: *Number Operations*
Module 3: *Shapes and Sizes*
Module 4: *Algebraic Processes*
Module 5: *Solving Equations*
Module 6: *Data Handling*

A MESSAGE FROM THE COMMONWEALTH OF LEARNING

The Commonwealth of Learning is grateful for the generous contribution of the participating Ministries of Education. The Permanent Secretaries for Education played an important role in facilitating the implementation of the 1998-2000 project work plan by releasing officers to take part in workshops and meetings and by funding some aspects of in-country and regional workshops. The Commonwealth of Learning is also grateful for the support that it received from the British Council (Botswana and Zambia offices), the Open University (UK), Northern College (Scotland), CfBT Education Services (UK), the Commonwealth Secretariat (London), the South Africa College for Teacher Education (South Africa), the Netherlands Government (Zimbabwe office), the British Department for International Development (DFID) (Zimbabwe office) and Grant MacEwan College (Canada).

The Commonwealth of Learning would like to acknowledge the excellent technical advice and management of the project provided by the strategic contact persons, the broad curriculum team leaders, the writing team leaders, the workshop development team leaders and the regional monitoring team members. The materials development would not have been possible without the commitment and dedication of all the course writers, the in-country reviewers and the secretaries who provided the support services for the in-country and regional workshops.

Finally, The Commonwealth of Learning is grateful for the instructional design and review carried out by teams and individual consultants as follows:

- Grant MacEwan College (Alberta, Canada):
  General Education Courses
- Open Learning Agency (British Columbia, Canada):
  Science, Technology and Mathematics
- Technology for Allcc. (Durban, South Africa):
  Upper Primary Technology
- Hands-on Management Services (British Columbia, Canada):
  Junior Secondary Technology

Dato' Professor Gajaraj Dhanarajan
President and Chief Executive Officer

ACKNOWLEDGEMENTS

The Mathematics Modules for Upper Primary and Junior Secondary Teachers in the Southern Africa Development Community (SADC) were written and reviewed by teams from the participating SADC Ministries of Education with the assistance of The Commonwealth of Learning.

CONTACTS FOR THE PROGRAMME

The Commonwealth of Learning
1285 West Broadway, Suite 600
Vancouver, BC V6H 3X8
Canada

Ministry of Education
Private Bag 005
Gaborone
Botswana

Ministry of Education
Private Bag 328
Capital City
Lilongwe 3
Malawi

Ministério da Eduação
Avenida 24 de Julho No 167, 8
Caixa Postal 34
Maputo
Mozambique

Ministry of Basic Education,
Sports and Culture
Private Bag 13186
Windhoek
Namibia

National Ministry of Education
Private Bag X603
Pretoria 0001
South Africa

Ministry of Education and Culture
P.O. Box 9121
Dar es Salaam
Tanzania

Ministry of Education
P.O. Box 50093
Lusaka
Zambia

Ministry of Education, Sport and
Culture
P.O. Box CY 121
Causeway
Harare
Zimbabwe

COURSE WRITERS FOR JUNIOR SECONDARY MATHEMATICS

Ms. Sesutho Koketso Kesianye: *Writing Team Leader*
Head of Mathematics Department
Tonota College of Education
Botswana

Mr. Jan Durwaarder: Lecturer (Mathematics)
Tonota College of Education
Botswana

Mr. Kutengwa Thomas Sichinga: Teacher (Mathematics)
Moshupa Secondary School
Botswana

FACILITATORS/RESOURCE PERSONS

Mr. Bosele Radipotsane: Principal Education Officer (Mathematics)
Ministry of Education
Botswana

Ms. Felicity M Leburu-Sianga: Chief Education Officer
Ministry of Education
Botswana

PROJECT MANAGEMENT & DESIGN

Ms. Kgomotso Motlotle: Education Specialist, Teacher Training
The Commonwealth of Learning (COL)
Vancouver, BC, Canada

Mr. David Rogers: *Post-production Editor*
Open Learning Agency
Victoria, BC, Canada

Ms. Sandy Reber: *Graphics & desktop publishing*
Reber Creative
Victoria, BC, Canada

TEACHING JUNIOR SECONDARY MATHEMATICS

Introduction

Welcome to *Data Handling*, Module 6 of Teaching Junior Secondary Mathematics! This series of six modules is designed to help you to strengthen your knowledge of mathematics topics and to acquire more instructional strategies for teaching mathematics in the classroom.

The guiding principles of these modules are to help make the connection between theoretical maths and the use of the maths; to apply instructional theory to practice in the classroom situation; and to support you, as you in turn help your students to apply mathematics theory to practical classroom work.

Programme Goals

This programme is designed to help you to:

- strengthen your understanding of mathematics topics
- expand the range of instructional strategies that you can use in the mathematics classroom

Programme Objectives

By the time you have completed this programme, you should be able to:

- develop and present lessons on the nature of the mathematics process, with an emphasis on where each type of mathematics is used outside of the classroom
- guide students as they work in teams on practical projects in mathematics, and help them to work effectively as a member of a group
- use questioning and explanation strategies to help students learn new concepts and to support students in their problem solving activities
- guide students in the use of investigative strategies on particular projects, and thus to show them how mathematical tools are used
- guide students as they prepare their portfolios about their project activities

How to work on this programme

Congratulations on reaching Module 6!

Data Handling, or Descriptive Statistics, is a relatively new concept in Maths teaching, basically a response to the need for Maths literacy in an information-rich society. As we have said in earlier modules, do the Exercises and Assignments yourself, transfer some concepts into your own classroom, and interact with your colleagues about this newer aspect of school Maths.

ICONS

Throughout each module, you will find the following icons or symbols that alert you to a change in activity within the module.

Read the following explanations to discover what each icon prompts you to do.

	**Introduction**	Rationale or overview for this part of the course.
	**Learning Objectives**	What you should be able to do after completing this module or unit.
	**Text or Reading Material**	Course content for you to study.
	**Important—Take Note!**	Something to study carefully.
	**Self-Marking Exercise**	An exercise to demonstrate your own grasp of the content.
	**Individual Activity**	An exercise or project for you to try by yourself and demonstrate your own grasp of the content.
	**Classroom Activity**	An exercise or project for you to do with or assign to your students.
	**Reflection**	A question or project for yourself—for deeper understanding of this concept, or of your use of it when teaching.
	**Summary**	
	**Unit or Module Assignment**	Exercise to assess your understanding of all the unit or module topics.
	**Suggested Answers to Activities**	
	**Time**	Suggested hours to allow for completing a unit or any learning task.
	**Glossary**	Definitions of terms used in this module.

CONTENTS

Module 6: Data Handling

Module 6
Data Handling

Introduction to the module

Over the past decade terms such as ‘data handling’, ‘exploratory data analysis’ and ‘data visualisation’ have replaced the use of the term ‘statistics’ in the secondary curricula. The current meaning of data handling emphasises collecting, organising, representing, analysing and interpreting data closely connected to the world of the pupils. The visual representation of data is of major importance at the secondary level. Data handling is seen as going beyond statistics (the science of data collection, representation, analysis and interpretation for decision making). Statistics can be said to be the content, data handling is the whole learning environment in which data are explored.

Aim of the module

The module aims at providing you with:

(a) ideas and materials to support the learning of pupils in data handling

(b) materials to extend and consolidate your knowledge on the topic of data handling

(c) knowledge on common misconceptions related to the topic of data handling

(d) ideas for pupil centred approach to the coverage of the concepts

(e) questions to make you reflect on your present practice as a teacher when covering the topic data handling and how to relate your present practice to the ideas presented in this module.

Structure of the module

This model first looks into the reasons to include data handling in the secondary curriculum and at misconceptions as to what data handling is. Unit 1 also looks at the basic concept of ‘data’, the various types of data and its distinction from information. Unit 2 covers the three basic methods for data collection, i.e., (a) surveys (b) experiments and (c) simulation. Attention is paid to questionnaires and interviews as means to collect data in surveys. Unit 3 covers the multiple ways in which data can be represented graphically. The major emphasis is on the question of which representation is most appropriate in the given context and less emphasis is on the techniques for the producing the various graphical representations. Unit 4 covers the measures for central tendency. The emphasis is again on ‘what is the most appropriate measure to use in the given context’ rather than on the techniques of computation. Unit 5 covers the measures for spread or dispersion and uses box plots, introduced in Unit 4, as a visual representation of spread.

Objectives of the module

When you have completed this module you should be able to create with confidence, due to enhanced background knowledge, a learning environment for your pupils in which they can:

(i) apply various techniques to collect, represent and analyse data

(ii) demonstrate a critical approach to data represented in the media

(iii) justify choices for representing and analysing collected data in a specific format

Prerequisite module

No special modules need to be covered prior to this module. The knowledge on data handling you have and experience as a teacher is all that is needed.

Unit 1: What is data handling?

Introduction to Unit 1

In this unit you will learn about the basic concept involved in data handling, i.e., What is data? You will also reflect on why this topic is included in the junior secondary school syllabus. Pupils generally do not have problems in drawing charts, calculating means and other techniques but fail to give meaning to these activities. The misconceptions as to what data handling is are discussed. Data handling serves a purpose: making of decisions. This unit looks at how pupils (people) make decisions.

Purpose of Unit 1

The main aim of this unit is to look at the basic questions: What is data handling? Why is it included in the junior secondary syllabus? What is data? How can types of data be distinguished? What erroneous ideas do exist with respect to data handling? What is the most appropriate method to use in the classroom to facilitate the learning of data handling?

Objectives

When you have complete this unit you should be able to:

- state reasons for including data handling in the curriculum
- defend the use of the term data handling versus the use of the word statistics
- state the purpose of data handling
- discuss the way people make decisions
- differentiate between data and information
- differentiate between qualitative and quantitative data
- differentiate between discrete and continuous quantitative data
- differentiate between nominal and ordinal qualitative data
- differentiate between descriptive and inferential statistics
- state reasons for the need of data collection
- sets activities to pupils to acquire knowledge on type of data
- set activities to make pupils aware of the base on which they make decisions
- list common misconceptions related to the question “what is statistics?”
- state the four stages of a statistical investigation

Time

To study this unit will take you about five hours.

Unit 1: What is data handling?

Section A: Data handling or statistics?

The term data handling has replaced the more traditional term statistics in the primary and secondary school curricula. The shift from statistics to data handling is more than a mere play on words. It signals a different methodology to be used in the classroom. One can identify two different approaches to the teaching / learning of data handling

(1) Statistical ideas and techniques are to be covered and the task of the teacher is to find suitable examples and activities to illustrate these.

(2) There are in the world around us a lot of questions and situations we like to understand, to describe, to explore and to access, and a range of statistical techniques will be an appropriate tool to do this.

Write down your view on why data handling should be taught at the junior secondary school level.

What method do you presently use to cover data handling?

Do you consider data handling to be part of mathematics or do you feel it should stand on its own?

Compare what you wrote with the information below and see how it agrees or disagrees with what you wrote.

It is the second view which is implicitly or explicitly expressed in the syllabus for the Junior Secondary School level in many countries. It is the view that data handling is **most effectively learned by pupils in the context of projects** which explore issues seen as relevant by the pupils rather than a set of skills and processes which have familiar illustrations. Projects such as "Smoking and Health", "Beer cans and pollution", "Teenage pregnancies" all start with a situation or an issue, and the need for relevant data handling techniques should develop when the need arises. In this way pupils are motivated to learn certain data handling concepts because they need them to analyse, describe and represent their collected data to answer the question they set at the start of their project.

This is also expressed in the Cockcroft Report (Cockcroft 1982) that states:

Statistics is essentially a practical subject and its study should be based on the collection of data, wherever possible by pupils themselves. It should consider the kinds of data which is appropriate to collect, the reasons for collecting the data and the problems in doing so, the ways in which the data may legitimately be manipulated and the kind of interference which may be drawn. (p 234)

The challenge is to produce a range of realistic activities and small projects that will allow the development of all the data handling work considered to be appropriate and useful at lower secondary level. If for a certain technique or concept no suitable activity can be found, it most likely does not deserve a place in the syllabus. Pupils need data handling tools to (i) understand and

critically look at data presented in the media and (ii) present and analyse data collected by themselves to answer questions related to their own world.

There is an ongoing debate whether or not data handling / statistics should be included in mathematics or whether it should stand on its own. The nature of data handling is rather different from pure mathematics. Mathematics deals with developing a logical system based on axioms through theorems. It leads to 'true' knowledge within the system. Statistics is concerned with making sense of data, representing and summarising the data and making decisions based on this data. However the outcomes are 'probable' not 'sure' as in the case with mathematical knowledge. In order to make this distinction clear to pupils, some educators plead for a separation of mathematics and data handling as two rather distinct disciplines. Others feel that at secondary school level both mathematics and data handling are attempts to describe, explore and understand patterns in numbers and shapes in the world around us and hence both should be under the umbrella of mathematics.

Data handling at the secondary level mainly covers the part of statistics called **descriptive statistics**. This part of statistics deals with the collection, representation (in tables, charts and diagrams) and analysis (calculation mean, measures of spread, etc.) of data. The part of statistics dealing with drawing conclusions, testing of hypothesis is called inferential statistics. This part is not covered rigorously in Module 6.

Section B: Reasons for including data handling in the curriculum

One of the main objectives, universally accepted, for secondary education is to prepare the learners for the world of work, economy, politics they are to participate in after completing their education. The reliance of the society on data analysis is obvious: decisions for the future, and for further development in the social, political, economic and other realms are taken based on statistics. In the learning and teaching of data handling the purpose should not be lost sight of. The aim is: **decision making**. Statistics (data handling) is all about exploring data in order to answer questions, and is the study of the variation of that data. The basic idea is: looking for patterns in the variation and trying to understand them. Some authors state that mathematics is about rigour, abstraction and certainty, while statistics is about uncertainty, investigation and estimation within a context. Others will not make such a distinction and feel that both mathematics and statistics are trying to describe the world around us by using models.

Data handling has decision making as its aim. For example:

A shoe factory will be interested in the most common shoe sizes in order to make a **decision** on the production process.

The Ministry of Education will be interested in the trend in the number of pupils starting each level of education in order to make **decisions** related to building of schools, training of teachers, etc.

Nearly all sciences, e.g., economics, business studies, political science, geography, biological sciences (including medicine), psychology, etc., all make use of statistics in order to make decisions based on the available data.

Section C: Misconceptions as to what data handling / statistics implies

That data handling is aimed at **decision making** is frequently lost sight of and some of the misconceptions among learners about data handling are:

1. Statistics is surveys

   Statistics is seen as restricted to questionnaire based data collection, representation and analysis.

   Data collection through **experiments** (scores on tossing a dice 500 times, length of maize plants grown using three different types of fertiliser, comparing two different teaching methods using an experimental and a control group) and through **simulation** (using random number devices to model a real life situation that is not available for experimentation or too dangerous for experiments or too costly, e.g., spread of a disease) form two other major methods which should be included in the school curriculum.

2. Statistics is art

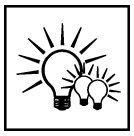
   A common misconception among younger pupils. They will produce pages of colourful neatly drawn charts with no interpretation.

3. Statistics is sums

   A common idea among older pupils, who have learned how to compute various values from raw data. However there is no rationale (why would you calculate a mode, median or mean in this particular case?) nor interpretation (what is this calculated value now telling us?).

4. Statistics is data collection

   Data are collected for the sake of it, without real purpose.



Practice task 1

The objective of this task is to find out pupils' ideas about data handling / statistics.

Ask pupils to write down (in small groups) what they understand by "data handling" (or if they are more familiar with the term statistics use statistics).

Categorise the responses. Are some responses reflecting the misconceptions listed above?

Using guided discussion discuss with pupils what data handling entails and what its purpose is.

Write an evaluative report on this class activity. Questions to consider are: What are pupils' ideas about data handling? Did the discussion go well? Did all participate? What ideas were brought forward? What ideas for questions or projects did come up? Why do you think it is important for a teacher to be aware of the ideas of the pupils?

Section D: How do people make decisions?

There is clearly a need to include the **purpose** in the learning of data handling. Textbooks frequently concentrate on the mechanical computation of statistics (number crunching) and production of charts. However data handling is at all times related to answering a question, to make a decision. Each and every person makes decisions—but frequently in an unscientific way. Decisions of people are often based on personal, informal data collection in cases where a more scientific approach would be more appropriate.

Write down how you make your decisions in various situations. For example when buying clothes, shoes or household items or when deciding how to spend a free weekend, which friends to invite to a party, etc.

Compare what you wrote with the information below and see in how far it agrees or disagrees with what you wrote.

For example:

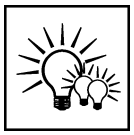
Lesego is going to buy a new pair of takkies.

She has read a report in a consumers magazine that compared takkies from three different makes on good fit, how long they last, how healthy they are for the feet, price, etc. From the three brands A, B and C, C is recommended as the best buy.

Lesego also asked some friends—they all are wearing brand A (the type considered to be 'cool', C is considered to be 'square'—the 'out' thing).

Which brand will she buy?

Research indicated that Lesego will go for brand A! People base their choices and decisions on the personal, informal data they collect—not on scientific (statistical) data analysis. People are influenced by friends, peers, advertisement (on radio and television, in the news papers) and make decisions based on this very selective and frequently one sided (biased) information.



Practice task 2

The objective of this activity is to make pupils aware of how they make decisions.

1. Ask pupils to discuss and write down (in small groups) how they would make decisions in the following situations, what facts they would take into account and how they would collect these 'facts' (add some more ideas appropriate for your class, or ask pupils to add their own situation requiring decision taking).

   (a) The type of drinks to buy for a class party.

   (b) The make of sport shoes to buy.

   (c) The type of clothes to wear to a party.

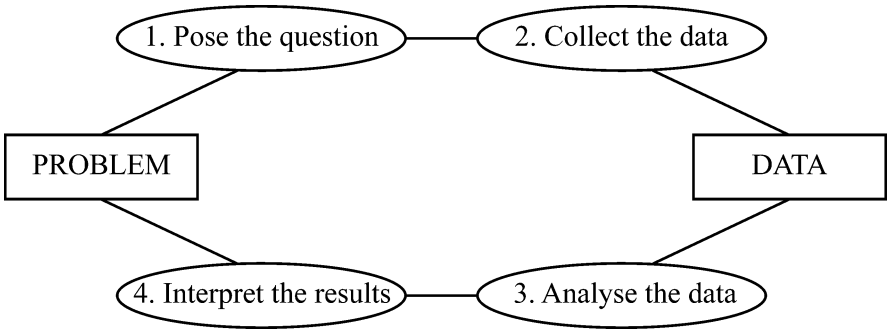
   (d) The make of school bag to buy.

   (e) The boy/girl to go out with to a disco.

2. Categorise the responses and analyse whether the pupils based their decisions on personal and informal data or whether they used a scientific method to collect and analyse data.
3. Use guided discussion to discuss with pupils what the most reliable method is to obtain data on which to base decisions in a specific situation.
4. Write an evaluative report on this class activity. Questions to consider are: What data do pupils use to make decisions? Did the discussion go well? Did all participate? What ideas were brought forward?

Section E: Stages in statistical investigation

A scientific data analysis goes through four distinct stages. The starting point is a question / problem one wants to find an answer to, or wants to make a well founded decision based on the available facts.



Posing the 'correct' question is important. If you do not pose a clear, well-defined question you might not know what data to collect or the question might be such that you cannot find data to answer it. Ill posed questions often lead to data that are impossible or very difficult to analyse and hence will lead to unreliable results.

Section F: Data and information

Asking pupils in your school their favourite colour gives you a list of the colours favoured by pupils. This is (raw) data. Measuring the height of each pupil gives you a collection of measures—again data. Facts, numbers, measures are collected on a certain target group of objects or people. The target group under study is called a population. A population is defined as the entire collection of objects having at least one similar characteristic. The pupils in your class can form a population that share the characteristics of being in the same class. The teachers in the district could form a target population, as do the maize plants growing in the school farm, the bolts produced by a machine in a factory, etc.

Once a set of data has been collected one can try to summarise, combine, compare the data—this leads to information. The height of all the pupils in your school is data, stating that the mean height is 1.62 m is information obtained from the data. This information can become data again if the mean height of pupils in each of the 24 classes in the school are collected.

A misconception is that data is 'number'. 5 is a number (actually a numeral representing the number), but "5 pupils in your class failed the mathematics tests" is data. Data is 'numbers' with a context. The word 'numbers' is in inverted commas as data can also be qualitative, i.e., not expressed in number form.

As teachers we have to encourage pupils to ask the following questions (and they will only do so if we do it ourselves).

a) How was this data collected?

b) What was the source of the data?

c) Is it accurate data or is it estimates?

d) Are they qualities, quantities (counts, measurements, rates, proportions, averages)?

Data is frequently presented in such a way as to favour a certain view or idea. For example, a factory wants to make their product look better than similar products from another factory. A government might want to play down the unemployment figures, a school might want to show how well their students are performing in an examination, a factory might want to show that their wages are high by selecting a specific type of average. By using selective data and representing it in a specific way the general public might be given the wrong impression. It is therefore very important to be critical towards data and the presentation of data in the media.

Self mark exercise 1

1. Practising statisticians claim that statistics is not mathematics. Why do you suppose they think so?
2. Data handling / statistics is included in all syllabi for the lower secondary school. Do you think pupils need to study data handling / statistics? Justify your stand.

Suggested answers are at the end of this unit.

Section G: Type of data

Data can be categorised in different ways. The type of data is very important as it determines the way the data can be represented and analysed. The different types of data are discussed in the following sections.

Section G1: Qualitative versus quantitative data (variables)

Qualitative data: Data that cannot be described by a number is referred to as qualitative data. For example: sex, region of the country, examination grades A, B, ..., preferred drink, political alliance, make of car, a person's blood group (O, A, B, or AB), etc.

In qualitative data the 'values' are words to identify defining categories. Qualitative data is also referred to as categorical data and leads most often to frequency counts for the categories.

Qualitative data are frequently given a numerical code, but any arithmetic done with these codes is meaningless. For example if gender is coded as 1 for male and 2 for female, then averaging to get 1.5 is meaningless. The numbers are only used to ease tabulating of results, but the numbers themselves have no numerical meaning.

Quantitative data is expressed by numbers: age, height, shoe size, income are examples of quantitative data.

As data is made up of a collection of **variables** you can also speak of qualitative variables and quantitative variables.

A **variable** is a factor which describes or characterises some aspect of the population and has different values (numerical or words/categories).

If you want to find out who is helping the pupils with their assignments, the variable is those people helping the pupil. The 'values' this variable can take are, for example: friend, mother, father, sister, uncle, ...

If you want to find the arm span of the pupils in a class, the variable is 'armspan'—a length taking all possible values within a certain range.

Section G2: Ordinal versus nominal qualitative data (variables)

Qualitative data can be ordinal or nominal. Data is **ordinal** when there is an underlying continuity between the words. For example, the response to "How often do you use games in the teaching of mathematics" might be rated on a scale as very often, often, sometimes, never. These words are clearly graded in order but their position on the continuum scale is incompletely defined. The 'distances' between these labels are not clear.

Similarly pupils might be asked to respond to the question "School uniforms should be abolished" by ticking one of the options: strongly agree, agree, indifferent, disagree, strongly disagree. These replies are the values taken by the ordinal variable.

Nominal data is data that names the categories, for example, favourite type of food of pupils in your class, number of different types of animals found in a wildlife park, a person's blood group, brands of rice.

Section G3: Discrete versus continuous data (variables)

Discrete data that can take only specific values or fall in a specific category. Discrete data is frequently the result of counting (quantitative data) or classifying (categorical data). Examples: shoe size, number of pupils in each form, number of songs on a CD, salary of the teacher, score on a dice, favourite drink.

It needs to be emphasised that both quantitative data and qualitative data can be discrete. It is a common misconception to equate qualitative and discrete data.

Discrete data can be displayed in bar charts (most appropriate for categorical data), bar-line diagrams (most appropriate for discrete qualitative data) or pie charts (discrete data with limited number of categories or values).

Continuous data. Data that can take any value within a certain range is called continuous data. Continuous data results in most cases from measuring, for example: pupils' mass, height, arm span, time spend on homework.

Continuous data is best displayed in histograms. In a histogram the frequencies are proportional to the area of the bar. When cases with bars of

the same width are considered the histogram becomes a bar graph with the bars touching each other. Details will be discussed in Unit 3, especially the problem of where the boundaries between two bars are to be exactly located.

Note that it is rather common to display certain discrete data (for example scores on a test, number of children in a family, i.e., numerical data that can be ordered) in a bar graph with the bars touching each other. This strictly speaking is not correct; to show discrete data properly, the bars should be separated.

Self mark exercise 2

State whether the following data is (a) qualitative or quantitative (b) discrete or continuous

1. The number of pips in oranges
2. The mass of oranges
3. The taste of oranges
4. The colour of cars in a car park
5. The number of desks in each classroom
6. The number of goals scored by the football teams that played on Saturday
7. The brands of toothpaste on sale in supermarket
8. The size of the dresses of the girls in school
9. The length of the sentences in this module (i) in words (ii) in centimetres
10. The attitude of pupils towards mathematics
11. The names of the teachers in the school
12. The age of the persons in your family
13. Number of voters that voted in the last elections
14. Salaries earned by teachers
15. Most popular music track during a school disco
16. The geographic districts (provinces) in the country
17. The audibility of teachers when speaking in class
18. Flavours of ice-cream sold in a super market
19. The grades (A, B, C, ...) by candidates obtained in an examination
20. The smell of different bathing soaps

Suggested answers are at the end of this unit.

You have come to the end of this unit. As a last assignment you are to bring what you have learned into the classroom.

Practice task 3

The objective of this activity is to make pupils aware of different types of data.

1. Write a lesson plan with these objectives (a) to introduce pupils to the different types of data (b) to practice and consolidate classifying of different types of given data (c) to illustrate different types of data with pupil generated examples.
2. Write an evaluative report on the lesson. Questions to consider are: Did pupils find difficulties in classifying data as discrete-continuous, qualitative-quantitative? Were the pupils able to generate their own examples? Were pupils well motivated to work on the activity? Were the objectives achieved? Did you meet some specific difficulties?

Present the lesson plan and report to your supervisor.

Summary

This unit introduced the subject of statistics by stressing **project-based learning**, **interpretation** of results, and **taking informed decisions**. These themes will be given more attention in subsequent units. Watch for an ongoing theme or motif about the true purpose of learning (and doing) statistics.

Unit 1: Answers to self mark exercises

Self mark exercise 1

1. In the view of some statisticians, mathematics deals with rigour and certainty. Statistics on the other hand deals with stochastic processes, i.e., processes related to chance. Hence the statistics lead to conclusions that are valid with a (stated) degree of probability.
2. Making decisions in today's world frequently requires the ability to analyse and interpret data, e.g., advertising claims, weather reports, statements of politicians, environmental impact reports, reports on population growth and spread of diseases. As a result data handling skills are increasingly fundamental to individuals in order to participate as well informed citizens in the society.

Self mark exercise 2

N.B. For some qualitative data, you can argue whether it is discrete or continuous. For example like or dislike for a subject: you can see these as discrete, but you might also look at them as the ends of a continuous scale with extreme ends: very great liking and very great dislike with all possible shades of like to dislike in between. In the table these arguable cases are indicated with √?

Question	Qualitative	Quantitative	Discrete	Continuous
1		√	√	
2		√		√
3	√		√?	√?
4	√		√	
5		√	√	
6		√	√	
7	√		√	
8		√	√	
9		√	(i) √	(ii) √
10	√		√?	√?
11	√		√	
12		√	√ (in years)	√ (exact time)
13		√	√	
14		√	√	
15	√		√	
16	√ (if named)	√ (if number)	√	
17	√		√?	√?
18	√		√	
19	√		√	
20	√		√?	√?

Unit 2: Methods of data collection

Introduction to Unit 2

In this unit you will look at methods of collecting data. There are three basically different ways to collect data: (a) using a survey (b) carrying out an experiment (c) carrying out a simulation. Data collection in the form of a survey frequently uses questionnaires and / or an interview. The unit will look at the characteristics of a well structured questionnaire. When collecting data, the question of the number of items needed for analysis in order to make well founded decisions will arise. This relates to concepts such as population, randomness of the sample and how well the sample represents the population. These will be looked at in this unit.

Purpose of Unit 2

The aim of this unit is to look at methods of data collection and questionnaire construction. The issue of sampling is introduced without going into details of the various methods of sampling.

Objectives

At the end of this unit you should be able to:

- distinguish between population and sample
- state reasons for sampling
- explain what is meant by a random sample
- explain common misconceptions in descriptive statistics
- select the most appropriate method of data collection (survey, experiment and simulation) for a given problem
- list four different types of surveys: a) questionnaire survey, b) interview, c) administering tests and d) structured observations
- explain and illustrate each of the above four types of surveys
- select the most appropriate type of survey given a problem requiring a survey method
- list and illustrate five characteristics of a well designed questionnaire
- explain the importance of a well designed questionnaire
- design a questionnaire for a given questionnaire survey
- illustrate and explain advantages and disadvantages of open or closed questions in a questionnaire
- list four points requiring special attention in a structured interview
- design a data collection sheet for a specified survey
- explain what a frequency distribution and a grouped frequency distribution is

- state reasons for using grouped frequency distributions
- explain disadvantages of using grouped frequency distributions
- design appropriate pupil centred activities for pupils to collect data

Time

To study this unit will take you about eight hours.

Unit 2: Methods of data collection

Section A: Data collection with a purpose

This unit is going to look at data collection. Although the emphasis in this unit is on methods of data collection, it should not be forgotten that prior to the collection of the data, objectives for the collection are to be set. The question: What is the purpose of collecting this data? is to be kept clearly in mind and answered before starting to collect data. The data that needs to be collected in order to answer the question is to be clearly identified. It is not uncommon for pupils to collect data NOT needed to answer the questions set or NOT to collect data that is needed. Pupils have the tendency at times to start to collect data in an ill organised way. This leads to situations where they might end up with data that is impossible or difficult to analyse or the omission of data needed to answer the question they set initially.

For example:

Suppose you want to find out what the favourite sport is of boys and girls in your school. Your objective is to make a list of the rank order of sports favoured by boys and a similar list for the girls in your school.

Pupil A made a questionnaire with one question:

Please write down what your favourite sport is.
My favourite sport is _____.

This pupil cannot answer the set question as the gender of the respondent is not included in the questionnaire. Data needed to answer the question was NOT collected.

Pupil B made a questionnaire as follows:

Please complete this questionnaire.

Are you male or female? (give a tick) Male ____ Female ____

In which Form are you? I am in Form _____

What is your favourite sport? My favourite sport is _____

This pupil collected data not needed to answer the question set in the objective, by including a question asking which Form the respondent is in.

It is very important that the data collected corresponds to the objective(s) set.

Self mark exercise 1

1. You want to find out how much time pupils in your class, on average, spend on preparing for a science test and whether there is a difference between boys and girls in this respect.
   a) What data would you collect?
   b) Give two examples of data that would be inappropriate to collect to answer the question.
2. You want to find out what is the favourite colour of pupils in your school.
   a) What data would you collect?
   b) Give two examples of data that would be inappropriate to collect to answer the question.
3. You want to find out to what occupation the pupils in your class aspire to.
   a) What data would you collect?
   b) Give two examples of data that would be inappropriate to collect to answer the question.

Suggested answers are at the end of this unit.

Section B: Population, sample and random sampling

For any question to be answered, a target group under study has to be identified before definition of variables, i.e., the data to be collected. The target group under study is called a population. A population is defined as the entire collection of objects with at least one similar characteristic. For example, if you want to find the proportion of the pupils at the school who regularly drink Coca Cola, you can ask each pupil in the school a question such as: "Do you drink Coca Cola at least once a week?" and record the answers. But talking with every pupil in the school would be time consuming and probably quite difficult to arrange. In this case, you might then try the question on just a portion of the pupils and based on the responses of that portion of the pupils make generalisation about all the pupils in the school on whether they regularly drink Coca Cola or not. This portion of the pupils is called the "sample." A sample is defined as a portion of the entire collection of objects of similar characteristics (the population). In the above example, the pupils in the school form the population and the portion of the pupils that were asked to respond to the question is called the sample.

The 'real' statistician will not see the physical objects as population and sample but the **collection of observations** obtained or **measurements** taken on the physical objects.

Need for Sampling

In order to get accurate data, on which decisions can be based, one would have to consider the whole population under study. However, in many cases it is not possible to obtain information about all members of a population for the following reasons:

1. The collection of information may destroy the sample, e.g., testing the life span of electric bulbs or electric fuses.
2. The population may be infinite, as in the results of throwing two dice (the experiment can be continued indefinitely).
3. It may be impracticable to make a measurement for every member of the population, e.g., finding the shoe size of all people in a country or the length of rice grains in a 50 kg bag. Even if a measurement could be made for each member of a population, considerations of time and expense usually dictate otherwise.

Due to these constraints it becomes inevitable in most cases to consider samples in obtaining information about a population. But how large should the sample size be in order for the sample to be representative of the population? The underlying idea to this question is how many measurements or observations should be taken so that the sample displays the true characteristics of the parent population? There is no definite answer that can be given to this question. It seems obvious that "the larger the sample size, the more accurate are the results," since the large sample is likely more closely an approximation to the population. If the sample size is small, there is a high chance that the results may not give a true picture of the parent population, because the sample size is too small to be representative of the population. However it is a misconception to think that the sample size is to be proportional to the population. A relatively small sample (say 150) can give reliable information on the whole population provided the sample is **representative**.

Purpose of Sampling

Since it is not possible at times to consider the whole population, a sample which is representative of the population is drawn from the population. That means the characteristics of the sample are the same as that of the population. The sample is linked to its population through **estimation theory**.

This theory states that if **statistics** such as mean, median, variance, standard deviation, etc., are calculated from the sample, it is possible to use these sample statistics to estimate **parameters** such as mean, median, variance, standard deviation etc., for the entire population from where the sample was drawn. The theory tries to generalise the characteristic of the sample to that of the parent population. This is a topic for advanced statistics and will not be discussed here.

Note the use of the word statistics in the above: a sample is described by measures such as mean, mode, median, variance, range—these measures are called statistics.

The word parameter refers to the same measures for the population. The mean of a sample is a statistic, the mean of a population is a parameter.

Estimation theory explains how statistics can be used to obtain parameters.

Random Sampling

In order for a sample to be representative of the whole population, each member of the population must have an equal chance of being chosen. A sample chosen in this way is called a random sample. To take a sample of 30 from all the pupils in your school you could place the names of all pupils on separate pieces of paper and put these in a box. You then pick from the box 30 slips. The names on these slips form the pupils that are going to be in your sample.

If your question to be answered requires pupils from each form to be in the sample and there are 18 forms in the school, you might make a sample by randomly picking two pupils from each form (giving a sample of size 36).

Anything that distorts data so that it does not fairly represent the population is called **bias**. The way in which data is collected and samples are formed must avoid bias so that any results produced are reliable. There is a famous example of an opinion poll giving the wrong prediction due to bias, when in 1948 an extensive telephone poll predicted that the new president of the USA was to be Dewey. However it turned out that Truman won the election. The incorrect prediction was due to bias: the telephone sample was not representative for the voting population as—at that time—many people did not have phones (apart from higher-paid voters who tended to vote for Dewey).

Pupils should understand that results obtained on a sample **may** apply to the whole population provided the sample was representative for the population.

Self mark exercise 2

What is wrong with the way that each of these samples is selected?

1. The head of the school wants to choose a new school uniform that pupils will be happy to wear. She asks 5 pupils from form 1 for their ideas. The new uniform is to be based on the responses of these pupils.
2. Batteries are produced by a machine. Every 50th battery is tested to ensure production standards are maintained.
3. A take away sells pies. Pies are made throughout the day. A sample from the first batch of pies produced is tested every day.
4. To decide what drinks to buy for the end of year party for all forms (200 pupils) you ask the first 10 pupils of your class entering the classroom for their choice. Based on their responses you buy the drinks.
5. A pupil is doing a survey into pupils' opinions about the school tuck shop. She picks pupils at random from those coming to the tuck shop to buy something.
6. You want to find out pupils' attitude in your school towards the learning of mathematics. You ask the 15 members of the mathematics club for their opinion.

Suggested answers are at the end of this unit.

Section C: Misconceptions in inferential statistics

The part of data handling / statistics dealing with the collection, representation and analysis (calculation of mean, etc.) is called **descriptive statistics**. The part dealing with the interpretation, testing of hypotheses and the drawing of conclusions is called **inferential statistics**. There are a number of well documented misconceptions in inferential statistics. It is important for a teacher to be aware of these so as to plan lessons such that the misconceptions are avoided—or when noted are addressed.

1) Concepts "population" and "sample."

   In data handling population refers to the complete group of persons, objects or measurements in which we are interested.

   A sample is part of that population, implicitly meant to be a representation of that population.

   These meanings differ from what the lay person understands by these words. A population is seen as a group of people in a certain region whether it be as large as a continent or as small as a village. The word sample has a wider general interpretation met in a context such as 'a free sample'.

2) There is no variability in the 'real world'.

   There is the idea that a sample accurately represents the population, i.e., that there is no variability (sampling errors). Pupils erroneously believe that, for instance, sample and population must have the same mean.

   For example: 1000 pupils were measured and the average height was found to be 168 cm. A sample of 10 pupils is taken from the 1000 randomly. The first pupil selected has a height of 162 cm. What will be the expected average height of the 10 pupils in the sample? The common error response is: 168 cm, i.e., the sample is thought to be identical to the population, so the given 162 cm is completely ignored.

3) Unwarranted confidence in small samples.

   Findings based on small samples are thought to be representative for the population, ignoring the fact that small samples, due to variability, might differ substantially from the population.

4) Insufficient respect for small differences in large random samples.

   Small differences in large random samples can be highly significant.

5) The size of a sample should be directly related to the population size.

   It is not so much the size as the fact whether or not the sample is representative that matters. If you are to test batteries on their life span, a well selected sample of 100 batteries will yield results reliable enough to make statements on the whole population of millions of batteries of that type. A sample of 200 would not yield more reliable results.

For students' misconceptions that are common in **descriptive statistics**, see Unit 3 Section F.

Section D: Methods of data collection

Data can be collected through

- Surveys (Section D1)
- Experiments (Section D2)
- Simulations (Section D3)

Surveys form a good starting point for pupils for gathering their own data and representing and analysing the data gathered. Experiments require more data gathering than surveys do, since they use both test and control subjects, or before-and-after testing. Simulations are similar to experiments but still more involved since they use random number devices—dice, spinners, random number tables or Random-key on calculator—to model real-world situations. In the following sections each method of data collection is discussed in some more detail.

Section D1: Surveys

What is a survey?

A survey is a method of collecting existing data to find the answer to a question. You collect the data from people (their mass, their favourite holiday activity, their view on the quality of the postal service) or objects (the content or mass of packaged cereals, different brands of cooking oil available in different supermarkets). You might want to know what activity pupils are involved in outside their school work, a survey, i.e., collection of data of pupils' responses to questions, might help you to answer the question. A survey based on the entire population is called a **census**. Surveys always have to start with a well defined **purpose**. The purpose of the survey is frequently formulated in the form of a **hypothesis**: a statement that might be true or false. To test the hypothesis a survey can be carried out.

Here are some examples of hypotheses:

1) Girls and boys perform equally well in mathematics in our school.
2) Pupils in the school are of the opinion that the amount of homework is about right.
3) Pupils feel that everybody in the school—including teachers—should participate in at least one sport.

Several factors have to be considered when setting a hypothesis:

a) Can you test it?

b) Can enough data be collected to give a valid outcome?

c) Can the type of data to be collected be analysed?

Self mark exercise 3

Are the following correctly formulated hypotheses, i.e., can they be answered? Justify your answer. If not restate the hypothesis in a correct form.

1. Group work in mathematics is better than whole class teaching.
2. There is life after death.
3. There will be an outbreak of malaria in Botswana every five years.
4. Pupils often participate in sporting activities.
5. Pupils are given too much homework.

Suggested answers are at the end of this unit.

Section D 1.1: Types of surveys

There are various types of surveys. We will look at the following:

- Questionnaire survey
- Interview
- Tests
- Structured observations
- Secondary data

Questionnaire survey

Questionnaires are collections of printed questions to be answered in writing by persons from the target group. Questionnaires are frequently used to collect data on opinions of people. A business might seek information about how the buyers view their products, a political party might want to know the opinion of voters in a region on their wishes for educational facilities in the region. In both cases a questionnaire presented to a well selected group of the population might be used to collect the required data.

Questionnaire design

The design of a questionnaire is most important. The way questions are phrased, the way answers are expected to be given and the overall layout of the questionnaire have an impact of the validity of the data collected.

Be very specific about what type of data you are looking for and why, to avoid questions not needed for the survey to be included in the questionnaire or questions needed for the survey to be (erroneously) omitted. The data collected in the questionnaire should be relevant to the research question(s) / objectives and it should be possible to analyse the data collected.

If a questionnaire is used to collect data:

a) Make it as short as possible. Long questionnaires lead to a poor response rate and/or unreliable responses (e.g., people quickly tick randomly, skip questions).

b) Questions should be phrased clearly and unambiguously. For example don't use two-in-one questions.

Examples: How often do you go to a sport activity and how do you go there?

This questions asks for two things and should be split. In addition the question is not clear: How often - do you mean per week, per month? Responses to such a question may be: not so often, from time to time, etc., which would be difficult to analyse.

c) Language and vocabulary must be at the level of the respondents i.e., as simple as possible so that everybody can understand what the question wants.

d) Do not include questions bearing no relevance to the research question, i.e., focus on questions directly relevant to the research objectives only. Each question should be directly related to one of the set objectives. If a question does not relate to an objective it is clearly out of place and should not be included.

For example if sex, school, number of years in teaching are NOT used as variables in the research there is no need to collect this data.

e) Avoid leading questions, i.e., be neutral. For example:

(i) Why do you use teaching aids? - implying that one IS using teaching aids, but perhaps the respondent is NOT using teach aids.

(ii) How would you improve your handball skills? - implying that they need improvement, the respondent might feel that his/her skills are inferior and do need further improvement.

(iii) Would you not agree that there is too much violence among pupils? - implying that the person setting the questionnaire thinks so and seeks support for his/her view. Better would be to ask:

Do you think that pupils are too violent with each other?
YES ☐ NO ☐

f) Ensure a neat, well typed and well spaced layout of the questionnaire with sufficient space for answering the questions.

g) Be aware that responses do not necessarily (i) reflect the real view of the respondent (ii) reflect the actions of the respondent especially when 'sensitive' questions (income or sex habits of person, for example) are asked.

h) Ask (in general) short questions which can be answered precisely.

i) Ask questions in a logical sequence.

j) In a closed questionnaire provide tick boxes to allow respondents to tick their choice. This makes it easier to complete and to analyse.

Questions that look rather straightforward can still cause problems. For example:

Are you a regular church goer? YES ☐ NO ☐

Are you in favour of trading on Sundays?
YES ☐ NO ☐ UNDECIDED ☐

A person might wonder whether a mosque or synagogue is considered as a church. A person might be against trade on Sundays but still wants to be able to buy 'take-away' or petrol on Sundays. These problems are generally overcome when an interview is used and the interviewer can clarify questions on the spot.

Administering the questionnaire

There are various ways in which the questionnaire can be administered.

a) Given personally to (random selected) individuals from the target population and to be filled in in your presence.

b) Given personally to (random selected) individuals from the target population and to be collected later. The individual can complete the questionnaire in his/her own time.

c) Given personally to a (random selected) group from the target population and to be filled in in your presence.

   A class of pupils can be asked to fill in a questionnaire related to menu and organisation of meals in the dining hall in your presence.

d) Posting questionnaires to the randomly selected individuals forming the sample.

Always when administering a questionnaire, in whatever form, explain—in an introduction or orally—the purpose of the questionnaire, and inform the respondents of what you are going to do with the data collected.

Self mark exercise 4

1. The following questions were included in a questionnaire on radio listening habits of pupils in the school. Explain why each question is unsuitable and rewrite the question so that it could be included in the questionnaire.

   a) When do you listen to the radio?

   b) What do you like about radio programmes?

   c) Don't you agree that the radio gives the best news reports?

   d) Shouldn't there be more educational programmes on the radio?

   e) How could the news readers improve their presentation of the news?

   f) How big are you?

   g) What do you think of the new improved menu in the dining hall?

2. The following questions were included in a questionnaire on the use of the school library. Explain why each question is unsuitable and rewrite the question so that it could be included in the questionnaire.

   a) Do you use the library frequently?

   b) Are you in the final year of school?

   c) Which books do you read?

   d) Don't you think that more books on science are needed in the library?

e) The journal section in the library is not up to date. Do you agree?

f) How could the library be improved?

3. Above were listed four ways to administer questionnaires. List advantages and disadvantages of each method mentioned.

4. a) Which of the following questions do you think are biased?

   b) Write down what makes them biased.

   c) Write a better question to replace the biased ones.

   (i) Most people think that the bus services are rather poor. Do you agree?

   (ii) Do you think that the bus services are better than they used to be?

   (iii) Do you think that smoking should be banned on buses?

   (iv) Do you agree that the buses should run more frequently?

   (v) Secondary school students often behave badly on buses. Do you think special buses should be provided for them?

   (vi) Don't you agree that beauty contests are degrading?

   (vii) Don't you think that the use of make-up is unnatural and bad for the skin?

   (viii) Do you agree that capital punishment should be brought back for murder and rape?

Suggested answers are at the end of this unit.

Format of the questions

Questions can be presented in two formats: closed or open. We look at them in more detail below.

I. Closed-ended: the respondents are to select responses from pre-specified answers

Advantages:

- A large sample can be used.
- Anonymity of respondents (ticking answers, handwriting will not 'detect' respondent).
- Responses can be easily (computer) analysed.
- Responses from different respondents are comparable.

Disadvantages:

- No guarantee that questions will be interpreted in the same way by all respondents.
- Questions might not be understood by respondents.
- Bias of researcher might come in: researcher chooses the questions and the responses, the respondents might not find 'their' questions and 'their' responses and just choose from what is presented. For example, only a

NO /YES option is offered but respondent feels "it depends on ..." or a scale does not offer a neutral position (strongly agree - agree - disagree - strongly disagree; the person might be neutral and neither agree nor disagree with the statement).

- Low return rate might affect reliability and validity of the research as the non respondents might differ in opinion from the respondents.
- Some respondents might be 'questionnaire tired' and random tick some answers.
- Generally respondents do not answer ALL questions—there are some blanks in the majority of cases. This might affect the outcome of the study.

To make a valid reliable closed-ended questionnaire a **preliminary pilot** in open format should be presented to a representative group. The responses given will indicate what questions (and how phrased) to include and what to include as possible responses in the closed-ended questionnaire. The closed questionnaire needs also to be piloted to reveal ambiguities, poorly worded questions or questions that are ill understood by respondents.

One should check a questionnaire on the following:

- Responses to questions should be consistent with the wording of the question.

  Question: Do you prefer hot or cold drinks? YES ☐ NO ☐

  The responses do not 'fit' the question. HOT ☐ COLD ☐ could have been given as responses.

- Responses should cover all possible answers to a question.

  Question: Do you discuss your grades with your friends?

  Responses ALWAYS ☐ SOMETIMES ☐ NEVER ☐

  If you did discuss your marks with most friends except for one or two, it will be difficult to answer the question. A box with USUALLY ☐ would improve the question.

- Responses offered should be balanced: the number of positive and negative responses should be the same.

  Example:
  The responses OUTSTANDING VERY GOOD GOOD POOR are unbalanced since there are three responses at the 'positive' side and only one at the 'negative' side. A more balanced scale would be

  VERY GOOD GOOD FAIR POOR VERY POOR

II. Open-ended: Each question represents a topic on which the respondent can freely comment.

Advantage

- Respondents can really bring out their point of view and present what they feel is most relevant.
- Good information for decision-taking often arises from thoughtful comments.

Disadvantages:

- Data is difficult to analyse.
- Responses of different respondents are difficult to compare.
- Bias of researcher can come into the categorisation of the responses.
- Answers might be scanty and lacking the data the researcher is looking for.

The main source of bias is in the analysis of the open-ended data. The researcher will have to use judgement as to how to categorise various responses. To safeguard against this bias the data should be analysed by at least TWO independent individuals.

Self mark exercise 5

1. Suggest 8 questions to be included in a questionnaire to be included in a survey of

   a) homework habits of the pupils in the school

   b) reading habits of the pupils in the school

   c) eating habits of the pupils in the school

   d) pupils' views on school rules

2. Design a questionnaire to test each hypothesis:

   a) Pupils in school are less superstitious than their parents / guardians.

   b) Girls, given a free choice, would hardly ever choose to wear a dress in preference to something else.

   c) Most cars these days use unleaded petrol.

   d) Pupils in form 3 are better in estimating length than pupils in form 1.

   e) Left-handed pupils are better in mathematics than right-handed pupils.

   f) The more time you use to prepare for a test the higher your score.

3. Which of the following questions included in various questionnaires are likely to get honest responses?

   a) I help enough at home. YES ☐ NO ☐

   b) Are you a kind person? YES ☐ NO ☐

   c) My family is poor. YES ☐ NO ☐

4. The following questions are likely to get responses that would not be very useful. Rewrite these statements so the responses would give you more useful information.

   a) I get 9 hours of sleep each night.
   ALWAYS ☐ SOMETIMES ☐ NEVER ☐

   b) I do my maths assignments.
   ALWAYS ☐ SOMETIMES ☐ NEVER ☐

   c) I go out during the weekend.
   ALWAYS ☐ SOMETIMES ☐ NEVER ☐

Suggested answers are at the end of this unit.

Interview

An interview is an oral, in-person administration of a standard set of questions that is prepared in advance. A structured interview (the questions have to be prepared as for a questionnaire and can be structured—respondent to choose from alternatives provided by the interviewer—or semi-structured) is generally a more reliable source of data collection than a questionnaire. All responses are coded, tabulated, and summarised numerically. The feedback received can be improved by probing and follow up questions. The quantity of data that can be collected is less than when using a questionnaire. Combining questionnaire design with interview is an option worthwhile to consider.

Advantages of using an interview to collect data

1. high response rate
2. more detailed information can be collected (follow up questions / clarifying question if misunderstood)

Disadvantages of using an interview

1. expensive and time consuming
2. possible bias in the way questions are asked and responses recorded

Administering tests

The term **"tests"** refers to the use of test scores as data. This technique involves subject response to either written or oral questions to measure knowledge, ability, aptitude, or some other traits that describe a characteristic of the subject.

For example:

(i) You want to find out pupils' skills in problem solving. You set a test with a variety of problems each testing a particular problem solving strategy. The scores on the test are your data.

(ii) You want to find out the skills of pupils in estimating the length of line segments in the range 0 cm to 50 cm. You prepare a test with various line segments drawn on it and pupils are asked to write down the estimated length. This will give you the raw data.

(iii) You want to find out pupils' skills in making models of 3D objects. The test set to pupils is to make a prescribed 3D object.

Tests require careful construction to ensure they will test what you want them to tests (i.e., the test is to be valid). Each test item must relate clearly to an objective to be tested. As tests require scoring a scheme is to be prepared as to how scores will be awarded to the responses (particularly how to allocate scores to partial correct responses).

Structured observations

Data can be collected by observation, visually and / or auditory, and the observations systematically recorded. For example, if you want to find out whether a pedestrian crossing should be made at a particular place you could use an observation sheet to tally the number of people crossing at or near that

point. If you want to find out whether boys or girls ask more questions in a classroom you can sit in the classroom and tally whenever a question is asked.

Data collection sheet are frequently used when collecting data by counting. When we want to find the favourite colour of pupils in the class the responses of each pupil could be recorded in a simple record sheet.

Data collection sheet

Colour	Tally	Frequency
red	| | |	3
blue	𝍸 | |	7
black	..	..
green	..	..
white	..	..
silver	..	..
orange	..	..
.....	..	..

The colour mentioned by each pupil is recorded in the **tally** column by a single stroke. To make counting easier groups of 5 are recorded as 𝍸.

The total number of times each colour is mentioned is called the **frequency**.

A table for discrete data with the totals included is called a **frequency distribution**.

For large amounts of discrete or continuous data the data is organised into **groups** or **classes**. Data collected in groups with totals included is called a **grouped frequency distribution**.

Grouped frequency distribution table: mass of boys in a class

Mass m kg	Tally	Frequency
$50 \leq m < 55$	|	1
$55 \leq m < 60$	𝍸 |	6
$60 \leq m < 65$	𝍸 𝍸 | |	12
		..
		..
		..
		..
.....	..	..

The masses of boys in the class are grouped in **class intervals** of equal width (5 kg). $50 \leq m < 55$ means 50 kg or more, but less than 55 kg.
The boundaries between classes are at 55 kg, 60 kg, etc. This should be made clear to pupils by asking questions such as the following.

Questions to ask are:

1. Elliot has a mass of 54.9 kg; in which class interval is his mass recorded?
2. Tumisang has a mass of 55.0 kg; in which class interval is his mass recorded?

In presenting an exercise to pupils on making distribution tables ensure you cover all the cases:

(a) frequency distribution of discrete data

(b) grouped frequency distribution of discrete data

(c) grouped frequency distribution of continuous data

An example of each is given in the following self mark exercise.

Self mark exercise 6

1. The colour of 40 cars in a car park are listed.

red	blue	red	silver	red
red	red	white	blue	silver
red	black	green	red	blue
green	blue	black	white	yellow
white	white	red	blue	silver
green	blue	green	yellow	black
white	white	green	blue	red
red	white	black	silver	green

   a) Make a frequency distribution for the data.

   b) Which colour of car is most popular? (This colour is called the mode.)

2. The ages of the 40 teachers in a school are listed.

27	35	26	33	24
43	34	20	47	56
42	49	57	34	54
29	39	50	21	37
58	30	28	26	20
34	33	27	41	59
47	62	52	29	30
25	37	29	44	22

   a) Make a grouped frequency distribution using classes 20 - 29, 30 - 39, 40 - 49, 50 - 59 and 60 - 69.

   b) In which class interval are most of the teachers? (This is called the modal class.)

3. The height of 36 girls in a class was recorded to the nearest cm.

148	155	156	175	160	165
159	179	161	173	167	154
163	158	160	172	147	170
158	155	165	178	168	157
166	171	157	159	172	162
155	172	165	168	157	173

   a) Make a grouped frequency table taking the class intervals for the height h as $145 \leq h < 150$, $150 \leq h < 155$, etc.

   b) How many girls are less than 160 cm?

   c) How many girls are 155 cm or taller?

4. a) Make a data collection sheet to record the month in which the pupils in the class were born.

   b) Make a frequency table for the data.

   c) In which month did most births occur?

Suggested answers are at the end of this unit.

Secondary data

The above listed methods to collect data through surveys are called **primary data** collection. The required data is directly obtained by the person collecting the data. The data is the result of direct observation, tests, questionnaires and interviews. At times one might make use of published data, for example, government statistics on number of pupils attending the various levels of the education system, the number of teachers teaching each subject, the number of children born in each district, etc. This type of data is called **secondary data** as it was not directly collected by the researcher. Secondary data need to be approached with great caution. Are the secondary data up-to-date (i.e., is the data accurate)? How was the data collected (i.e., is the data reliable)?

In your assignment you are required to facilitate the learning of pupils on one (or more) of the following issues related to data collection:

a. What data is to be collected in a given situation
   (See self mark exercise 1)

b. How is a sample to be taken to ensure the outcomes are reliable
   (See self mark exercise 2)

c. How to formulate correct hypotheses (See self mark exercise 3)

d. How to formulate correct questions to be included in a questionnaire
   (See self mark exercise 4)

e. Designing a questionnaire to collect data to answer a specific question or hypothesis (See self mark exercise 5)

f. How to tabulate various types of data (Self mark exercise 6)

You might decide to cover several of the above by setting a small project to groups (4 per group is recommended). However prior to pupils embarking on data collection, preparing a questionnaire, setting hypotheses, recording the data in tables, they have to be made aware of and have to discuss the issues covered in the self mark exercises 1 to 6.

Practice task 1

1. Choose one (or more) of the topic listed above (a to f).
2. Write a lesson plan with clearly stated objectives. Prepare worksheets for the pupils to work in groups.
3. a) Write an evaluative report on the lesson. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did you meet some specific difficulties in preparing the lesson or during the lesson?

   b) Present the lesson plan and report to your supervisor.

Two-way tables to tabulate data

Data can at time be collected and represented in **two-way tables**. These tables can be used when two different variables are involved. Reading and interpretation of these tables will be one of the objectives.

Self mark exercise 7

1. The table displays data on pupils with spectacles in a school.

Pupils wearing glasses in the school

	Wear glasses	
	YES	NO
Girls	82	273
Boys	86	295

a. How many girls wear glasses?

b. How many pupils do not wear glasses?

c. How many pupils were in the survey?

d. Do the results prove or disprove the hypothesis "More boys than girls wear glasses"? Justify your answer.

2. The two-way table displays the results of a survey to test the hypothesis "More girls are left-handed than boys."

Pupils being left-handed in the school

	Left-handed	
	YES	NO
Girls	10	60
Boys	15	90

Do the results prove or disprove the hypothesis? Justify your answer.

Suggested answers are at the end of this unit.

Guidelines for designing and using an observation sheet

In the following assignment you are to guide pupils in the designing of a data collection / data observation sheet. Next pupils in groups collect the data and present the data in tabulated form to the class, explaining their design and pointing out any valid conclusions that can be drawn from the data.

The following information is to be given and explained to pupils:

Draw up your observation sheet after you have decided on the categories you are going to use.

Make some initial decisions on how you hope to organise and analyse the data.

Decide when and where to collect the data.

Practice task 2

1. Set to groups of pupils the task to design an observation / data collection sheet and to collect data on topics of their choice. Suggestions could be given if needed, for example (see below) or add some other ideas from pupils or yourself.

   a. pupils coming late for the first class in the morning

   b. pupils leaving the classroom (with permission) during lessons

   c. how frequently pupils use toilets

   d. places pupils group during morning break

   e. clubs in which pupils are participating

   f. favourite magazine

   g. favourite type of music

   h. commonest shoe size

For all tabulations collect data for boys and girls separately.

Pupils present their results to the class and explain the ways they collected their data and the ways they tabulated the data. They are also to mention valid conclusions they draw from their tabulated data.

2. a. Write an evaluative report on the activity. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did you meet some specific difficulties in co-ordinating the activity?

   b. Present the report to your supervisor.

Section D2: Experiments

The scientific method is used to collect data resulting from an experiment. Unlike in a survey where existing data is collected in an experiment, a situation is experimentally manipulated and the data resulting from the experimentation is collected. This frequently takes the form of collecting data on an experimental group and a control group. To find out whether or not pupils learn expansion of algebraic expression better using concrete manipulatives or using a multiplication table model, the researcher will use two comparable groups of pupils (a pre-test establishes that existing groups are comparable on algebra knowledge at the start of the project, or two comparable groups are formed). Next the two groups cover the topic using one of the two methods. Data from a post-test is used to determine whether or not one of the two methods is superior.

These types of situations are very common: Is medicine X more effective than medicine Y to cure a disease?

Do batteries of factory A last longer than batteries produced by factory B?

At other times the purpose of the experiment might not be comparing but testing, for example a product: Under what pressure do gas bottles explode?

Will cables marked that they safely can be used up to a load of 500 kg indeed not break under a lighter load?

The outcomes of the data analysis of these experiments is for decision taking. Which batteries to buy? Which medicine to bring on the market? How to set the machines making the gas bottles in a factory? What cables to use for a lift so people can use it without fear that the cables might break?

In experiments, with the exception of the variable being changed all, other conditions should remain the same throughout the experiment. For example, if you are looking at the effect of fertiliser on the growth of sorghum plants the seedlings should be of the same height and strength at the beginning of the experiment. The trays with seedlings should be in the same environment (same amount of light, same temperature condition) and only the amount (or type) of fertiliser should vary—ensuring that one tray with seedlings has no fertiliser applied at all (the 'control' tray).

Practice task 3

The objective of this activity is for pupils to design an experiment, and to collect and tabulate data from the experiment.

1. Give different experiments to groups of pupils to enhance discussion when the outcomes are presented to the class.

   Some suggestions are:

   a) What is the reaction time of boys / girls to catch a ruler between thumb and finger when the ruler is released? Is there difference between using left hand or right hand?

   b) How many 'nonsense' words can pupils recall after studying them for x minutes?

   c) How accurately do pupils measure the height of a desk?

   d) How many times does a drawing pin land head up?

   e) How strong a pillar can you make with an A4 sheet of paper?

   f) How high does a table tennis ball rebound from different surfaces?

   g) How fast do pupils react (pressing a buzzer) after seeing a red light (a green bulb and a red bulb are lighted randomly at various intervals)

2. a) Write an evaluative report on the activity. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did pupils control factors that might lead to unreliable results? Did you meet some specific difficulties in facilitating this activity? What did you learn yourself from the activity?

   b) Present the lesson plan and report to your supervisor.

Section D3: Simulations

There are questions people like to be answered for decision making but the data cannot be obtained by surveys or experiments. For example: the spread of a disease among people, animals, trees or plants; the number of the impalas in a wild life park in relation to the number of lions; the (average) length of a queue or the (average waiting) time at a public phonebooth; the waiting time of cars to find a parking place in the car park near a shopping centre.

Data to investigate these questions is difficult or impossible to obtain. Data can be generated in a simulation (using random numbers from calculator or table) as all data involves aspects of randomness.

A note on random numbers.

Random numbers can be found in tables or on calculators having a RANDOM key. Random numbers consist of lists of digits 0, 1, 2, 3, ... 9 which are such that each digit has an equal probability of appearing (p = 0.1 for each digit). The numbers might be listed individually or grouped in some way (for example in groups of three).

The RANDOM key on a calculator produces random numbers. Pressing the key produces a number between 0 and 1 to 3 decimal places (end zeros are not displayed!). The decimal part is used to make a list of random numbers (do not forget the 0 if this is to be the third digit).

Here is a list as generated by the calculator.
817 806 553 734 790 376 668 379 735 916

Suppose you have a group of 70 pupils and you want to select 8. The people can be given the numbers 01 02 03 ... to 70

You change the random number list so the digits are grouped in pairs:
81 78 **06** **55** **37** **34** 79 **03** ..

The person with the bolded numbers will be in the sample, because these pairs of digits lie between 1 and 70.

Another way to select the persons in your sample is multiplying the random number by 70 and rounding to the nearest whole number. The whole number identifies the person to be included. Continue this process until you have the size of sample required.

For example:

A car park has 50 parking lots. It opens at 8.30 am at the same time as the shops near the parking place open. What will be the average time to fill the car park?

Random digits are now used to represent the number of cars arriving each minute.

Digit	Number of cars arriving each minute
0 or 1	0
2 or 3	1
4 or 5	2
6 or 7	3
8 or 9	4

Using the random numbers listed on the previous page:
817 806 553 734 790 376 ...

The first random digit is 8 and this corresponds with 4 cars arriving.

The second random digit is 1, this corresponds with 0 car arriving.

Next digit is 7, corresponding with 3 cars arriving, etc.

So the number of cars arriving are as follows:
4, 0, 3, 4, 0, 3, 2, 2, 1, 3, 1, 2, 3, 4, 0, 1, 3, 3, ...

At the same time a dice is used to determine the number of cars leaving the car park but in the first 15 minutes no cars depart.

Number on dice	Number of cars leaving
1 or 2	0
3 or 4	1
5 or 6	2

A table can now be used to record how many cars are parked each minute until there are 50 cars and the park is full. For example, the table for the first 20 minutes could look as illustrated.

Time (min after opening)	Arrival	Departure	Total cars in park
0	0	0	0
1	4		4
2	0		4
3	3		7
4	4		11
5	0		11
6	3		14
7	2		16
8	2		18
9	1		19
10	3		22
11	1		23
12	2		25
13	3		28
14	4		32
15	0	1	31
16	1	2	30
17	3	2	31
18	3	0	34
19	3	1	36
20	3	1	38
...			

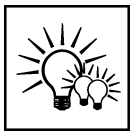
Continuing the table until 50 cars are in the car park gives the time needed to fill the park.

The simulation is repeated several times, or results of pairs of pupils (ensure they start with a different random number) are pooled together.

Averaging the times obtained gives the mean time needed to fill the car park. The model is clearly based on assumption: number of cars arriving per

minute in the range 0 to 4, number of cars leaving is 0, 1 or 2. Early in the morning this might be realistic, towards closing time of the shops obviously hardly any cars will arrive in the park and more will be leaving—arrival and departure will have to be simulated in other forms.

Similar methods as described above can be used to determine the average waiting time in a queue. In each time period (3 or 5 minutes for example) take the last person to join the queue and determine how long the person has to wait. The numbers of people attended to per time period can be simulated by numbers on a dice, the number of people arriving in the queue by random numbers. If the queue exceeds say 30 people, the 31st (or more) will leave since they don't want to wait for so long.



Practice task 4

Objective of this activity is for pupils to collect and tabulate data from a simulation.

1. Give the same simulation to different groups of pupils.You can choose from the suggestions in the above section or design your own. Ensure each group uses different start digits for the random numbers they use. Outcomes of the groups are to be presented and discussed with the whole class.
2. a) Write an evaluative report on the activity. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did pupils bring up ideas to improve on the simulation to make it more realistic? Did you meet some specific difficulties in facilitating this activity? What did you learn yourself from the activity?

   b) Present the lesson plan and report to your supervisor.

Section D3.1: Pupils simulation activity: Dice and disease in the classroom

Below is an outline for a classroom activity with pupils to simulate the spread of a disease. What follows are the lesson outline and notes for the teacher. Work through them before trying out the activity with your pupils in assignment 5.

Assumptions:

- Pupils are aware that germs spread disease.
- You can avoid catching some diseases by avoiding risky encounters.
- The definition of a risky encounter varies with illness. For example, such illnesses as a common cold may be spread by an activity as common as shaking hands, whereas AIDS is frequently spread by sexual contact. Cures do not exist for either illnesses.

Goal of activity:

To model the exponential growth of the common cold, AIDS, or any other communicable disease. The activity underscores the effect that a friend's or partner's previous behaviour may have on a current relationship and on a society at large. The activity works best with thirty or more pupils.

A risky encounter

If the sum is less than or equal to the **cut-off number** 5, the encounter has been risky.

A **risky encounter** means that if either of the pupils is a carrier of the disease, it is passed to the other pupil.

	1	**2**	**3**	**4**	**5**	**6**
1	**2**	**3**	**4**	**5**	6	7
2	**3**	**4**	**5**	6	7	8
3	**4**	**5**	6	7	8	9
4	**5**	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The probability p that an encounter will be risky (cut-off number less than or equal to 5) is

$$p = \frac{10}{36} \approx 0.28$$

If, say, ten infected individuals are in the population at a particular time, those ten individuals will interact with ten other individuals and a fraction, p, of those encounters, approximately three encounters, will be risky and result in new infections.

It is possible to model the effect of an **education campaign**. Suppose that for three stages the cut-off number is set to five, as described previously, and in subsequent stages the cut-off number is reduced to three. This change lowers the infection rate, which could happen if the population were educated about the dangers of risky encounters.

The class activity

The activity is introduced by the teacher reminding the pupils of the points listed under "Assumptions."

The pupils represent a group of people interacting socially. A possibility exists that members of the group are infected with a communicable disease. During interaction between individuals the disease is passed along if an individual has contact with an infected person and has a risky encounter. The aim is to track the spread of the disease and to observe how many people become infected.

- Pupils are NOT informed when encounters will be risky and when not.
- Pupils are given an ID number, an encounter sheet and a pair of dice.

- Pupils walk around the room to ‘encounter’ friends.
- On the sheet, pupils enter the ID number of the person they ‘encounter’ and the score on the pair of dice thrown during this encounter.
- Both pupils involved in the encounter write the information on their data sheet.
- The stage should be timed (approximately 2 minutes) such that pupils can make two to four encounters.
- Teacher announces end of stage 1 and start of stage 2. In stage 2 the same procedure as stage 1 is followed. This is once more repeated in stages 3 to 5. (More stages can be included.)

My ID number

Stage 1

ID number				
Dice score				

Stage 2

ID number				
Dice score				

Stage 3

ID number				
Dice score				

Stage 4

ID number				
Dice score				

Stage 5

ID number				
Dice score				

After the three stages data is going to be collected and analysed. All the ID numbers are on the board.

(i) The teacher announces (and explains the reason for) the cut-off point for risky encounters.

(ii) Pupils identify (circle) on their encounter sheet whether or not their encounter was risky.

(iii) Randomly one ID number is chosen: this is the person carrying the disease (say ID xxx).

(iv) Systematically the teacher guides pupils through the 'encounters' looking at column 1, stage 1.

a) Who met with ID xxx?

b) Was the encounter risky or not?

NO: go to column 2 and repeat a. Who met with ID xxx? and look at whether the encounter was risky or not.

YES: write down (on the board) the ID yyy of the person also infected and go to column 2. Who met with ID xxx or ID yyy? Where the encounters risky or not?

Step by step the first stage is analysed and the total number of pupils infected by the end of stage 1 is noted.

Next analyse stage 2 and 3 in a similar manner. Present the data graphically.

The data can be analysed again (or the simulation can be repeated) with different scenarios. For example, increase or decrease probability of a risky encounter; simulate the effect of an information campaign by lowering the cut-off point for encounters in stage 3 and higher.

Practice task 5

Objective of this activity is for pupils to collect and tabulate data from the simulation.

1. Carry out the simulation dice and disease activity with your class.
2. a) Write an evaluative report on the activity. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did pupils bring up ideas to improve on the simulation to make it more realistic? Did you meet some specific difficulties in facilitating this activity? What did you learn yourself from the activity?

   b) Present the lesson plan and report to your supervisor.

OPTIONAL

Section D3.2: Analytical model for dice and disease (for background knowledge of teacher)

Analytical, or pencil-and-paper, models can also be devised to describe the activity.

Assumption:

The disease spreads through the population in stages and each individual interacts with exactly one other individual in each stage.

Our goal is to produce a sequence of numbers that we denote

$$\{\mathbf{I_0}, \mathbf{I_1}, \mathbf{I_2}, \mathbf{I_3}, \mathbf{I_4}, \mathbf{I_5}, \ldots\ldots\}$$

in which I_n stands for the number of infected individuals at the n-th stage.

The initial population of infected individuals can always be taken as $\mathbf{I_0} = 1$.

We look for a rule that has the form

$\mathbf{I_{n+1}}$	=	$\mathbf{I_n}$	+	$\mathbf{N_n}$
Number of infected at stage n		Number of infected at stage n		Number of new infected at stage n

If I_n infected individuals are involved in stage n, they will interact with I_n healthy individuals, producing

$$p \times \mathbf{I_n}$$

newly infected individuals.

Therefore

$\mathbf{I_{n+1}}$	=	$\mathbf{I_n}$	+	$p\mathbf{I_n}$
Number of infected at stage n+1		Number of infected at stage n		Number of new infected at stage n
	=	$(1+p)\mathbf{I_n}$		

$\mathbf{I_0} = 1$

$\mathbf{I_1} = (1+p)\mathbf{I_0} = (1+p)$

$\mathbf{I_2} = (1+p)\mathbf{I_1} = (1+p)(1+p) = (1+p)^2$

$\mathbf{I_3} = (1+p)\mathbf{I_2} = (1+p)(1+p)^2 = (1+p)^3$

$\mathbf{I_n} = (1+p)\mathbf{I_{n-1}} = (1+p)(1+p)^{n-1} = (1+p)^n$ for n=0, 1, 2, ..

Refining the model

Remove the assumption that infected individuals interact only with healthy individuals.

The number of newly infected people we called N_n is really proportional to the number of interactions between healthy individuals and infected individuals.

Let **T** be the total population.

Assuming no births, deaths or immigration, it follows that the number of healthy individuals at stage n is $\mathbf{T} - \mathbf{I_n}$.

The number of interactions between infected individuals and healthy individuals is proportional to the product

$$\mathbf{I_n} \times (\mathbf{T} - \mathbf{I_n})$$

So we can write

$\mathbf{I_{n+1}}$	=	$\mathbf{I_n}$	+	$\mathbf{N_n}$
Number of infected at stage n+1		Number of infected at stage n		Number of new infected at stage n
	=	$\mathbf{I_n}$	+	$a\mathbf{I_n} \times (\mathbf{T} - \mathbf{I_n})$

Note that the number of newly infected individuals is zero for $\mathbf{I_n} = 0$ or $\mathbf{I_n} = \mathbf{T}$

Section E: Choice of data collection method

The method of data collection to be used depends on the nature of the question one wants to answer, i.e., the hypothesis to be tested.

For example:

a) You want to test the hypothesis that there is no difference in the leisure time activities of boys and girls whatever their age.

You will in this case collect data, using a questionnaire or interview, of a sample of boys and girls of different ages and ask them what their leisure time activities are. Hence you collect data using a survey.

b) You think that the height that pupils can jump depends on the length of their legs. You will collect data by using an experiment: measuring length of legs and the height of the jump of a random sample of pupils (separating boys and girls).

c) Waiting time in queue during lunch time.

Although you could obtain data by observation, this would take a long time to gather sufficient data. A simulation might be the most appropriate way to collect data.

Self mark exercise 8

Which form of data collection would you use in the following cases:

1. Germination rates of seeds.
2. Pupils are better at estimating mass than volume.
3. Litter: causes and possible solutions.
4. Teenage pregnancies: causes and possible solutions.
5. Blind-folded short people can walk in a straight line for a greater distance than blind-folded tall people.
6. Girls have better memories than boys.
7. What is the most popular magazine among pupils in your school?
8. Reaction time with right hand versus left hand.
9. Preferred type of video watched.
10. People are of the opinion that animals should not be used for drug testing.

Suggested answers are at the end of this unit.

Practice task 6

Objective of this activity is for pupils to carry out a small project by setting a question or hypothesis to be researched, collecting and tabulating the relevant data, and stating any valid conclusion from the tabulated data. Allow pupils to choose their own research question.

1. a) Write an evaluative report on the activity. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did pupils bring up ideas to improve on the validity of the data to be collected? Where the research questions realistic? Did you meet some specific difficulties in facilitating this activity? What did you learn yourself from the activity?

   b) Present the lesson plan and report to your supervisor.

Summary

This unit has ranged widely over the topic of how to gather "good" data from a sample. Your teaching outcome should be (eventually) a few classroom projects which engage your students in collecting real data. Don't expect that the projects which your students can tackle will convey all, or even most, of the concepts taught in this unit. And conversely, don't try to teach the unlearned concepts by lecturing. Your students will better grasp the reality of data handling from doing the projects and learning from them.

A good practice is to assign a different data-gathering project to each grouping of two or three students. Each group can add new conceptual layers to its project as the class progresses through the stages of data gathering, representation, and so on.

Unit 2: Answers to self mark exercises

Self mark exercise 1

1. a) Gender and time spend on preparation.

   b) Any other from 1a. such as age, favourite leisure time, etc.

2. a) Using a sample of pupils in which pupils of all classes are represented and about an equal number of boys and girls collect their favourite colour.

   b) Anything different from 2a.

3. a) As the occupation aspired to might depend on the gender, data to collect is: gender and occupation he/she wants to take up.

   b) Anything other than in 3a, e.g., form of the pupil – as all pupils are in your form.

Self mark exercise 2

1. Biased sample as pupils from all classes are not involved. Sample too small.
2. Machines frequently make systematic errors, i.e., every tenth or hundredth item might have a defect. Hence a sample of products produced by a machine should not take equally spaced items.
3. Biased sample as the first production is not representative for the product produced throughout the day. The first batch might be very outstanding as workers are just starting, the ingredients have just been freshly made, etc. A sample should include pies produced throughout the day, i.e., from each batch.
4. Biased sample as the pupils are not representative for the all the pupils in the school.
5. Biased sample. Those not going to the tuckshop are not included!
6. Biased sample as the pupils participating in the mathematics club are likely to have positive attitudes towards mathematics. A sample representative for the whole school should be taken.

Self mark exercise 3

1. The statement is not specific enough as it is not clear what is meant by "better." Leading to better results? Or "better" in the opinion of pupils?

   Hence correct forms could be:

   Pupils prefer group work in mathematics over whole class teaching.

   Pupils' achievement in mathematics is significantly better when they work in groups than when a whole class teaching method is used.

2. No data can be collected to prove or disprove the statement.

   What could be investigated, for example, is pupils' opinion in junior secondary on whether or not they believe that there is life after death.

   The hypothesis could then be that pupils in junior secondary school believe that there is life after death.

3. Not enough data can be collected to give a valid outcome. Data over a very long period 50 – 100 years would be needed to see whether or not a five-year cycle can be discovered.

4. Not specific enough as the word 'often' is open to multiple interpretation.

   Pupils participate in sporting activities more than three times a week.

5. Not specific enough as 'too much' is very relative and needs to be specified.

   Pupils are given homework that will take them daily 2.5 hours to complete according to the teachers.

   Pupils spend between 2 – 4 hours daily to complete their homework.

Self mark exercise 4

1. a) Question is not specific enough. Responses could be "At the weekend" "At 6 o'clock" "When I am working on my homework."

   b) The responses will be so varied that no proper analysis will be possible as the question is not specific enough.

   c & d) Are leading questions

   e) The open question will lead to a wide variety of responses that are difficult to analyse and is biased: it suggests that the news reader need to improve the presentation but people might feel that the news is read all right.

   f & g) Irrelevant questions in relation to the topic: radio listening habits.

   Suggestions for questions and their format to include in the questionnaire are below (need to be worked out in more detail):

   a) Indicate the times and days you generally listen to the radio by completing the table below.

DAY	TIMES (for example 6.00 am to 6.30 am, 5.00 pm to 8.00 pm)
Monday	
Tuesday	
Wednesday	
Thursday	
Friday	
Saturday	
Sunday	

b) What radio programmes do you listen to? Complete the table.

Programme	Listen ALWAYS	MOST OF THE TIME	SOME-TIMES	NEVER
News				
Request programmes				
Top 10				
Educational programmes				
Church services				
… list here the programmes offered on the local radio service ..				

c) How do you rate (tick) the level of the news supplied by the following media?

MEDIA RATING

	VERY GOOD	GOOD	FAIR	POOR	VERY POOR
Daily Times					
Guardian					
Sun					
Radio					
TV					
…. (others)					

d) Indicate whether you feel that more or less time should be given to the following programmes or whether it is OK.

PROGRAMME Time allocated

	SHOULD INCREASE	IS OK	SHOULD REDUCE
News			
Sport			
Request programmes			
Popular music			
Classical music			
Educational programmes			
Religious programmes			
…			

e) How do you rate the reading of the news? (tick your choice)

(i) NEEDS IMPROVEMENT _____

(ii) IS FAIR _____

(iii) IS VERY GOOD _____

If you ticked (i) suggest what improvement is needed.

_____

2. a) Not specific enough to lead to useful data.

How often do you use the library during a week?

Never _____

1 or 2 times _____

3 or 4 times _____

5 or more times _____

b) Why only final year?

In what Form are you? I am in Form _____

c) The question is too general to give data that can be analysed meaningfully.

Better would be to ask:

Which of the following books do you read or borrow from the library? Complete the table.

	NEVER	SOMETIMES	FREQUENTLY
Novels			
Science fiction			
Study books			
Picture books			
Dictionaries			
Encyclopaedia			

d) Biased question suggesting that more science books are needed.

Better to formulate as:

How do you rate the following book sections in the library? Complete the table.

BOOKS	MORE NEEDED	SUFFICIENT	TOO MANY
Novels			
Science fiction			
Picture books			
Dictionaries			
Encyclopaedia			

Study books in

Science			
Mathematics			
History			
Geography			
Art			
Languages			

e) Biased question, suggesting that the journals are outdated.

Rephrase in unbiased format:

The Journals available in the journal section are

Current (latest issues are available) _____

Slightly out of date (latest issue is not available, the issues available are less than a year old) ___

Greatly outdated (issues available are more than a year old) _____

f) Question is too general and generates a wide range of responses, is difficult to analyse and/or does not give the data one is looking for.

Rephrase to make more specific.

How do you rate the following aspects of the library?

Complete the table. If you feel improvement is needed give a suggestion as to how this might be done.

	SATISFACTORY	SHOULD IMPROVE	BY
Sitting space			
Opening hours			
Catalogue			
Service at borrowing desk			
Shelving of books			
Access to books			

3. a) Given personally to (randomly selected) individuals from the target population and to be filled in in your presence.

Advantages

- High return rate.
- Questions not understood can be explained as the person filling the questionnaire can ask for clarification.

Disadvantages

- Time consuming. Hence likely only a relatively small sample can be used.
- Those participating might not feel free to honestly respond as the person is present and knows who filled in the questionnaire (lack of confidentiality).
- Person might quickly fill in spaces (tick randomly responses, skip open questions), as the person is waiting to receive the filled in questionnaire. This might lead to unreliable responses and incomplete questionnaires.

b) Given personally to (random selected) individuals from the target population and to be collected later. The individual can complete the questionnaire at his/her own time.

Advantages

- High return rate.
- Individuals have more time to complete questionnaire (do not feel under pressure) hence higher reliability of the responses.
- If collection not completed method a. can be used.

Disadvantages

- Time consuming. Hence likely only a relatively small sample can be used.
- Confidentiality not guaranteed.

c) Given personally to a (randomly selected) group from the target population and to be filled in in your presence.

Advantages

- High return rate.
- Economical use of time, high number of respondents can be reached at once.
- Confidentiality can be ensured.
- Questions on questionnaire can be explained if respondents ask for clarification.

Disadvantages

- Presence of researcher might reduce reliability (people tend at times to answer the way they think the person asking the question is expecting them to; people might feel under pressure to fill form in quickly).

d) Posting questionnaires to the randomly selected individuals forming the sample.

Advantages

- Economical use of time.
- Confidentiality can be ensured.

Disadvantages

- Can be expensive.

- Often a low return rate.
- Those responding are the ones interested in the topic, those not interested might not respond. This makes the sample biased.
- Questions might be misunderstood by respondents and there is no way to correct this.

4. Biased are

   (i) (ii) (iii) (iv) (vi) (vii) (viii) as an opinion is suggested
   (v) expresses an opinion not all people will agree with

   (i) How do you rate the bus services in the country?

   Very good ___ Good ___ Fair ___ Poor ___ Very poor ___

   (ii) Compare the present bus services with those 5 years ago.
   Do you think that they have
   improved ___ are still the same ___ deteriorated ___

   (iii) What is your opinion on smoking on buses? Tick your opinion.

   Should be allowed _____

   Should be restricted to special areas in the bus _____

   Should be banned completely _____

   (iv) What do you think about the frequency of the bus services?

   Should increase ___ Is sufficient ___ Could be cut back ___

   (v) Should secondary school students travel on special buses?
   YES ☐ NO ☐

   If YES, please explain _____

   (vi) What is your opinion of beauty contests? Tick your choice and give reasons.

   _____ Should be held, because _____

   _____ Should be banned, because _____

   (vii) What is your opinion on the use of face make-up? Tick your choice.

   _____ Should be used, because _____

   _____ Should not be used, because _____

   (viii) What is your opinion on capital punishment?

   _____ Capital punishment should be completely banned

   _____ Capital punishment should be used for the following types of crimes_____

Self mark exercise 5

1/2. A wide variety of responses is possible. Keep the points mentioned on good questionnaire design in mind. Avoid bias and leading questions. The Self mark exercise 4 has given you examples of properly structured questions.

3. None of them, as most people will feel (in their own opinion) that they do enough at home, that they are (reasonably) kind and that their family should not be described with the word 'poor'.

4. a) In general I sleep each night (tick the most appropriate)

   Less than 4 hours _____ 4 – 5 hours _____ 6 – 7 hours _____

   7- 8 hours _____ more than 8 hours _____

   b & c) add a box for USUALLY

Self mark exercise 6

1.

Colour	Tally	Frequency
Red	𝍸 𝍸	10
Blue	𝍸 ||	7
Green	𝍸 |	6
Silver	||||	4
White	𝍸 ||	7
Black	||||	4
Yellow	||	2

Mode: RED

2.

Age	Tally	Frequency
20 –29	𝍸 𝍸 ||||	14
30 – 39	𝍸 𝍸 |	11
40 – 49	𝍸 ||	7
50 – 59	𝍸 ||	7
60 - 69	|	1

Modal class 20 –29

3.

Height	Tally	Frequency
$140 \le h < 150$	||	2
$150 \le h < 155$	|	1
$155 \le h < 160$	𝍸 𝍸 |	11
$160 \le h < 165$	𝍸	5
$165 \le h < 170$	𝍸 ||	7
$170 \le h < 175$	𝍸 ||	7
$175 \le h < 180$	|||	3

b. 14 c. 33

Self mark exercise 7

1. a) 82 b) 568 c) 736

   d) No, 23.1% (1 dp) of the girls and 22.6% (1 dp) of the boys wear glasses.

   To the nearest percent both 23%. There seems to be no significant difference at all.

2. No. For girls and for boys 1 out of 7 is left handed.

Self mark exercise 8

Collect data by experiment for 1, 2, 5, 6 and 8

Use a survey to collect data in cases 3, 4, 7, 9 and 10

Unit 3: Data representation

Introduction to Unit 3

In this unit you will look at different ways to represent data in tables, charts, graphs and diagrams. The emphasis is not on the techniques to produce these representations, but on the question of whether or not the representation best represents the data.

Purpose of Unit 3

The aim of this unit is to look at a variety of ways to represent data and to compare these for the best representation of the data given. The unit will look at : frequency tables, pictograms, bar charts, line bar charts, histograms, pie charts, line graphs, frequency polygons, stem-leaf plots, scatter plots.

Objectives

At the end of this unit you should be able to:

- organise data
- describe data
- read and interpret displays of data
- construct appropriate displays of data: frequency table, pictogram, bar chart, line bar chart, histogram, pie chart, line graph, frequency polygon, stem-leaf plots, scatter plots
- justify the choice of display used for given data
- critically analyse data displays
- state common pupil errors in data representation
- illustrate methods to misrepresent data
- use appropriate project work in the classroom to assist the pupils in their learning of data representation

Time

To study this unit will take you about 10 hours.

Unit 3: Data representation

Section A: Represent or model data

What is the best way to represent the collected data? Is the data discrete or continuous, is the data qualitative or quantitative, how does one change from one form of representation to another, what is the effect of changing scale? These are questions to be considered. Too frequently the emphasis is on operational understanding, on the techniques of drawing a bar chart, a pie chart, a cumulative frequency curve while questions as to why to use the (most of the time) stated representation in the given circumstances (are they valid? are they appropriate?) are hardly considered. It should be left to the pupils to decide what is the most appropriate way to represent their data (and that is the difficult part—the actual drawing of chart is not the problem in general, and could be done by a computer). One way to do this is by comparing different forms of representing the data.

Graphical displays should:

- show the data
- induce the viewer to think about substance rather than about methodology
- represent large data sets in a relatively small space
- make large data sets coherent
- encourage pupils to make comparisons between different pieces of data
- reveal the data at several levels of detail
- serve a reasonably clear purpose
- be integrated with the statistical and verbal descriptions of the data

Section B: Tables

Data collected is generally first tabulated in frequency distribution tables. These tables might contain data that is grouped or ungrouped. Sometimes two-way tables are used.

These were covered in the previous Unit 2, section D1.

Section C: Nature and format of data

The type of representation that can be used depends on

a) the nature of the data, i.e., discrete or continuous data
b) the format in which the data is given ungrouped or grouped

Discrete data

Discrete data can be displayed in bar charts (categorical data), bar-line graphs (discrete quantitative data) or pie charts (categorical data / discrete quantitative, provided the number of categories or discrete values is not large).

In a **bar graph** or **bar-line graph** the height of the bar or line is proportional to the frequency.

Bars are to be drawn separated equally, with same width. The discrete value or category is placed at the centre of the bar. The frequencies, along the vertical axis, are placed against the lines (NOT the spaces). Bar-line graphs are very appropriate with discrete data (number of children in the family, shoe size of pupils, etc.), bar graphs (also called frequency diagrams) are more appropriate for grouped discrete data or for categorical data.

In a **pie chart** the angle at the centre of each sector is proportional to the frequency. Therefore the radius of the pie chart is not relevant. The number of sectors should, generally, not exceed 6 - 8 to make the presentation meaningful and allow comparison between the various sectors.

Continuous data

Continuous data is best displayed in histograms. In a histogram the frequencies are proportional to the area of the bar. In cases where bars of the same width are considered the histogram becomes a bar graph, but the bars touch each other. Details will be discussed below.

N.B. It is rather common to display certain discrete data (for example, scores on a test, number of children in a family, i.e., numerical data that can be ordered) in a bar graph with the bars touching each other. This is strictly speaking not correct, but you should not try to make the distinction with students of this age.

Independent vs. dependent variables

An **independent variable** is presumed to have an effect on another variable. It is the variable that is manipulated or changed by the researcher to investigate the effect on a **dependent variable**. It is also known as the **manipulated** or **experimental variable** that we have discussed above. The effect of the manipulation is observed on the dependent variable. The independent variable is a variable that by itself does not necessarily give rise to the behaviour of interest except if manipulated.

The dependent (or outcome) variable is that variable which occurrence or frequency of occurrence depends on the conditions and the manipulation of the independent variable. It is called the dependent variable because its value depends on and varies with the value of the independent variable.

The independent variable is commonly plotted along the horizontal axis and the dependent variable along the vertical axis.

Write down the different data representations (charts, graphs, diagrams) you remember.

1. What type of data is most appropriately represented by each of the representations you listed above?
2 What type of data cannot be represented by each of the representations you listed?

Section D: Graphical representations

Data can be represented in various ways, and in the following sections you are going to look at the following representations of data.

- Bar charts (Section D1)
- Line/stick graphs (Section D2)
- Histograms (Section D3)
- Pie charts (Section D4)
- Pictograms (Section D5)
- Line graphs/charts (Section D6)
- Frequency polygons (Section D7)
- Stem-leaf diagrams (Section D8)
- Scatter diagram (Section D9)

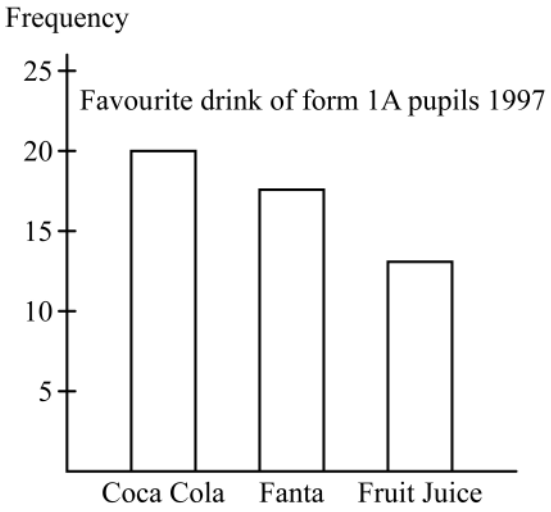
In each section special attention will be given to the type of data that can be represented in that particular way.

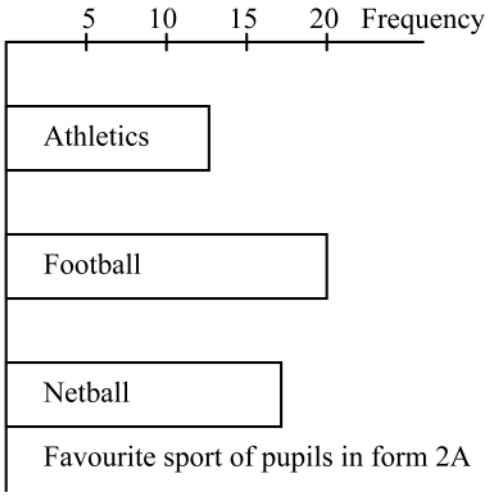
Section D1: Bar charts (bars horizontal or vertical)

Most appropriate use: to compare categories (qualitative data, the independent variable is non-numerical) and grouped discrete quantitative data (scores on a test, amount spend by customers in a shop)

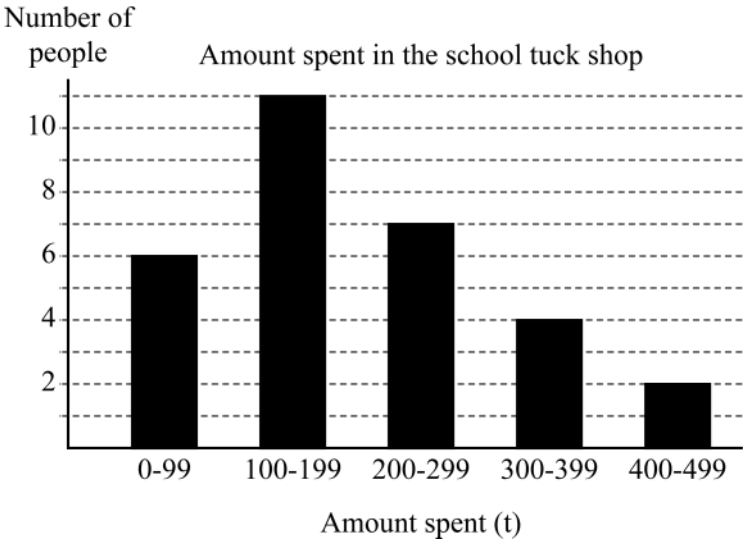
How to draw: Rectangles with equal width are used. The height/length represents the frequency of the category. Do not draw the bar adjacent. Label the diagram as a whole (title), the bars and the frequency axis. Indicate scale on the frequency axis.

Examples of qualitative data display

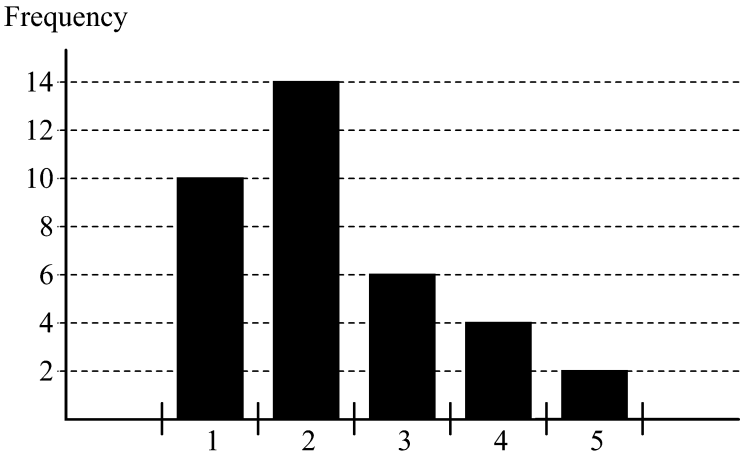




Examples of quantitative discrete (grouped) data display.



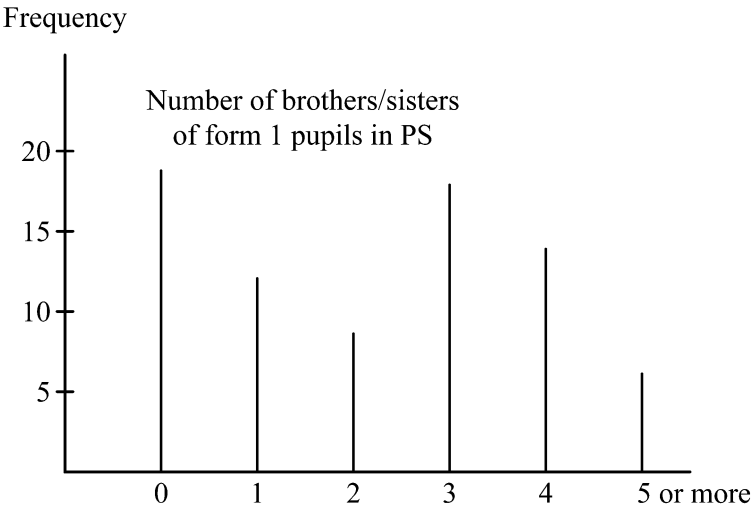
Number of customers in a group



Section D2: Line/stick graphs (can be horizontally or vertically displayed)

Most appropriate use: to compare discrete variables

How to draw: lines/sticks of length proportional to the frequencies.Labelling as with the bar graph.



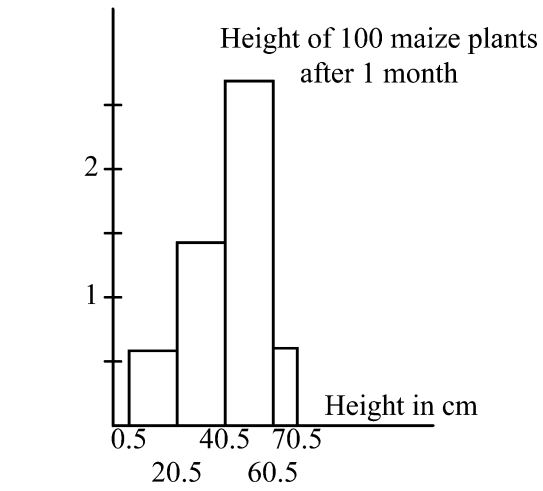
Section D3: Histograms

Most appropriate use: to represent **grouped continuous** variables. Always depicts frequency (or count) versus a continuous or nearly continuous variable.

How to draw: Rectangles whose areas are proportional to the frequencies. The rectangles are adjacent (that is, the rectangles touch each other.) The axes are labelled, the graph has a title.

Example: The height of 100 maize plants was measured, to the nearest cm, one month after planting.

Height of maize plants	Frequency	Frequency density
1 - 20 cm	12	0.6
21 - 40 cm	28	1.4
41 - 60 cm	54	2.7
61 - 70 cm	6	0.6



N.B.

(i) Different notations for the classes are in use, 1 - 20 standing for heights from 1 to 20 both inclusive in the above case. Also the notation [1, 20] or $1 \leq \text{height} \leq 20$ can be used.

Some books use as the first class 0 - 20 to mean $0 \leq \text{height} < 20$ and write the next class as 20 - 40 to imply $20 \leq \text{height} < 40$, etc. The notation to be used is a matter of agreement.

(ii) Attention is to be paid to the upper and lower boundaries. The context dictates how they have to be taken.

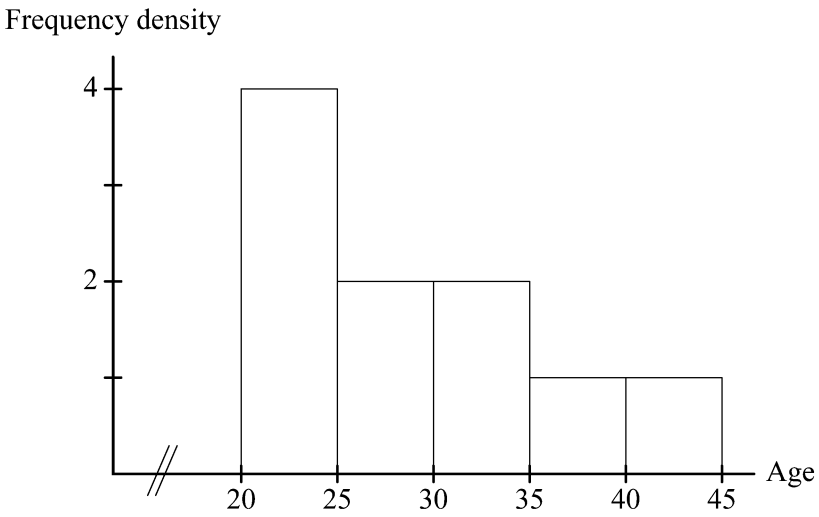
In the above example of measurements to the nearest cm, the boundaries are half way between two classes. The rectangles are to be drawn—in the above example—from 0.5 (the lower boundary) to 20.5 (the upper boundary), from 20.5 to 40.5 and the last one from 60.5 to 70.5.

In case the variable is age (a continuous variable) the situation is different. Ages are given in completed years, not to the nearest year. A person of 20 years and 11 months and 25 days is still considered to be 20. Consider the following example.

The ages of applicants for a teaching post have the following distribution.

Age	Frequency
20 - 24	4
25 - 29	2
30 - 34	2
35 - 44	2

In the class 20 - 24 fall all applicants with age $20 \leq \text{age} < 25$, in the class 25 - 29 fall all applicants with ages $25 \leq \text{age} < 30$, etc. The class boundaries are 20, 25, 30, 35, 45.



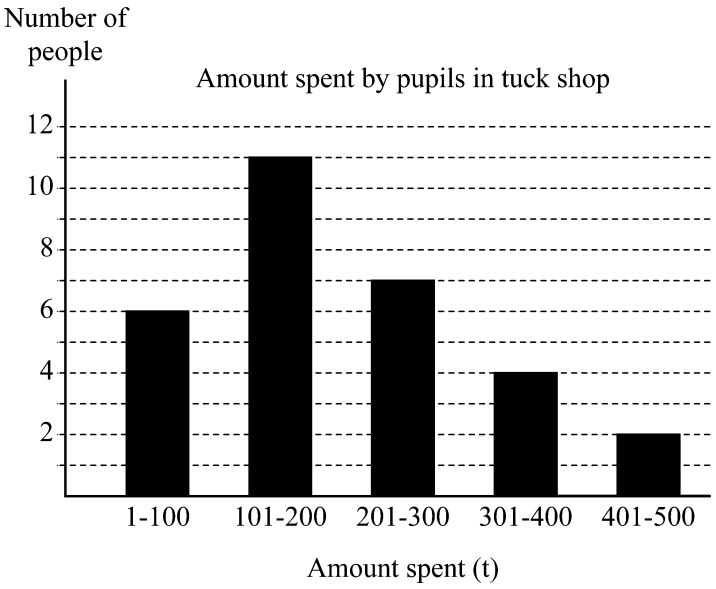
(iii) The frequency density is the frequency divided by the class width (upper boundary – lower boundary of the class). This is a fine point, probably one that you should not teach.

(iv) In cases that classes have all the same width, the frequency density and frequency are directly proportional and some authors will label the axis: "frequencies" in that case.

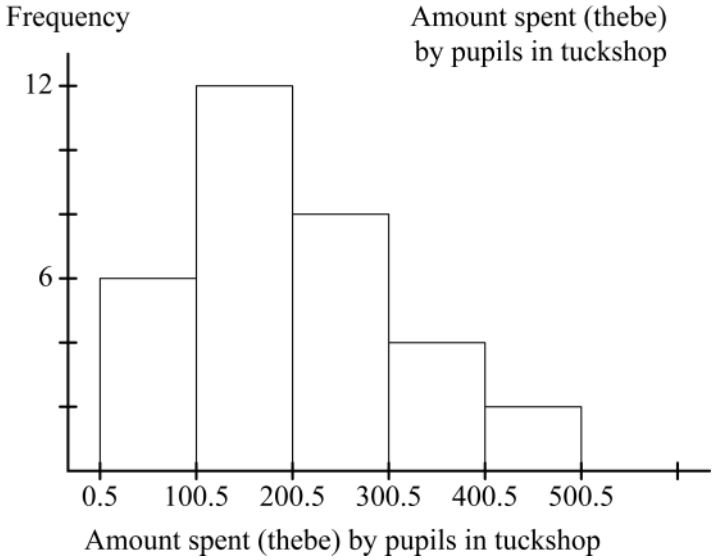
(v) Discrete grouped data (test scores, amount of money spent in a shop) should be displayed in bar graphs. However it is rather common practice to display grouped discrete data as if continuous.

Amount spent by pupils in the tuckshop (thebe)	Frequency
1-100	6
101 - 200	11
201 - 300	7
301- 400	4
401- 500	2

The data is best displayed on a bar chart.



However the data is also displayed in histograms - taking the data as if it is continuous. Class boundaries 0.5 /100.5/200.5, etc. are then correct, but for students of this age, whole-number boundaries on a histogram would be "close enough."



(vi) Grouping data is a means to summarise the raw data. Be aware that by grouping some of the original information is lost.

If, for example, in a test marked out of 10 the scores were:

Mark	0	1	2	3	4	5	6	7	8	9	10
Frequency	3	3	3	2	3	3	4	3	3	2	1

Then by grouping:

Mark	Frequency
0 -1	6
2 - 3	5
4 - 5	6
6 - 7	7
8 and more	6

some of the original information can no longer be found in the grouped frequency table. This has implications for calculation of mean / median / mode. These measures obtained from the raw data will differ from (approximated) values obtained from the grouped frequency table. Changing the class width will again lead to different approximations for the measures of central tendency. Grouping results in what is referred to as "grouping error." The error is reduced by using small class intervals. If the class intervals are increased so does the 'grouping error' in the approximation for the mean and median obtained from the grouped frequency table.

Self mark exercise 1

1. Represent the following data in a bar chart.

   The month in which form 1 pupils in a school were born are tabulated in the frequency table below.

Mon.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Freq.	3	33	21	6	18	30	24	18	54	21	6	9

2. Represent the following data in a bar chart.

   The amount (in thousands of litres) of petrol sold at a petrol station during a month was

Type of petrol	Leaded	Unleaded	Diesel
Number of litres (x 1000)	45	35	20

3. Represent the following data in a bar chart.

   The percent of pupils obtaining a certain grade in a test are listed:

Grade	A	B	C	D	E	F	G
% of pupils	2	5	36	24	15	9	4

4. Use the following raw data of the length (mm) of nails found in packets of 'assorted nails'.

11	48	53	32	28	15	17	45	37	41
55	31	23	36	42	27	19	16	46	39
41	28	43	36	21	51	37	44	33	40
15	38	54	16	46	47	20	18	48	29
31	41	53	18	24	25	20	44	13	45

   a) Make a grouped frequency table taking class intervals 10 -14, 15 - 19, etc., and draw a histogram.

   b) Make a grouped frequency table taking class intervals 10 - 19, 20 -29, etc., and draw the histogram.

      Compare the two representations of the data.

Suggested answers are at the end of this unit.

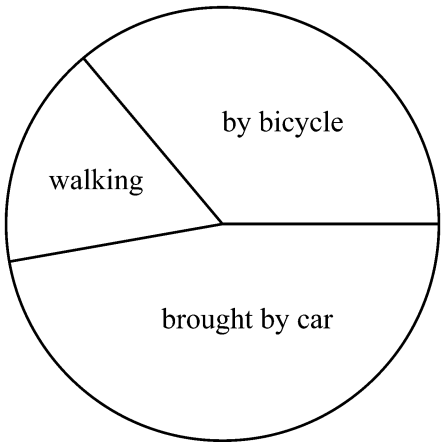
Section D4: Pie charts

Most appropriate use: to represent data as part of a whole, to illustrate differences in categories (qualitative or discrete variables) provided the number of categories is limited (generally between 2 and 8).

How to draw: Measure of the angle at the centre of the circle is proportional to the frequency

(measure of the angle at the centre = $\frac{\text{frequency of the category}}{\text{total of all frequencies}} \times 360^{\circ}$)

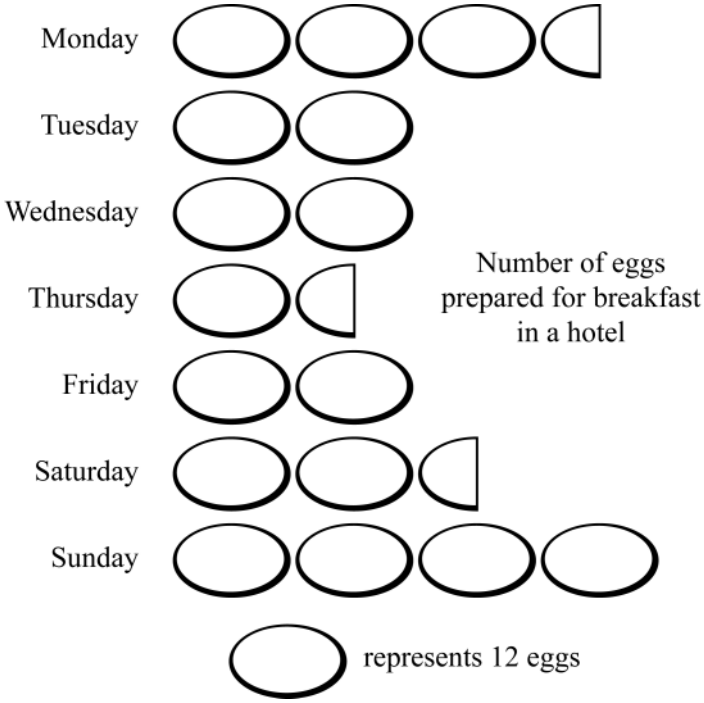
How the form 1 pupils come to Marua Pula



Section D5: Pictograms

Most appropriate use: to illustrate broad differences between categories (qualitative and discrete variables).

How to draw: Draw simple pictures (instead of bars) to represent the frequency. A key is to be added to show what each picture represents.



Self mark exercise 2

1. Display the following data in a pie chart and pictogram.

   The total world wool production was distributed over various countries as follows in 1994:

Country	% of wool world production produced
Australia	30%
USSR	30%
New Zealand	20%
Argentina	10%
SA	10%
Others	10%

2. Display the following data in a pie chart and pictogram.

   The type of vehicles coming to a petrol station during one day are tabulated below

Person cars	26
Lorries	12
Busses	8
Combis	14

3. Display the following data in a pie chart and pictogram.

   The sizes of T-shirts sold during a month in a shop were

Size	S	M	L	XL
Number sold	12	34	56	18

4. Comparing the two representations, pie chart and pictogram, list some advantages and disadvantages of each.

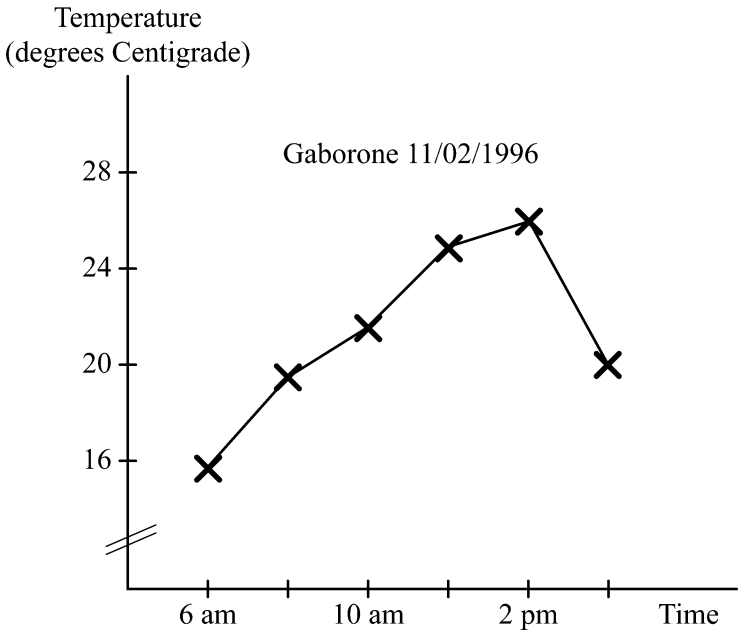
Suggested answers are at the end of this unit.

Section D6: Line graphs/charts

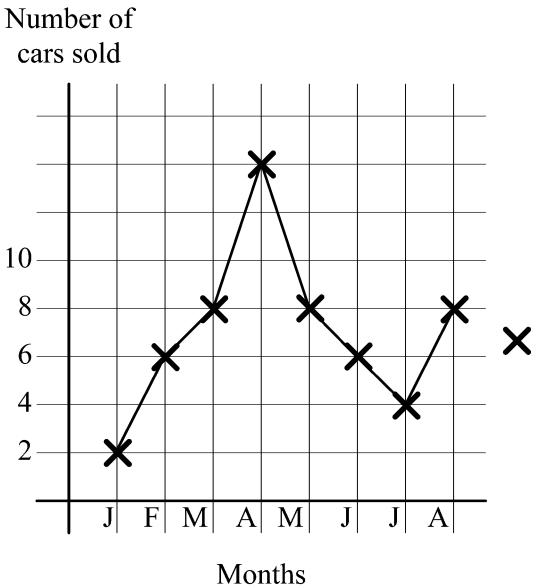
Most appropriate use: to illustrate changes of continuous variables (over time)

How to draw: plot the given corresponding pairs of data as points and join consecutive points by line segments.

Example



If **trends**—changes over time—are looked for, **line graphs** can be used. Line graphs are used for both discrete and continuous data. For example, in the line graph is displayed the number of cars sold in a certain garage over the first 8 months of a year. Although 'in between' values such as $2\frac{1}{2}$, $3\frac{1}{4}$, etc., do not exist the points are joined with straight lines to show the trend. For trend lines not all 'in between' values have to make sense. Trend lines should not be confused with linear graphs. In linear graphs the 'in between' values must exist, otherwise it would be inappropriate to draw the line.



Self mark exercise 3

1 Represent the following data in a line graph and comment on the trend.

Number of pupils in a primary school 1993 - 1999

Year	1993	1994	1995	1996	1997	1998	1999
Number	450	435	465	478	490	510	524

2 a) Represent the following data in a line graph and comment on the trend.

Infant mortality rate per 1000 live births

Year	1971	1981	1991
Infant mortality rate	100	71	45

b) If the trend is continuous what do you expect the infant mortality rate to be in 2001?

3 The number of teacher trained for the Senior Secondary School in Botswana are tabulated:

Year	1996	1997	1998	1999
Number trained	81	105	115	184

a) Represent this data in a line graph and comment on the trend.

b) If the trend is continuous what number of Senior Secondary school teachers do you expect to be trained in the year 2000?

Suggested answers are at the end of this unit.

Section D7: Frequency polygons

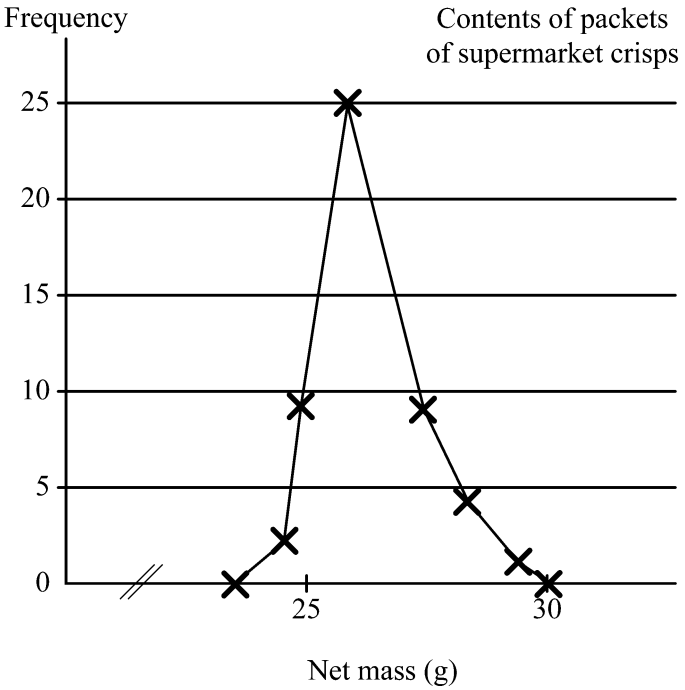
Most appropriate use: to *compare* **grouped continuous** variables, for example, distribution of the height of girls (in grouped form) and distribution of the height of boys (in grouped form). Drawing the frequency polygons on the same axes allows easy comparison. This is far superior to drawing two histograms on one grid!

How to draw: Plotting the points with co-ordinates (mid-interval value, frequency) and connecting these with straight line segments. Adding points on the horizontal axis at both sides.

Example: The following grouped frequency table summarises the net content of 50 packets of a supermarket's brand of potato crisps labelled 25 grams.

Net mass ($\times$ g)	Mid-point	Frequency
$23.5 \leq \times < 24.5$	24.0	2
$24.5 \leq \times < 25.5$	25.0	9
$25.5 \leq \times < 26.5$	26.0	25
$26.5 \leq \times < 27.5$	27.0	9
$27.5 \leq \times < 28.5$	28.0	4
$28.5 \leq \times < 29.5$	29.0	1
	TOTAL	50

A frequency polygon is drawn by plotting the frequency for each interval against the mid-point of the interval and joining the points with straight line segments. The points (24.0, 2), (25.0, 9), ... (29.0, 1) are plotted and joined. The polygon is continued to the axis so the points (23.0, 0) and (30.0, 0) are included.



Self mark exercise 4

1. A factory producing hats measured the circumference of the heads of 100 people in centimetres to the nearest centimetre. The following results were obtained.

Head circumference c (cm)	Frequency
$50 \le c < 52$	8
$52 \le c < 54$	12
$54 \le c < 56$	30
$56 \le c < 58$	44
$58 \le c < 60$	6

   a) Represent the data in a frequency polygon.

   b) If you are to give advice to the factory manager as to the size of hats to be produced what advice would you give? Explain.

2. a) Represent the following data in a frequency polygon. The data gives the mass of apples from one tree.

Mass of apple (g)	20 –30	30-40	40-50	50-60	60-70
Frequency	6	18	34	30	12

   b) Using the same scale and axes, represent the following data giving the mass of the same type apples from another tree in the orchard.

Mass of apple (g)	20 –30	30-40	40-50	50-60	60-70
Frequency	3	14	26	36	21

   c) Comparing the two polygons what conclusion(s) can you make? Explain.

Suggested answers are at the end of this unit.

Section D8: Stem-leaf diagrams

Most appropriate use: To represent ungrouped quantitative data. Also to compare two sets of ungrouped quantitative data (male and female data on the same variable for example). Stem-leaf diagrams are the only graphical representations that also display all the original data values.

How to draw: Part of the number, often the whole number part, is used as the stem (placed vertically under each other), the other part of the number forms the leaves. Place an explanatory legend beneath the diagram and a title above it.

Example:
Potato crisps come in packets marked 25 grams. The mass of 25 packets was found to the nearest 0.1 g.

26.4	25.2	26.3	26.0	24.1
25.3	25.6	26.2	27.8	24.5
25.0	27.5	25.8	26.0	25.7
25.5	26.4	25.5	24.7	26.9
27.3	25.3	25.1	27.7	26.8

Data can be organised in stem-leaf diagram. The whole number part of the mass can be used to form the stem shown at the left of the vertical line and the decimal part of the masses forms the leaves on the right. The leaves on each level or row in the diagram increase in value outwards from the stem.

Contents of packets of crisps (g)

24	157
25	0123355678
26	00234489
27	3578

$n = 25$ 24 | 1 represents 24.1 gram

The 'scale' is very important as 24 | 1 could mean 241, 24.1, 2.41, 0.241, etc., depending on the quantities displayed. The diagram has a title and the sample size $n = 25$ is noted.

The stem-leaf diagram can be stretched by choosing the stem to represent more levels.
For example
24.0 to 24.4
24.5 to 24.9
25.0 to 25.4
etc.

Contents of packets of crisps (g)

24	1
24	57
25	01233
25	55678
26	002344
26	89
27	3
27	578

$n = 25$ 24 | 1 represents 24.1 gram

The stem-leaf diagram can also be made 'double' allowing comparison.
For example, the following stem-leaf diagram gives the height of pupils in a class. The girls are on the left, the boys on the right.

Height of pupils in Form 2X

Girls		Boys
443310	15	2
9865	15	579
43220	16	12234
865	16	5566889
42	17	044
	17	58

$n = 41$ 16 | 8 represents 168 cm

The diagram makes visual that on the whole the girls are shorter than the boys. Or does it?

Self mark exercise 5

1. A sample of eggs from an one day's production has mass in grams:

40	50	72	51	60
55	67	46	57	53
55	42	51	59	49
52	46	64	43	66
54	64	48	58	52

   Draw a stem-leaf diagram using stems 4, 4, 5, 5, 6, 6, 7

2. Represent the following data in (a) a grouped frequency table (b) a histogram.

   Contents of packets of crisps (g)

24	14
24	57
25	0122233
25	55556778888
26	0000112233344
26	56789
27	001134
27	5678

   $n = 50$ 27 | 0 represents 27.0 gram

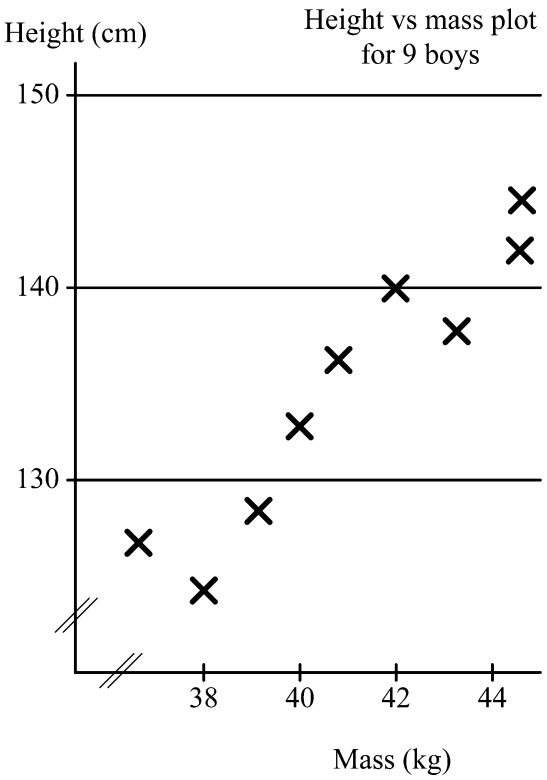
Suggested answers are at the end of this unit.

Section D9: Scatter diagrams

Most appropriate use: When looking at statistical data it is often observed that there are connections between sets of data. For example the mass and height of persons are related: the taller the person the greater his/her mass. To find out whether or not two sets of data are connected **scatter diagrams** can be used.

How to draw: In a scatter diagram each plotted point represent a pair, for example a (mass, height) pair of one person.

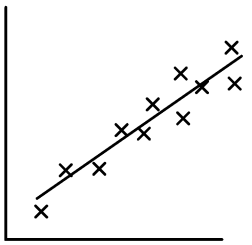
Example



The scatter graph illustrates that generally taller boys have greater mass.

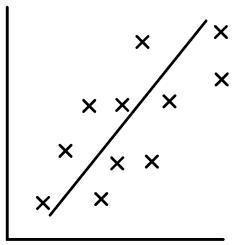
The relationship in a scatter diagram between the two sets of variables is described with the word **correlation**.

Strong positive correlation



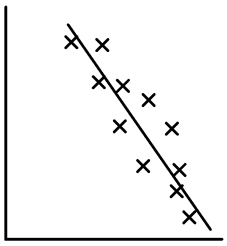
Points are clearly clustered around a line with positive gradient

Weak positive correlation



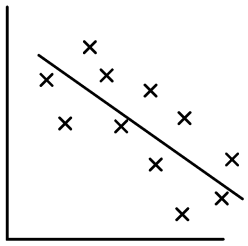
Points are roughly clustered around a line with positive gradient

Strong negative correlation



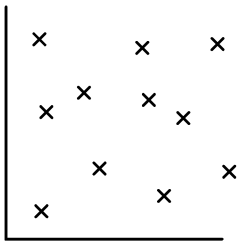
Points are clearly clustered around a line with negative gradient

Weak negative correlation



Points are roughly clustered around a line with negative gradient

No correlation



Points DO NOT cluster around a line

Positive correlation between two variables x and y can be described in words as: if x increases then y will also increase or if x decreases then y will decrease (x and y are directly proportional). Negative correlation between two variables x and y can be expressed in words as: if x increases y will decrease or if x decreases then y will increase (x and y are inversely proportional).

Strong positive or negative correlation between two sets of data *does not* prove that the two variables are **causal** related. For example, the length of a spring and the mass attached to it are likely strongly positively correlated and a greater mass attached causes the spring to extend more. If in a scatter diagram a positive correlation was found between the scores in mathematics of the pupils in a class and the distance they stay from the school (those staying close to school score low, those staying far from school score high) then it is very unlikely that there is any causal relationship (if a pupil moves to a place far from school his/her marks in mathematics are unlikely to increase!). This type of non causal correlation is called **spurious**

correlation, and is surprisingly common.

Estimating values:

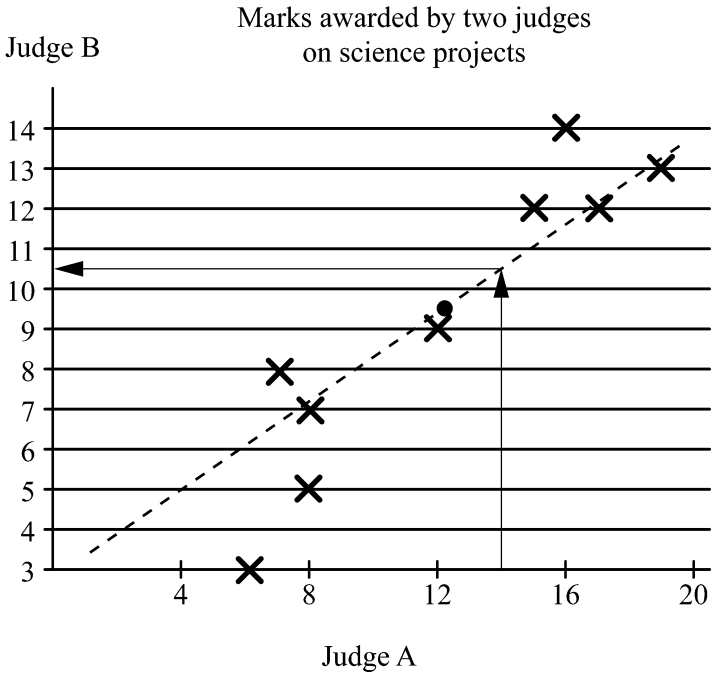
If two sets of data show correlation you can use your scatter graph to estimate missing values. You draw the 'best fitting' line through the point with co-ordinates (mean value of x, mean value of y).

Example:

Two judges awarded marks in a science fair for projects. Judge A scored out of 20 and judge B scored out of 15.

Judge A	15	12	8	19	7	6	17	8	15	16
Judge B	12	9	7	13	8	3	12	5	12	14

Plotting these data in a scatter graph gives the following graph.



The line of best fit is drawn through the point (12.3, 9.5) as the mean score of judge A is 12.3 and the mean score of judge B is 9.5 and such that about the same number of points is at each side of the line.

Judge A scored a project 14 marks but the project was not seen by judge B. You can now use the line of best fit to obtain an estimate for the score of judge B. You find 14 on the Judge A axis. Follow the arrows in the diagram to find the estimate for the mark judge B most likely should have given: 10.4 (or rounded to 10 as only whole marks were awarded).

Self mark exercise 6

1. Do you expect positive (strong or weak), negative (strong or weak) or no correlation between the following variables? Justify your answer.
   a) Number of days absent from school and the mark in the examination
   b) Number of boys and number of girls in a family
   c) Level of education and income
   d) Income and time spent in bars
   e) Age and number of pages read in one hour
   f) Number of rooms in a house and the number of doors
   g) Arm length and length of javelin throw
   h) Circumference of head and intelligence
   i) Length of spring and mass attached to its end
   j) Amount of pocket money and number of friends
   k) Age of a car and its second-hand selling price
   l) The shoe sizes of pupils and the distance they travel to school
   m) Depth of tread on a car tyre and the distance travelled
2. During the term 10 pupils obtained the following scores in mathematics and science (out of 50).

Maths	29	38	22	27	32	41	20	32	36	31
Science	27	34	26	34	27	42	21	28	34	29

   a) Draw a scatter graph to represent the data.
   b) Are pupils' score in mathematics and science correlated?
   c) Find the mean of the maths scores and the science scores to draw the line of best fit.
   d) A pupil scored 35 in mathematics. Work out an estimate for the pupil's score in science.
   e) Another pupil scored 40 in science. Work out an estimate for the pupil's score in mathematics.

3. A supermarket asked randomly some of their customers how many times they come to the supermarket in a three month period and how far away they live from the supermarket.
   The results are tabulated below:

Number of visits	Distance from supermarket (km)
9	10
7	8
12	6
11	7
14	4
12	5
6	12
8	11
15	3
13	4
5	13
10	8
4	15
9	11
2	16

   a) Plot a scatter graph to represent this data.
   b) Describe the correlation of the data.
   c) Find the mean number of visits and the mean distance. Use these to draw a line of best fit.
   d) Estimate the number of visits made by a customer living 10 km from the supermarket.
   e) How reliable do you think the estimate in (d) is?

Suggested answers are at the end of this unit.

Section E: Representing data for understanding

Representing data in diagrams is to enhance the understanding of the data. The question to be asked in each situation is: What kind of representation(s) would help you to make sense of the given data? Too frequently questions are set that prescribe to transform data (frequency table data for example) into a given format (bar chart for example). However it is important that pupils learn to decide what might be the most appropriate format to present their data. Different formats should be considered and within a given format (bar chart for example) the effect of re-scaling or changing class width. Transforming from one format to another is also a skill to be developed.

Practice task 1

1. Find some realistic raw data (3 different sets) and represent the data using several of the above-mentioned forms of representation. Discuss and justify which of the representations is the most appropriate to represent your raw data.
2. The table below gives different forms of representing data and the skill required to transform from one format to the other.

To / From	Verbal	Table	Graph/Chart/ Leaf-Stem/ Pictogram	Formula
Verbal	reformulating, expressing in own words	analysing, extracting data	modelling	analytical modelling
Table	reading	reorganising, regrouping	plotting	fitting
Graph/Chart/ Leaf-Stem/ Pictogram	interpretation	reading off	Re-scaling, using different class width	curve fitting
Formula	explaining	computing	sketching	changing subject

   Illustrate each transformation with an appropriate example.

3. Class based activity

   Split your class into two groups A and B. Present to each group the same data sheet.

   Group A is to use the data to make the country look as good as possible when compared with the other countries.

   Group B using the same data is to present the data such that it shows that the country has much to do to catch-up with the other countries.

   Groups are to be encouraged to use any diagram: bar charts, pie-charts, pictograms, histograms, frequency polygons, scatter graphs, stem-leaf plots.

   The data sheet is on the following page.

   Write an evaluative report on the activity.

Data sheet					
	Botswana	Namibia	Zimbabwe	Zambia	Mozambique
Population density (#/km^2)	2	2	29	52	20
Annual pop. growth	2.9%	3.1%	2.8%	2.7%	3.3%
Children per woman	5.2	6.0	5.5	6.5	6.5
Under 5 mortality	85/1000	120/1000	88/1000	200/1000	292/1000
Doctors	1/5000	1/4600	1/7000	1/10 000	1/38000
Safe water available to *x*% of population	54%	72%	66%	60%	24%
Access to health services	80%	52%	82%	50%	22%
Literacy, male	84%	72%	74%	81%	45%
Literacy, female	65%	48%	60%	65%	21%
Secondary enrolment	54%	41%	48%	20%	10%
University (students/inhabitants)	30/10 000	28/10 000	43/10 000	19/10 000	-
TV sets per 1000 inh.	16	21	26	26	3
Radios per 1000 inh.	122	127	84	81	47
Inflation rate	13%	12%	14%	48%	38%
Food import dependency	75%	31%	5%	7%	22%
Economic growth	6.1%	-1.0%	-0.9%	-2.9%	-3.6%

Pop - population
Inh - inhabitants

Representing data for understanding (continued)

An activity for use in the classroom

Use of visual representations of written text is frequently very helpful to understand the text. The 'standard' representations as used in statistics are not the only way data or text can be represented.

Practice task 2

1. Below are descriptions of five situations. Present these to pupils working in groups and ask them to come up with at least two diagrams to clarify the situation described. Each group could be given one or two situations to represent in diagrams, pictures, charts, etc.
2. After the groups have worked on the activity they are to present their work to the class for discussion. Some of the questions to be asked could be: What is the strength of the suggested representation? What is the weakness? How could it be improved? Are there other alternatives?

   The instruction given to pupils in each of the situations is:
   What kind of representation(s) would help you to make sense of each of the following passages (situation 1 – 5)?
3. a) Write an evaluative report on the activity. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did you meet some specific difficulties in preparing the lesson or during the lesson?

   b) Present the report to your supervisor.

Situation 1: Kidnapped

One of the most influential educationalists in Botswana, Cees, was kidnapped from outside his Gaborone home this morning by masked armed men. Although he was seized in broad daylight on one of the main streets leading to the station fly over, only two eye witnesses have been found by the police and they have been of little help.

Mr. Cees left his home in DO IT street just before 7 am. His driver saw him into the back of the UB INSET van and was driving down the road towards the station flyover when he was forced to pull out to overtake a champagne-coloured Toyota station car that seemed to be vary badly parked.

Immediately opposite was a minibus double-parked and therefore well out from the pavement. The car was forced into what seemed to be an innocent narrow passageway. As the driver was negotiating the gap, a motor cyclist pulled in front of him forcing him to stop. Two masked men jumped out of the back of the Toyota and the motor cyclist pulled a gun.

The men knocked out the driver who was thrown into the minibus that drove off towards Molepolole where he was found dumped along the road a few minutes later. One of the kidnappers jumped into the driver's seat of the UB INSET van and with another holding Mr. Cees at gun point in the back, drove off into the traffic heading for the station fly over. The car was found later at Tlokweng.

The kidnapping was all over in seconds and the witnesses have been able to give only vague descriptions of what they saw.

Situation 2: Ladybirds to the rescue

The Australian cottony cushion scale insect was accidentally introduced into America in 1888 and increased in number until it seemed about to destroy the Californian citrus orchards where it lived. Its natural predator, a ladybird, was artificially introduced in 1889 and this quickly reduced the scale insect population. Later, DDT was used to try to cut down the scale insect population still further. However the net result was to increase their number as, unfortunately, the ladybird was more susceptible to DDT than the scale insect! For the first time in fifty years the scale insect again became a problem.

Situation 3: What foods contain which vitamins?

Vitamin A is found mainly in fats and the fatty parts of some foods, so plenty of milk and butter will help to provide it. Other valuable sources are fish-liver oils and certain vegetables, especially carrots, tomatoes and dark green leafy vegetables. Vitamin D is also found in butter, cheese, milk and eggs but the richest source is fish-liver oils. Sunlight acting on the skin produces vitamin D in the tissue. Vitamin C is found in fresh fruit and vegetables, so plenty of these should be served. In addition orange juice should be given every day. The B vitamins are found in whole meal bread, oats, yeast, liver and dairy foods. One pint of milk a day will supply all the riboflavin (B2) that a child under five needs.

Situation 4: What are you doing in your holiday?

Well it depends. My mum and dad might buy a new car or a colour television or nothing. If we have a car, dad says we will just go for a few day-trips. If we buy nothing we might go to Harare or perhaps Cape Town to stay with Grandma and Grandpa. If we have a colour television, dad says he will only have enough money to travel to Durban and stay for free in the caravan of uncle Sam.

Situation 5: How time works

Emulo gave a description of how to work out what time it is in different parts of the world and completed a table of cities, their longitudes and hence their local times. She wrote:

How time works

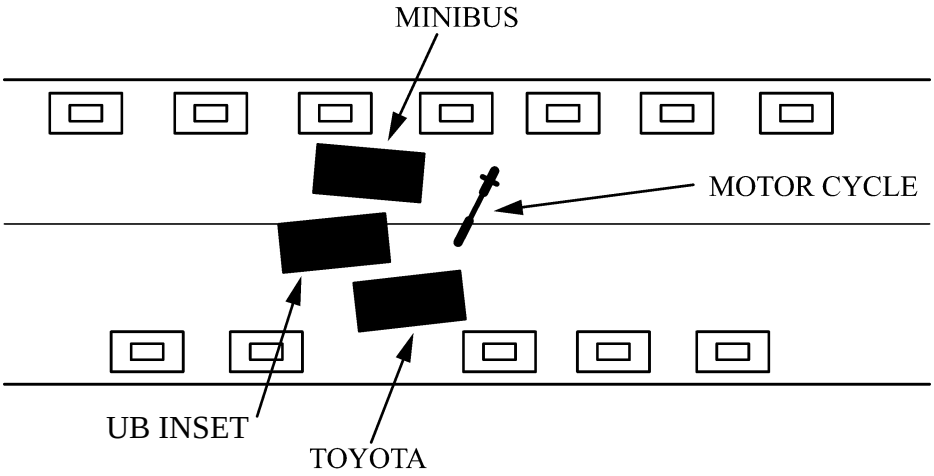
If you want to know what time it is in another place you see what the time is and then whether the country is east or west of Botswana. If it is east every 15° of longitude you pass over you add one hour and if it is west of Botswana you take an hour off Gaborone time as you pass over 15° of longitude. So if it is 10.00 am in Gabs and you want to know what time it is somewhere 60° west of Botswana, fifteen goes into 60 four times so this place is four hours behind Gabs time which it 6.00 am in somewhere 60° west of Botswana.

The verbal description process is rather cumbersome. Is not there an easier way to represent the situation?

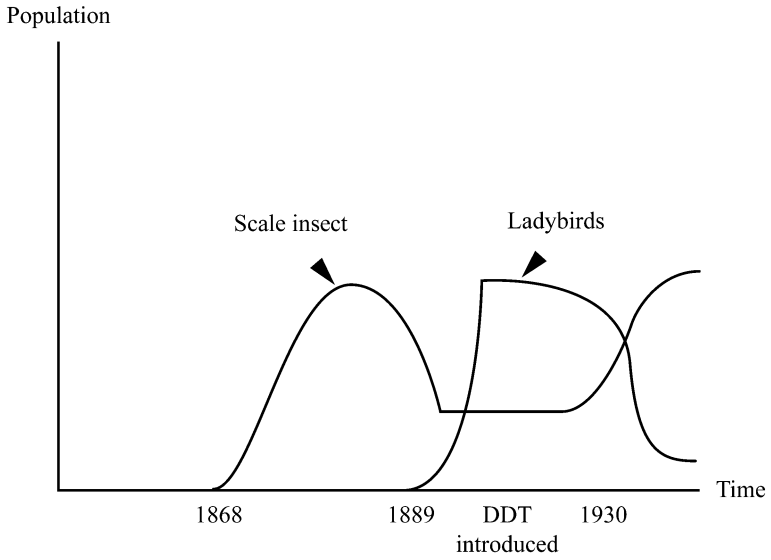
Information for the teacher on the above situations.

Here are a few possible suggestions, but pupils will come up with many others.

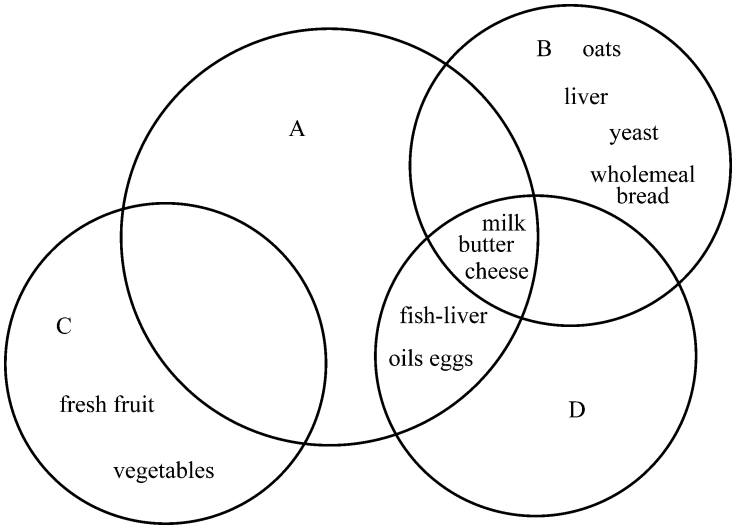
Cees kidnapped



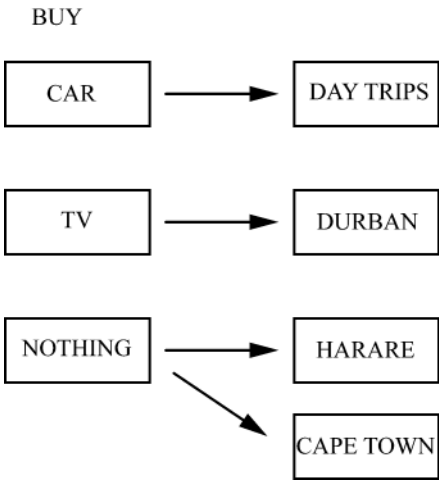
Ladybirds to the rescue



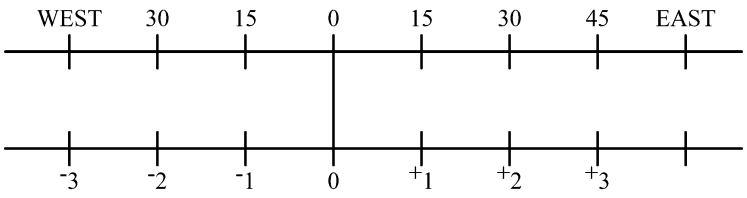
What foods contain which vitamins?



Where are you going for your holidays?



How time works



Section F: Misconceptions of pupils in descriptive statistics

Note: Points 1) through 5) in this list covered common misconceptions in inferential statistics. They were listed in Unit 2 Section C. Below are misconceptions in descriptive statistics.

6) Histograms and bar charts (categorical data: qualitative data) / line charts (discrete numerical data: the possible numerical values are separated from each other by impossible values)

   In a histogram the area is proportional to the frequency, while in a bar chart or line chart the height (or length) is proportional to frequency. The width of the blocks in a histogram need not be equal.

7) Confusing the observations with the frequencies (count) of the observation. Pupils often have a 'surface' concepts of mean, median and mode. Superficially the median is thought of as the 'middle value' and the mode as 'the most frequent value'. But the middle value of what? The most frequent value of what? In a frequency distribution pupils might erroneously take the 'middle' or the 'largest' *frequency* instead of the appropriate *observation*.

8) Difficulties in the interpretation of the meaning of something to *represent* another thing. There are two concepts of 'represent' to distinguish:

   a) the accurate representation. A histogram, for example, represents the sample accurately.

   b) the probabilistic representation. A sample represents a population probabilistically. The confidence in the representation will depend on randomness (being unbiased) of the sampling technique and the size of the sample.

   Pupils tend not to distinguish between these two types of representation. This results in the idea sample = population and if it differs they think the experimenter must have made a mistake.

Self mark exercise 7

In each of the following question 1 & 2:

a) Identify the error / misconception of the pupil

b) Develop activities using realistic data that will

- confront pupils with (common) misconceptions
- resolve the conflict
- consolidate the correct concept

1. An examination was taken by 1000 students and the overall average was 80%. A random sample of 10 examination scripts was taken from the 1000. The first script picked randomly had a score of 60%.

   What do you expect the average of the sample to be?

   Pupil answer: 80%

2. The following data are given:

   Height of 133 plants in cm

Height	160	161	162	163	164	165	166
Frequency	10	15	29	28	24	21	6

   Pupils are to find the mode and the median.

   Pupil answers: mode 29

   median 28

3. Which graphical representation(s) would you use in each of the following situations? Choose from: pie chart, bar chart, histogram, stem-leaf plot, scatter graph. Justify your answer.

   a. Testing the reaction speed of people after drinking a number of cans of beer.

   b. Representing the amount of money Government spends on health, education, armed forces, etc.

   c. Comparing the prices for the same brand of shoes in different shops.

   d. Comparing the salaries of workers in a factory.

   e. Comparing the height of boys and girls in your class.

Suggested answers are at the end of this unit.

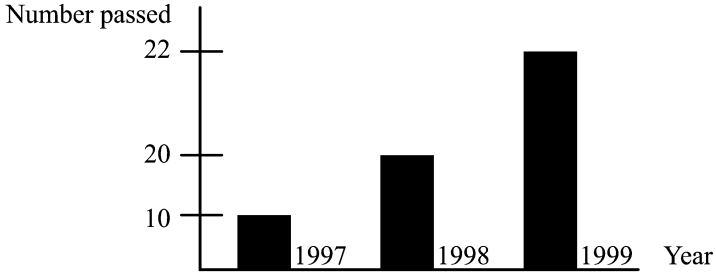
Section G: Making nonsense of statistics

Pupils are to learn to look critically at data presentations. Data is at times presented so as to carry misleading information to a careless observer. The irreverent name for this is "How to lie with statistics."

Common misleading techniques are:

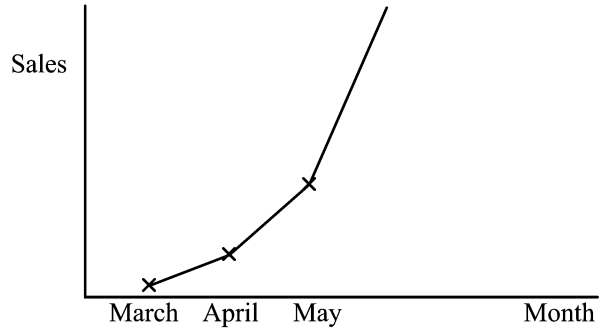
- vertical axis labelled unevenly

More learner drivers are passing in our driving school year after year.



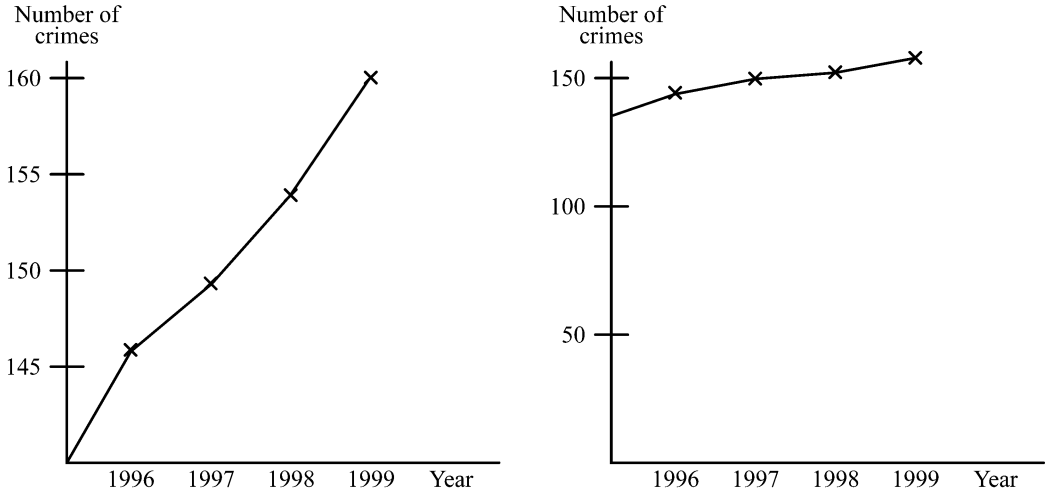
- no scale provided

The sales of our newspaper goes up. Join our readers!



- vertical scale not starting at 0, without clearly indicating this ("squeezed" line)

Crime rate in town has increased very rapidly over the past 5 years.

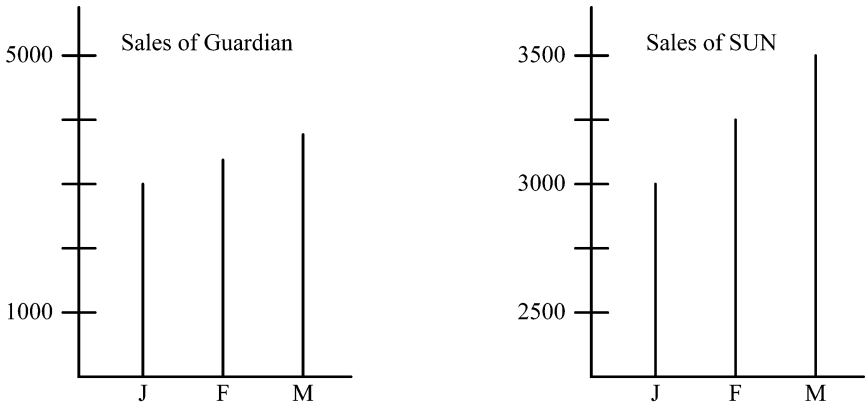


Because the vertical axis does not start at zero the number of crimes appear to be increasing quickly.

In the other graph the data has been redrawn using a different scale and starting from zero on the vertical axis. This illustrates that the number of crimes have very slightly increased.

- different scales used in two displays to 'show' that a certain company or product is doing better

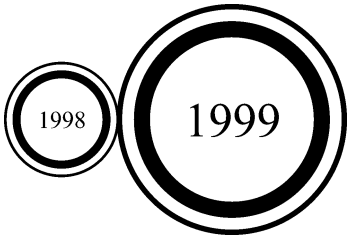
Sales of two newspapers "Guardian" and "SUN" over the first 3 months of 1996 are illustrated below. The SUN claims: Our readership is increasing faster than the readership of the Guardian.



As the scales in the two graphs to be compared differ, a false impression is created: both increased by the same amount from 3000 to 3500.

- using 2D- area or 3D- solid diagrams. To display 'doubling', for example, all dimensions are doubled and hence in the area case the area is 4x the original and the solid is 8x the original.

Our sales of tyres have doubled from 1998 to 1999.



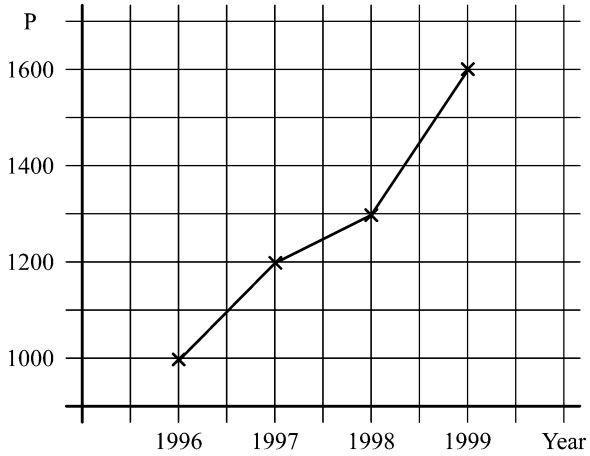
10 000 tyres 20 0000 tyres

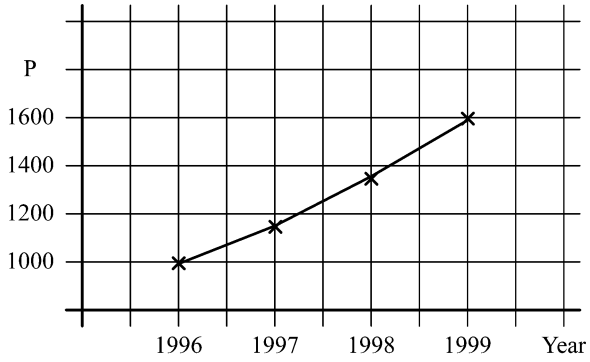
In 'picture graphs' it is the area (or volume that represents the quantities). The radius for the 1999 tyres sales is double that of the 1998 one. However that implies that the area of the circle representing the 1999 sales is 4-times the area of the circle representing the 1998 sales!

Self mark exercise 8

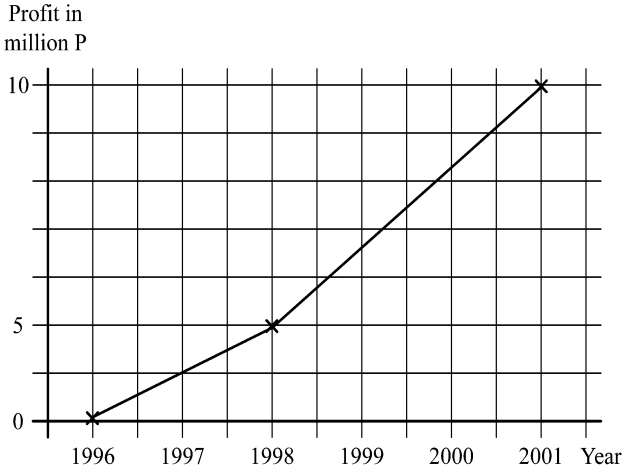
1. Look at the following graphs and explain why they are misleading. Give a better presentation of the data.

   a) "Invest with FAST GROWTH, your investment will grow faster than with any other company"



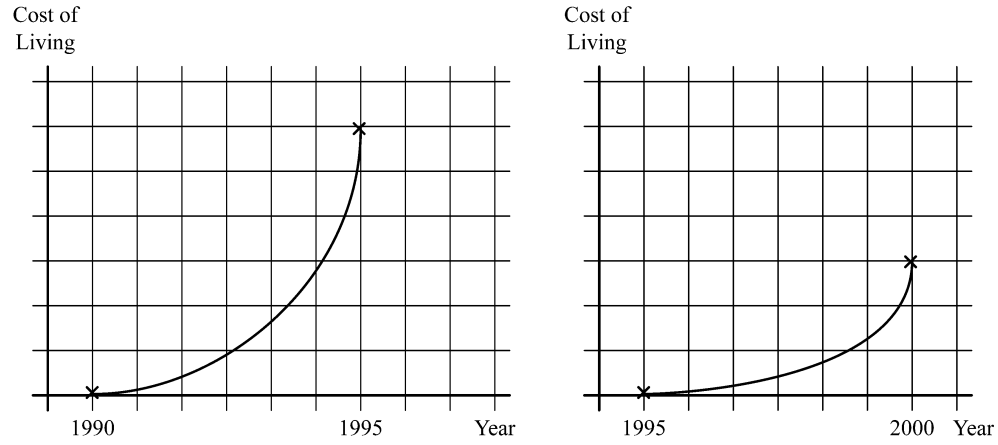


   b) "Our profits have increased faster over the past three years"



Continued on next page

c) "Cost of living increase has been reduced"



2. Chocco Sweets Company wants to impress their shareholders that their sales have increased from 20 000 kg in 1997 to 40 000 kg in 1998 to 60 000 kg in 1999.

   a) Represent the data in a bar chart.

   b) Design a picture graph to create an impression of much greater increase than actually is the case.

3. The following line graph was found in the newspaper with the heading "Unemployment figures have gone down rapidly over the past four years." Do you agree with the newspaper? Justify your answer.

4. The following table gives the number of cars sold in a garage during the first three months since a new sales manager took over.

Month	August	Sept	Oct
Number sold	80	74	64

The sales figures are clearly going down.

   a) You are the sales manager and are to present the figures to the board of directors in a bar chart. As much as possible you want to disguise the dropping sales figures. Draw the bar chart you would present to the board.

   b) You are the supervisor of the sales manager and want to impress on her that since she took over sales figures are dropping. Draw a bar chart you would use to get your message powerfully across.

Continued on next page

5. The following table displays the percent of households in some countries having television sets.

Country	Botswana	Namibia	Zimbabwe	Zambia
% of households with TV	20	26	32	18

a) You live in Zimbabwe and you want to make your country look wealthier than all the other countries. Draw a bar chart with a scale that will help you to make your point.

b) You are living in Zambia and want to make your country look as wealthy as possible as compared to the other countries. Draw a bar chart with a scale that will help you to make your point.

Suggested answers are at the end of this unit.

Section H: Interpreting data

In magazines and newspapers you frequently come across data representations in a variety of forms. Data representations need to be looked at critically. Reading and interpreting graphical representations of data is not a trivial task. Many diagrams, charts or graphs in newspapers and magazines have been designed to magnify differences or to emphasise minor points. You are to ask yourself questions such as: How was the data collected? Does the representation give a fair picture of the data? Are the data reliable? What purpose do the presenters of the data have?

Self mark exercise 9

For each of the following seven data representations answer the following questions:

1. What is the representation about?
2. What type of data is represented (qualitative, quantitative; discrete, continuous; grouped, ungrouped)?
3. What type of graph is it?
4. How do you think the data was collected?
5. What conclusions can you draw from the diagram?
6. What other representation would you consider appropriate for the data? Justify.

Suggested answers are at the end of this unit.

Fig 1:

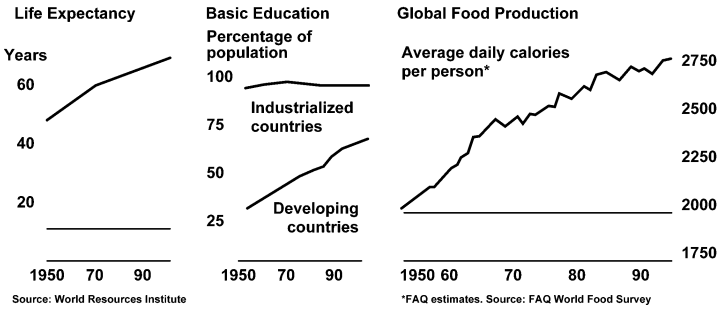


Fig 2:

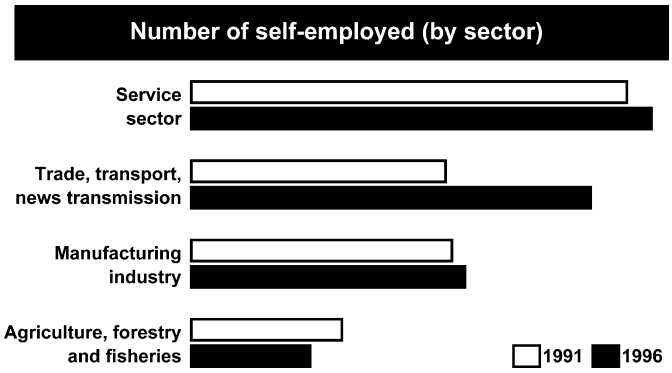


Fig 3:

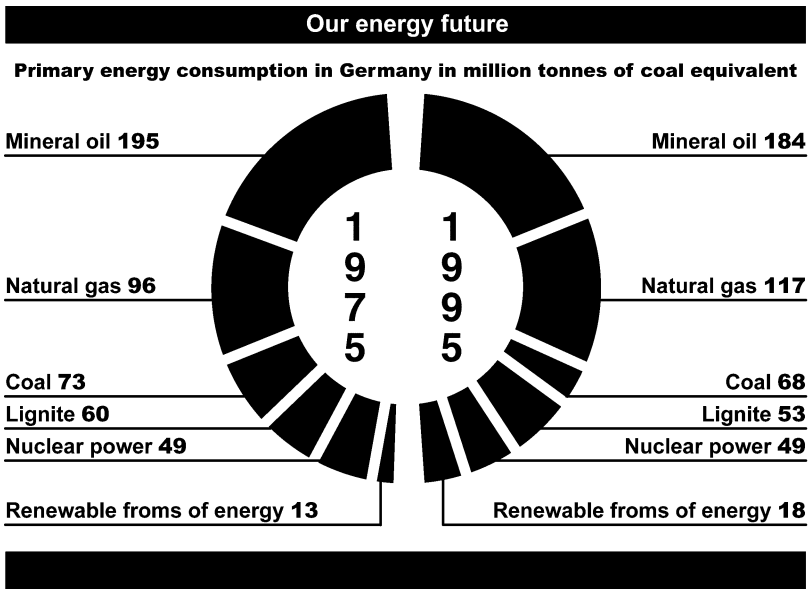
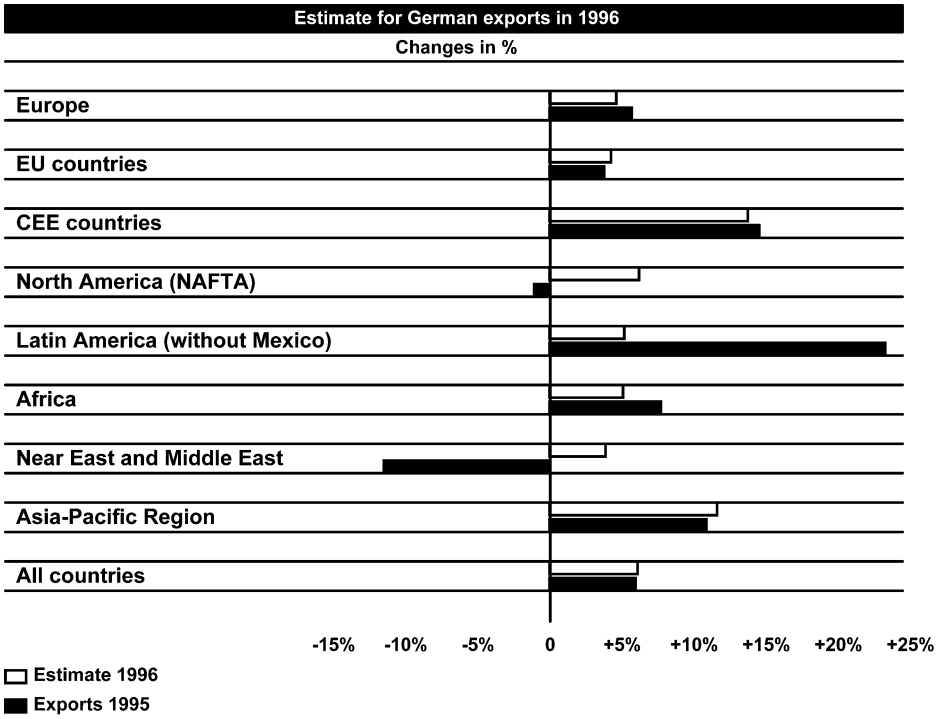


Fig 4:



Source: DIHT

Fig 5:

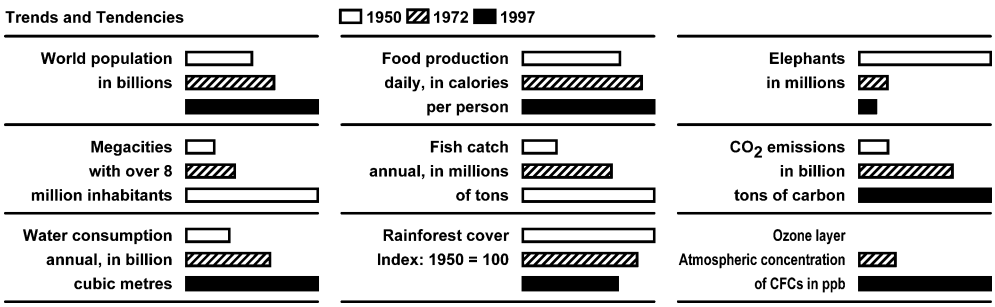


Fig 6:

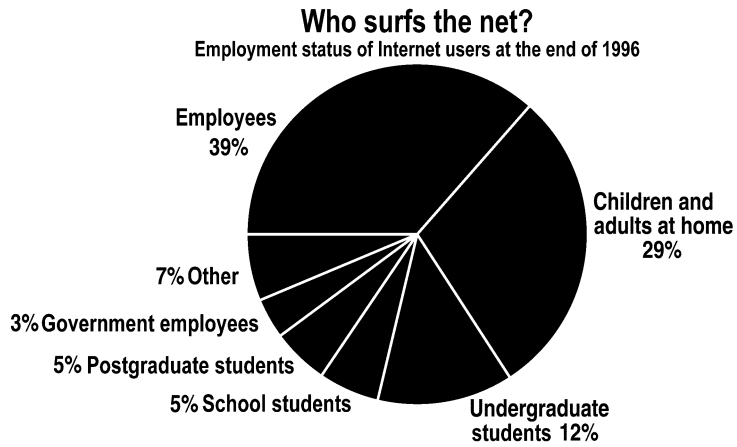
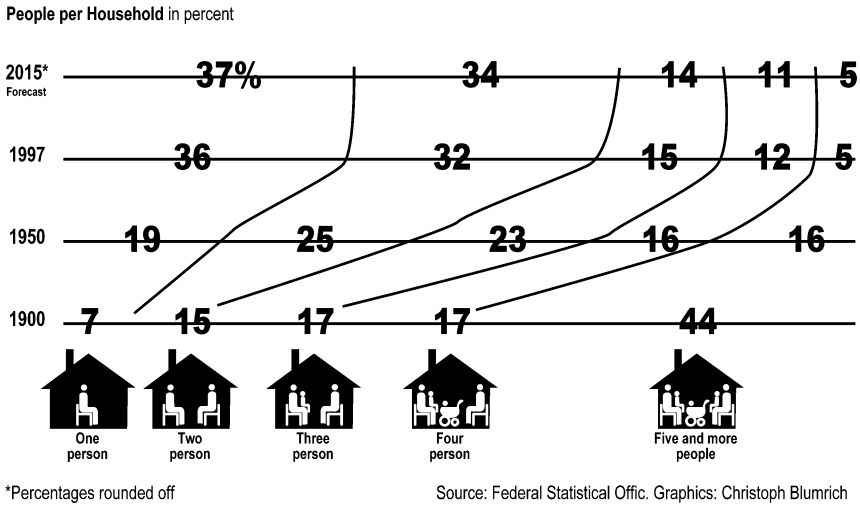


Fig 7:



Practice task 3

1. Choose one (or more) of the data representations in Section D, or an activity related to Section G or H.
2. Write a lesson plan with clearly stated objectives. Prepare worksheets for the pupils to work in groups.
3. a. Write an evaluative report on the lesson. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did you meet some specific difficulties in preparing the lesson or during the lesson?

   b. Present the lesson plan and report to your supervisor.

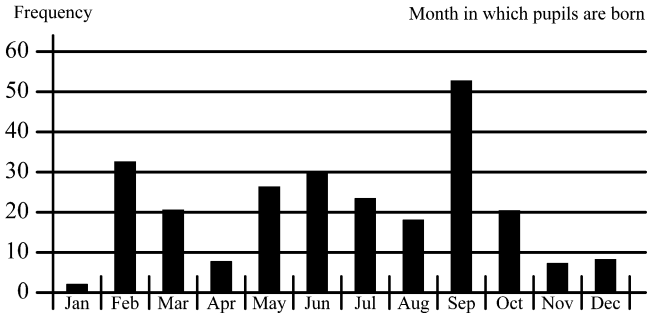
Summary

Data handling and interpretation is one of the few topics in Secondary Mathematics that impacts the daily lives of most thinking people. Graphs, like those in Section H and like the biased ones in Section G, abound in the media. Students should also produce representations of the data they gathered in their own projects—and, if possible, recommend a decision that could be taken on the basis of their work. Groups will learn a great deal from presenting their work to the class as a whole.

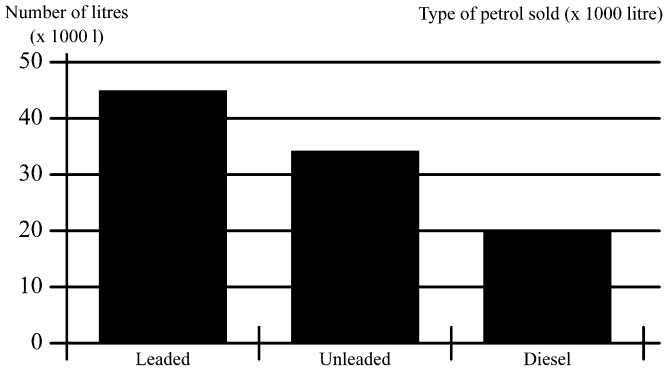
Unit 3: Answers to self mark exercises

Self mark exercise 1

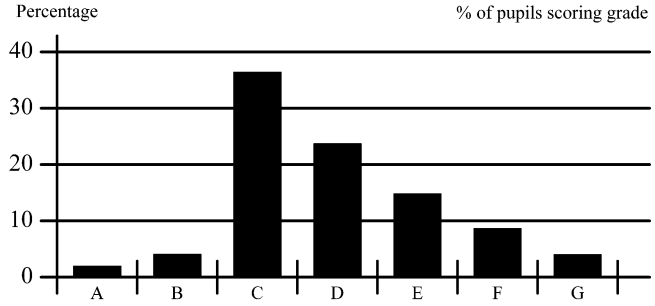
1.



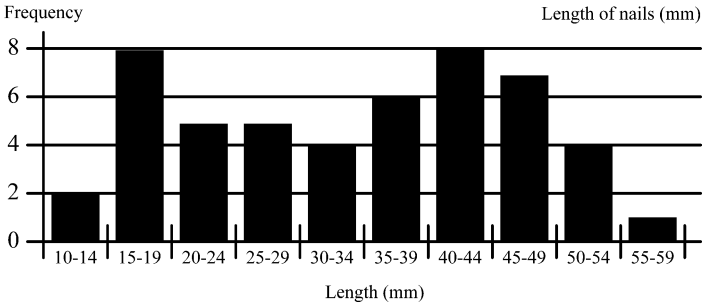
2.



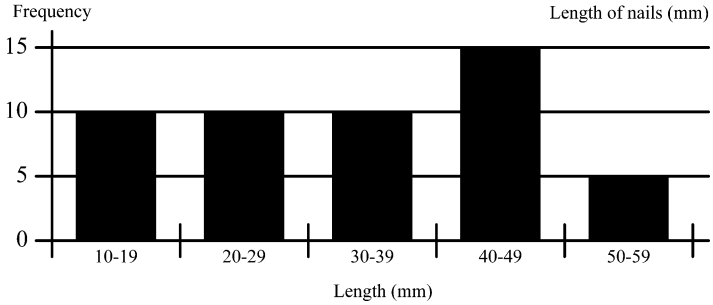
3.



4a Class interval	Frequency	4b Class interval	Frequency
$10 \le l \le 14$	2	$10 \le l \le 19$	10
$15 \le l \le 19$	8		
$20 \le l \le 24$	5	$20 \le l \le 29$	10
$25 \le l \le 29$	5		
$30 \le l \le 34$	4	$30 \le l \le 39$	10
$35 \le l \le 39$	6		
$40 \le l \le 44$	8	$40 \le l \le 49$	15
$45 \le l \le 49$	7		
$50 \le l \le 54$	4	$50 \le l \le 59$	5
$54 \le l \le 59$	1		



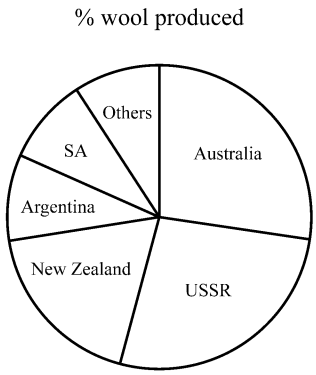
4b.

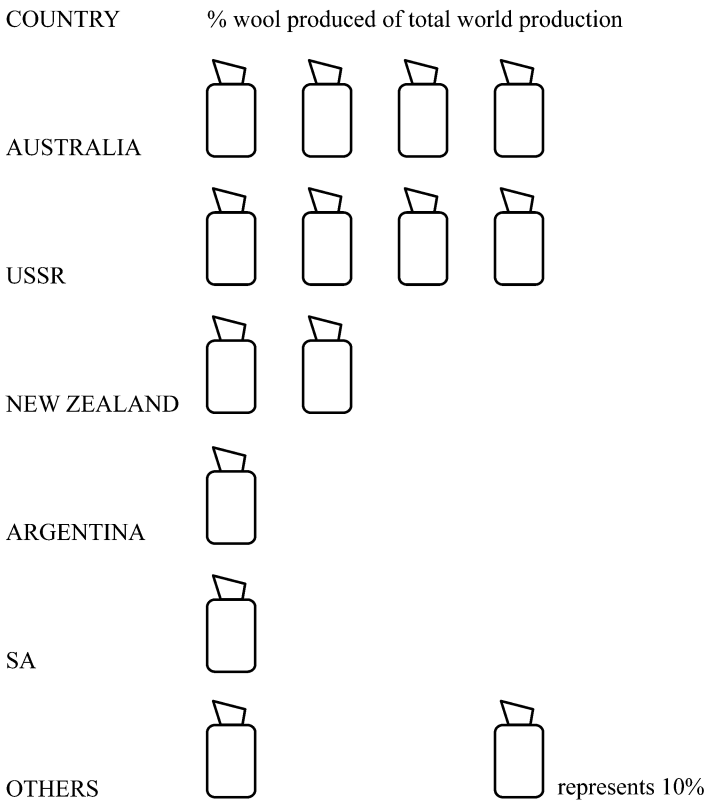


Changing the class width to 10 (as in 4b) gives a less informative representation of the data, as some information is lost. On the other hand, as the number of bars is reduced, the diagram becomes easier to read.

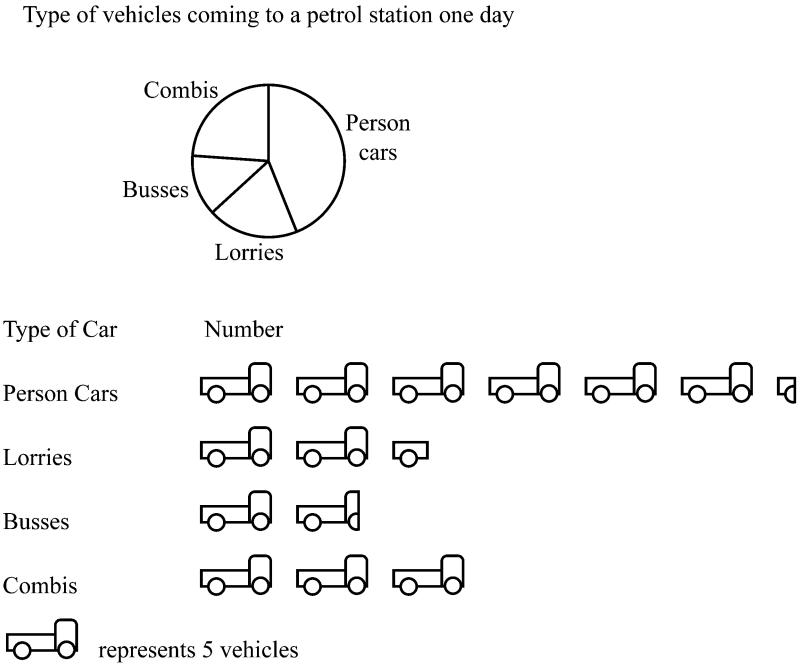
Self mark exercise 2

1.

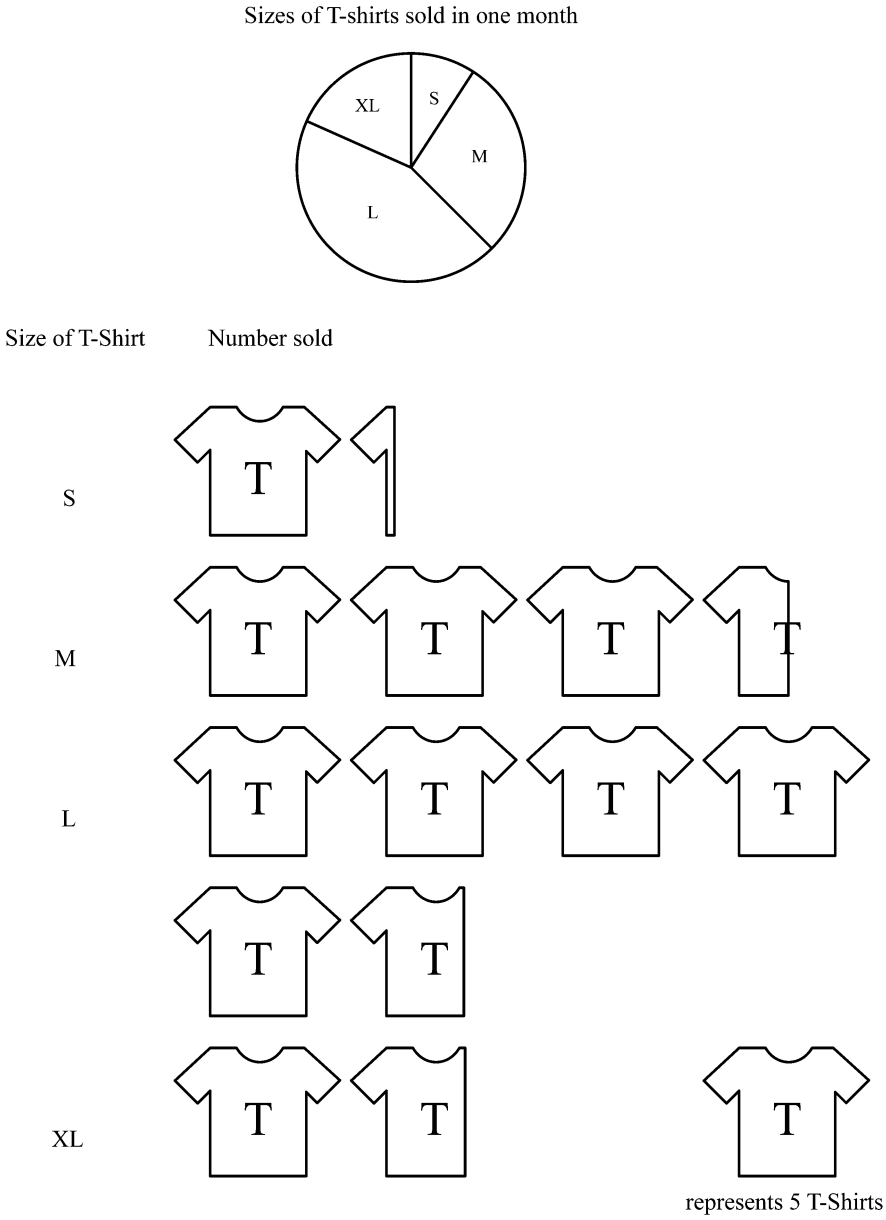



COUNTRY
% wool produced of total world production
AUSTRALIA
USSR
NEW ZEALAND
ARGENTINA
SA
OTHERS
represents 10%

2.


Type of vehicles coming to a petrol station one day
Combis
Person cars
Busses
Lorries
Type of Car
Number
Person Cars
Lorries
Busses
Combis
represents 5 vehicles

3.



4. **Pie Chart**

Advantages

- Allows easy comparison of parts with whole

Disadvantages

- At times tedious to calculate the sector angles
- The actual frequencies are not shown and need to be obtained by interpreting the chart

Pictogram

Advantages

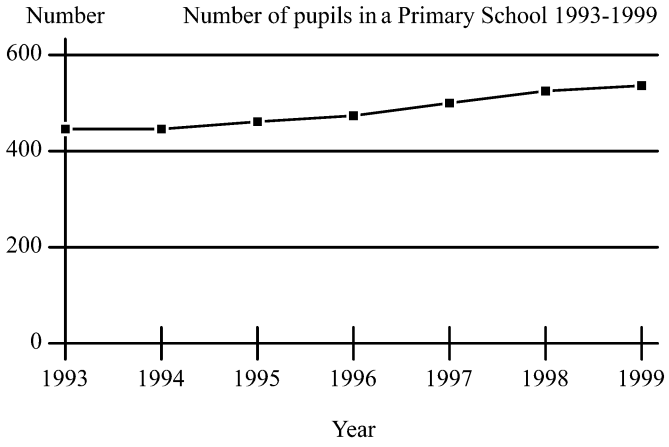
- Can be made visually attractive
- Pictures make 'topic' clear

Disadvantages

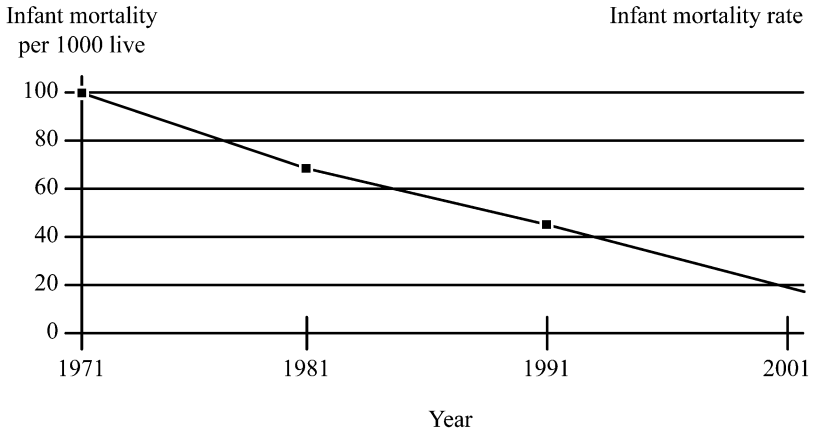
- Hard to draw
- 'Fractional' pictures difficult to interpret

Self mark exercise 3

1.

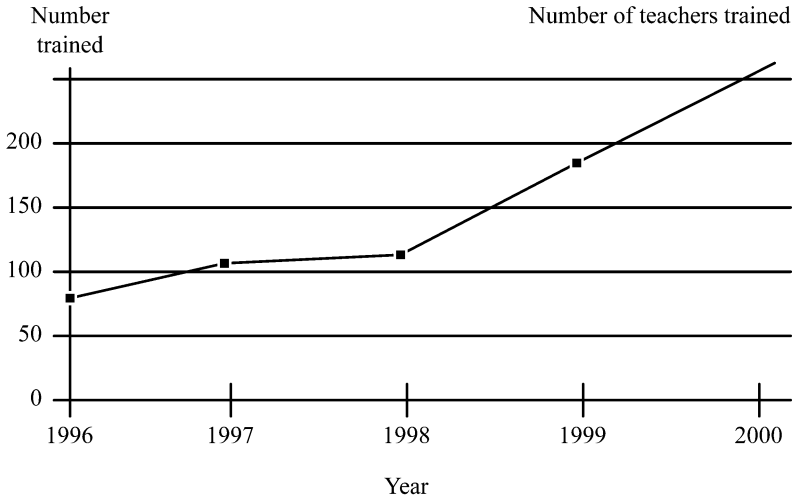


2a)



2b) 24

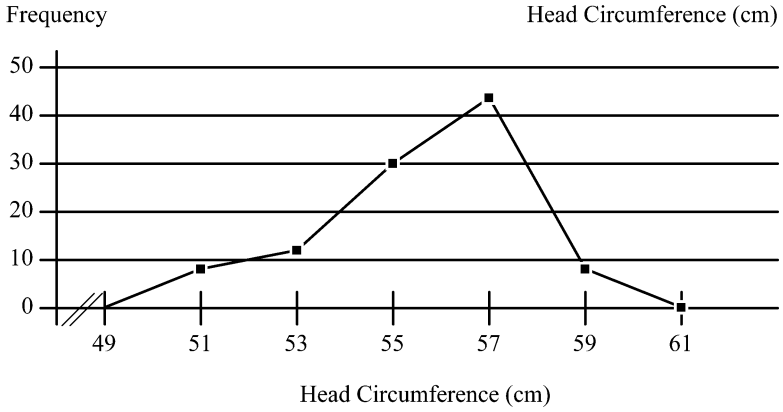
3a)



3b) 260

Self mark exercise 4

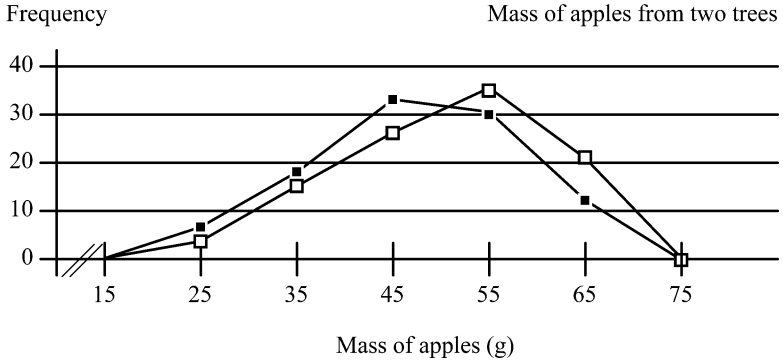
1. Plot the points ((49, 0), 51, 8), (53, 12), (55, 30), (57, 44), (59, 6), (61, 0)



1b) Produce 45% size 56 - 58, 30% of size 54 - 56, 15% of size 52 - 54, and 55 each of size 50 - 52 & 58 - 60.

2a) Plot (15, 0), (25, 6), (35, 18), (45, 34), (55, 30), (65, 12), (75, 0)

2b) Plot (15, 0), (25, 3), (35, 14), (45, 26), (55, 36), (65, 21), (75, 0)



2c) Both trees produced 100 apples. The apples from the second tree on average have a greater mass. The 'top' of the graph of the masses of the apples from the second tree is more to the right.

Self mark exercise 5

1.

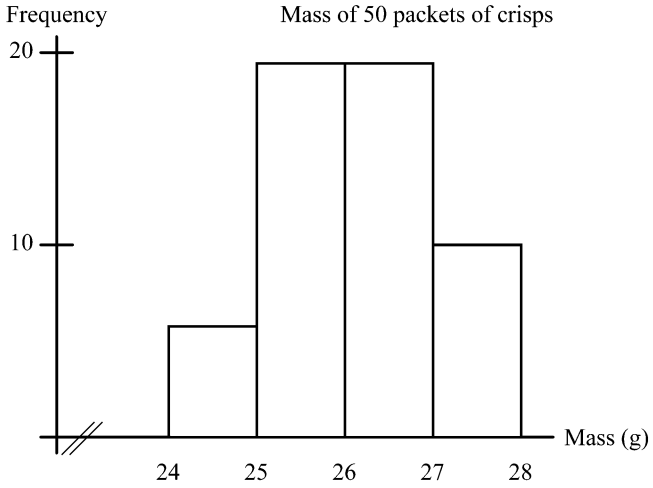
4	023
4	6689
5	0112234
5	55789
6	044
6	67
7	2

n=25 4 | 6 represents 46 g.

2a)

Mass	frequency
$24 \leq m < 25$	4
$25 \leq m < 26$	18
$26 \leq m < 27$	18
$27 \leq m < 28$	10

2b)



Self mark exercise 6

1a) Weak / strong positive correlation (as it will depend on the achievement level of the pupil. A high achiever will not suffer much from being absent, but an average or low achiever will be affected by missing school days).

b) No correlation.

c) Weak / strong correlation. People with higher levels of education generally will get better paid positions. However in business, people without much formal education can become business tycoons.

d) Difficult to predict. One might argue for different types of correlation. E.g., positive correlation - people with higher incomes have more money and can spend more time in bars.

Negative correlation - people with low incomes spend the little they have on drinks in bars, while the people with higher incomes prefer to drink in their houses and not in public places.

No correlation - a group of people with low incomes as well as a group of people with high incomes spend time in bars.

e) No linear correlation. Children take a long time to read a page and with an increase in age they might start to read faster, but at an older age the speed might go down again.

f) Strong positive correlation. Each room will need (at least one) doors to enter. More rooms leads to more doors.

g) No correlation expected. Technique and strength of muscles might be more relevant than arm length.

h) No correlation.

i) Strong positive correlation. The more mass attached the more the spring will extend.

j) No correlation.

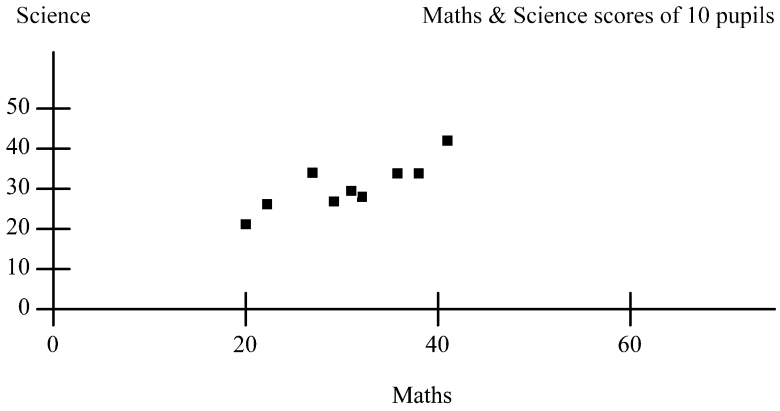
k) Negative correlation. The older the car the less its value will be.

l) No correlation.

m) In the form the question is set: no correlation as a car might have covered many thousands of kilometres but just have new tyres.

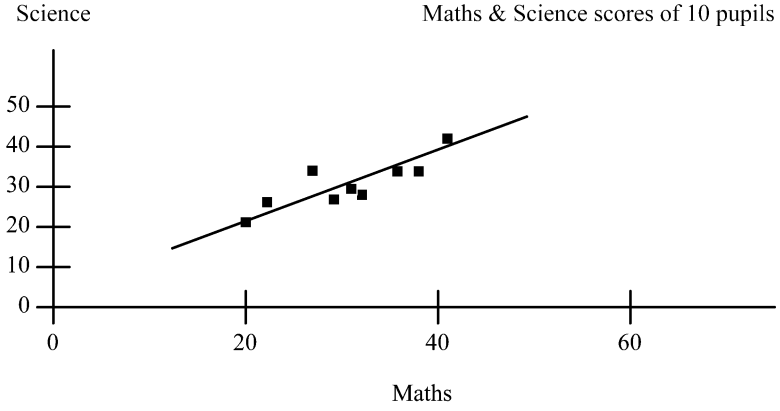
If the question is intending to say: distance travelled with those tyres, then there is a strong negative correlation. Depth of tread reduces if number of kilometres travelled with those tyres increases.

2a)



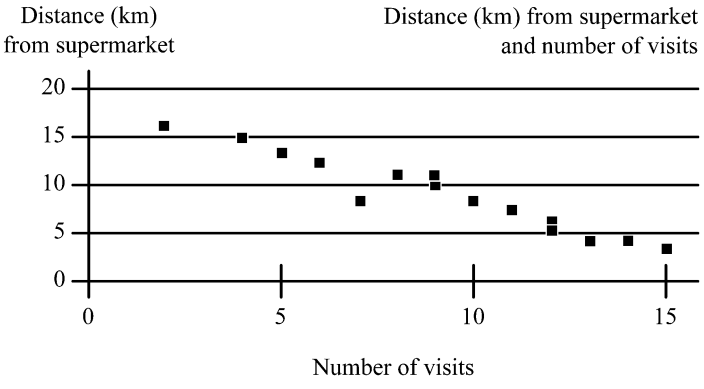
b) A positive correlation exists between the scores.

c) Mean Maths 30. 8, mean science 30.2.



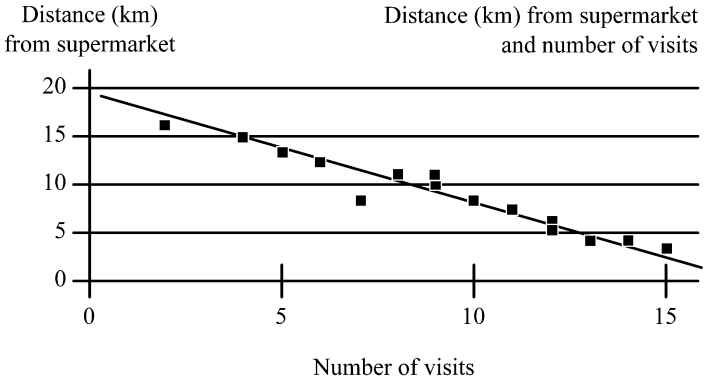
d) 35 e. 40

3a)



b) The data are negatively correlated, i.e., the number of visits to the super market decreases when the distance the person is away from the supermarket increases. Or the number of visits increases if the distance the person is away from the supermarket decreases.

c) Mean number of visits 9.13, mean distance 8.87



d) 8

e) As the correlation is rather strong the estimate is reasonable reliable.

Self mark exercise 7

1a) The pupil thinks that the sample must have the same mean as the population.

b) (i) Ask pupil to explain how (s)he worked the question.

(ii) Create mental conflict. Suppose only 2 scripts were taken at random. The first has a score of 55%. What is the expected mean of the sample of two?
Here the pupil will discover it cannot be 80% because then the second script would need a mark of 105%, more than is possible.

(iii) Guided questioning should bring out:
One script has score 60%, the other nine scripts, we don't know, but looking at the mean we have as our only option to assume that those nine will have an mean of 80%. Hence the expected mean of the sample will be $(60\% + 9 \times 80\%) \div 10 = 78\%$.

(iv) Consolidate the 'new' knowledge of the pupil by setting similar questions.

2a) Pupil confuses observations with frequencies. Mode and median are observation related, not frequency.

b) (i) Ask pupil to explain how (s)he worked the question.

(ii) Create mental conflict by looking at data set
160, 160, 160, 161, 161, 162, 163

Ask pupil for mode and median.

Next ask pupil to place the data in a frequency table thus:

Height	160	161	162	163
Frequency	3	2	1	1

Ask again for mode and median.

Ask pupil what is to be looked at for mode and median, the height row or the frequency row.

(iii) Guide the pupil through the set problem, which will be not so hard after step (ii). The pupil is now aware of the error and will work with the correct row.

(iv) Set similar questions—given a distribution table of discrete data, to find median and mode—for consolidation purposes.

3a) Line graph to see whether or not there is a trend.

b) Bar graph if the number of categories turn out to be large (8 or more). Otherwise use pie chart to compare the amounts spend on each category.

c) Bar chart or bar line chart, the height of the bars giving the prices.

d) Pictogram / bar chart if comparing discrete salaries (or range of salaries) of different groups of workers. If the data is to display how many workers earn a salary in a particular range a histogram might be appropriate (unequal classes).

e) Frequency polygons on the same axes will allow easy comparison.

Self mark exercise 8

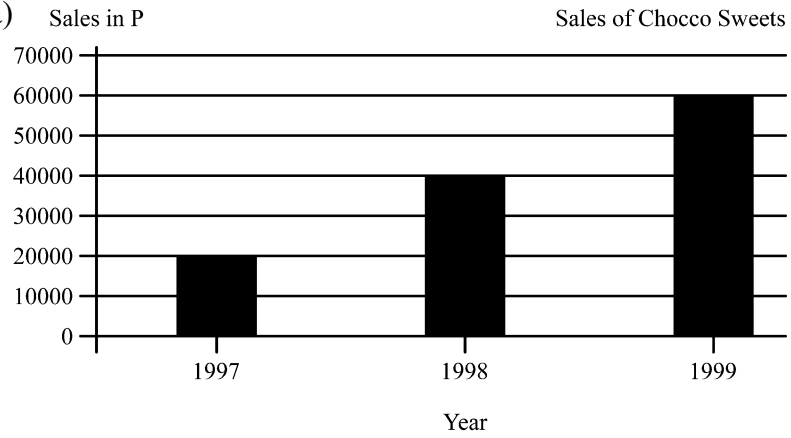
1a) The two line graphs use different scales on the vertical axis (that they do not start at zero is less serious although the 'squeeze' should have been indicated).

Draw the two line graphs on the same axes system (or if on different then use the same scale on the vertical axis). The FAST GROWTH company has the same trend as the others.

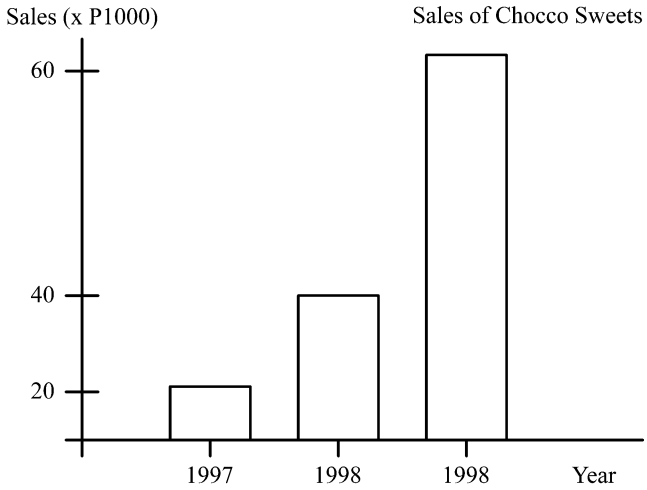
b) The scale along the vertical axis is uneven. First two squares represent P5 million, next 5 squares are used to represent the same amount. Draw the graph using consistent scale along vertical axis and note that the increase is the same for both 3 year periods.

c) No scale is available, so no conclusion can be drawn from the graphs. They should be drawn on the same axis—but due to missing scale you cannot do it.

2a)

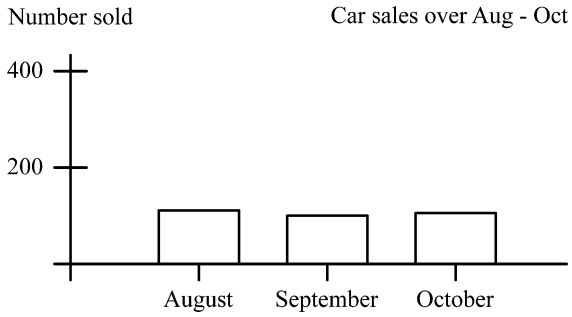


2b)

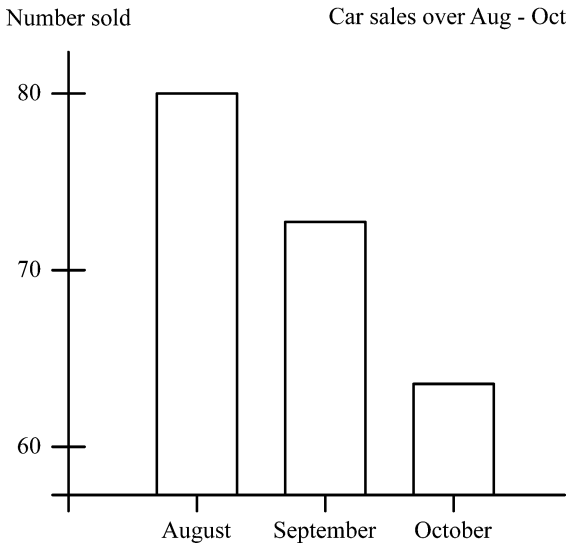


3) The vertical axis is not labelled. Hence no valid conclusions can be made.

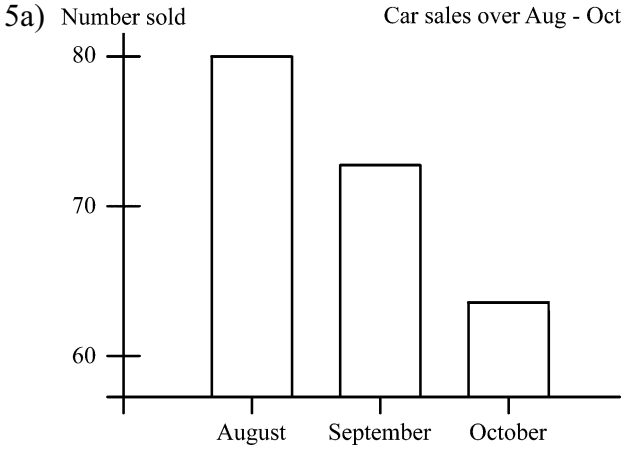
4a)



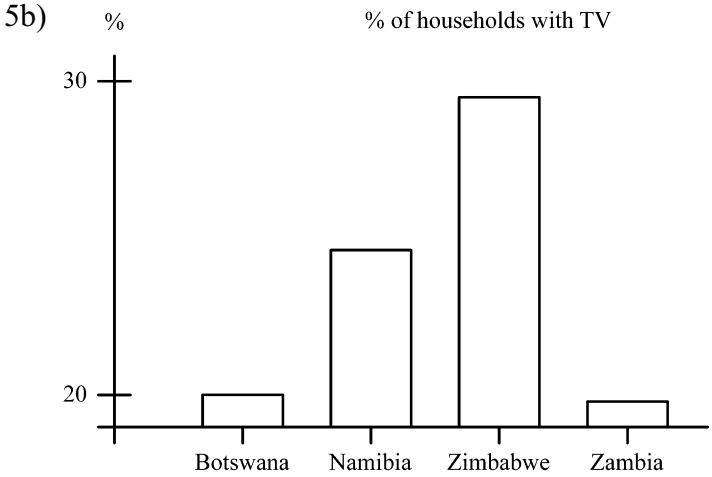
4b)



5a)



5b)



Self mark exercise 9

	Fig. 1	Fig. 2	Fig. 3	Fig. 4	Fig. 5	Fig. 6	Fig. 7
2. What type of data? (Answers may vary)	Quant, contin., ungrouped	Qual, discrete, grouped	Quant, contin., grouped	Quant, contin., grouped	Quant, mixed, grouped	Qual, contin., grouped	Qual?, contin., grouped
3. What type of graph?	Line	Bar	Bar*	Bar	Bar	Pie	Pie/Bar**

* Not Pie, because (a) two sets of data are displayed and a pie chart can show only one, and (b) each bar is separate from its neighbour, as in a bar chart but not in a pie chart. Another description of this is a "stacked bar" chart consisting of two bars, except that the bars are not vertical. Incidentally, the graph shows bias! Compare the two bars for "renewable forms" and for "lignite": they are clearly not sized correctly for the numbers they represent.

** This clever graph could be considered a line graph turned on its side, or as four pie charts which each add up to 100%. Although not mentioned in this course, this graph is also a "stacked bar" graph turned on its side, since stacked bar graphs often have tie-lines that connect a level in one bar to the same level in the next bar. It is *not* a pictogram, since the pictures of houses do not represent counts or frequencies—only categories.

Unit 4: Measures of central tendency

Introduction to Unit 4

The use of the mind should precede the use of the mean. This is to say that one has to look further than the mere calculation of averages.

There are four aspects to consider when it comes to giving meaning to the represented data:

1. Look behind the data
   Data sets are related to a context and gathered and presented by someone who might have a particular agenda. It is therefore important to look behind the data. Questions that are to be considered are: is there bias, attempts to disguise some data, attempts to mislead with data, attempts to present the data from only one point of view? Misuse and abuse of statistics are to be an important aspect of pupils' data handling experiences.
2. Look at the data
   This covers computational and representation aspects: which statistics are meaningful to compute and what is the best way to represent the data in chart or diagram.
3. Look between the data
   This is the comparison aspect of the analysis: looking for differences and similarities.
4. Look beyond the data
   This is to cover the inference part of the analysis: what conclusion can be (safely) drawn from the results.

Purpose of Unit 4

In this unit you are going to look at different averages: mean, mode and median and how to calculate them when data is grouped or ungrouped, as well as which average is most appropriate to use in a given situation. The unit begins with reasonably standard material on central tendency. However, it includes classroom assignments with a twist: students determine the central tendency of their own understanding of central tendency! Thus, an underlying purpose of this unit is to help you teach statistics by means of statistics.

Objectives

After completing this module, you should be able to:

- find the measures of central tendency (mean, median and mode) of ungrouped data (frequency tables)
- find an estimate of mean, median and mode of grouped data given in a frequency table and /or histogram or cumulative frequency curve
- justify which measure of central tendency is most appropriate to use in a given context

- investigate the effect of increasing/decreasing all data by a constant, or increasing/decreasing all data by a given factor on the mean, median and mode
- investigate the effect of data value of 0 on mean, median and mode
- set activities for pupils to enhance their understanding of measures of central tendency
- state the common misconceptions of pupils related to measures of central tendency

Time

To study this unit will take about 10 hours.

Unit 4: Measures of central tendency

You must be familiar with the three averages: mean, median and mode. Write down how you have been teaching these concepts to your pupils. Illustrate with the examples you generally use.

When reading through the next section refer to what you wrote down.

Section A: Averages: mean, mode, median

Many questions related to mean, mode and median merely test a pupil's ability to recall a formula, to substitute the values into the formula and to compute an arithmetically correct answer (operational or instrumental understanding) while pupils lack a relational or functional understanding of these measures of central tendency. Many questions in data handling deal only with **arithmetical aspects** and not real statistical questions. How to compute measures for central tendency is of limited value when the pupil does not know how to interpret the values. Interpreting the meaning of a stated measure of central tendency should be part of the activities presented to pupils.

A question such as: Find the mean of 6.3, 5.4, 4.9, 4.3 and 0.8 does nothing to test a pupil's understanding of the functional characteristics of the mean as a representative statistic, or model of the given data, and so is a bad question. It can also be criticised because it is based on a small sample of meaningless figures. Data is to be interpreted in a context.

For example:
A train is to leave the station at 8.30 each morning. The departing time on Monday was 35 minutes late due to a fire in the restaurant car. On Tuesday the train was delayed 5 minutes, on Wednesday 3 minutes, Thursday 3 minutes and Friday 4 minutes.

What was the average number of minutes that the train was late leaving? (10 minutes)

Is this 'average' a good figure to use to represent the week's data? (No, because the delay on Monday is an exceptional instance, and hence it should NOT be included in the average; the 10 minutes are not a reflection of what passengers might expect.)

The problem of when to consider an outlier as an outlier and when as an extreme (but possible) value is seldom considered by teachers. In the first example based on meaningless data, pupils with real understanding of the statistical concept 'average' face serious problems: is 0.8 an outlier or not? There is no context to provide clues.

Data handling has to relate to a context to make it meaningful. Providing pupils with meaningful and realistic problems is essential for developing understanding of measures of central tendency or more generally for the understanding of data handling.

Section B: The concept of the mean

The following aspects need attention.

1) The mean is located between the extreme values.

   For example: The number of pupils present in class during a week are as follows:

Monday	26	Tuesday	18	Wednesday	24
Thursday	29	Friday	28		

   The mean is $(26 + 18 + 24 + 29 + 28) \div 5 = 25$

   This is between the extreme values of 18 and 29.

2) The sum of the deviations from the mean is zero.

   Using the same data as in the example above, the deviations from the mean 25 are

   $+1, {}^{-}7, {}^{-}1, +4, +3$. The sum being zero.

   This property is important for estimating the mean of a set of numerical data. It can be taught effectively by reversing the traditional order: instead of giving data and asking to compute the mean, the teacher can give the mean (say 25) and ask to find a data set (of say 10 data) with that mean. Pupils quickly discover the "balance strategy", i.e., if I include 23 (two below the mean of 25), it is to be balanced by say 27 (two above the mean).

3) The average is influenced by values that deviate from the average.

   Using the above data and assuming that on Saturday 28 pupils attended, the mean attendance over the six days will be different from 25. However if on Saturday 25 pupils attended (the mean over the first five days) the mean over the six days will remain 25.

4) The average does not necessarily equal one of the values that was summed.

   The mean over the six days is $(26 + 18 + 24 + 29 + 28 + 28) \div 6 = 25.5$. However 25.5 pupils can never attend. The mean is not a value equal to any of the values averaged.

5) The average can be a fractional value with no counterpart in reality.

   The example in 4 shows that the mean (25.5) can be a decimal with no real object it can refer to in reality: 25.5 pupils do not exist.

6) The average value is representative of the values that were averaged.

   This is an important property used when interpreting data. The mean represents all the data in the data set.

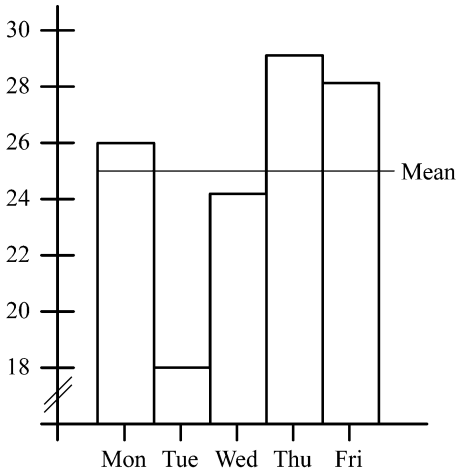
7) In computing an average, a value of zero, if it appears, is to be taken into account.

   Some pupils have the misunderstanding that 0 is 'nothing' and hence need not to be included in calculation of the mean. However 0 is a legitimate numerical value. For example somebody might have 0 brothers and sisters, the temperature might be 0^{o} C.

8) Relating the mean to the representations of data.

In a bar chart the mean frequency will be the height of a rectangle enclosing an area equal to the area of the bars.

For example: The bar chart represents the number of pupils present in class from Monday to Friday.



The mean number of pupils in class during the five days was 25. The area from the "mean" line down to the lower axis equals the area of the five bars.

Section B1: The average or mean has three meanings

Take for example the statement that "On average a box of matches contains 35 matches."

1) measure of location

   Words such as 'around', 'near', 'about', 'close to' were used by pupils to explain the above statement. These responses (the great majority) indicate that pupils understand mean as a measure of location.

2) representative number

   This notion is usually lacking among 13 -14 year olds. They do not understand the mean as a measure resulting from a stochastic process, i.e., a random process based on chance.

3) expected value

   Words such as 'normally' 'the usual amount' express the idea of expectation. That one might expect 35 matches in the box.

Section B2: Misconceptions and pupils' errors

The following are some misconceptions that can be expected.

1) any difference in the means between two groups is significant

   If the average height of pupils (12 - 13 years old) in Form 1A is 158 cm and in Form 1B the average is 160 cm, this difference need not be significant. It would be incorrect to conclude that the pupils in Form 1B are taller than the pupils in Form1A (additional information would be needed to make such a statement). It could be that there is one very tall girl in Form 1B causing the mean to go up, while all other pupils might in fact be below the height of 158 cm.

2) a lack of awareness of regression to the mean in everyday life

   If a process is repeated many times (throwing a dice) the number of times a six is thrown will come closer to the theoretical mean.

3) errors due to inconsistent notation used in different textbooks

   The conventions used in different textbooks are not the same and are problematic to pupils that are not fluent in the use of mathematical symbols. As project work implies consulting different textbooks a teacher should be aware of this possible source of errors. To avoid errors caused by differences in conventions a teacher should make it a rule NEVER to present concepts exclusively in symbolic algebraic format. A word format should be presented as well.

Section C: Mean, median and mode for ungrouped data

I(a) The mean of n listed values:

The mean of n numbers $a_1, a_2, a_3, \ldots a_{n-1}, a_n$ is (sum of all the data values) ÷ (number of data values) . In formula form

$$\frac{a_1 + a_2 + a_3 + \ \ldots \ + a_{n-1} + a_n}{n} = \frac{\sum_{i=1}^{n} a_i}{n}$$

Σ(pronounced: sigma) is the Greek letter for S. It is used to mean "the sum of ."

$\sum_{i=1}^{n} a_i$ means the sum of all the a_i where I take values from 1 to n.

It is short for $a_1 + a_2 + a_3 + \ldots + a_{n-1} + a_n$

I(b) The mean of numbers given in a frequency distribution

Number	a_1	a_2	a_3	.....	.....	a_{n-1}	a_n
Frequency	f_1	f_2	f_3	....	.....	f_{n-1}	f_n

The mean of $a_1, a_2, a_3, \ldots a_{n-1}, a_n$ if the numbers have frequencies of respectively $f_1, f_2, f_3, \ldots f_{n-1}, f_n$ is $\frac{\sum_{i=1}^{n} f_i a_i}{\sum_{i=1}^{n} f_i}$

The mean can be obtained from discrete or continuous quantitative data. Qualitative data *cannot* be summarised by a mean. If you have data on the favourite type of music of the pupils in your class it does not make sense to average 'gospel music' with 'reggae'.

II The **mode** is the observation with the highest frequency. The mode uses the frequencies and hence a mode can be obtained for both quantitative and qualitative data.

II(a) The mode of n listed observations.

For example:

(i) 10 people were asked for their favourite drink and the responses were tea, fruit juice, milk, fruit juice, fizzy drink, fruit juice, fizzy drink, milk, fruit juice, tea.

The mode being: fruit juice.

(ii) The height of ten plants was 1.64 m, 1.58 m, 1.67 m, 1.71 m, 1.65 m., 1.68 m, 1.65 m, 1.60 m, 1.65 m, 1.71 m.

The most frequent height is 1.65 m. The mode is 1.65 m. Both qualitative and quantitative data can have a mode.

II(b) The mode read from a frequency distribution.

Number of words in a sentence in a nursery rhyme.

No. of words	5	6	7	8	9	10
Frequency	3	3	6	5	4	4

The highest frequency is 6. The mode is 7. Seven words per sentence occurred most frequently.

III The **median** is the middle observation if the number of observations is odd or the mean of the two middle observations if the number of observations is even provided the data is written in increasing (or decreasing) order. Median can be obtained for quantitative data; qualitative data has no median.

III(a) The median of *n* listed values:

The scores of the school's netball teams during a tournament were
Team A: 2, 4, 5, 3, 4, 1, 7 Team B: 3, 6, 4, 2, 4, 5, 8, 6.

The median number of goals scored by the school's netball teams requires ordering the given data first:

Team A: 1, 2, 3, 4, 4, 5, 7

Team B: 2, 3, 4, 4, 5, 6, 6, 8.

The first team A had a median score 4, the middle of the seven ordered scores.

The middle value of *n* observation, when *n* is odd, is the $\frac{1}{2}(n+1)$th observation.

The median of team B is the mean of 4th and 5th observation (as the number of observation is even there are two numbers in the middle).

The median is therefore $\frac{4+5}{2}=4.5$.

The middle value of *n* observation, when even, is the mean of the $\frac{1}{2}n$th and the $\frac{1}{2}(n+1)$th observation.

III(b) The median of quantitative data given in a frequency distribution.

Using the same data as above

No of words	5	6	7	8	9	10
Frequency	3	3	6	5	4	4

The median is found by first finding the total number of observations $(3 + 3 + 6 + 5 + 4 + 4) = 27$.

The median is therefore the 13^{th} observation: the 3^{rd} observation is 5, the sixth $(3 + 3)$ is 6, the 12^{th} $(3 + 3 + 6)$ is 7, the 13^{th} is the first of the five, 8. The median is 8.

Self mark exercise 1

1. The heights of three friends are 1.56 m, 1.67 m and 1.61 m.
   Find the mean height.

2. The frequency distribution table shows the number of pips in 50 oranges.

number of pips	8	9	10	11	12	13	14	15
frequency	2	5	12	10	6	7	5	3

   Find the mean number of pips in the 50 oranges.

3. A survey in the class on the number of hours spent by pupils to study for the end of year Science examination gave the following data (to the nearest hour).

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

   Find the mean, mode and median of this data.

   Which of the three averages best represents the data? Justify your answer.

4. Number of days pupils were absent during one week in form 2A4.

Days absent	1	2	3	4	5	6
Frequency	6	0	2	2	1	1

   Find the mean, mode and median of this data.

   Which of the three averages best represents the data? Justify your answer.

5. The stem-leaf diagram represents the height of boys and girls in a class.

   Height of pupils in form 2X

Girls		Boys
443310	15	2
9865	15	579
43220	16	12234
865	16	5566889
42	17	044
	17	58

   $n = 41$ 16 | 8 represent 168 cm

   Find the mean, mode and median of (i) girls (ii) boys.

Continued on next page

Compare the averages of boys and girls and make some statements based on the averages calculated.

Represent the data in one bar chart (150-154, 155 - 159, 160-164, etc.) placing the bars for boys and girls next to each other (double bar chart).

6. Tashata collected data on the number of people in a car, parking in front of a shopping centre between 09 00 h and 10 00 h.

Number of people	1	2	3	4	5	6
Number of cars	42	20	14	12	10	2

a) Find the mean, mode and median.

b) Represent the data in a graph/chart. Justify your choice.

7. Packets of crisps are marked 30 g. A sample of packets was taken and the mass of crisps determined to nearest gram. The results are as listed:

Mass (g)	29	30	31	32	33
Frequency	20	50	45	10	25

a) Find the mean, mode and median.

b) Represent the data in a histogram.

8. Two running teams A and B took part in a competition. The times, to nearest 0.1 of a second, of the members of team A and team B in the 100 m are listed below.

Team A	Team B
11.9	12.0
12.4	12.5
13.1	13.2
13.3	13.8
13.6	14.1
12.9	12.6
13.8	13.5
13.3	12.7
12.8	12.1
12.4	13.7
12.0	11.8

a) Find the mean, mode and median.

b) Which average best represents the data? Explain.

c) Which is the ‘better’ team? Justify your answer.

Suggested answers are at the end of this unit.

Section D: Which is the best average to use?

Which is the best average to use depends on the situation and what you want to use the average for. The mean is the most commonly used measure of central tendency as it is the only one of the three averages using all the data. It takes all the data in the distribution into account. The mean—being arithmetically based—can be combined with the means of other groups on the same variable. For example: If you found that the average score on a mathematics test in class 1A was 68% and in 1B the average score was 71%, the average score of the combined class 1AB can be computed. The median and the mode, not being arithmetically based, do not have such a property. If the modal height in form 1A is 165 cm and in 1B is 167 cm, you cannot draw any conclusion with regards to the mode of the combined class 1AB.

However using the mean can give a rather distorted picture of the data if there are outliers, or if the mean is not meaningful in the given context.

Examples:

1. The leader of a youth club can get discounts on cans of drinks if she buys all one size. She took a vote on which size the members of the club wanted.

Size of can (ml)	100	200	330	500
Number of votes	9	12	19	1

   Mode = 330 ml, median = 200 ml and mean = 245.6 ml (1 decimal point)

   Which size should she buy?

   The mean is clearly of no use—cans of size 245.6 ml do not exist. The median would be possible as 200 ml cans are for sale. However only 12 out of the 41 club members want this size. In this case the mode is the best average to use as it is the most popular one among the club members.

2. A teacher wants 50% of his pupils to get a credit in a test and wants to set the minimum mark for the credit. Which average should the teacher use? In this case the median is the only appropriate average as this is the middle score—50% will be above and 50% below this score. (N.B. the median mark can only be set after the test is written).

3. In a small butchery the four labourers earn each R 400 per month, the supervisor earns R 1200 and the manager R 2600. Which average best represents the monthly wages earned?

   The mean (R 900) is misleading as it is more than twice the salary earned by most workers. The median (R 400) is representative. The other appropriate average to use is the mode: it gives the wages of most of the workers. When datasets are skewed to one side, like wages or house prices, the median and mode are more realistic than the mean.

4. The time taken (in hours) by 6 pupils to complete their project was 20, 25, 31, 35, 87, 87.

   The mean is 47.5 but most pupils worked less than that on their project. The mode is 87 but is also misleading. The median is

$33\left[\frac{1}{2}(31+35)=33\right]$ which is the best to use as it tells us that half of the number of pupils needed less than 33 h and half needed more.

5. The number of border crossings at 5 border posts between Botswana and Zimbabwe on a certain day were 40, 60, 60, 80 and 810. The median is 60, the mode is 60 and the mean is 210. The outlier 810 makes the mean move to 210, a value atypical for the data. The median would be a better value to represent the data.

6. Occasionally, distributions arise for which none of the averages is particularly informative. For example:

   The table shows the number of cigarettes smoked per day by 50 persons.

Number of cigarettes	Number of people
0	30
1	10
2	5
3	3
4	2

   The mean is 7.4, the mode is 0 and the median is 0. None of these averages represent the data well. In this case it would be better to state that 60% are non-smokers and that the smokers smoke on average (mean) 18.5 cigarettes a day.

The *mean* is generally used if the data is more or less symmetrically grouped about a central point, i.e., the data do not contain outliers. If further calculation is required (e.g., measures of dispersion) or comparison with a similar measure on another group is intended, or the (sample) mean is to be used in estimating parameters of the population then the mean is to be used as mode and median cannot be used in 'further' calculations.

A distribution with outliers is frequently best described by using the *median*.

The *mode* is used when the context suggests 'most usual' or 'typical' value.

Self mark exercise 2

1. Decide which of the following averages—mean, mode or median—is the most appropriate to use to summarise the following data. Justify your answer.

   a) number of children in the family
   b) number of letters in pupils' surnames
   c) number of pupils born in a certain month
   d) shoe size of the boys in the class
   e) favourite subject in school
   f) number of days each pupil in a class was absent during the term
   g) method of payment for goods bought in a furniture shop
   h) most popular activity of pupils during a long weekend

Continued on next page

2a. Calculate mean, median and mode for each of the following sets of data.

b. Decide and justify which of the three best represents the data.

(i) Modise scored the following number of goals during the eight matches of the school's football team:

1, 1, 0, 6, 2, 1, 3, 0

(ii) A pupil scored the following percent grades in geography:

75%, 72%, 68%, 57%, 62%, 10%, 75%.

3. There are six possible ways of listing the three averages in ascending order of size. For example: mean, mode, median; mean, median, mode, etc. Write down all six and then try to find sets of numbers which will fit each arrangement.

For example, to make the order mode < median < mean you could choose the four numbers 1, 1, 3, 11 with mode 1, median 2 and mean 4.

4. Averages are meant to represent the data. List advantages and disadvantages of the mode, median and mean and give examples when you would use one instead of the others.

5a. (Challenge question for the strong mathematicians.) In mathematics you might also meet the **geometric mean** of numbers. The geometric mean of p and q is $\sqrt{(pq)}$.

Show that always for two numbers (arithmetic mean) ≥ (geometric mean).

b. The harmonic mean H of two numbers p and q is defined as $\frac{1}{H} = \frac{1}{p} + \frac{1}{q}$.

Express H in terms of the geometric and arithmetic mean.

Suggested answers are at the end of this unit.

Section E: Mean, mode and median: classroom lessons

The following pages give suggestions for an investigative approach to mean, mode and median. It is assumed that pupils have basic knowledge of what each of these measures stands for. The sequence of activities starts with a diagnostic test. This will give feedback to the teacher as to the ideas of the pupils and help to plan the lessons: which points need to be emphasised. At the end of the lesson sequence the teacher might decide to discuss some of the questions in the diagnostic test with the pupils.

The main objective is to make pupils aware of the effect on the three measures (mean, mode and median) when values are added to the group. For example, a new pupil, age 12 years 3 months, height 1.58 m, mass 58 kg joins the class. How will the mode, median and mean age / length / mass of the class change? The lessons also aim at developing in pupils an awareness as to which measure is most appropriate to use in a given context. The lessons use an investigative method, with pupils working in groups, to discuss points with

each other before (if needed) whole class discussion is used to round up the activities.

Practice task 1

1. Work through the diagnostic test and lesson outlines by yourself. Write down any problems you encounter.
2. Administer the diagnostic test to your class and analyse the results. Write a report on your findings.
3. Based on the outcomes of the diagnostic test make, if necessary, changes in the lesson outline and work out the lessons in detail. Prepare detailed lesson plans and notes. Prepare the worksheets for the pupils. You might need different versions for different levels of achievement of your pupils (differentiated worksheets to meet 'mixed ability' of your class).
4. Try out the planned activities / lessons and write an evaluative report. Cover questions such as:

   Were the objectives attained? How do you know?

   Did pupils enjoy the activities?

   What needs changing in the material?

   Was the method different from what you used to use?

Present the assignment to your supervisor.

A diagnostic instrument

Answer the following questions. Give reasons for your answers.

1. Suppose you have calculated the average age of a family. Afterwards a baby is born in that family. If you were asked to re-calculate the average age of that family, your task is very simple. Since the baby is 0 years old the average age will be exactly the same.

   TRUE / FALSE (circle the correct answer)

   Reason:

2. Somebody has calculated the average of a set of numbers. She tells her friend that if you take the total of all differences between the average and the numbers in the set you will always find zero.

   TRUE / FALSE (circle the correct answer)

   Reason:

3. A mode of a set of scores is the score with the highest frequency. What is the mode of the following set of scores: 60, 58, 33, 98, 58, 60, 42, 58

   A) 60 B) 58 C) 33 D) 98 E) 42 (circle correct answer)

   Reason:

4. Give 5 different numbers such that their average is 21.

   The numbers are:

   I found these numbers by:

5. A median of a set of scores is the value in the middle when the scores are placed in order. In case there is an even number of scores there is not a single 'middle score'. In that case you must take the middle two scores and calculate their average.

   Four students scored the following results for a test: 19, 20, 17, 11. Find the median.

   Median =

   This is how I found my answer:

6. a) If you add a number to a set of numbers, the mean changes

      ALWAYS / SOMETIMES / NEVER (circle the correct answer)

      Reason:

   b) If you add a number to a set of numbers, the median changes

      ALWAYS /SOMETIMES / NEVER (circle the correct answer)

      Reason:

   c) If you add a number to a set of numbers, the mode changes

      ALWAYS /SOMETIMES / NEVER (circle the correct answer)

      Reason:

Lesson outline 1: Mean, median, mode

Time: 80 minutes

Prerequisite knowledge: Pupils have met mean, median and mode before and know the arithmetic involved in computing these measures of central tendency.

Objectives: Pupils should be able to

a) find the mean, median and mode of a set of data in context.

b) make statements about the effect on mean / median / mode if values are added to the data set (adding zero value, adding two values with equal but opposite deviation from the central measure, adding values equal to the central value).

Review of mean, median and mode

Exposition - discussion strategy (15 minutes)

Teacher presents the question:

A test was scored out of 20 (only whole marks were given) and 12 pupils scored: 19, 20, 17, 11, 19, 19, 15, 8, 15, 20, 17 and 18.

Pupils are asked to compute the mean, median and mode. (Give sufficient time to pupils to do the working.)

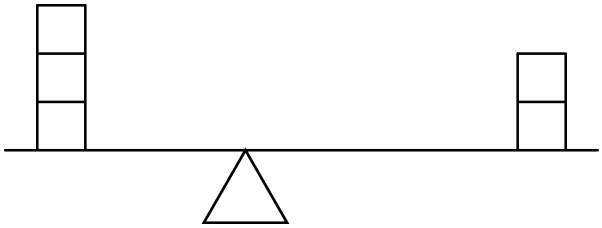
Answers: mean 16.5 / mode 19 / median 17.5

Teacher calls on pupils to explain how they obtained their results.

Expected to result in:

Mean = (sum of all scores) ÷ (number of pupils).

Illustrate the *mean* as the balancing point:



"Forces" (deviations) at one side balance the "forces" (deviations) at the other side.

Mode is the score with the highest frequency (the value that is 'in fashion', the most popular).

Median is the value in the middle when the scores are placed in order (if odd number of observations) OR average of the middle two scores (if even number of observations).

Half the number of observations are to the right of the median the other half to the left.

Questions: Which of these three—mean, median or mode—do you feel can be used best to represent the set of scores? Justify your answer.

DO NOT ANSWER the question at this stage; only make an inventory of the pupils' opinions and their reasons, without further comment.

Write the results on the chalkboard:

Best measure to use	number of students in favour	because
mean		
median		
mode		

Inform pupils that they are going to investigate how mean, mode and median behave, so as to make a decision on which measure might be best used in a certain context.

Investigating (40 minutes)

The following are covered in the pupil's worksheets (Worksheet for pupils is on a following page – seven pages ahead)

a) Is mean, median, mode necessarily a value belonging to the set and/or a value that could be taken in reality?

b) The effect on mean, median, mode of adding a zero value to the value set.

c) The effect on mean / median/ mode of adding two values with equal but opposite deviations or unequal deviations from mean, mode, median.

d) The effect on mean / median / mode of adding values equal to mean / median / mode.

Pupils' activity
Teacher gives worksheets to pupils.

In small groups pupils are to answer the questions individually, then next compare and discuss the following questions.

A test was scored out of 20 (only whole marks were given) and 12 pupils scored: 19, 20, 17, 11, 19, 19, 15, 8, 15, 20, 17 and 18.

Using the above or other pupils' scores (using real scores obtained by the class for example) answer the following question:

Q1) Must the mean / median / mode be a score attained by one of the pupils in the class?
Justify your answer. Illustrate with examples and non examples.

Note to the teacher:

(i) mean
The mean represents the scores but need not be one of scores itself, it might even be a 'score' that is impossible ever to get.

(ii) median
The median will be a score of one of the pupils if number of scores is odd. If the number of scores is even the median will be a score nobody did get or even nobody ever can get. If the median is half way between 16 and 18, then 17 is a possible score although nobody did score 17; if the median is between 17 and 18 the median is 17.5, a score nobody can ever get as it is not a whole number.

(iii) mode
The mode is necessarily a score attained by several pupils. If all scores are different there is no mode. If certain scores have the same frequency a set of scores can have more than one mode (bimodal , trimodal, etc., distribution).

Q2) Investigate how the mean / median / mode changes when a zero score is added to the following set of scores.

a) 19, 20, 17, 11, 19, 19, 15, 8, 15, 20, 17 and 18.
Mean is 16.5; median is 17.5 and mode is 19.

b) 19, 20, 17, 11, 19, 19, 15, 8, 15, 20, 17, 18 and 11
Mean is 16.1; median is 17; mode is 19

c) 0, 19, 20, 17, 11, 0, 19, 19, 8, 15, 20, 17 and 18.
Mean is 14.1; median 17; mode 19

Make a correct statement:

1. If to a set of scores a zero score is added the mean changes
   ALWAYS / SOMETIMES / NEVER
2. If to a set of scores a zero score is added the median changes
   ALWAYS / SOMETIMES / NEVER
3. If to a set of scores a zero score is added the mode changes
   ALWAYS / SOMETIMES / NEVER

   Does the size of the number of observations matter? Are the changes (if any) the same whether you considered 20 observations or 2000?

   (Answer: Mean / mode / median all change sometimes. If a large number of observations is involved, the change in the mean is very small (the first decimal place might not change at all) or when the mean is zero, adding a zero will not change the mean. Median changes are likely to be smaller in a large population than in a smaller, but even there changes are generally small. The nature of the observations (do observations have close to the same frequency) determines whether or not changes in mode occur.)

Q3) A set of scores has a mean of 16.

Without calculating the new mean state how the mean changes if two more scores are to be taken into account.

a) the two scores are 14 and 18

b) the two scores are 15 and 17

c) the two scores are 14 and 17

d) the two scores are 12 and 20

e) the two scores are 12 and 19

Make a general statement about when the mean will change and when it will not change.

(Answer: the mean will not change if two values with equal but opposite deviations from the mean are added, or if the added value equals the mean; otherwise it will change.)

Q4) A set has a median score of 16.

Without calculating the new median state how the median changes if two more scores are to be taken into account.

a) the two scores are 14 and 18

b) the two scores are 15 and 17

c) the two scores are 14 and 15

d) the two scores are 8 and 20

e) the two scores are 12 and 19

f) the two scores are 18 and 19

Make a general statement: when will the median change, when will it not change?

(Answer: median will not change whatever values are added as long as one is to the left and one to the right of the median; if the added values are both to the right or the left the median might change.)

Q5) A set has a mode score of 16.

Without calculating the new mode state how the mode changes if two more scores are to be taken into account.

a) the two scores are 14 and 18

b) the two scores are 14 and 15

c) the two scores are 18 and 19

Make a general statement: when will the mode change, when will it not change?

(Answer: No statement can be made as the added values might make the distribution bimodal or trimodal. For example 14, 14, 16, 16, 16, 18 has mode 16 adding 14 and 18 makes it a bimodal distribution with modes 14 and 16. If the original set was 14, 14, 16, 16, 16, 18, 18 the adding of 14 and 18 makes it a trimodal distribution with modes 14, 16 and 18.)

Q6) A set of scores has a mean of 16. Without calculating the new mean state how the mean changes if two more scores equal to the mean are added.

Q7) A set of scores has a median of 16. Without calculating the new median state how the median changes if two more scores equal to the median are added.

Q8) A set of scores has a mode of 16. Without calculating the new mode state how the mode changes if two more scores equal to the mode are added.

Q9) Answer question 6, 7 and 8 if only ONE value equal to mean /median /mode respectively were added.

(Answer Q6/ Q7/ Q8/ NO changes in mean, median and mode; Q9/ only the median might change.)

Q10) Write down a data set of the ages of 12 people travelling in a bus with

a) mean 24

b) median 24

c) mode 24

Compare the data sets each member in your group has written down.

Are all the same?

Why are there differences? How can different data sets have the same mean (median / mode)?

Which set is the best? Why?

A baby is born in the bus, making now 21 passengers the last one with age 0.

Each pupil is to compute the change in mean / median / mode of her /his data set.

The grand-grand parents (age 90 and 94) of the newborn enter the bus, making up a total of 23 passengers.

Each pupil is to compute the change in mean / median / mode of her /his data set.

The 0 and the 90 / 94 are called outliers—they are 'far' from the mean / mode/ median.

Comparing your results how do outliers affect the mean / median / mode?

Make a correct statement:

1. If to a set of ages outliers are added the mean changes
   ALWAYS / SOMETIMES / NEVER
2. If to a set of ages outliers are added the median changes
   ALWAYS / SOMETIMES / NEVER
3. If to a set of ages outliers are added the mode changes
   ALWAYS / SOMETIMES / NEVER

Which of the three measures is most affected? (Answer: In general the mean is most affected by outliers as compared to median and mode.)

N.B. The above outlined activity would be more powerful if carried out on a computer using spreadsheets. In the summary the teacher could use a computer (provided the screen can be projected) to illustrate the effect of certain changes on both large and small data sets.

Reporting, summarising of findings, setting assignment (25 minutes)

Groups report / discuss / agree. Teacher summarises in table (outline already on the (back) of board before start of lesson).

a) Mean and median need not be observed values (values included in the observation set). They might even have a value that can never be an observed value. The mode (if it exists) always is an observed value.
b) Effect on mean / median / mode if one or two observations are to be included.

CHANGE	EFFECT ON		
	MEAN	MEDIAN	MODE
Adding zero value(s)	S	S	S
Adding two values with equal but opposite deviations	N	N	S
Adding two values with opposite unequal deviations	A	N	S
Adding two values with deviations both positive (negative)	A	S	S
Adding two values equal in value to the central measure at the top of each column	N	N	N
Adding one value equal in value to the central measure at the top of each column	N	S	N

A indicates will always change
S indicates will sometimes change
N will never change

c) Effect of the number of observations involved (small sample or large sample)

In the case of a large number of observations, adding of observations (not equal to the central measure) will ALWAYS change the mean—but the change will be (very) small. Outliers have a great impact on the mean of a small data set, but very little on a very large data set.

The median and mode are more likely to remain the same in the case of large numbers of observations, but can change.

Now come back to the original question:

Questions: Which of these three—mean, median or mode—do you feel can be used best to represent the set of scores? Justify your answer.

The tabulated answers of the pupils.

Best measure to use	number of students in favour	because
mean		
median		
mode		

Ask whether or not pupils want to change their previous opinion based on the increased insight on behaviour of the measures. If a pupil wants to change he/she is to justify the decision.

The discussion should lead to the decision that the median is most appropriate: half of the pupils scored below / above 17.5. The mean is less appropriate as it does not give any information as to how many pupils scored above / below the average of 16.5 (as mean is affected by outliers).

Pupils' assignment
(or take some questions for discussion in class if time permits)

In each of the following cases decide, giving your reasons, whether the mean, median or mode is the best to represent the data.

1. Mr. Taku wants to stock his shoe shop with shoes for primary school children. In a nearby primary school he collects the shoe sizes of all the 200 pupils (one class group from class 1 to class 7). Will he be interested in the mean size, median size or modal size?

   Answer: mode

2. In a small business 2 cleaners earn P340 each, the 6 persons handling the machinery earn P600 each, the manager earns P1500 and the director P3500 per month.Which measure—mean, median or mode— best represents these data?

   Answer: mode

3. An inspector visits a school and want to get an impression of how well form 2X is performing. Will she ask the form teacher for mean, median or mode?

   Answer: median

4. A pupil did 4 small projects in mathematics on the topic of number patterns during the term scoring (out of 20) in order : 4, 16, 15 and 16. Which represents best the overall attainment level of the pupil on project work on number patterns—mean, median or mode?

   Answer: median/mode

   Discuss: Is using the mean score to represent the work done in mathematics during a term a fair measure for the attainment of the pupil?

5. A house building company wanting to find out what type of houses they should build most often in a region carried out a survey in that region to find out the number of people in a family. Will they use mean, median or mode to decide what type of houses should be build most?

   Answer: mode

6. A car battery factory wants to give a guarantee to their customers as to the lifetime of their batteries, i.e., they want to tell the customer if you have a problem with the battery in the next ??? months we will replace your battery with a new one. They checked the 'lifetime' of 100 batteries. Will they use mean, median or mode to decide on the number of months to guarantee their batteries?

   Answer: mean

Worksheet for Pupils

Investigation

Question 1:

A test was scored out of 20 (only whole marks were given) and 12 pupils scored: 19, 20, 17, 11, 19, 19, 15, 8, 15, 20, 17 and 18.

Using the above or other pupils' scores (using real scores obtained by the class for example) answer the following question:

Has the mean, median or mode to be **a score attained by one of the pupils** in the class?

Justify your answer. Illustrate with examples and non examples.

Question set 2:

Investigate how the mean, median and mode change when a zero score is added to the following set of scores.

a) 19, 20, 17, 11, 19, 19, 15, 8, 15, 20, 17 and 18.

b) 19, 20, 17, 11, 19, 19, 15, 8, 15, 20, 17, 18 and 11

c) 0, 19, 20, 17, 11, 0, 19, 19, 8, 15, 20, 17 and 18.

Make a correct statement:

1. If to a set of scores a zero score is added the mean changes
   ALWAYS / SOMETIMES / NEVER
2. If to a set of scores a zero score is added the median changes
   ALWAYS / SOMETIMES / NEVER
3. If to a set of scores a zero score is added the mode changes
   ALWAYS / SOMETIMES / NEVER

Question 3:

A set of scores has a mean of 16. Without calculating the new mean state how the mean changes if two more scores are to be taken into account.

a) the two scores are 14 and 18

b) the two scores are 15 and 17

c) the two scores are 14 and 17

d) the two scores are 12 and 20

e) the two scores are 12 and 19

Make a general statement about when the mean will change and when it will not change.

Question 4:

A set has a median score of 16. Without calculating the new median state how the median changes if two more scores are added to the set as follows:

a) the two scores are 14 and 18

b) the two scores are 15 and 17

c) the two scores are 14 and 15

d) the two scores are 8 and 20

e) the two scores are 12 and 19

f) the two scores are 18 and 19

Make a general statement about when will the median change, and when will it not change.

Question 5:

A set has a mode score of 16. Without calculating the new mode state how the mode changes if two more scores are added to the set as follows:

a) the two scores are 14 and 18

b) the two scores are 14 and 15

c) the two scores are 18 and 19

Make a general statement about when will the mode change, and when will it not change.

Question 6:

A set of scores has a mean of 16. Without calculating the new mean, state how the mean changes if two more scores equal to the mean are added.

Question 7:

A set of scores has a median of 16. Without calculating the new median, state how the median changes if two more scores equal to the median are added.

Question 8:

A set of scores has a mode of 16. Without calculating the new mode, state how the mode changes if two more scores equal to the mode are added.

Question 9:

Answer question 6, 7 and 8 if only ONE value equal to mean, median and mode respectively were added.

Question 10:

Write down a data set of the ages of 12 people travelling in a bus with

a) mean 24

b) median 24

c) mode 24

Compare the data sets each member in your group has written down.
Are all the same?

Why are there differences? How can different data sets have the same mean (median / mode)?
Which set is the best? Why?

A baby is born in the bus, making now 21 passengers the last one with age 0. Compute the change in mean / median / mode of your data set.

The grand-grand parents (age 90 and 94) of the newborn enter the bus making up a total of 23 passengers.

Compute the change in mean / median / mode of your data set.

The 0 and the 90 / 94 are called outliers—they are 'far' from the mean / mode/ median.

Comparing your results how do outliers affect the mean / median / mode?

Make a correct statement:

1. If to a set of ages outliers are added the mean changes

   ALWAYS / SOMETIMES / NEVER

2. If to a set of ages outliers are added the median changes

   ALWAYS / SOMETIMES / NEVER

3. If to a set of ages outliers are added the mode changes

   ALWAYS / SOMETIMES / NEVER

Which of the three measures is most affected?

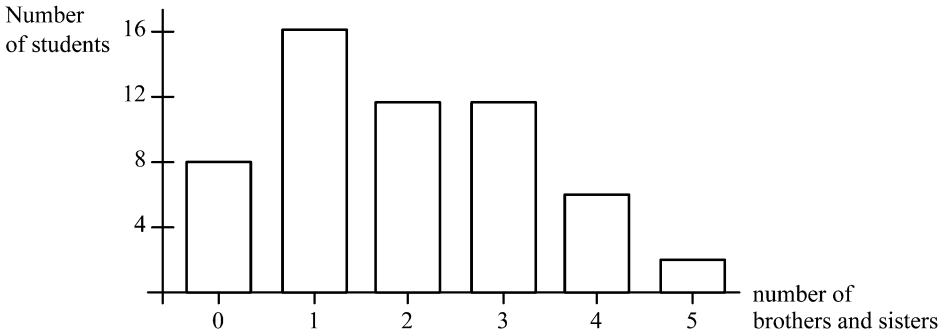
Recording sheet for students

Effect on mean, median and mode if one or two observations are to be added. (Place an A, S, or N in each empty box. A indicates 'will always change', S indicates 'will sometimes change', N will 'never change'.)

CHANGE	EFFECT ON		
	MEAN	MEDIAN	MODE
Adding zero value(s)			
Adding two values with equal but opposite deviations			
Adding two values with opposite unequal deviations			
Adding two values with deviations both positive (negative)			
Adding two values equal in value to the central measure at the top of each column			
Adding one value equal in value to the central measure at the top of each column			

DISCUSS IN YOUR GROUP

1. If you collect data on the pupils in your class will the data collected always have a mode? a median? a mean? Justify your answer, illustrate with examples and non examples.
2. Averages are meant to be representative for the data. List advantages and disadvantages of the mode, median and mean and find examples when you would choose one rather than another.
3. How 'real' are averages? Consider the following:
   "The average number of children in a family is 2.58, so each child in the family will always have 1.58 other children to play with."
   "If you have no money you join a group of 4 friends who have P2.50 each. Now you are a group of 5 and on average each person in the group has P2.00. You suddenly have P2.00."
4. A bar graph illustrates the number of brothers and sisters of a group of students.



   From the bar graph find the mean, median and mode. Which of the three measures is easiest to find?

Lesson outline 2: Mean, median, mode

A look at the average wage
The scenario for this lesson idea is a manufacturing and marketing company, in which the notion of "average" wage is considered from different points of view. The purpose is to show how a selection of the mean, the median, or the mode gives different answers to the same question.

Materials
The question "A look at average wage" below can either be photocopied or copied onto the blackboard.

Students are allowed to use a calculator.

Classroom organisation
Students can work individually, in pairs or in small groups.

The following information is to be given to pupils:

Question: "A look at average wage"

The Head of the Union Mr. Motswiri in the Matongo Manufacturing and Marketing Company was negotiating with Ms. Kelebogile Matongo, the president of the company. He said, "The cost of living is going up. Our workers need more money. No one in our union earns more than P9000.- a year."

Ms. Matongo replied, "It's true that costs are going up. It's the same for us—we have to pay higher prices for materials, so we get lower profits. Besides, the average salary in our company is over P11000.-. I don't see how we can afford a wage increase at this time."

That night the union official conducted the monthly union meeting. A sales clerk spoke up. "We sales clerks make only P5000.- a year. Most workers in the union make P7500.- a year. We want our pay increased at least to that level."

The union official decided to take a careful look at the salary information. He went to the salary administration. They told him that they had all the salary information on a spreadsheet in the computer, and printed off this table:

Type of job	Number employed	Salary	Union member
President	1	P125 000	No
Vice president	2	P65 000	No
Plant Manager	3	P27 500	No
Foreman	12	P9 000	Yes
Workman	30	P7 500	Yes
Payroll clerk	3	P6 750	Yes
Secretary	6	P6 000	Yes
Sales Clerk	10	P5 000	Yes
Security officer	5	P4 000	Yes
TOTAL	72	P796 750	-

The union official calculated the mean:

$$\text{MEAN} = \frac{\text{P796 750}}{72} \text{ P11 065,97}$$

"Hmmmm," Mr. Motswiri thought, "Miss Matongo is right, but the mean salary is pulled up by those high executive salaries. It doesn't give a really good picture of the typical worker's salary."

Then he thought, "The salary clerk is sort of right. Each of the thirty workmen makes P7500.- That is the *most common* salary—the mode. However, there are thirty-six union members who don't make P7500.- and of those, twenty-four make less."

Finally, the union head said to himself, "I wonder what the *middle* salary is?" He thought of the employees as being lined up in order of salary, low to high. The middle salary (it's called the *median*) is midway between employee 36 and employee 37. He said, "employee 36 and employee 37 each make P7500.- , so the middle salary is also P7500.-."

Questions:

1. If the twenty-four lowest salaried workers were all moved up to P7500.-, what would be

   a) the new median?

   b) the new mean?

   c) the new mode?

2. What salary position do you support, and why?

Activity 1: Presentation of the scenario

Review with students the problem setting and the salary information. Ask students to identify how the mean, median, and the mode are used in the problem description. Use question 1 to review one way in which pay raises may be distributed.

Answers: a) New median P7 500

b) New mean P11 812.50

c) New mode: P7 500

Pose these questions:

- Which measures of central tendency stayed the same?
- Which measures of central tendency changed? Why?
- If you changed only one or two salaries, which measure of central tendency will be sure to change? [The mean, since its calculation includes all values.]
- If you changed only one or two salaries, which measure of central tendency will be most likely to stay the same? [The mode is most likely to stay the same, because it is the most frequently occurring salary, and only one or two salaries are being changed.]
- If you change only one or two salaries, how likely is the median to change? [It depends. If the median is embedded in the middle of several

salaries that are the same, it won't change. If the median is close to a different level of salary, it is not likely to change.]

Activity 2: Using a Spreadsheet (optional)

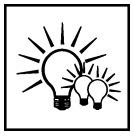
Having students enter the employees' salaries into a spreadsheet on a computer or demonstrating the use of a spreadsheets to students may clarify the role of a computer in solving real life problems. Column A could list the number of employees of each type, and column B could list the salary of that type of employee. Display the mean salary for all employees in a cell at the bottom of the spreadsheet labelled "mean salary." [Define the cell as the total salary value (payroll) divided by the total number of employees.] After using the spreadsheet to display the new salaries and calculating the new mean for each situation, pose the following inquiries:

- Predict the mean if the twenty-four lowest paid employees have their salaries increased to P7 500. Make the changes in the spreadsheet to find the actual mean.
- The president gave himself a raise that resulted in increasing the mean salary by P500. Predict what you think his new salary was. Use the spreadsheet to experiment and find the new salary.
- Two new employees were hired by the company: a plant manager and a foreman. Predict whether the mean salary will increase, decrease, or stay the same. Explain your prediction. Check it out with the spreadsheet.

Activity 3: Developing an argument

Use question 2 to initiate a discussion on drawing conclusions from the information. Small groups of students can develop position statements and report back to the class. There is no single correct answer to the discussion question. Management would naturally favour the mean; the union leader, the median; and the lower-paid members the mode.

Evaluation: This problem has more than one reasonable solution. However, many students expect problems in the mathematics class to have only one correct solution. Teachers can promote student consideration of multiple solutions by asking students to write up or present at least two reasonable alternatives. At first, students may simply take ideas from one another without much reflection, but if the teacher continues to value creative, reasonable alternatives, students will begin to enjoy actively looking for multiple solutions.



Practice task 2

1. Try out the lesson outline 2 in your class: A look at the average wage

2a) Write an evaluative report on the lesson. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did you meet some specific difficulties in preparing the lesson or during the lesson? Was discussion among pupils enhanced?

b) Present the lesson plan and report to your supervisor.

Section F: Mean, median and mode for grouped discrete data

A 60 item multiple choice test was tried in a class with 43 pupils. The results are represented in the following frequency distribution.

No. of correct answers	1-10	11-20	21-30	31-40	41-50	51-60
Frequency	4	5	11	9	8	6

A mode or median cannot be obtained from this frequency table. You can only read off the class interval that contains the mode and the class interval that contains the median. The class interval with the highest frequency is 21 - 30: the **modal class**. The median is the 22nd observation, i.e., the score of the 22nd student; that falls in the class interval 31 - 40.

If the number of data is large but discrete (for example the scores of 2000 pupils in an examination marked out of 60) or continuous (the time taken to run 100 m) data is best placed in groups or intervals.

The scores could be grouped in five intervals: 1 - 10, 11 - 20, 21 - 30, 31 - 40, 41 - 50, and 51 - 60 as in the distribution table above for 43 students only. By grouping data some information is lost. For example in the class 1 - 10 there are 4 students, and we no longer can see what their actual scores were (did all score 10?).

For calculation purposes the 'mid-interval value' (average of lower bound and upper bound value of the interval) is used.

From a grouped frequency table you can find:

– the modal class (the class with the highest frequency)

– the interval in which the median is found

– an estimate of the mean

A calculation of the mean using mid-interval values.

$$\text{Estimate of the mean} = \frac{\text{sum of } [\text{mid interval value} \times \text{frequency}]}{\text{sum of the frequencies}}$$

Example
The table gives the end of year examination mark of 200 students (maximum mark was 50) and the calculation to obtain an estimate of the mean.

Mark	Frequency	Mid-value	Mid-value × Frequency
1-10	10	5.5	55
11-20	20	15.5	310
21-30	60	25.5	1530
31-40	90	35.5	3195
41-50	20	45.5	910
TOTAL	200		6000

An estimate for the mean is $\frac{6000}{200} = 30$. The estimated mean is 30.

The modal class is 31 - 40. Most students scored in the range 31 - 40.

The median is the $\frac{1}{2}(200 + 1) = 100\frac{1}{2}$ th term, i.e., the average of the 100th and 101st term. As the actual value of these terms is unknown you can only give the interval in which these values are: the interval 31 - 40.

Calculator

Your calculator can help you with these and longer calculations in statistics. Actual procedures vary with the brand of calculator, but if yours has statistics capability, the general way to use it is as follows. For more details, consult the instructions that came with the calculator.

1. Place it in Statistics Mode (if it has such a mode).

2. Clear out any previously stored statistical data from the memory registers.

3. Now enter individual data values:

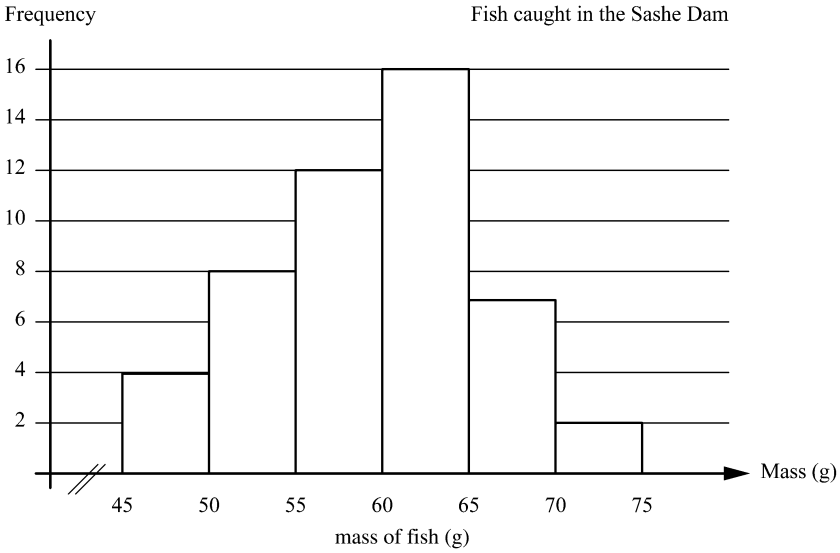
   a) for single values… key in each value, followed by a press of the [DATA] or [ENTER] or [$\sum+$] key.

   b) for grouped values… some calculators allow the entry of grouped data by having separate entry keys for the group frequency (or count) and for that group's average value. Consult your documentation.

4. Once all data points have been entered, a press of the [$\overline{X}$] key (or equivalent) will display the mean of all the entered data. In most calculators there are other keys for the sample and population standard deviations.

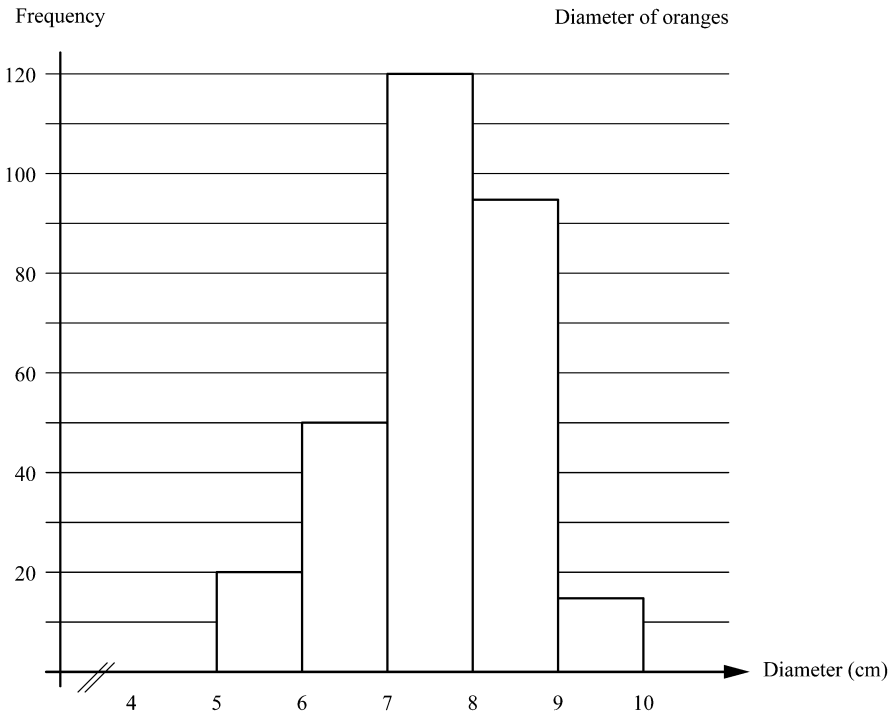
Self mark exercise 3

1. The histogram illustrates the mass of fish caught in the Sashe dam one day.



   a) Make the grouped frequency table.
   b) How many fish were caught in all?
   c) What is the modal class?
   d) In what class interval is the median mass?
   e) Calculate an estimate for the mean mass of fish caught (use your frequency table).

2. The diameters of a batch of oranges were measured and the results displayed in a histogram.



Continued on next page

a) How many orange had a diameter d cm in the range $8 \leq d < 9$?

b) Make the grouped frequency table corresponding to the histogram.

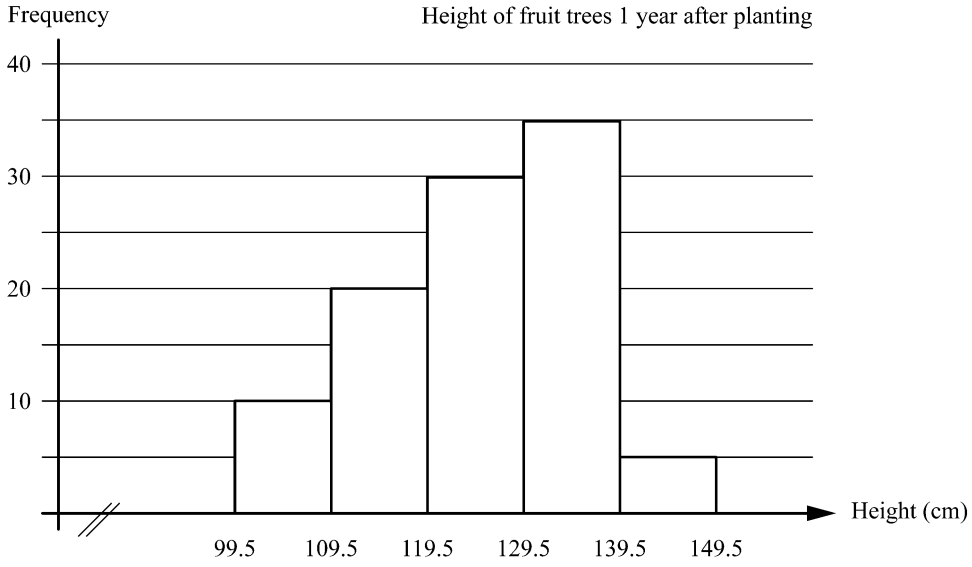
c) How many oranges in total were measured?

d) What is the modal class interval?

e) Calculate an estimate of the mean length of the diameter of the oranges (use your frequency table).

3. The height of fruit trees was measured one year after planting. The result is displayed in the histogram below.

   As the data, length of the trees, is continuous and taken to the nearest cm the first class 100 - 109 has as class boundaries 99.5 and 109.5. The class boundaries are taken half-way between the class intervals.



a) How many trees had a height h cm in the range 110 - 119 cm?

b) Make the grouped frequency table corresponding to the histogram.The first class interval is 100 - 109, the second 110 - 119, etc.

c) How many fruit trees in total were measured?

d) What is the modal class interval?

e) Calculate an estimate of the mean height of the fruit trees from the grouped frequency table.

4. The ages of pupils in Sefhare CJSS were represented in a histogram. Although age can be considered to be continuous, you are 14 from the day you turn 14 until the day you turn 15. The class interval has as lower bound 14 and as upper bound 15.

Continued on next page

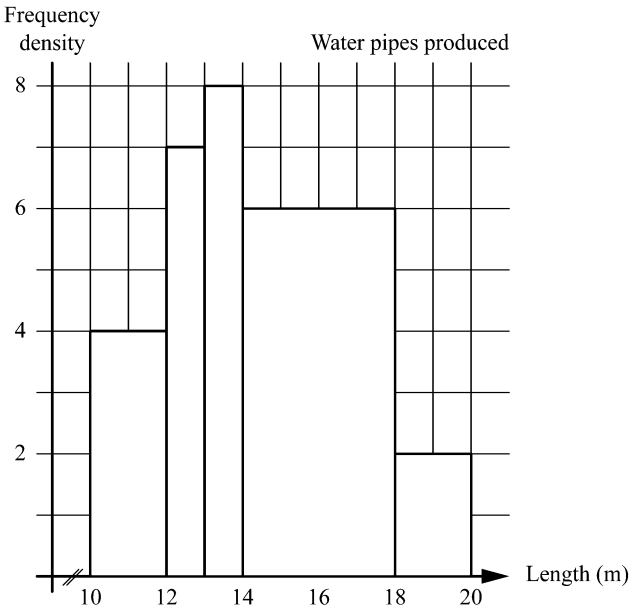
a) What are the modal classes? (a bi-modal distribution)

b) How many pupils are in their 16th year?

c) Make a grouped frequency table, first class $11 \leq \text{age} < 12$.

d) Calculate an estimate of the mean age of the pupils in the school (use the frequency table).

5. The number of water pipes of different lengths made in a factory during a month are shown in the histogram below.



a) How many pipes with length L m, $14 \leq L < 18$ were made?

b) What is the modal class?

c) In which class is the median length of pipe?

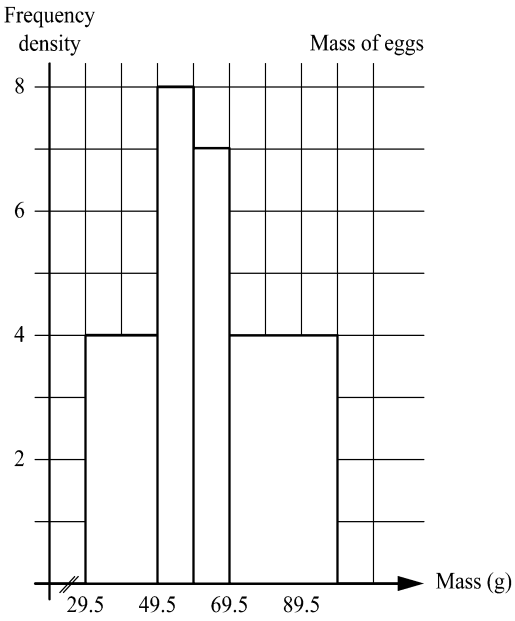
d) How many pipes were produced during the month?

e) Make an estimate for the total length of pipe produced.

f) Make an estimate for the mean length of pipe produced.

Continued on next page

6. The histogram illustrates the mass of eggs collected on a poultry farm during a day.



a) What is the modal class interval?

b) Complete the grouped frequency table

Mass (g)	30 ≤ m < 50			
Frequency				

c) Calculate an estimate of the mean mass of an egg.

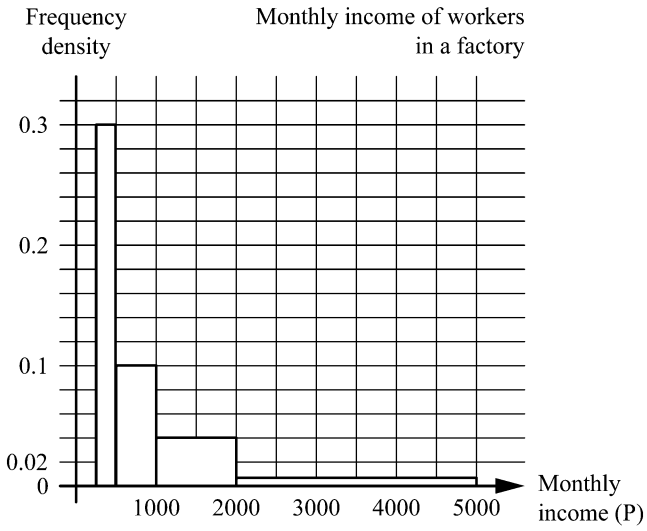
7. The monthly salary of the workers in a factory are illustrated in the histogram.

a) In what interval falls the income of most workers?

b) Complete the grouped income table.

Income (P)	250 ≤ I < 500			
Frequency Density				0.004
Frequency				

Continued on next page



c) In which class is the median monthly income?

d) Calculate an estimate for the mean income.

e) Which of the three averages, modal class, median class or estimated mean, best represents the data? Justify your choice.

8. The height of pupils in a class was distributed as follows

Height (cm)	Frequency
151 -155	4
156 - 160	10
161 - 165	16
166 - 170	22
171 - 175	26
176 - 180	15
181 - 185	2

a) Draw a histogram to represent these data.

b) Calculate an estimate of the mean height.

9. The ages of participants in a fund raising walk were distributed as follows:

Age	Frequency
10 -14	28
15 - 19	65
20 - 24	82
25 - 34	76
35 - 44	54
45 - 59	43
60 - 74	12

a) Draw a histogram to represent these data.

b) Calculate an estimate of the mean age.

Suggested answers are at the end of this unit.

Section G: Estimation of the median, quartiles and percentiles

Median, quartiles and percentiles can be estimated. Estimations are based on some assumptions. In this case the assumption is that the data is evenly distributed over the class interval. There are three methods considered: estimation using the cumulative frequency curve (section G1), estimation by using linear interpolation (section G2), and estimation of the median using the histogram (section G3).

Section G1: Estimation of the median, quartiles and percentiles from cumulative frequency curves

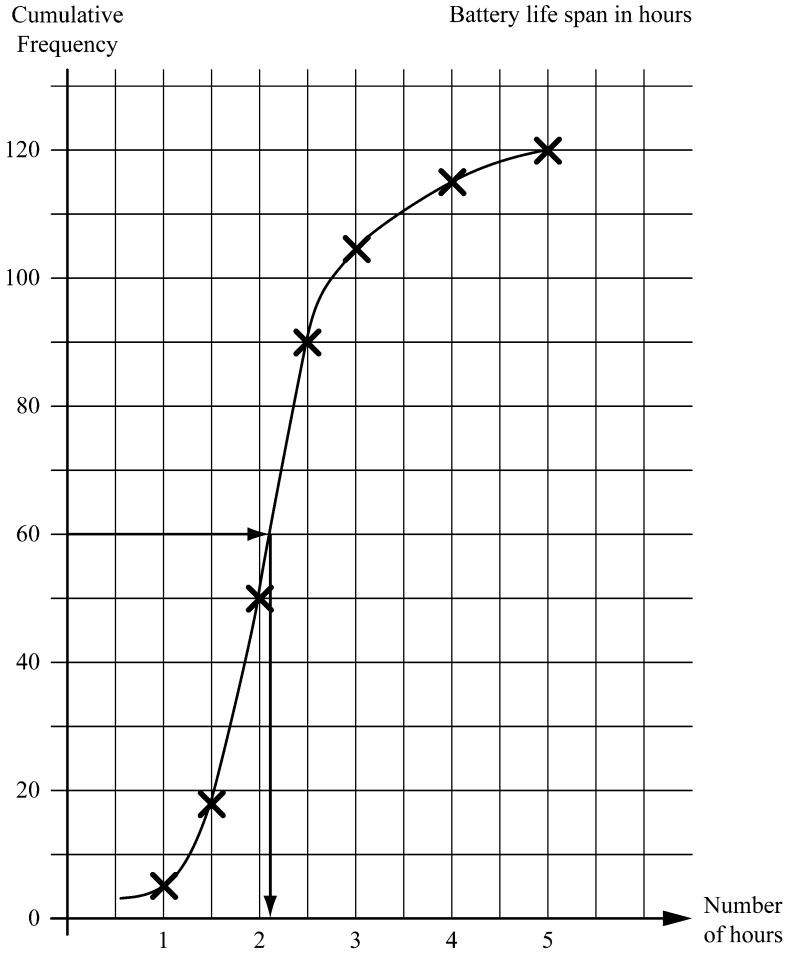
Making a cumulative frequency table and plotting the cumulative frequency curve:

A factory producing batteries might be interested in what percent of their batteries last up to 3 hours, what percent last more than 3.5 hours, etc. To obtain this information the data from the experiment on the time batteries lasted is to be represented in a cumulative frequency table and curve.

Cumulative frequency table

Battery life time t hours	Frequency	Cumulative frequency
$0 < t < 1$	5	5
$1 < t < 1.5$	12	17
$1.5 < t < 2$	32	49
$2 < t < 2.5$	40	89
$2.5 < t < 3$	16	105
$3 < t < 4$	9	114
$4 < t < 5$	6	120
TOTAL	120	

To plot the cumulative frequency curve the cumulative frequency is plotted at the end of each class. For example (1, 5), (1.5, 17), (2, 49) are points on the curve.



Reading from cumulative frequency curves

To obtain an estimate of the median battery life, start at the cumulative frequency axis at the 60th observation and follow the arrows to reach the time axis. (Half of 120, strictly speaking you are to take the average of the 60th and the 61st observation. However for larger number of observation generally for the median is just taken as half of the total).

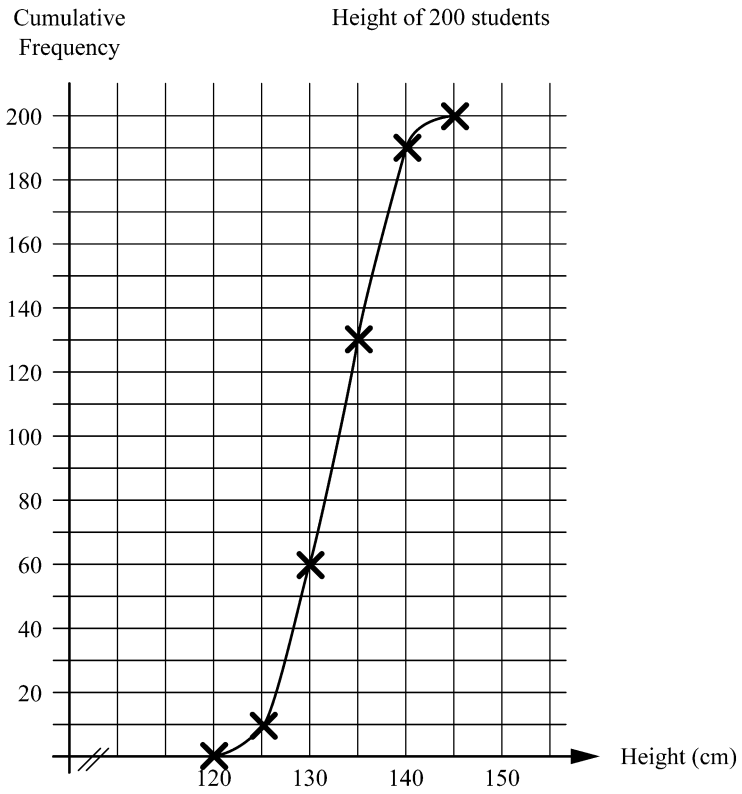
The estimate of the median is 2.1 hours. 50% of the batteries last for more than 2.1 hours (and 50% last less or equal to 2.1 hours).

Similarly you can read from the cumulative frequency curve the lower quartile LQ (at 30th observation being one-quarter of 120, giving an estimated 1.8 hours. Check this!) and the upper quartile UQ at 90th observation, being three-quarters of 120, giving an estimated value of 2.5 hours.

The interquartile range is defined as UQ - LQ. In the above example 2.5 - 1.8 = 0.7 h. The middle 50% of batteries last between 1.8 h and 2.5 h.

Self mark exercise 4

1. The cumulative frequency curve gives information on the height of 200 students.



a) From the cumulative frequency graph obtain:

(i) an estimate of the median

(ii) an estimate of the lower and upper quartile

b) Calculate an estimate of the interquartile range.

c) Estimate the number of students that are between 126 cm and 136 cm tall.

d) Estimate the number of student taller than 138 cm.

e) Complete the grouped frequency table.

Height (cm)	120-124	125-129				
Number of students						

f) Draw a histogram and a frequency polygon to represent the data.

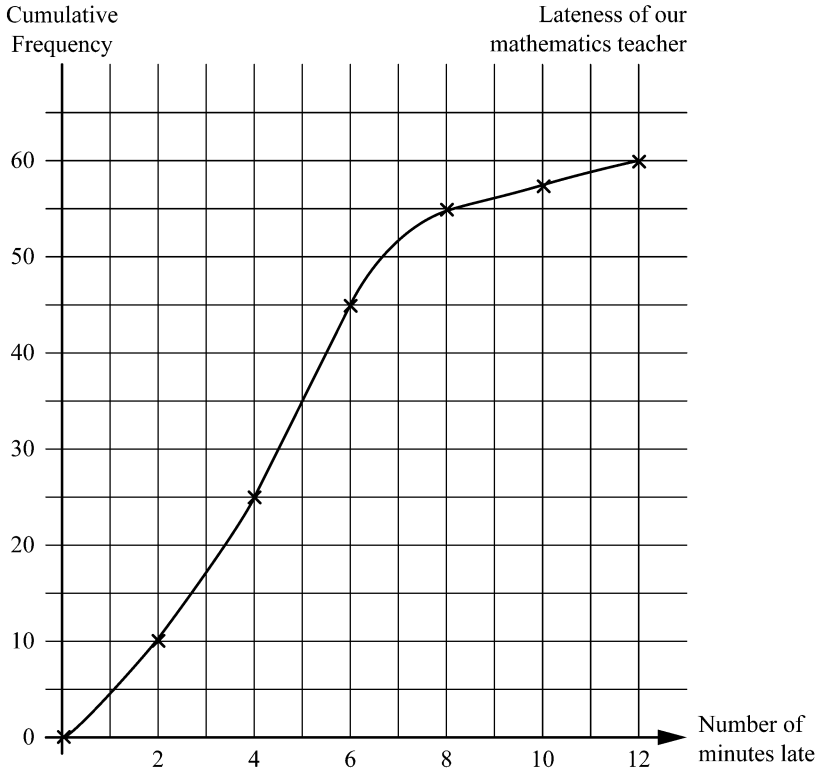
g) The frequency table, histogram, frequency polygon and cumulative frequency curve all display the same data.

(i) What are the advantages and disadvantages of each form of display?

(ii) When will you use which format?

(iii) Is one of the forms more useful than the others?

2. The cumulative frequency curve was made by students of Form 5 who kept record of the number of minutes their mathematics teacher came late to class on the 60 days of a term.



a) From the cumulative frequency graph obtain:

(i) an estimate of the median

(ii) an estimate of the lower and upper quartile

b) Calculate an estimate of the interquartile range.

c) How many times was the teacher between 5 and 10 minutes late?

d) Complete the grouped frequency table.

Number of minutes late	$0 < t < 2$	$2 < t < 4$				
Number of days						

e) Draw a histogram and a frequency polygon to represent the data.

f) Calculate an estimate of the mean number of minutes the teacher is late for class.

3. The table shows the length of time, in minutes, cars stayed in a parking lot in front of an office.

Time t	$0< t < 20$	$20< t < 40$	$40 < t < 60$	$60 < t < 90$	$90 < t < 120$
Frequency	12	42	78	22	6

a) Make a cumulative frequency table.

b) Draw a cumulative frequency curve.

c) Use your curve to obtain an estimate of

(i) the median (ii) the lower quartile (iii) the upper quartile

4. The table shows the times, in minutes, it took for patients to be treated in a clinic.

Time t (min)	$0 < t < 10$	$10 < t < 20$	$20 < t < 30$	$30 < t < 40$	$40 < t < 50$
Frequency	32	60	54	36	18

a) Make a cumulative frequency table.

b) Draw a cumulative frequency curve.

c) Use your curve to obtain an estimate of:

(i) the median (ii) the lower quartile (iii) the upper quartile

Suggested answers are at the end of this unit.

Percentiles

Quartiles divided the data into quarters, similarly **percentiles** divide the data into hundred parts.

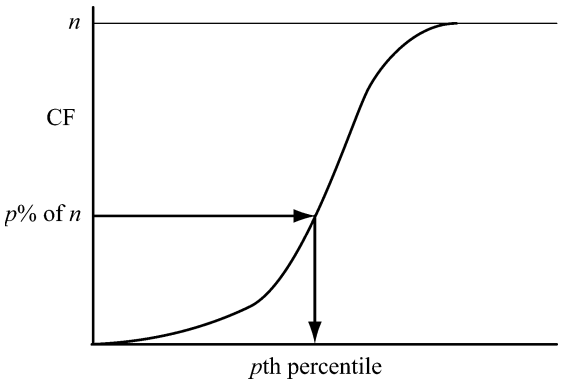
The median is the 50th percentile, the $\frac{1}{2}(n+1)$th value in the ordered sequence of the n values.

The lower quartile is the 25th percentile, the $\frac{1}{4}(n+1)$th value.

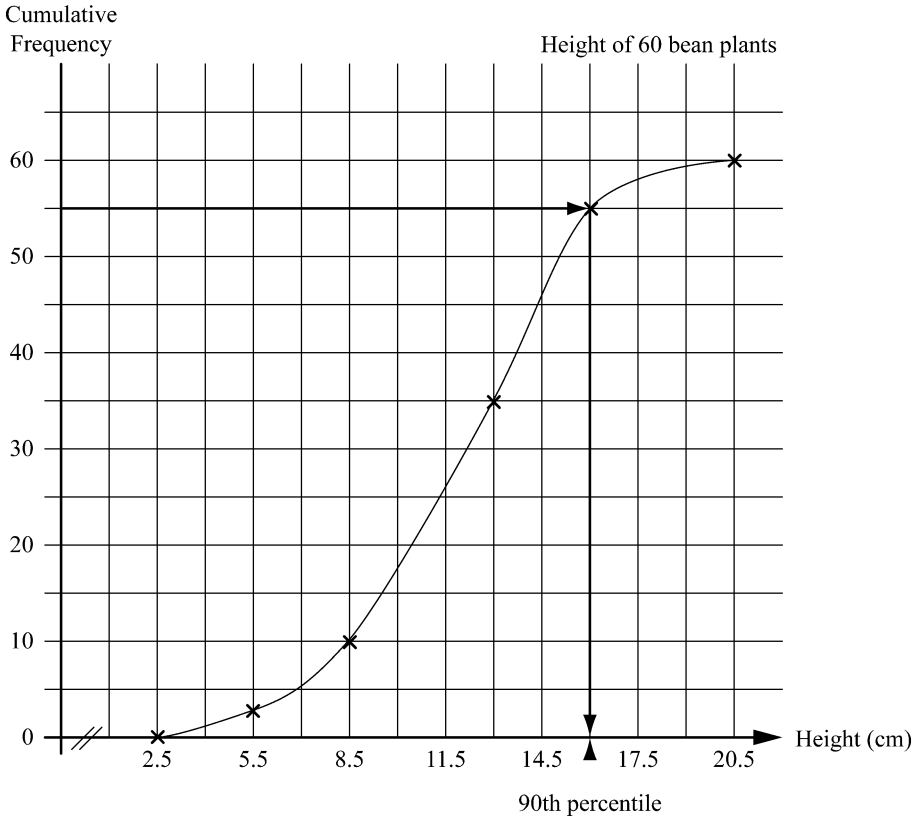
The upper quartile is the 75th percentile, the $\frac{3}{4}(n+1)$th value.

The pth percentile can be estimated from a cumulative frequency curve by taking (as an approximation) the $\frac{p}{100}$ of the total number of values: $p\% \times n$.

More exact it is the $\frac{p}{100}(n+1)$th value.



Example
The cumulative frequency curve shows the height of 60 bean plants.



Estimate the height of the tallest 10% of plants.

If 10% is taller than h cm, 90% will be below h cm. You are looking for the 90th percentile.

90% of (60 + 1) is 55. Following the arrows in the diagram: the 90th percentile is 16 cm. The tallest 10% of plants is between 16 cm and 20.5 cm.

Section G2: Estimation of median, quartiles and percentiles by linear interpolation

A multiple choice test was tried with 200 students. The number of correct responses are tabulated:

Number of correct answers	1 – 10	11 – 20	21 – 30	31 – 40	41 – 50
Number of students	12	43	71	49	25

From this grouped frequency table mean, median and mode cannot be calculated exactly.

You have seen that from the table you can obtain:

(i) an estimate of the mean (assuming all the values in the interval take the mid interval value)

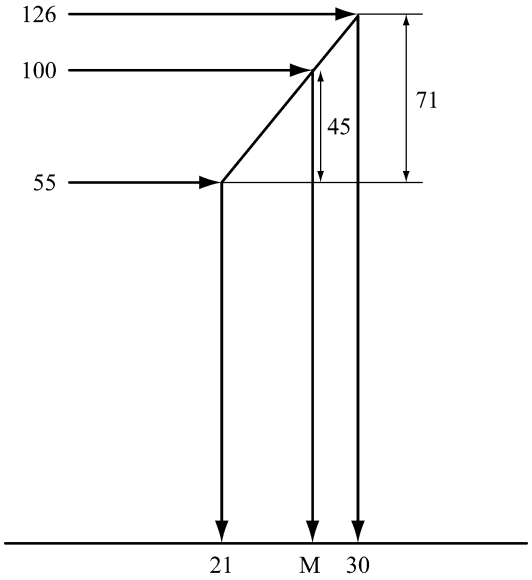
(ii) the modal class: 21 - 30 correct answers

(iii) The interval that contains the median, the $\frac{1}{2}(200+1)$th value, the average of the 100th and 101st value—both are in the interval with 21 - 30 correct answers. Generally you just take the $\frac{1}{2} \times 200$ which is the 100th value.

The cumulative frequency curve makes it possible to give an estimate of the median.

It is also possible to calculate an estimate of the median from the table by assuming that all the data values in the intervals are evenly (linear) distributed.

The class boundaries for the interval containing the median are 21 and 30 (discrete variable).



At the beginning of the interval you have covered already 12 + 43 = 55 values. So the interval starts at the point with co-ordinates (21, 55).

By the end of the interval you have covered the 71 values in the interval, so you end at the 55 + 71 = 126th value. Co-ordinates (30, 126).

Linear interpolation means that you assume these two points to be connected by a line segment.

You want the value M, corresponding with the 100th value i.e., 45 more than at the beginning of the interval. As the interval contains 71 values, you have to go $\frac{45}{71}$ of the way along the interval (which is 10 long).

So an estimate of the median is $20 + \frac{45}{71} \times 10 = 26$ to the nearest whole number.

The median number of correct questions is 26.

The process of estimation of the median from a grouped frequency table is called **linear interpolation**.

Quartiles and percentiles can be estimated by linear interpolation in a similar way.

Self mark exercise 5

1. The number of letters in the words in a newspaper article were counted. The result was

Number of letters	1 –3	4 –6	7 – 9	10 – 12	13 - 15
Frequency	57	46	28	6	3

   a) Calculate an estimate of the median word length.

   b) Calculate an estimate of the lower quartile and upper quartile word length.

2. The table gives the heights, to the nearest cm, of boys and girls in a class.

Height (cm)	151-155	156 - 160	161 - 165	166 - 170	171 - 175	176 - 180	181-185	186-190	191-195
Girls	3	8	9	16	12	2			
Boys	1	2	7	10	14	14	5	5	2

   a) Calculate an estimate of the median height of boys and girls.

   b) What height is exceeded by 80% of the (i) girls (ii) boys?

3. A multiple choice test was tried with 200 students. The number of correct responses are tabulated:

Number of Correct answers	1 – 10	11 – 20	21 – 30	31 – 40	41 - 50
Number of students	12	43	71	49	25

   a) Calculate an estimate of the number of correct answers of the top 10% of the pupils.

   b) Calculate an estimate of the number of correct answers of the bottom 10% of the pupils.

4. Use the following raw data of the length (mm) of nails found in packets of 'assorted nails'.

11	48	53	32	28	15	17	45	37	41
55	31	23	36	42	27	19	16	46	39
41	28	43	36	21	51	37	44	33	40
15	38	54	16	46	47	20	18	48	29
31	41	53	18	24	25	20	44	13	45

   a) Using the raw data calculate mean and median.

   b) Make a grouped frequency table taking class intervals 10 -14, 15 - 19, etc. Using the grouped frequency table calculate the estimate of the mean and the median.

   c) Make a grouped frequency table taking class intervals 10 - 19, 20 -29, etc. Using the grouped frequency table calculate the estimate of the mean and the median.

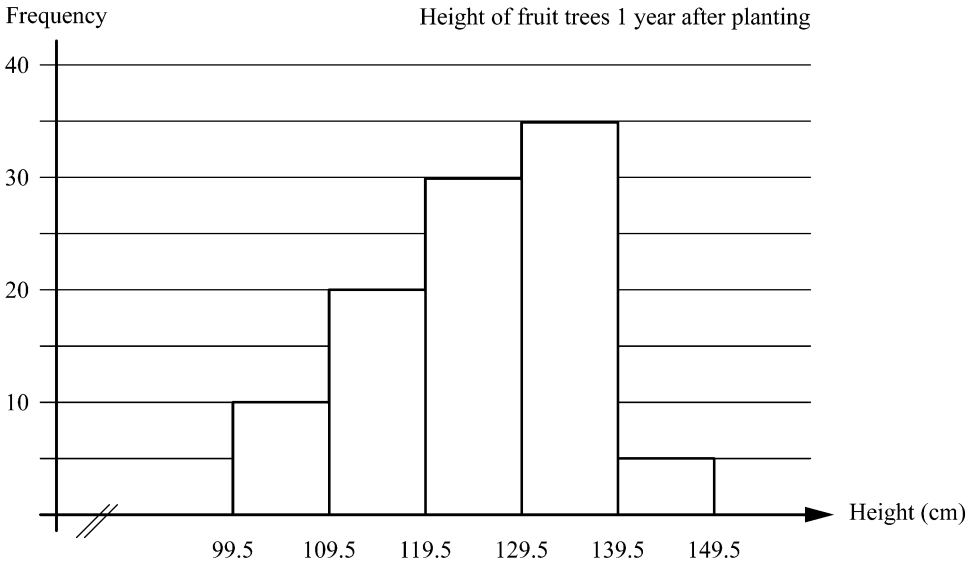
   d) What is the effect of changing class width on the estimate of (i) mean (ii) median?

Suggested answers are at the end of this unit.

Section G3: Estimation of the median from a histogram

In a histogram the area is proportional to the frequency. The median is the value in the middle and will therefore divide the area under the histogram into two equal parts.

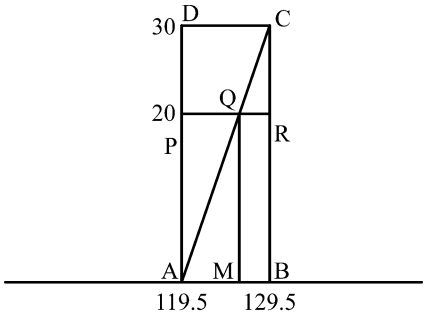
For example the histogram illustrates the height of 100 fruit trees 1 year after being planted.



Altogether there are 100 units of area contained in the histogram. We are looking for a line that will divide the area such that 50 units of area are to the left of the line and 50 units of area to the right.

The line is to be drawn somewhere in class 119.5-129.5 (containing 30 units of area). To the left of this class there are 10 + 20 = 30 units of area. We need 20 more units of area to make up the 50.

We need therefore to divide the 30 units of area of the class 119.5-129.5 in the ratio 20: 10 = 2 : 1.



To divide AB in the ratio 20 : 10 = 2 : 1 you first locate the point P on AD 20 unit in vertical direction.

Now AP : PD = 2 : 1.

Draw the diagonal AC. This diagonal meets the line PR at Q. Drop from Q a vertical to AB. The foot of this vertical (M) is the estimate for the median. This can be proved as follows using the similar triangles APQ and ACD. This implies AP : AC = PQ : CD = 2 : 3

But PQ = AM and AB = DC so also AM : AB = 2 : 3 or AM : MB = 2 : 1.

The estimate for the median is therefore read at the point M.

Self mark exercise 6

1. The times (to nearest tenths of a second) taken by pupils to run 50 m is tabulated in the following frequency distribution table.

Time(s)	Number of pupils
9.0-9.9	1
10.0-10.9	4
11.0-11.9	6
12.0-12.9	7
13.0-13.9	12
14.0-14.9	11
15.0-15.9	6
16.0-16.9	3

   a) Calculate an estimate for the median.

   b) Represent the data in a histogram.

   c) Use your histogram to obtain an estimate of the median.

2. Obtain an estimate for the median

   (i) by calculation

   (ii) from a cumulative frequency curve

   (iii) from a histogram

   The time taken by 110 pupils to complete a mathematics assignment (to nearest minute) is represented in the following frequency distribution table.

Time (min)	Number of pupils
5 - 14	10
15 - 24	14
25 - 34	40
35 - 44	31
45 - 54	5

Suggested answers are at the end of this unit.

Section H: Estimation of the mode from grouped data

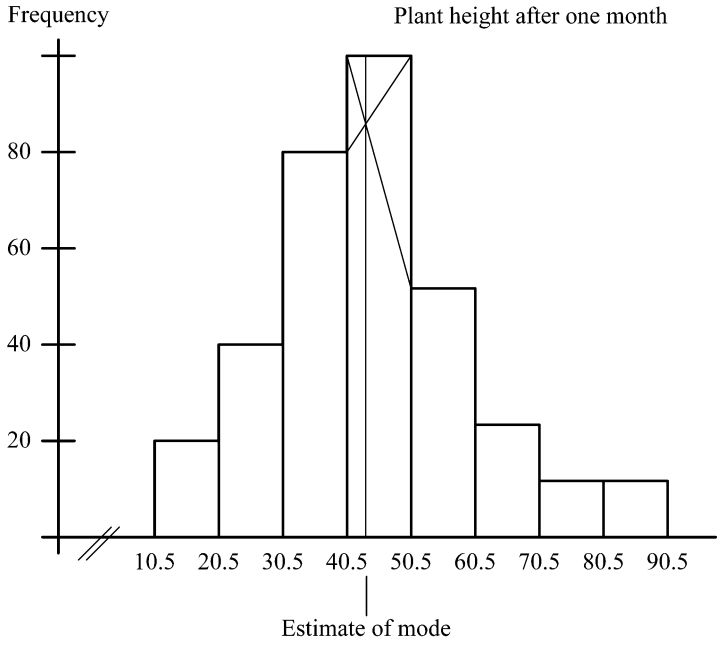
When data has been grouped into classes the class with the highest frequency can easily be identified: the modal class. An estimate of the mode can be made from the modal class.

1. Geometrical estimate of the mode from a histogram:

The length of plants (cm) on a plot one month after planting showed the following distribution.

Length	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90
Frequency	20	40	80	100	50	20	10	10

Representing the data in a histogram gives the following:



The modal class is 41 -50 cm.

An estimate of the mode can be found from the histogram by drawing the lines as illustrated in the diagram. This gives as estimated mode 43.

2. Estimation of the mode by calculation:

The modal class contains 20 more than the class below and 50 more than the class above the modal class. We therefore assume that the modal class is divided by the estimated mode in the ratio 20 : 50 = 2 : 5.

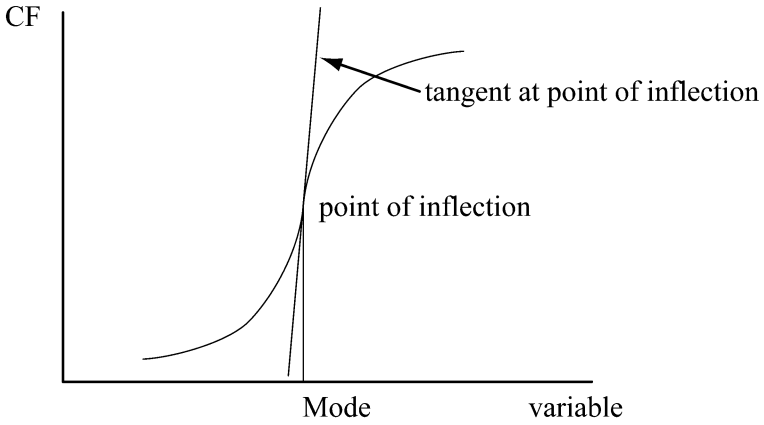
The calculated estimate is therefore $40.5 + \frac{2}{7}(10) = 40.5 + 2\frac{6}{7} \approx 43.4$

The modal length of the plants is 43.4 cm.

3. Estimate of the mode from the cumulative frequency curve:

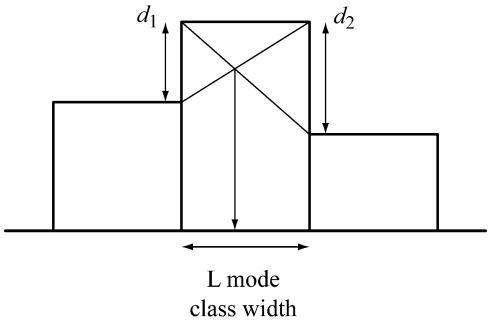
As the modal class is the class with the highest frequency on a cumulative frequency curve the cumulative frequency will increase fastest at the mode (the rate of increase of the cumulative frequency is highest at the mode). Calculus teaches us that in the case of a cumulative frequency type of curve

the mode must be located at the point of inflexion of the curve (the point where the direction of the curvature changes from concave to convex—or the other way round). Locating as best as possible the point of inflexion on a cumulative frequency curve will give an estimate of the mode. The tangent at the point of inflexion 'passes through the curve'.



Self mark exercise 7

1. For the calculation you need the modal class and the class below and above of it. Using the diagram below show that the calculated estimate of the mode is given by (lower class boundary L) + $\left(\frac{d_1}{d_1 + d_2}\right)$(class width)



2. As calculated estimate of the mode is frequently used:

   calculated estimate of mode =
   3 × calculated estimate of median - 2 × calculated estimate of mean

   Investigate the validity of this relation.

3. For the plant height data used above draw a cumulative frequency curve and use it to find an estimate of the mode.

   The length of plants (cm) on a plot one month after planting showed the following distribution.

Length	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90
Frequency	20	40	80	100	50	20	10	10

Continued on next page

4. The length cars remained in a parking lot of a supermarket was recorded during a day to the nearest minute. The results are tabulated below.

Length of stay (min)	6-25	26-45	46-65	66-85	86-105	106-125	126-145
Frequency	60	70	90	120	80	50	40

a) Represent the data in a histogram.

b) Use the histogram to obtain an estimate for the mode.

c) Calculate an estimate of the mode.

d) Make a cumulative frequency table and draw the cumulative frequency curve.

e) Use the cumulative frequency curve to obtain an estimate for the mode.

Suggested answers are at the end of this unit.

Section I: Boxplots or box and whisker diagrams

An average summarises all the collected data in a single value (mode, median or mean). This obviously leads to loss of information as conveyed by the original data. Reducing the data to five numbers chosen from across the range of values is more informative. A five number summary gives the minimum and maximum values (the extremities) together with the lower quartile, the median and the upper quartile. These five values can be illustrated in a box plot, also called box and whisker plot or diagram.

The stem-leaf diagram illustrated the marks scored in a maths test.

1	67
2	1368
3	1122335589
4	01122345678
5	0012667899
6	234566
7	01379
8	05
9	2

$n = 50$ 1 | 6 represent 16 marks.

To represent this data in a boxplot first find the five measures.

Minimum score 16

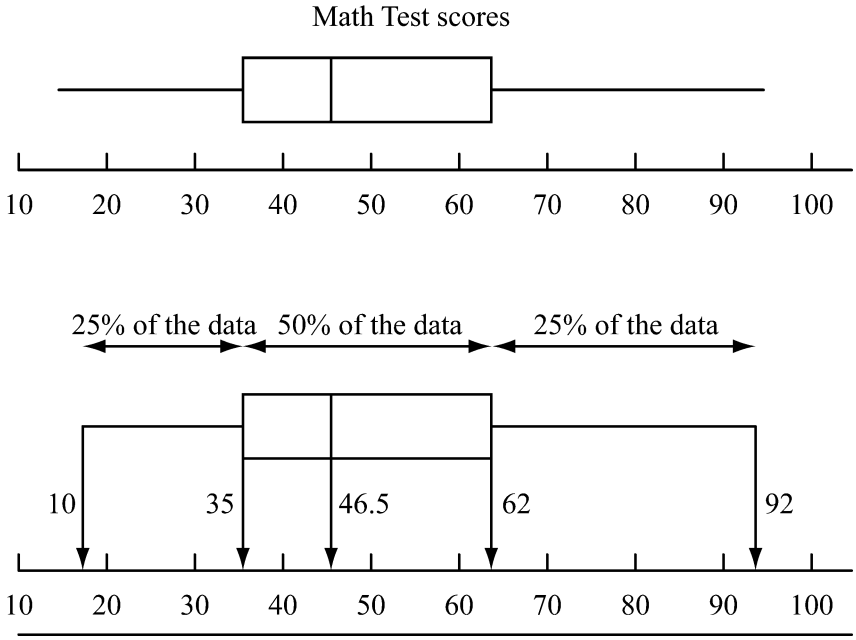
Maximum score 92

Median is mean of 25th and 26th score: $\frac{46+47}{2} = 46.5$ (Q_2)

The lower quartile is the median of the lower 25 observation, i.e., the 13th which is 35 (Q_1).

The upper quartile is the median of the upper 25 observations, i.e., the 38th which is 62 (Q_3).

Represented in a box plot:



Note that 25% of the pupils scored less than $Q_1 = 35$ (represented by the lower whisker).

50% of the pupils scored between $Q_1 = 35$ and $Q_3 = 62$ (represented by the box).

25% of the pupils scored more than $Q_3 = 62$ (represented by the upper whisker).

The diagram also illustrates the lowest (16) and the highest score (92).

The box illustrates that of the middle 50% of the pupils, 255 scored between 35 and 46.5 (the lower part of the box) and 25% between 46.5 and 62 (the upper part of the box).

The diagram not only illustrates the measures of central tendency but simple measures of the amount of spread (variability) can be obtained from the diagram:

the range	(maximum - minimum)
the interquartile range (IQR)	$Q_3 - Q_1$
semi-interquartile range (semi-IQR)	$\frac{1}{2}(Q_3 - Q_1)$

Self mark exercise 8

1. On a stretch of road with a 60 km/h limit the following speeds of cars were measured (in km/h).

57	53	53	71	73	54	69	56	58	49
56	53	52	82	62	61	60	71	75	60
57	61	58	78	64.					

   a) Represent the data in a stem-leaf plot.

   b) Use your stem-leaf plot to obtain the median speed, the upper quartile and the lower quartile speed.

   c) Represent the data calculated in a box plot.

   d) What does it imply that one whisker is longer than the other?

   e) Explain why the median is not in the centre of the box.

   f) What percent of the drivers was speeding over the limit?

Suggested answers are at the end of this unit.

Practice task 3

1. Discuss what you consider the most effective method to facilitate the learning of data handling. Illustrate with example activities.

2. a) Collect test data on the same topic from two parallel classes.

   b) Represent the data in (i) grouped frequency table (ii) histogram (iii) frequency polygon (both sets of data on the same axes) (iv) double stem-leaf plot.

   c) Which of the representations do you feel best represents the data? Justify your choice.

   d) Calculate (i) the exact value of the mean (ii) an estimate of the mean from the grouped frequency table (iii) the percent error in the estimated value.

   e) Which of the three averages, mean, mode or median, best represents the data? Explain.

   f) Represent the data of both classes in a box plot.

   g) What conclusions can you safely draw from the data?

3. a) Collect data for your school on the ages of the students by gender.

   b) Present your data in two frequency polygons, one for the girls and one for the boys, using the same axes.

   c) Calculate an estimate of the mean age of (i) boys (ii) girls in your school.

Continued on next page

d) Comparing the data for boys and girls comment on any differences and try to find an explanation for the differences.

4. Obtain the height of all pupils in your class. Investigate the effect of choosing different class intervals on the estimated mean height. Compare with the actual mean obtained from the raw data. What conclusion can you reach?

Summary

This unit began with a project-based approach to the teaching of central tendency. It ended with a number of self-marking exercises to teach you some lesser-known ways of representing quantitative data. It is hoped that you, and eventually your students, will benefit from this practical approach to statistics. Remember this caution from the Introduction to the unit: no set of projects could ever teach your students all, or even most, of the techniques we have covered. But since your students will benefit more (in later life) from the projects, a wise teacher omits many techniques of statistics in order to leave room for completion of interesting projects.

Unit 4: Answers to the self mark exercises

Self mark exercise 1

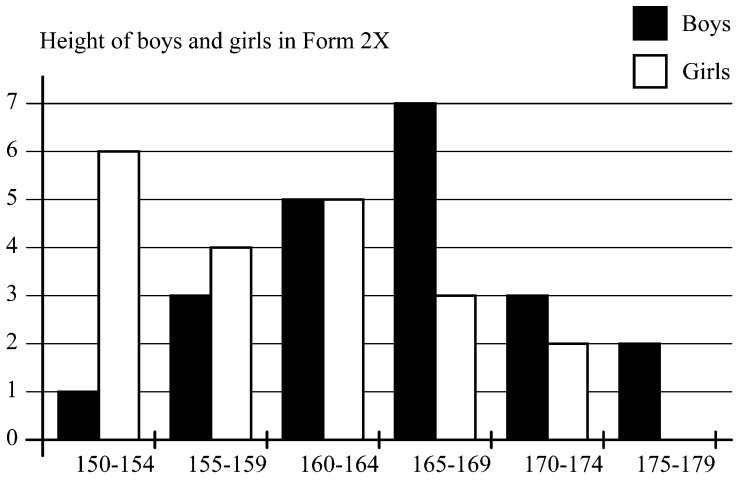
1. Mean height 1.61 cm (2 dp)
2. 11.4 pips (1 dp)
3. Mean 2.75 h, median 2.5 h, mode 2 h

   Use mean (majority of pupils 15 spend between 1 – 3 hours) or median (half of the pupils spend less than 2.5 h, half more)
4. Mean 2.6 days, median 2 days, mode 1 day.

   Half of the pupils that were absent were absent for one day. Most common is that if a pupil is absent it is just for one day. Mode best to use.
5. Girls: Mean height 160.0 cm (1 dp), Mode trimodal: 153 cm, 154 cm and 162 cm, median height 159.5 cm.

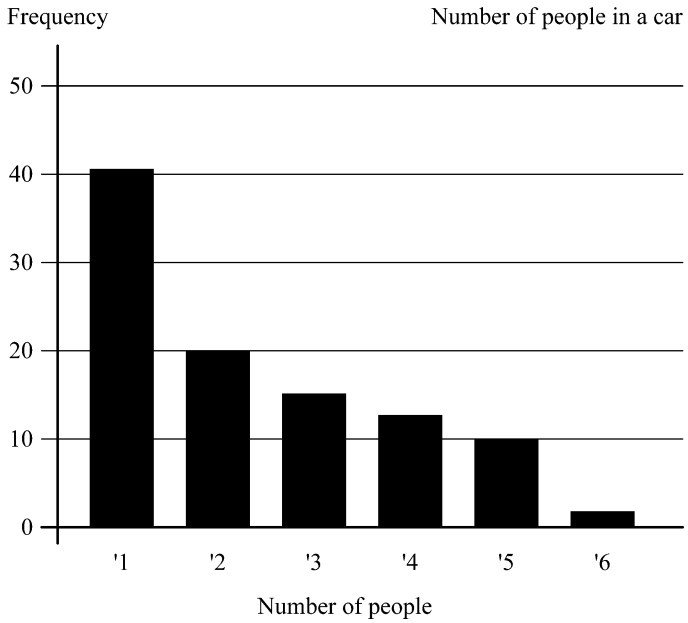
   Boys: Mean height 165.4 cm, multi modal 162 cm, 165 cm, 166 cm, 168 cm, median height 165 cm.

   Generally boys are taller than girls (higher mean and median). There is no 'most common' height (the distributions have no single mode).



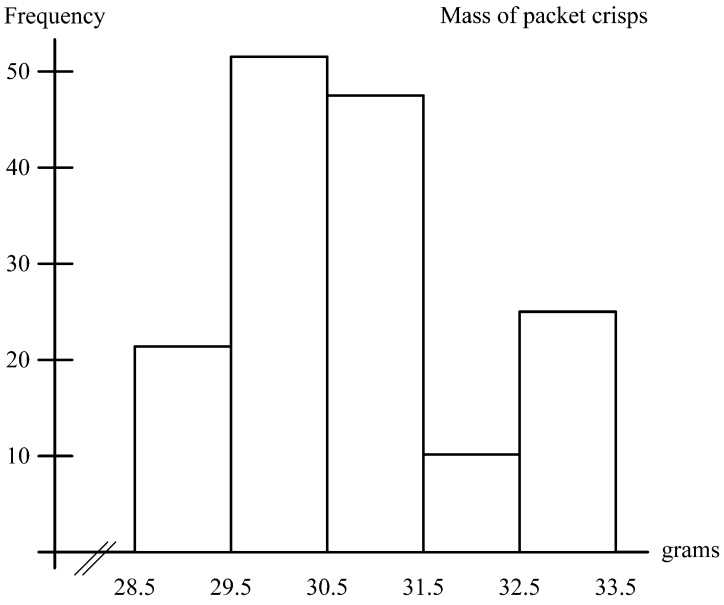
6. a) Mean 2.3 persons, mode 1 and median 2

   b) A bar or bar line graph would be appropriate (discrete data, ungrouped).



7. a) Mean mass 30.8 g, mode 30 g, median 31 g

   b) Frequency



8. a) Team A: mean 12.86 s, bimodal 13.3 s & 12.8 s, median 12.9 s

   Team B: mean 12.91 s, no mode, median 12.7 s

   b) Median or mean time for team A (two 1 dp they are the same)

   Median for team B 12.7 s.

   c) Team B with the lower median, half their runners run below 12.7 s, in team A this is 12.9 s.

Self mark exercise 2

1a. Mode, giving the ‘most common’ case.

b & c) Median (half of pupils' names are longer, half shorter than that number) or mean if there are no outliers.

d) Mode, the most common size will be of interest to traders.

e) Mode, the subject liked by most pupils. Median, mean non existing.

f) None of them might be very useful as most pupils will not have been absent. It makes sense to leave out those never absent (say 80%) and use the mode for those who were absent for one or more days.

g) Mode, giving the most common method. Mean and median non existing.

h) Mode is the only average available for this type of data.

2. (i) Mean 1.75, mode 1, median 1

   Mode / median as mean is influenced by outlier 6.

   (ii) Mean 59.9 % (1 dp), mode 75%, median 68%

   Median might be best reflection of pupils attainment. Mean is influenced by outliers. For few data mode does not make much sense.

3.

Arrangements	Examples	
mode<median<mean	1, 1, 4, 6	mode 1<median 2.5<mean 3
mean<median<mode	1, 3, 6, 9, 1	mean 6<median 8<mode 9
mode<mean<median	1, 1, 7, 8, 13	mode 1<mean 6<median 7
median<mean<mode	0, 1, 4, 10, 10	median 4< mean 5<mode 10
median<mode<mean	0, 3, 7, 8, 8, 28	median 7.5<mode 8<mean 9
mean<mode<median	$^{-}$28, $^{-}$8, $^{-}$8, $^{-}$7, $^{-}$3, 0	mean $^{-}$9<mode $^{-}$8<median $^{-}$7.5

4. **Mode**

   Advantages
   Simple to understand
   Not affected by extreme values (outliers)
   Only one that can be used for qualitative data
   Is an actual observation data

   Disadvantages
   Cannot be used in calculations or combined with mode of similar distributions
   Might not exist or distributions might be multiple modal

   Mean

   Advantages
   Includes all the values of the distribution
   Allows use in further calculations (e.g. SD)
   Allows combining with results from other similar groups

   Disadvantages
   Sensitive to outliers in the distribution. This might give a distorted picture.

Median

Advantages
Easy to understand
Not affected by extreme values

Disadvantages
Cannot be used in further calculations or combined with median of similar distributions

5. a) If p and q are positive integers:

$(p-q)^2 \geq 0$

$p^2 - 2pq + q^2 \geq 0$

$p^2 - 2pq + q^2 + 4pq \geq 4pq$

$p^2 + 2pq + q^2 \geq 4pq$

$(p+q)^2 \geq 4pq$

$p + q \geq 2\sqrt{(pq)}$

$\frac{p+q}{2} \geq \sqrt{(pq)}$

Arithmetic mean of p and $q \geq$ Geometric mean of p and q

b) Arithmetic mean A=

$\frac{p+q}{2}$ $\quad 2A = p + q$

Geometric mean = $\sqrt{pq}$ $\quad G^2 = pq$

$\frac{1}{H} = \frac{1}{p} + \frac{1}{q}$ $\quad \frac{1}{H} = \frac{p+q}{pq}$ $\quad \frac{1}{H} = \frac{pq}{p+q} = \frac{G}{2A}$

Self mark exercise 3

1. a)

Mass(g)	45≤m<50	50≤m<55	55≤m<60	60≤m<65	65≤m<70	70≤m<75
Frequency	4	8	12	16	5	2

b) 47

c) modal class $60 \leq m < 65$

d) Median in class $55 \leq m < 60$

e) Estimate of mean 59.2 g

2. a) 95

b)

Diameter (cm)	5≤d<6	6≤d<7	7≤d<8	8≤d<9	9≤d<10
Frequency	20	50	120	95	15

c) 3000 oranges d) Modal class $7 \leq d < 8$. e) 7.6 cm

3. a) 20

b)

Height (cm)	100≤h<109	110≤h<119	120≤d<129	130≤d<139	140≤d<149
Frequency	10	20	30	35	5

c) 100 d) $130 \leq h < 139$ e) 125 cm

4. a) $14 \leq a < 15$ and $15 \leq a < 16$

b) 80

c)

Age (years)	11≤a<12	12≤a<13	13≤a<14	14≤a<15	15≤m<16	16≤a<17
Frequency	10	90	100	150	150	80

Age (years)	17≤a<18	18≤a<19
Frequency	40	20

d) 14.8 years

5. a) 24 b) $14 \leq$ length < 18

c)

Length (m)	$10 \leq l < 12$	$12 \leq l < 13$	$30 \leq l < 14$	$14 \leq l < 18$	$18 \leq l < 20$
Frequency	8	7	8	24	4

Median (26th observation) in class $14 \leq$ length < 18

d) 51 e) 743.5 m f) 14.6 m

6. a) Modal class $70 \leq m < 100$

b)

Mass (g)	$30 \leq m < 50$	$50 \leq m < 60$	$60 \leq d < 70$	$70 \leq d < 100$
Frequency	80	80	70	120

c) 62.1 g

7. a) $250 \leq I < 500$

b)

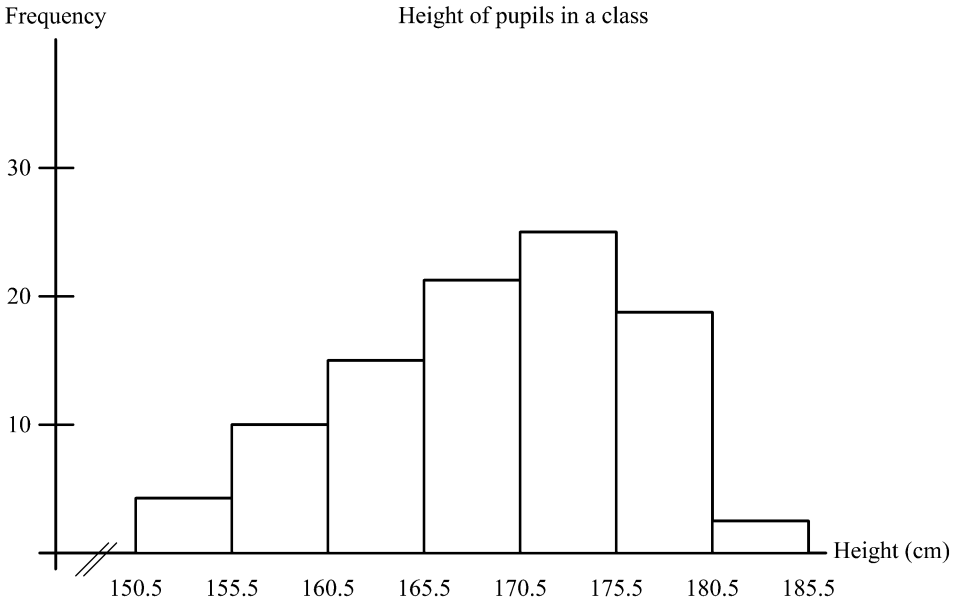
Income (P)	$250 \leq I < 500$	$500 \leq I < 1000$	$1000 \leq d < 2000$	$2000 \leq d < 5000$
Frequency density	0.3	0.1	0.04	0.004
Frequency	75	50	40	12

c) $500 \leq I < 1000$

d) P947

e) Mode, the salary earned by most people

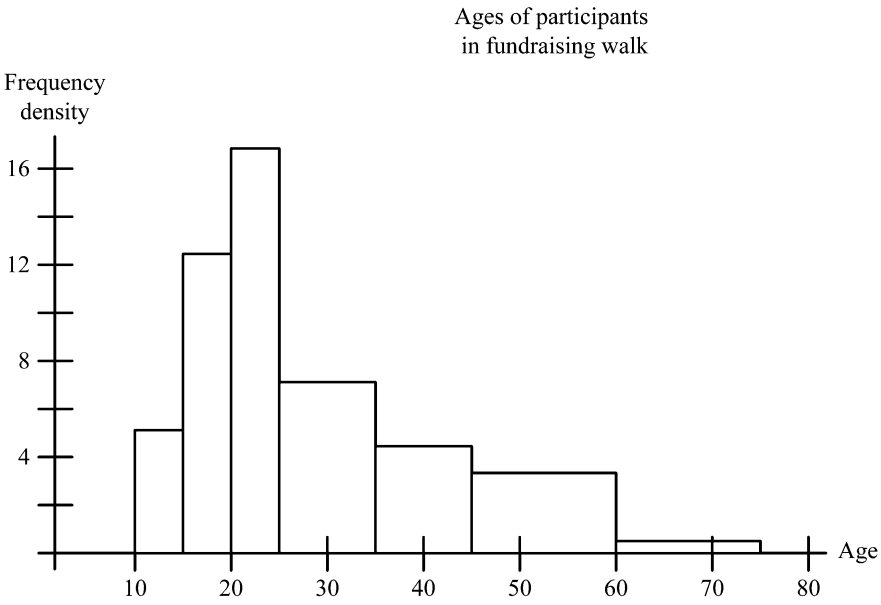
8. a)



b) 168.7 cm

9. a) Boundaries of bars 10, 15, 20, 25, 35, 45, 60 and 75

Frequency Density	5.6	13	16.4	7.6	5.4	2.9	0.8



9. b) 30.1 years

Self mark exercise 4

1. a) (i) 133 cm (ii) LQ 129 cm, UQ = 137.5 cm

b) IQR 8.5 cm

c) 120

d) 40

e)

Height	120 – 124	125 – 129	130 – 134	135 – 139	140 – 144
Frequency	10	40	70	65	15

f) Frequency polygon

g) (i) & (ii) (listed under advantages)

Frequency table

Advantages: Overview of data
Needed to draw various graphical representations
Calculations can be based on the table
Allows to obtain mean, median and mode (if grouped estimates of these measures can be obtained by linear interpolation procedures)

Disadvantage: Difficult to get an overall idea of the distribution.

Histograms

Advantages: For display of continuous grouped data (also used for discrete grouped data), especially if classes over of unequal width. Graphical estimates of median and mode can be obtained from the histogram.

Disadvantage: Difficult for pupils: where to take the class boundaries?

Frequency polygon:

Advantages: Gives impression of the distribution
Useful for comparison: more than one frequency polygon on the same axes (e.g. height of boys and girls).
Easy to plot using points with co-ordinates (midpoint of interval, frequency)

Disadvantage: No use for calculation of statistics

Cumulative frequency polygon

Advantage: Useful for obtaining estimates of median, quartiles, percentiles and mode

Disadvantage: Rather time consuming to draw

(iii) All have their specific use depending on what one wants to illustrate / calculate or estimate

2. a) (i) 4.8 minutes (ii) LQ 2.9 minutes, UQ 6.0 minutes

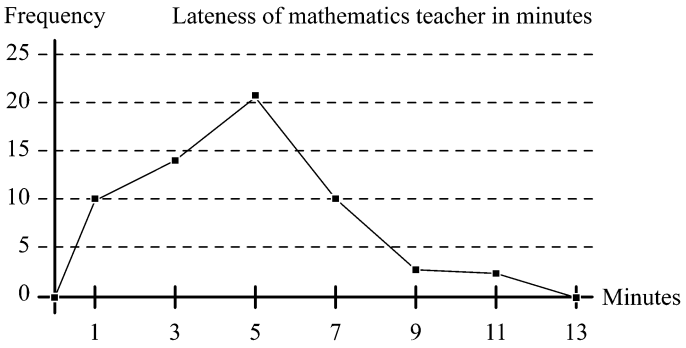
b) 3.1 minutes

c) 25

d)

Number of minutes late	$0 \le t < 2$	$2 \le t < 4$	$4 \le t < 6$	$6 \le t < 8$	$8 \le t < 10$	$10 \le t < 12$
Number of days	10	14	21	10	3	2

e)

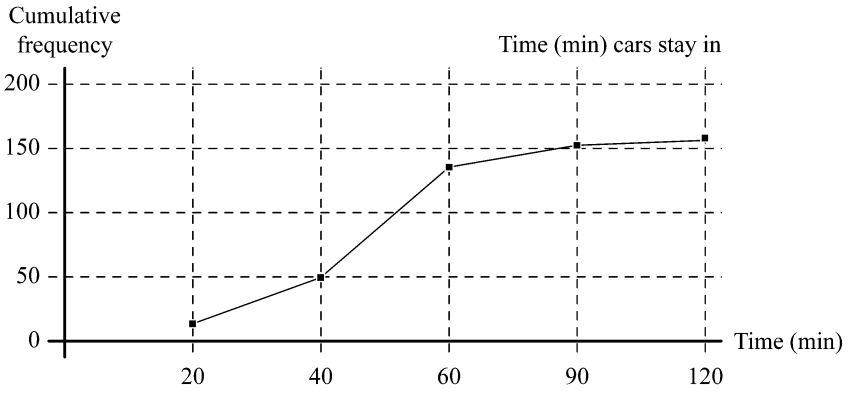


f) 4.6 minutes

3. a)

Time t	$0< t \le 20$	$20< t \le 40$	$40 < t \le 60$	$60 < t \le 90$	$90 < t \le 120$
CF	12	54	132	154	160

b)

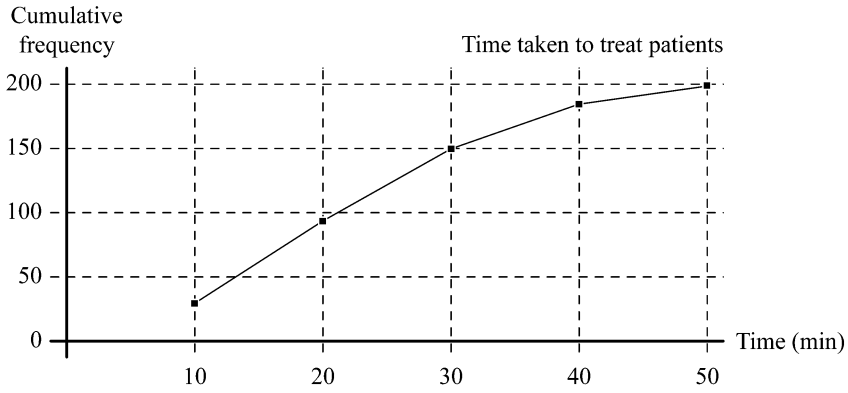


c) (i) 50 (ii) 35 (iii) 58

4. a)

Time t (min)	$0< t \le 10$	$10< t \le 20$	$20 < t \le 30$	$30 < t \le 40$	$40 < t \le 50$
CF	32	92	146	182	200

b)



c) (i) 22 (ii) 13 (iii) 30

Self mark exercise 5

1. a) 4.8
   b) LQ 2.8 UQ 7.2
2. a) boys 174,1 cm girls 167.1 cm
   b) boys LQ 168.3 cm, UQ 179.6 cm girls LQ 161.6 cm, UQ 171.3
   c) (i) 172.2 cm (ii) 180 5 cm
3. a) 43 or more b) less then 13
4. a) Mean 33.6 mm Median 36 mm

4b Class interval	Frequency	4c Class interval	Frequency
$10 \le l \le 14$	2	$10 \le l \le 19$	10
$15 \le l \le 19$	8		
$20 \le l \le 24$	5	$20 \le l \le 29$	10
$25 \le l \le 29$	5		
$30 \le l \le 34$	4	$30 \le l \le 39$	10
$35 \le l \le 39$	6		
$40 \le l \le 44$	8	$40 \le l \le 49$	15
$45 \le l \le 49$	7		
$50 \le l \le 54$	4	$50 \le l \le 59$	5
$54 \le l \le 59$	1		

   b) Mean 34.9 mm. Median 35.7 mm
   c) Mean 33.5 mm. Median 35.0 mm
   d) In this example both have reduced with increased class width. (It is a good investigation to find out whether or not that is always true!)

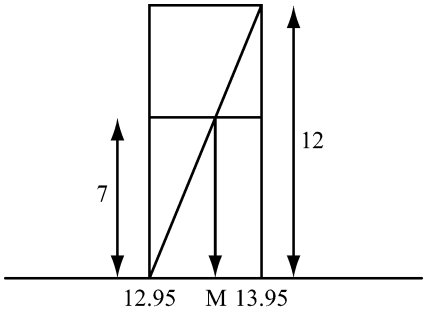
Self mark exercise 6

1. a) 13.53 s.

   c) Median in class with boundaries 12.95 – 13.95.

   Total area 50 units, to be divided into two.

   Left of the class is already 18 units, required 7 more.



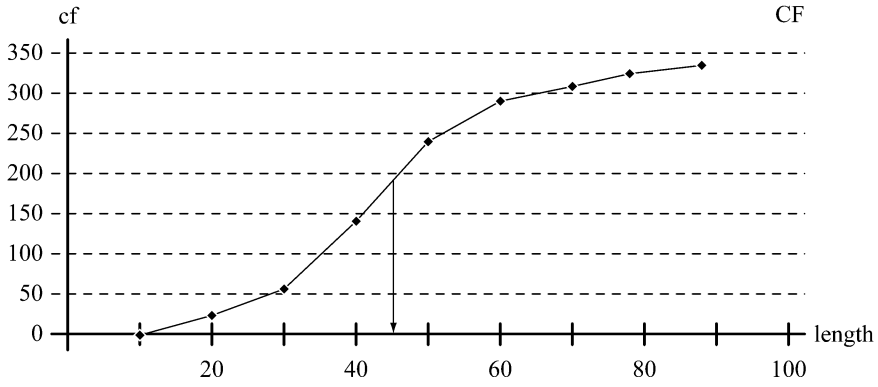
   Draw accurate diagram in your histogram as illustrated above. The median should be close to the calculated value of 13.53.

2. (i) 29.4 minutes

Self mark exercise 7

Hint. The class width is divided in the ratio $d_1 : d_2$.

3. Mode 45.



4. a) Use as class boundaries 5.5, 25.5, 45.5, etc. Class width is 20.

   Modal class 66 – 85. Use construction as illustrated in question 1.

   Estimated mode 74.1.

   d, e) Read length on horizontal axis at the point of inflexion as illustrated in question 3.

Self mark exercise 8

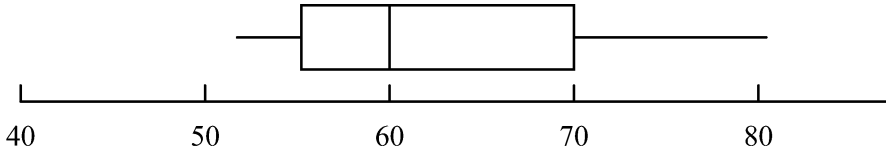
1. a)

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

$n = 25$ 6 | 7 represent 67 km/h

Median 60 km/h, LQ 55 km/h, UQ 70 km/h

c)



d) More drivers had a speed beyond the upper quartile speed than driver with a speed below the lower quartile speed.

e) The 25% of drivers with speed above the median speed of 60 km/h had a wider spread (60 to 70 km/h) than the 25% of drivers with a speed below the median speed (range 55 to 60 km/h).

f) 50%

Unit 5: Measures of dispersion

Introduction to Unit 5

In the previous unit you learned how to describe data sets using measures of central tendency. However a measure of central tendency cannot describe the data in sufficient detail. Look at the following two sets of data representing marks (out of 25) of two pupils on three different tests.

Pupil 1 scored 11, 12, 13 and pupil 2 scored 1, 12, 23. Both pupils have the same mean (12) and the same median (12), but can we say that they performed equally well? Pupil 1's marks are all very close together, while the marks of pupil 2 are widely spread (from 1 to 23). You could say that pupil 1 is more consistent in performance than pupil 2. It is the range or spread of marks which gives us this information. In this unit you are going to look at different measures of spread or dispersion.

Purpose of Unit 5

The main aim of this unit is to look at some basic measures of spread: how to calculate them and how to interpret them. This unit covers range, inter quartile range, variance and standard deviation. Box plots—as covered in Unit 4—are a useful graphical aid to visualise spread of data.

Objectives

When you have completed this unit you should be able to:

- calculate the range and inter quartile range of ungrouped data
- obtain an estimate of the inter quartile range of grouped data
- calculate the standard deviation and variance of ungrouped data
- calculate an estimate of the standard deviation and variance of grouped data
- use measures of central tendency and of spread to compare sets of similar data

Time

To study this unit will take you about five hours.

Unit 5: Measures of dispersion

Section A: Interquartile range for ungrouped data

The simplest measure to describe the spread or **dispersion** of values is the **range**. The range is the difference between the lowest and the highest values.

The problem with the range is that only two values are used and so it can give a wrong impression if one (or both) of the values is very high or very low.

The problem of distortion by extreme values can be overcome by calculating the range of the central half (middle 50%) of the values. This is called the **interquartile range**.

Example 1

7 students estimated the length of a book to the nearest cm. In order their estimates were: 25 cm 28 cm 30 cm 31 cm 32 cm 34 cm and 37 cm.

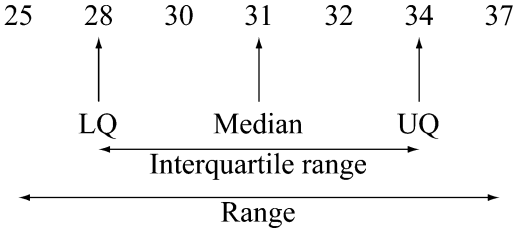
The **range** is 37 – 25 = 12 cm.

The **median** is the middle value, $\frac{7+1}{2}$ = 4th value, which is 31 cm.

The **lower quartile (LQ)**, the value $\frac{1}{4}$ away in the list of values, so the value of the $\frac{7+1}{4}$ = 2nd term, which is 28 cm.

The **upper quartile (UQ)**, the value $\frac{3}{4}$ away in the list of values, this is the $\frac{3}{4}(7+1)$ = 6th term, which is 34 cm.

The **interquartile range** is (UQ) - (LQ) = 34 - 28 = 6 cm.



Example 2

The estimate of the length of the book by 10 students was in order to the nearest cm: 24 24 26 28 29 31 32 33 34 36.

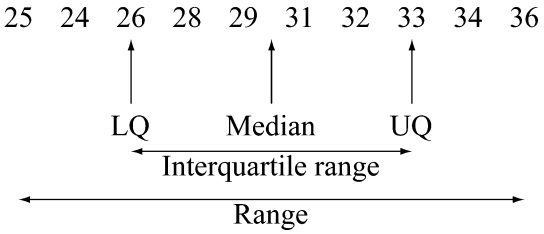
The **range** is 36 – 24 = 12 cm.

As $\frac{10+1}{2}$ = 5.5, the median is the average of the 5th and 6th term in the ordered sequence. **Median** is $\frac{29+31}{2}$ = 30 cm.

As $\frac{10+1}{4} = 2.75$, the **lower quartile** is the value of the third term which is 26 cm.

Similarly the upper **quartile value** is the value of the 8th which is 33.

The interquartile range is 33 - 26 = 7 cm.



Self mark exercise 1

1. During a term a pupil obtained the following percent marks.

Setswana	56	49	63	58	52	50	57	61	
English	61	70	53	60	57	52	48	79	65
Science	68	56	58	73	39	47	55	76	
Maths	45	46	42	48	40	45	44	41	47

   a) Find for each subject the range and interquartile range.

   b) In which subject is the pupil most 'consistent'. Explain.

   c) What is the pupil's 'best' subject? Explain.

Suggested answers are at the end of this unit.

Section B: Interquartile range for grouped data

If the data is grouped, you use either:

(i) a cumulative frequency curve to obtain estimates of the lower and upper quartile (see section G1on the cumulative frequency curve in unit 4)

(ii) linear interpolation to obtain estimates of the lower and upper quartiles (See section G2 on linear interpolation in unit 4)

From these an estimate for the inter quartile range can be obtained.

Self mark exercise 2

1. The height of a group of pupils is distributed as in the table below.

Height (cm)	151-155	156-160	161-165	166-170	171-175
Frequency	6	9	14	23	8

Continued on next page

a) Make a cumulative frequency table and draw a cumulative frequency curve of the data.

b) Use the cumulative frequency curve to obtain an estimate for the interquartile range.

2. A machine is to produce nails of 7 cm length. A sample is taken and measured to the nearest 0.1 cm. The results are tabulated:

Length of nail	6.7-6.8	6.8-6.9	6.9-7.0	7.0 - 7.1	7.1-7.2
Frequency	4	11	36	44	5

a) Make a cumulative frequency table and draw a cumulative frequency curve of the data.

b) Use the cumulative frequency curve to obtain an estimate for the interquartile range.

3. The ages of people attending a football match were distributed as in the following table.

Age	Frequency
5-9	8
10-14	26
15-19	74
20-24	90
25-29	124
30-34	142
35-39	86
40-44	54
45-59	26
60-74	15

a) Make a cumulative frequency table and draw a cumulative frequency curve of the data.

b) Use the cumulative frequency curve to obtain an estimate for the interquartile range.

4. Two types of batteries were tested on the number of hours they lasted.

Number of hours	5-10	10-15	15-20	20-25
Type A	8	37	43	12
Type B	16	30	32	22

a) Calculate estimates for the quartiles and obtain an estimate for the interquartile range of each type of battery.

b) Which type would you recommend a school to buy? Justify your answer.

Suggested answers are at the end of this unit.

Section C: Standard deviation of ungrouped data

There are three commonly used measures of dispersion. You know already two of them: **range** and **interquartile range**.

The disadvantage of both of these measures is that not *all* data are used. For the range you use only the two extreme values: the highest and the lowest and this can be misleading.

Using the interquartile range ignores the top and bottom quarter of the values.

A measure of spread using all the data is the **standard deviation**. It uses the differences (**deviations**) of the data from the mean.

To calculate the standard deviation you calculate

(i) the mean of the data

(ii) the deviations of the data from the mean

(iii) the mean of the squares of the deviations (which is called the **variance**)

(iv) the square root of the variance which is the standard deviation

Calculation of the standard deviation is best done using a table or an electronic device (calculator or computer). (Only the electronic method is worth remembering! Do not assess your students on their ability to calculate a standard deviation by hand.)

Disadvantage of using the standard deviation is that it is difficult to understand intuitively.

Example

The length of six leaves from a certain tree were measured to the nearest cm.

7 cm, 9 cm, 11 cm, 12 cm, 12 cm, 14 cm.

x	$x-\bar{x}$	$(x-\bar{x})^2$
7	-4	16
9	-2	4
11	0	0
12	1	1
12	1	1
15	4	16
$\sum x = 66$		38

$$\bar{x} = \frac{\Sigma x}{6} = \frac{66}{6} = 11$$

$\sum x$ means sum of all the data x. $\sum x^2$ means sum all the squares of the data x.

$$\text{Variance} = s^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{38}{6}$$

The standard deviation is s.d. = $\sqrt{\frac{38}{6}} = 2.5$ cm (1 dp)

A formula for the standard deviation is s.d. = $\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$

At times the form s.d. = $\sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$ is more convenient to use.

Practice task 1

Derive the form $\sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$ from $\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$. Remember that $\frac{\sum x}{n} = \bar{x}$

Suggested answer at the end of this unit.

Section D: Standard deviation of grouped data

The standard deviation of a set of data occurring with given frequencies can be found as illustrated in the following example.

Example

A sample of 60 batteries was tested as to how long (in hours) they lasted (the life span of the battery).

The results are in the table.

Battery life span (h)	12	13	14	15	16	17	18
Frequency	6	8	10	14	15	5	2

The working can be laid out as follows.

x	$x2$	f	fx	$fx2$
12	144	6	72	864
13	169	8	104	1352
14	196	10	140	1960
15	225	14	210	3150
16	256	15	240	3840
17	289	5	85	1445
18	324	2	36	648
$\Sigma x=$	$\Sigma x2=$	$\Sigma f=$	$\Sigma fx=$	$\Sigma fx2=$
105	1603	60	887	13259

The mean is $\dfrac{\sum fx}{\sum f} = \bar{x} = \dfrac{887}{60} = 14.78$ h (2 dp)

The standard deviation s.d. = $\sqrt{\dfrac{\sum fx^2}{\sum f} - \bar{x}^2} = \sqrt{\dfrac{13259}{60} - (14.78)^2} = 1.6$ h (1 dp)

If the data is grouped or continuous an **estimate** of the standard deviation can be computed by assuming that all the data in a particular class interval has the value of the mid-point of the class interval. If the mid-interval value is indicated by m, the relation for mean and standard deviation becomes:

$\dfrac{\sum fm}{\sum f} = \bar{m}$ and s.d. $= \sqrt{\dfrac{\sum fm^2}{\sum f} - \bar{m}^2}$

Self mark exercise 3

1. In a test (maximum marks 50) the distribution of the marks was as follows.

Marks	1 - 10	11 - 20	21 - 30	31 - 40	41 - 50
Frequency	2	14	22	26	16

   a) Copy the table below. Complete the column for the mid-interval values and the other columns.

Marks	Frequency f	Mid-interval values m	m^2	fm	fm^2
1 – 10	2	5.5	30.25	11	60.5
11 – 20	14				
21 – 30	22				
31 – 40	26				
41 – 50	16				
	$\Sigma f =$			$\Sigma fm =$	$\Sigma fm^2 =$

   b) Calculate an estimate of the mean mark and the standard deviation.

   c) Explain why mean and standard deviation are estimates and not the exact value.

2. The time spent by customers in a supermarket was measured and the distribution was as shown:

Time (min)	$0 < t \leq 10$	$10 < t \leq 20$	$20 < t \leq 30$	$30 < t \leq 40$	$40 < t \leq 50$
Frequency	22	86	62	21	9

   Calculate an estimate of the mean time spent by customers in the supermarket, and the standard deviation.

Continued on next page

3. The height of 12-year-old boys and girls in a school was measured. The heights were distributed as in the table below.

Height	$130 < h \leq 135$	135-140	140-145	145-150	150-155	155-160	160-165	165-170	170-175	175-180
Frequency for boys	0	2	5	17	24	31	36	28	6	1
Frequency for girls	1	6	8	20	32	32	28	18	4	1

   a) Using the same axes draw a frequency polygon for the heights of the boys and of the girls.

   b) Use the frequency polygon to compare the heights of boys and girls.

   c) Calculate an estimate of the mean height and standard deviation for both boys and girls.

   d) Compare the height of boys and girls using your estimated values of mean and standard deviation.

4. Two novels were compared with each other by counting the number of words in the sentences in a section of the novel.

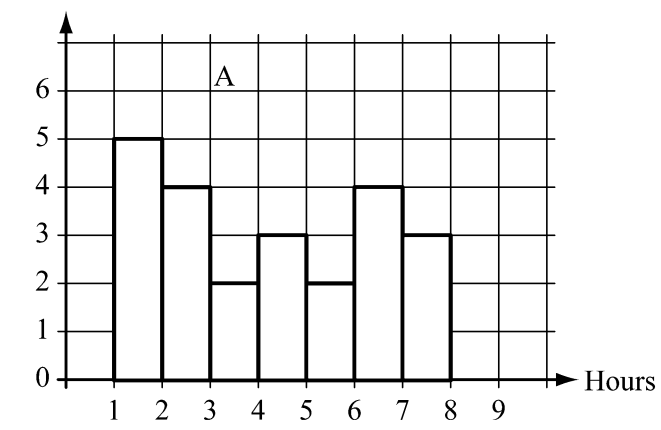
Number of words	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39
Novel A	12	15	10	26	14	8	4
Novel B	8	18	12	31	18	4	2

   a) Calculate for each novel an estimate for the mean number of words in a sentence and the standard deviation.

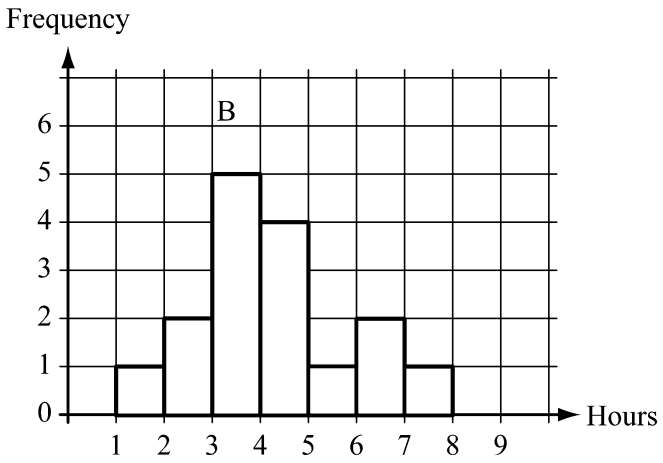
   b) Compare the results of a. Which novel do you think is easier to read? Justify your answer.

5. The number of hours spent on sports by two groups of pupils, group A and group B, are represented in the histograms below.

   Which of the data represented in the histograms has a greater standard deviation? Justify your answer.



Continued on next page



6. Investigate what happens to mean and standard deviation of a set of data if

   (i) to each value the same number N is added

   (ii) each value is multiplied by the same number k

7. a) Find the mean and standard deviation of 12, 15, 16, 14, 17, 13.

   Using the result of 5(i) write down the mean and standard deviation of

   b) 22, 25, 26, 24, 27, 23

   c) 83, 86, 87, 85, 88, 84

   d) 8, 11, 12, 10, 13, 9

8. a) Find the mean and standard deviation of 52, 61, 73, 68, 49, 67. Using your result from 5(ii) write down the mean and standard deviation of

   b) 5.2, 6.1, 7.3, 6.8, 4.9, 6.7

   c) 26, 30.5, 36.5, 34, 24.5, 33.5

   d) 208, 244, 292, 272, 196, 268

Suggested answers are at the end of this unit.

Practice task 2

1. a) Collect data on the height of the boys and girls in one or two of your classes.

   b) Calculate mean and standard deviation for boys and girls separately from the raw data.

   c) Make a grouped frequency table separating boys and girls. Use class widths of 5 cm and 10 cm.

   d) Use the grouped frequency tables to calculate an estimate for the mean and the standard deviation for boys and girls separately for both class widths.

   e) Calculate an estimate of the interquartile range from both the grouped frequency tables.

   f) Using your calculated data, compare and make some valid statements.

   g) Comparing the **exact values** of mean and standard deviation with the **estimated values** from the grouped frequency tables (with width 5 cm and width 10 cm), which of these three do you consider to represent the data best?

Unit 5: Answers to the self mark exercises

Self mark exercise 1

1a)

Subject	Range	IQR
Setswana	63 – 49 = 14	59.5 – 51 = 8.5
English	79 – 48 = 31	67.5 – 52.5 = 15.0
Science	76 – 39 = 37	70.5 – 51 = 19.5
Maths	48 – 40 = 8	46. 5 – 41. 5 = 5.0

b) Mathematics, smallest range and IQR

c) English with median 60%

Self mark exercise 2

1a)

Height	151 - 155	156 - 160	161 - 165	166 - 170	171 - 175
Frequency	6	9	14	23	8
Cumulative Frequency	6	15	29	52	60

b) Plot the points (155.5,6), (160.5, 15), (165.5, 29), (170.5, 52), (175.5, 60)

c) IQR ≈ 168.5 – 160.0 = 8.5 cm

2a) Plot (6.85, 4), (6.95, 15), (7.05, 51), (7.15, 95), (7.25, 100)

b) IQR ≈ 7.09 – 6.99 = 0. 1 cm

3a) Plot (10, 8), (15, 34), (20, 108), (25, 198), (30, 322), (35, 464), (40, 550), (45, 604), (60, 630), (75, 645)

b) IQR ≈ 36 – 23 =13 years.

4a) Use linear interpolation for the estimates (to 2 dp)

	LQ	UQ	IQR	Median
Type A	12.30	18.49	6.19	15.58
Type B	11.50	19.53	8.03	15.63

b) Type A , smaller IQR and more consistent in performance [as both types have the same median (to 1 dp)]

Practice task 1

$$\sqrt{\frac{\sum(x-\bar{x})^2}{n}} = \sqrt{\frac{\sum(x^2-2x\bar{x}+\bar{x}^2)}{n}} = \sqrt{\frac{\sum x^2}{n} - \frac{2n\bar{x}\sum x}{n} + \frac{n\bar{x}^2}{n}}$$

$$= \sqrt{\frac{\sum x^2}{n} - 2\bar{x}^2 + \bar{x}^2} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

Self mark exercise 3

1b) mean 30.5 SD 10.7

c) It is not known how the frequencies are distributed over each interval. Mid interval values were used, i.e., assuming that the mid interval value had the indicated frequencies.

2. Mean 20.5 minutes (1 dp) SD 9.7 minutes

3c) Boys Mean 158.6 cm, SD 8.0 cm (1 dp)

Girls Mean 156.0 cm, SD 8.6 cm

d) Boys are on average taller than girls. The heights of the girls have a wider spread around their mean than is the case for the boys.

4a) Novel A : Mean number of words per sentence 20.1, SD = 8.3 (1 dp)

Novel B : Mean number of words per sentence 19.9, SD = 7.2 (1 dp)

b) Novel B lower mean and less spread about the mean.

5. Group A, wider spread than group B

6. (i) Mean increases by N, SD remains unchanged.

(ii) Mean and standard deviation change by factor k.

7a) Mean 14. 5, SD 1.7 (1 dp)

b) Mean 14.5 + 10 = 24.5, no change in SD

c) Mean 14. 5 + 71 = 85.5, no change in SD

d) Mean 14.5 – 4 = 10.5, no change in SD

8a) Mean 61.67, SD = 8.67 (2 dp)

b) Mean and SD of 8a divided by 10: Mean 6.17, SD 0.87 (1 dp)

c) Mean and SD of 8a divided by 2: Mean 30.8, SD 4.3 (1 dp)

d) Mean and SD of 8a multiplied by 4: Mean 246.7, SD 34.7 (1 dp)

References

Cockcroft W. H., *Mathematics Counts*, 1982, HMSO London

Additional References

In preparing the materials included in this module we have borrowed ideas extensively from other sources and in some cases used activities almost intact as examples of good practice. As we have been using several of the ideas, included in this module, in teacher training over the past 5 years the original source of the ideas cannot be traced in some cases. The main sources are listed below.

Mathematics Teacher, Journal of the National Council of Teachers of Mathematics

Mathematics in School, Journal of the Association of Teachers of Mathematics

NCTM, *Dealing with Data and Change*, 1991, ISBN 087 353 3216

NCTM, *Data Analysis and Statistics*, 1992, ISBN 087 353 3291

Owens, D. T., 1993, *Research Ideas for the Classroom,. Middle Grades Mathematics*, NCTM ISBN 002 895 7954

Further reading

Bank, T. et al. 1999, *Mathematics for SEG GCSE Intermediate Tier*, Causeway Press Limited, ISBN 187 392 9870.

Maths in Action, 1999, Intermediate 1, Nelson 017 431 4973

Maths in Action, 1999, Intermediate 2, Nelson 017 431 4949

Maths in Action Statistics for Higher Mathematics, Nelson 017 431 4965

Glossary

Census	collection of data on the whole population
Class interval	the width of the groups used in grouped frequency tables
Continuous data	data that can take any value within a certain range (e.g., height / mass of persons)
Data	facts, numbers, measures collected on a population or sample
Descriptive Statistics	branch of statistics covering collecting, representing and analysing of data
Discrete data	data that can take only specific values (e.g., shoe size) or falls in specific categories (e.g., sex)
Estimation theory	theory that describes how the statistics obtained on a sample can be used to estimate the parameters of the population
Experimental data	data collected using a scientific experimental design, frequently in the form of an experimental group and a control group
Frequency table	a way of collating the information recorded on a data collection sheet
Hypothesis	a statement which may or may not be true
Inferential Statistics	branch of statistics dealing with drawing conclusions from data, testing hypotheses, etc.
Nominal	classification into categories using words /descriptions. Non numerical data
Ordinal data	data that can be placed in an order, e.g., taste of oranges from very sweet to sour
Parameter	a single fact (numerical or nominal) for the whole population
Population	the entire collection of objects with at least one similar characteristic also: set of all the possible observations
Qualitative data	data which can only be described in words
Quantitative data	data that has a numerical value
Questionnaire	a set of questions used to collect data in a survey
Random sampling	method of obtaining a sample such that each member of the population has an equal chance of being included

Sample	portion of the entire collection of objects of similar characteristics also: collection of data from a subset of the population
Simulation	method of collecting data using random number to model a real life situation
Statistics	a single fact (numerical or nominal) obtained from a sample
Survey	method of collecting data using, e.g., questionnaires, interviews, tests, observations, secondary sources
Tally	a way of recording each item of data on a data collection sheet
Variable	characteristic that varies over the population