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.)

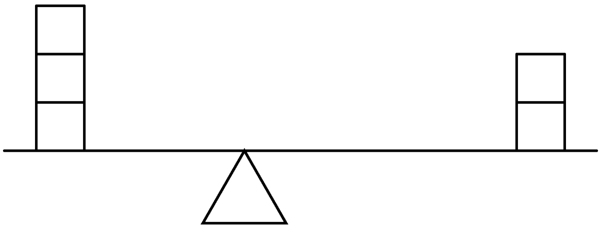
Answers: mean 16.5 / mode 19 / median 17.5

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'.)

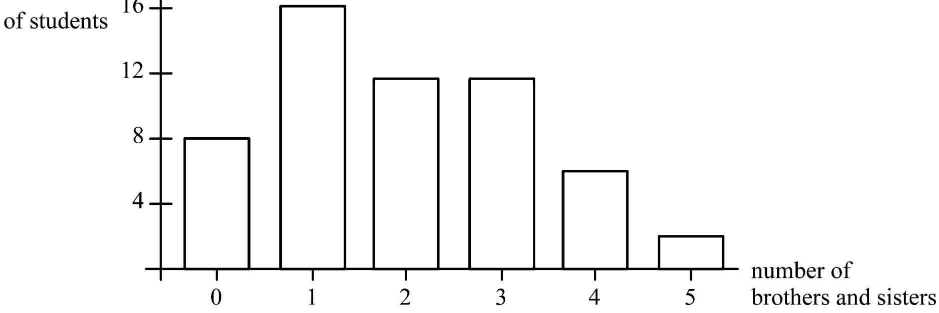
CHANGE	EFFECT ON		
	MEAN	MEDIAN	MODE
Adding zero value(s)			
Adding two values with equal but opposite deviations			
Adding two values with opposite unequal deviations			
Adding two values with deviations both positive (negative)			
Adding two values equal in value to the central measure at the top of 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:

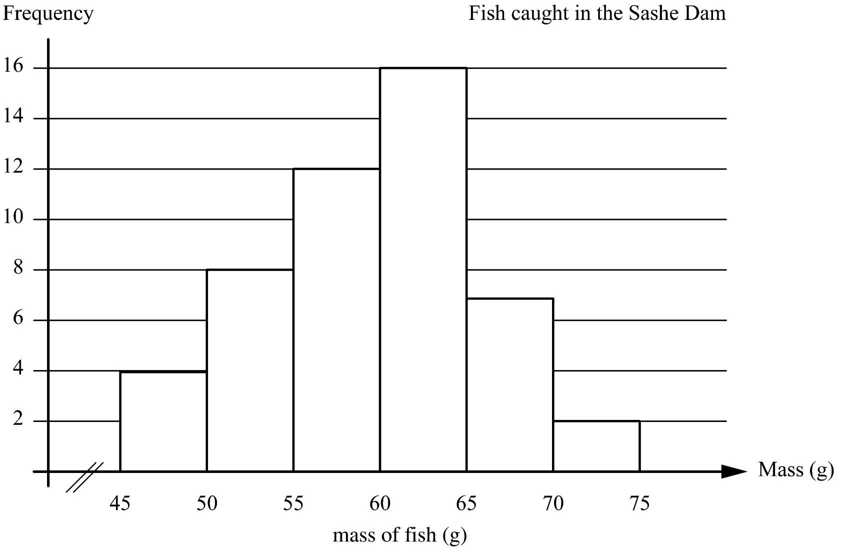
   a) for single values… key in each value, followed by a press of the $\boxed{\text{DATA}}$ or $\boxed{\text{ENTER}}$ or $\boxed{\sum+}$ key.

   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 $\boxed{\overline{X}}$ 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.

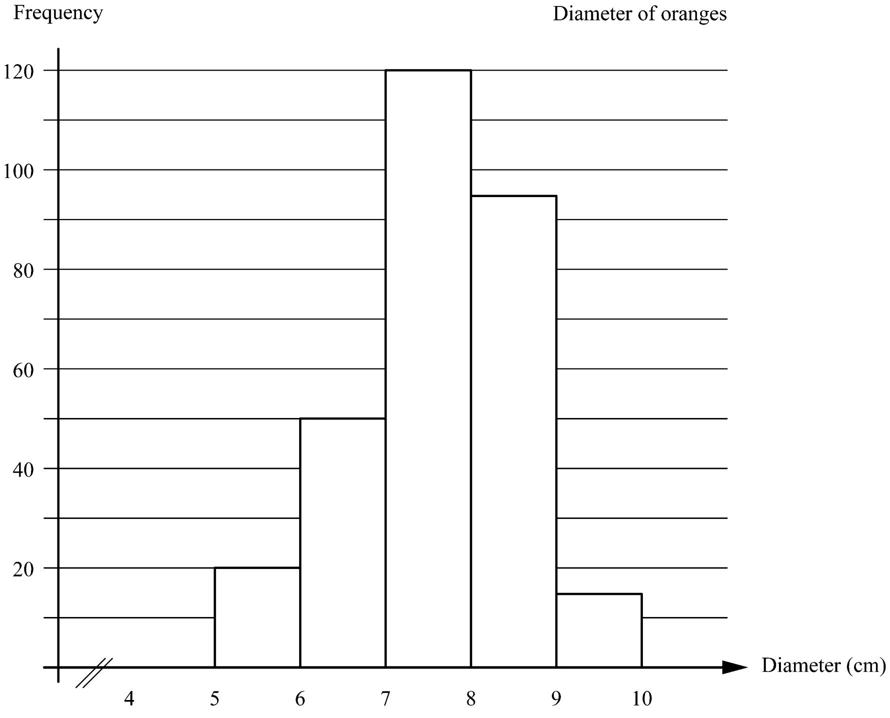
Self mark exercise 3

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.

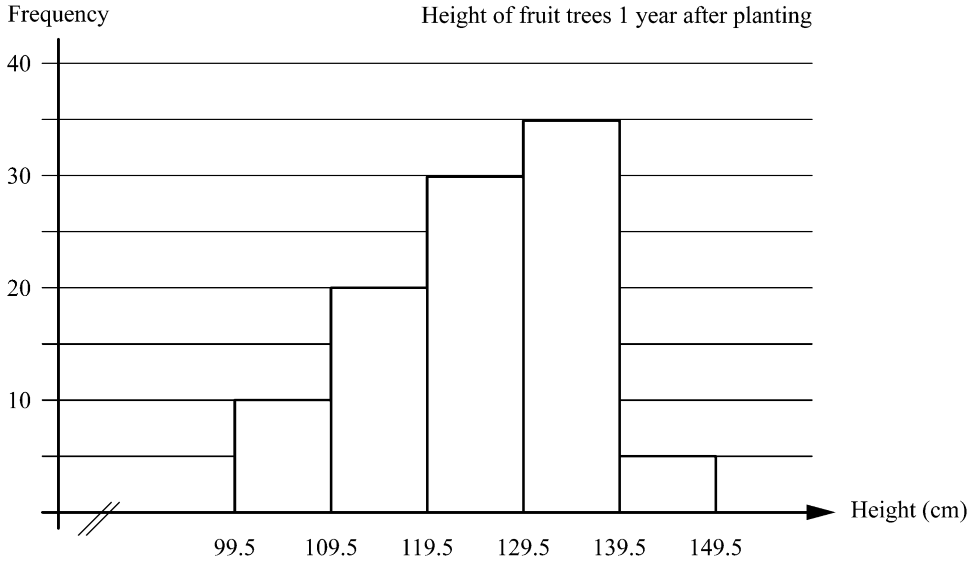
c) How many oranges in total were measured?

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.

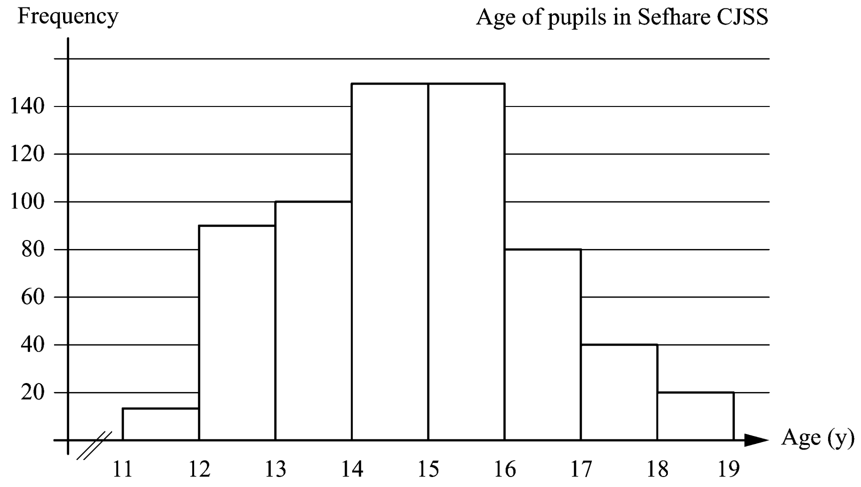
c) How many fruit trees in total were measured?

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



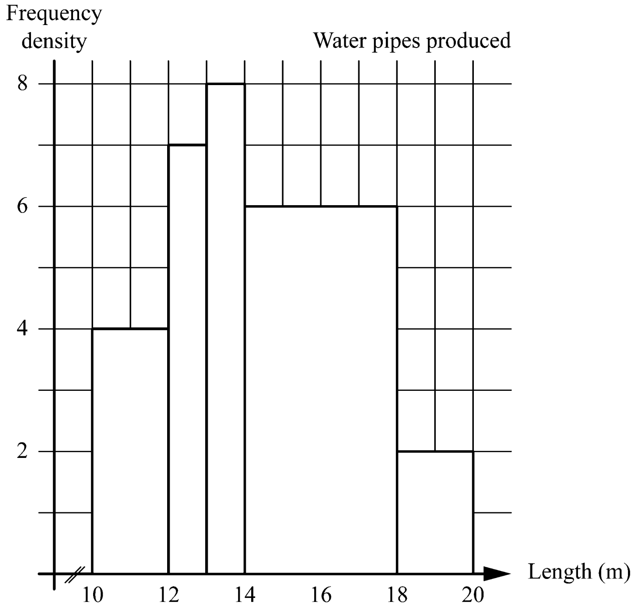
a) What are the modal classes? (a bi-modal distribution)

b) How many pupils are in their 16th year?

c) Make a grouped frequency table, first class 11 ≤ 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 ≤ L < 18 were made?

b) What is the modal class?

c) In which class is the median length of pipe?

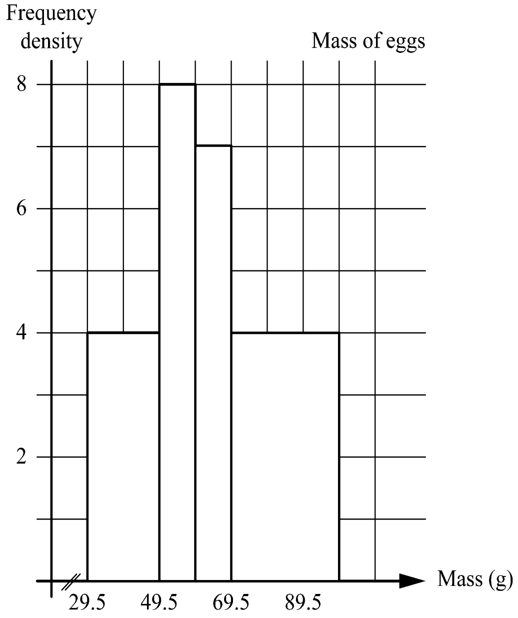
d) How many pipes were produced during the month?

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.

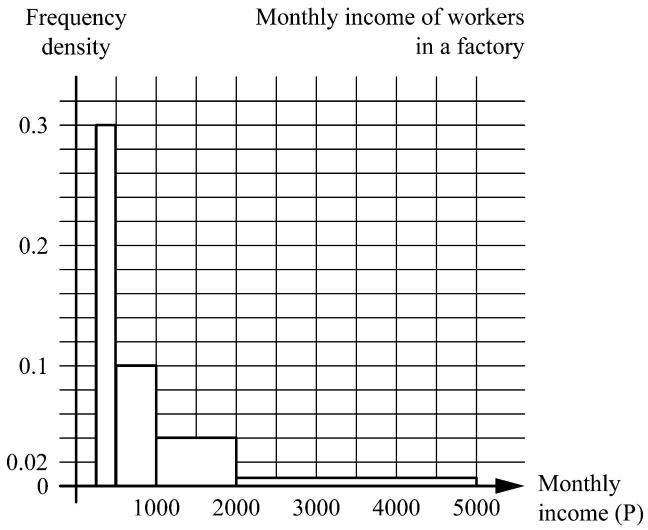
7. The monthly salary of the workers in a factory are illustrated in the histogram.

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

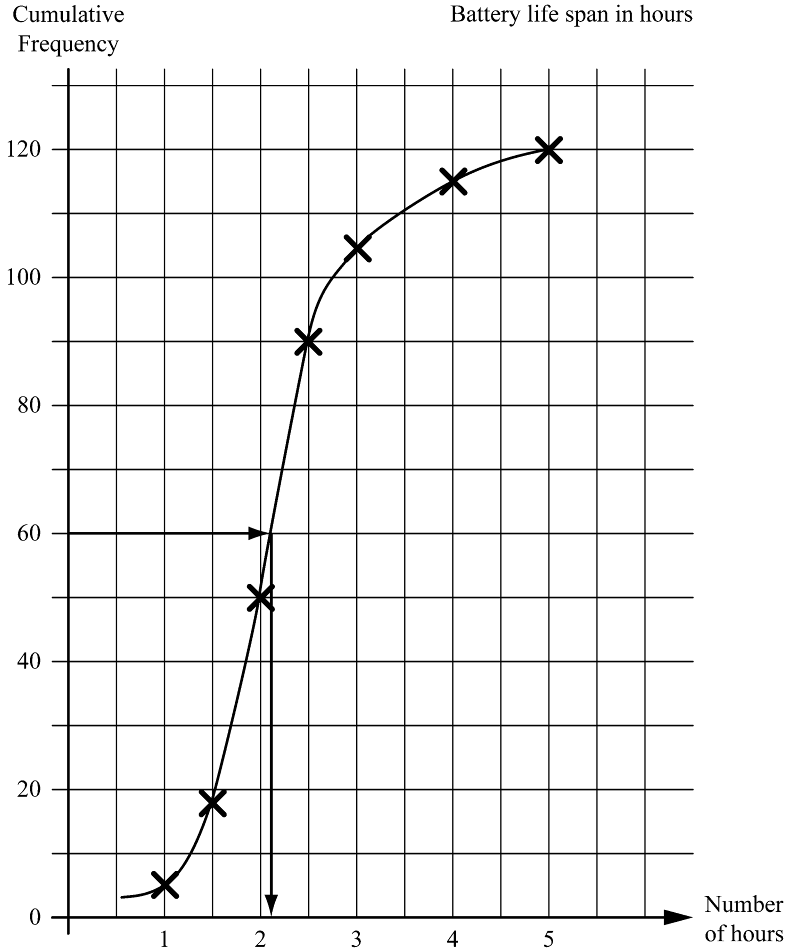
Making a cumulative frequency table and plotting the cumulative frequency curve:

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).

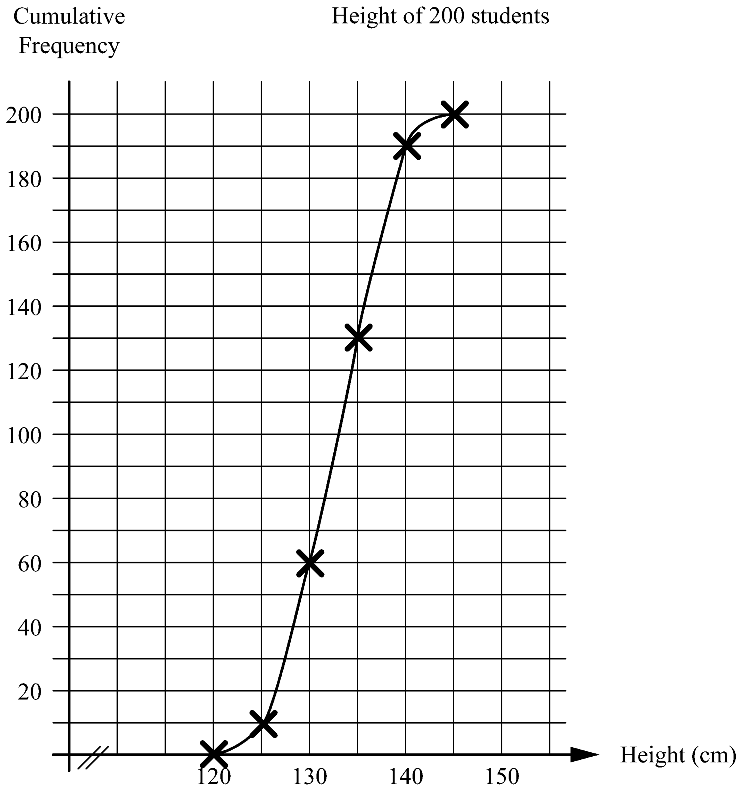
The estimate of the median is 2.1 hours. 50% of the batteries last for more than 2.1 hours (and 50% last less or equal to 2.1 hours).

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.

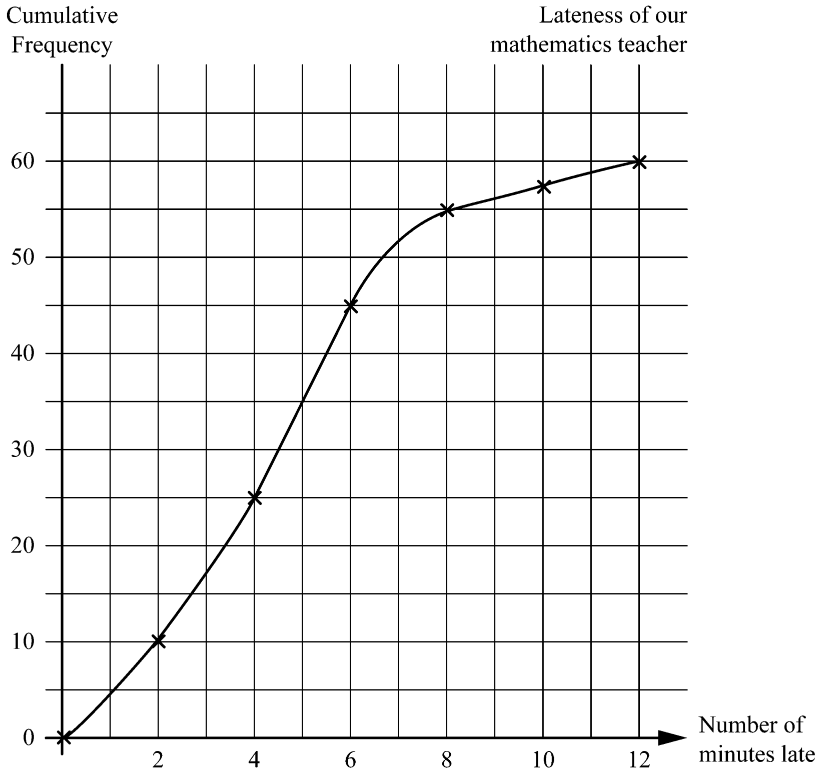
g) The frequency table, histogram, frequency polygon and cumulative frequency curve all display the same data.

(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.

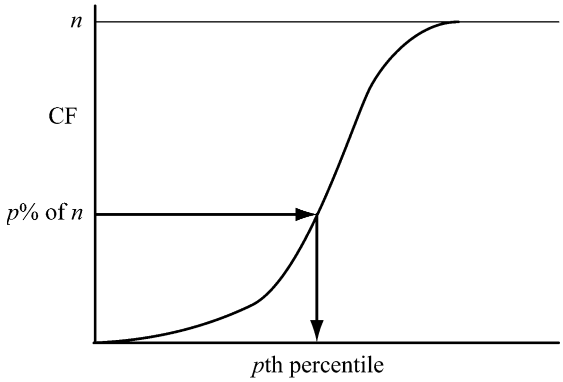
The median is the 50th percentile, the $\frac{1}{2}(n+1)$th value in the ordered sequence of the n values.

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.

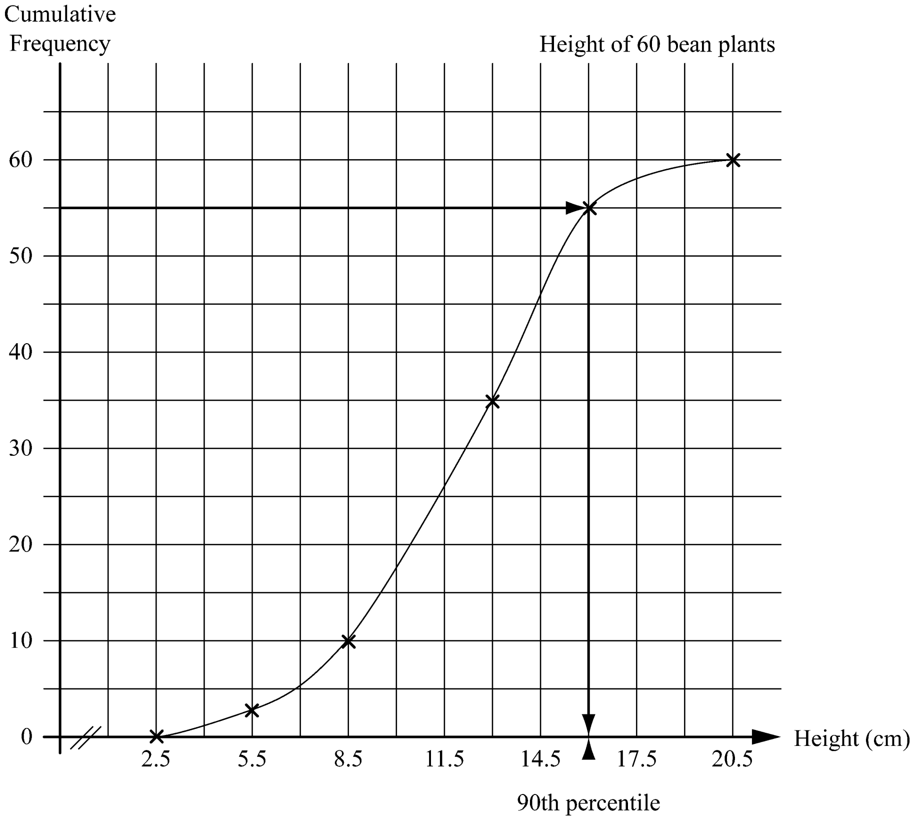
The pth percentile can be estimated from a cumulative frequency curve by taking (as an approximation) the $\frac{p}{100}$ of the total number of values: $p\% \times n$.

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)

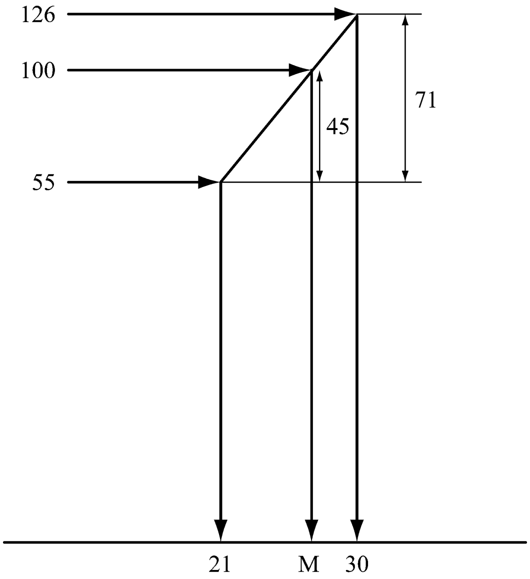
(ii) the modal class: 21 - 30 correct answers

(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.

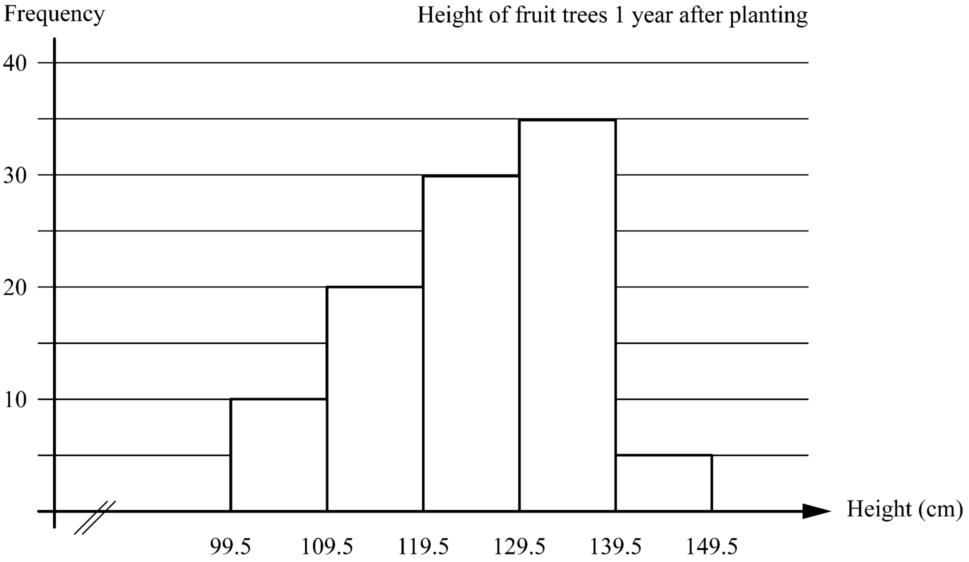
d) What is the effect of changing class width on the estimate of (i) mean (ii) median?

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.

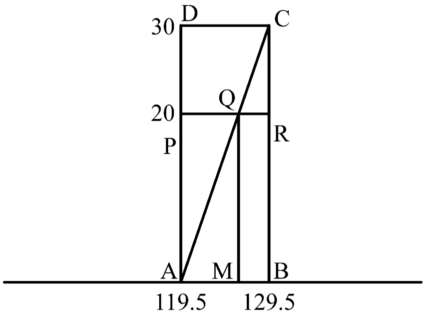
For example the histogram illustrates the height of 100 fruit trees 1 year after being planted.



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

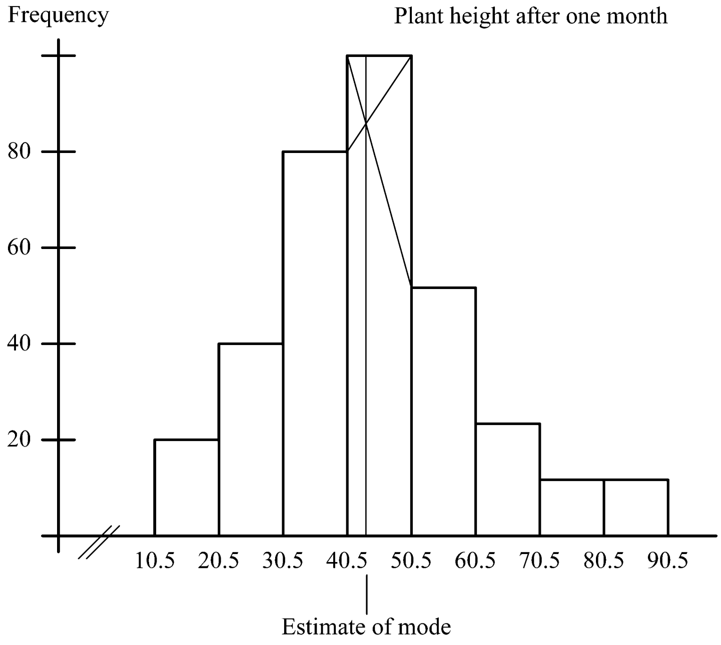
When data has been grouped into classes the class with the highest frequency can easily be identified: the modal class. An estimate of the mode can be made from the modal class.

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:

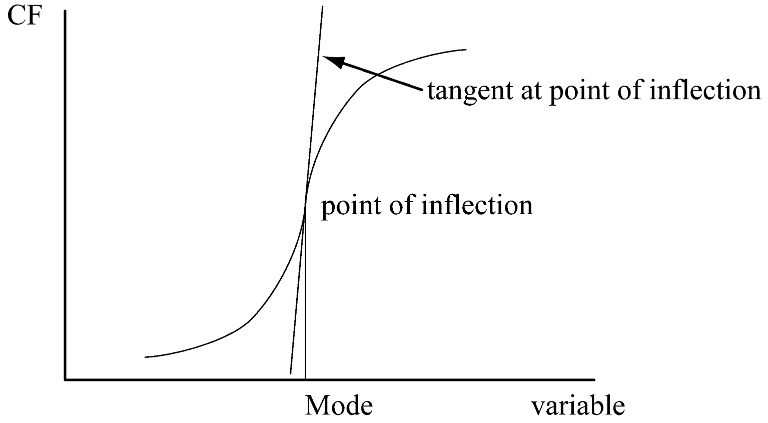
The modal class contains 20 more than the class below and 50 more than the class above the modal class. We therefore assume that the modal class is divided by the estimated mode in the ratio 20 : 50 = 2 : 5.

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.

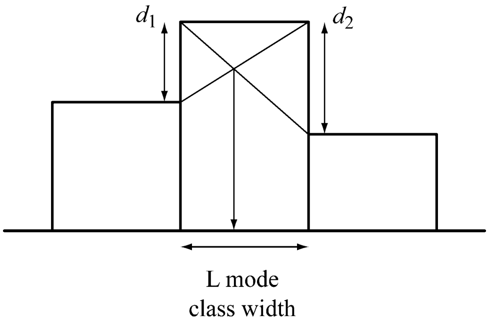
3. Estimate of the mode from the cumulative frequency curve:

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.

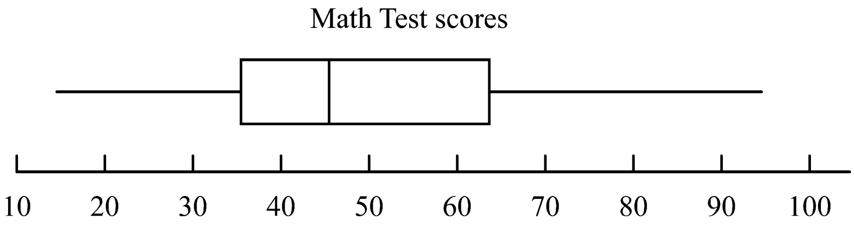
Minimum score	16
Maximum score	92

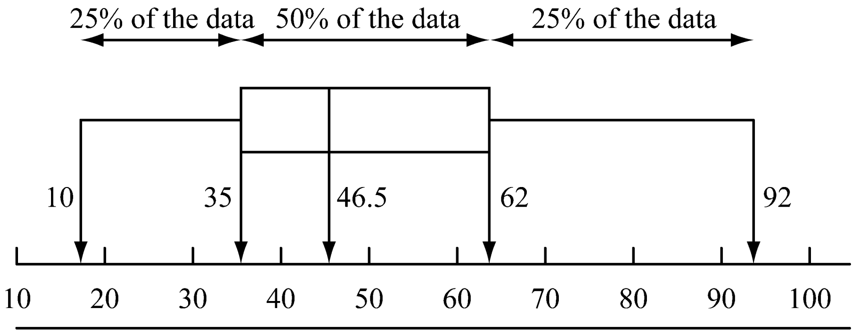
Median is mean of 25th and 26th score: $\frac{46+47}{2} = 46.5$ (Q_2)

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

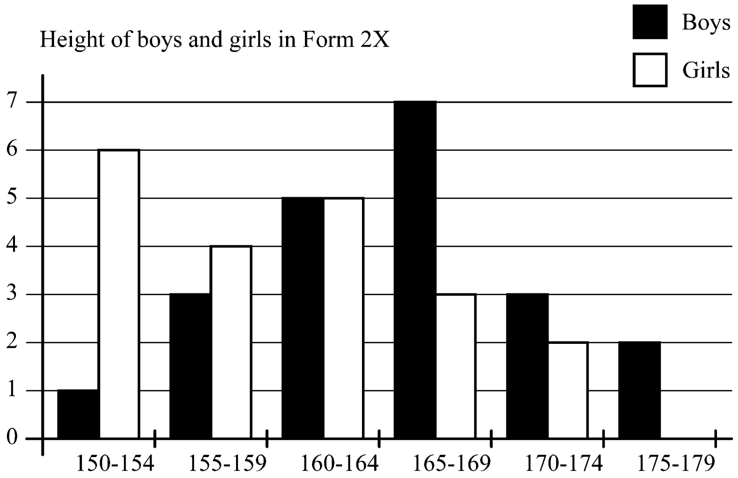
1. Mean height 1.61 cm (2 dp)
2. 11.4 pips (1 dp)
3. Mean 2.75 h, median 2.5 h, mode 2 h

   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.

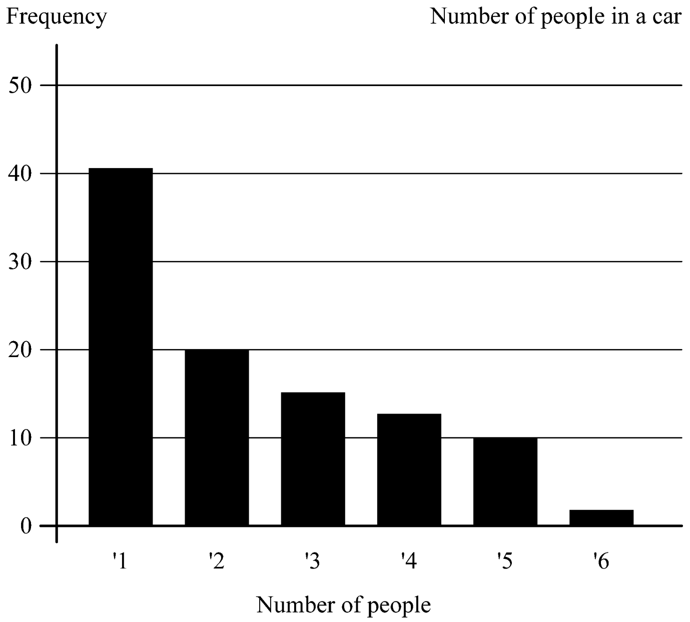
   Boys: Mean height 165.4 cm, multi modal 162 cm, 165 cm, 166 cm, 168 cm, median height 165 cm.

   Generally boys are taller than girls (higher mean and median). There is no 'most common' height (the distributions have no single mode).



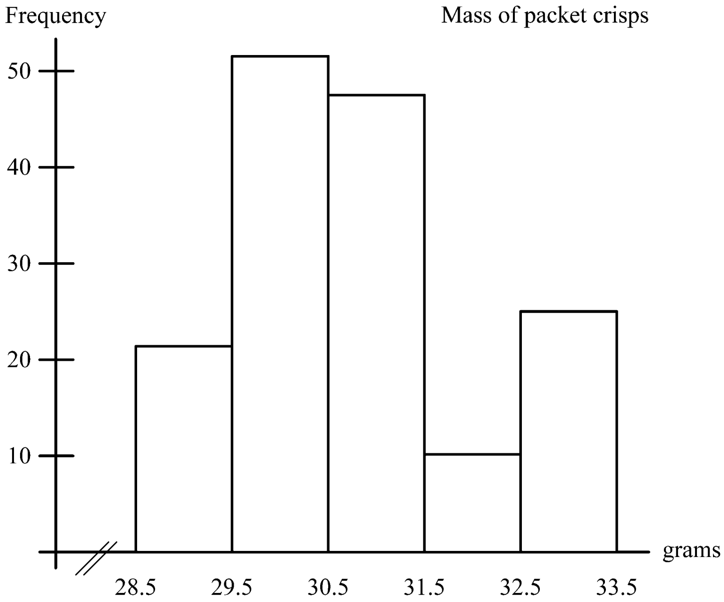
6. a) Mean 2.3 persons, mode 1 and median 2

   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 ≥ Geometric mean of p and q

b) Arithmetic mean A=

$\frac{p+q}{2} \qquad 2\text{A} = p + q$

Geometric mean = $\sqrt{pq} \qquad \text{G}^2 = pq$

$\frac{1}{H} = \frac{1}{p} + \frac{1}{q} \quad \frac{1}{H} = \frac{p+q}{pq} \quad \frac{1}{H} = \frac{pq}{p+q} = \frac{G}{2A}$

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 ≤ m < 65

d) Median in class 55 ≤ 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 ≤ 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 ≤ h < 139 e) 125 cm

4. a) 14 ≤ a < 15 and 15 ≤ 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 ≤ length < 18

c)

Length (m)	10 ≤ l <12	12≤ l < 13	30≤ l < 14	14≤ l< 18	18 ≤ l < 20
Frequency	8	7	8	24	4

Median (26th observation) in class 14 ≤ length < 18

d) 51 e) 743.5 m f) 14.6 m

6. a) Modal class 70 ≤ m < 100

b)

Mass (g)	30 ≤ m <50	50≤ m < 60	60≤ d < 70	70≤ d< 100
Frequency	80	80	70	120

c) 62.1 g

7. a) 250 ≤ I < 500

b)

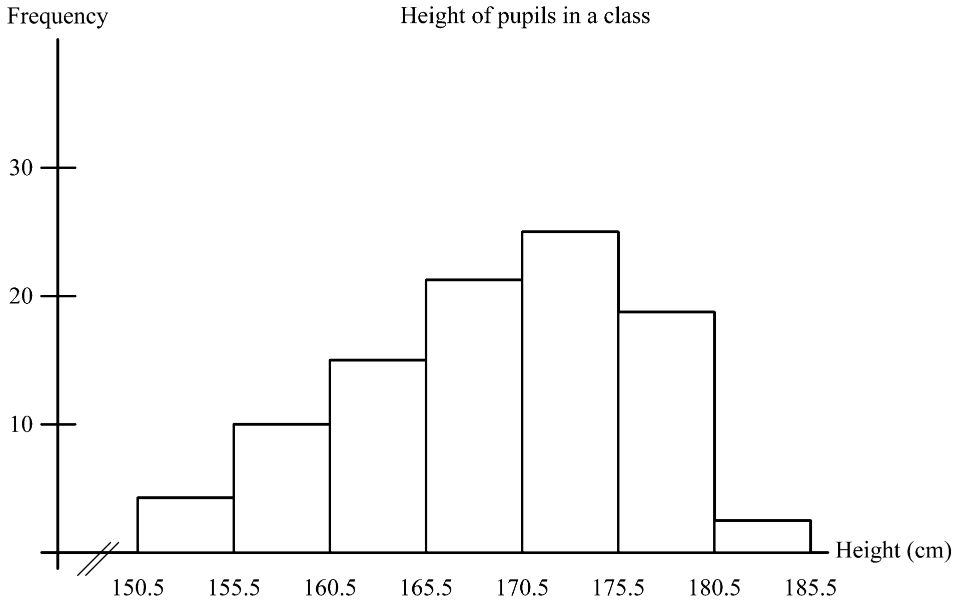
Income (P)	250 ≤ I <500	500≤ I < 1000	1000≤ d < 2000	2000≤ d< 5000
Frequency density	0.3	0.1	0.04	0.004
Frequency	75	50	40	12

c) 500 ≤ I < 1000

d) P947

e) Mode, the salary earned by most people

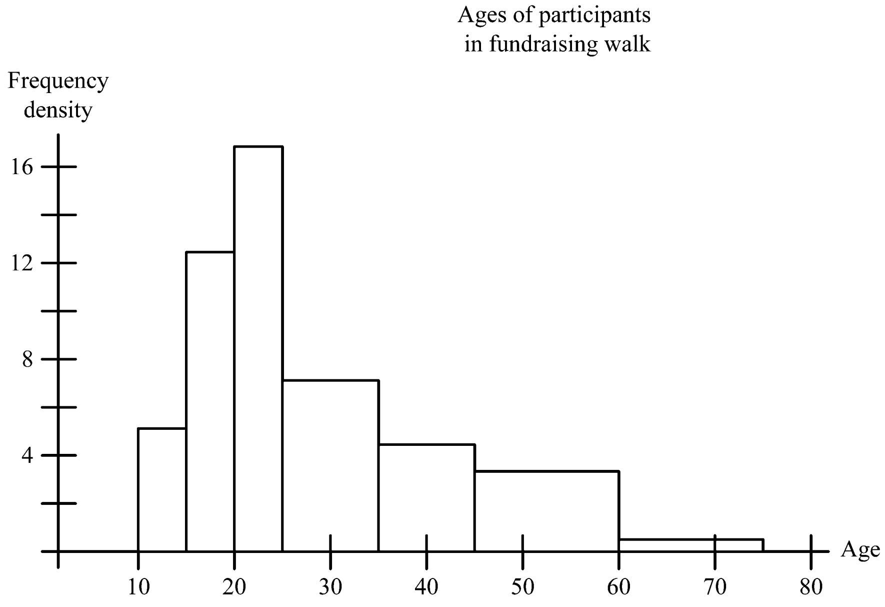
8. a)



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

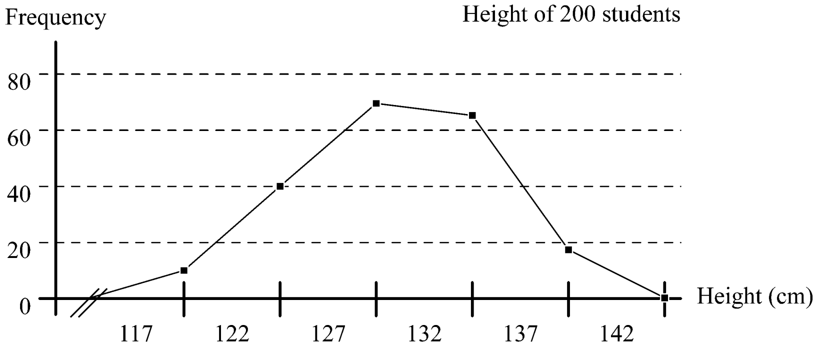
c) 120

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

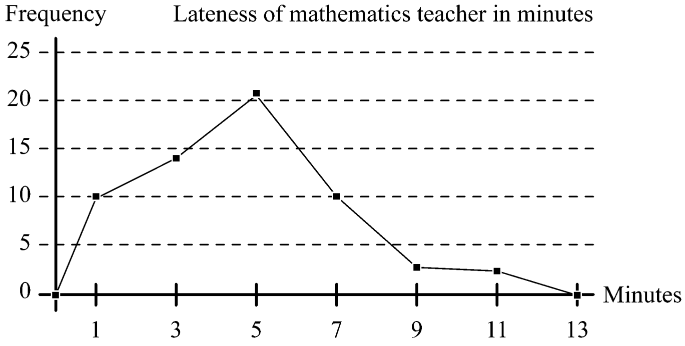
b) 3.1 minutes

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)

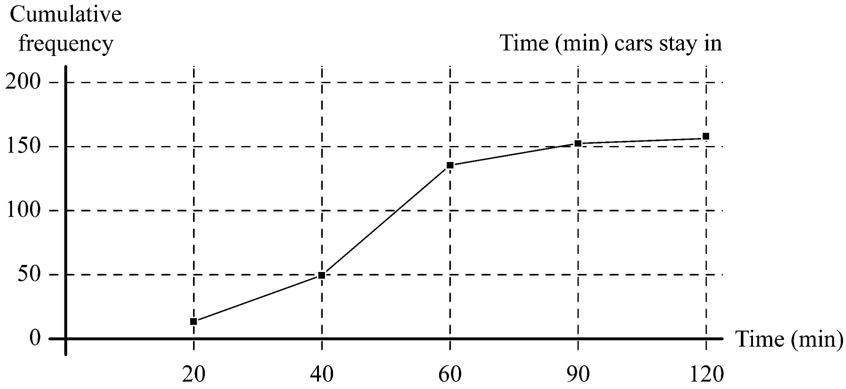


f) 4.6 minutes

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)

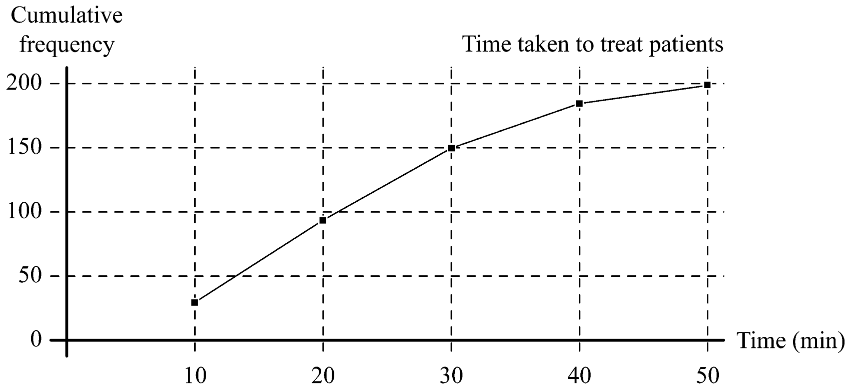


c) (i) 50 (ii) 35 (iii) 58

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 ≤ l ≤ 14	2	10 ≤l ≤ 19	10
15 ≤ l ≤ 19	8		
20 ≤ l ≤ 24	5	20 ≤ l ≤ 29	10
25 ≤ l ≤ 29	5		
30 ≤ l ≤ 34	4	30 ≤ l ≤ 39	10
35 ≤ l ≤ 39	6		
40 ≤ l ≤ 44	8	40 ≤ l ≤ 49	15
45 ≤ l ≤ 49	7		
50 ≤ l ≤ 54	4	50 ≤ l ≤ 59	5
54 ≤ l ≤ 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!)

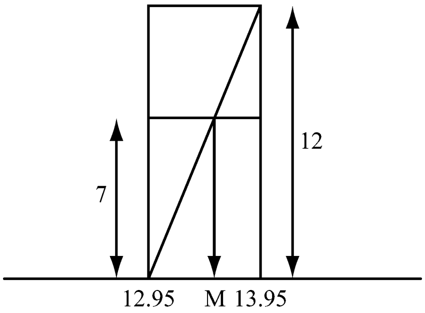
Self mark exercise 6

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.



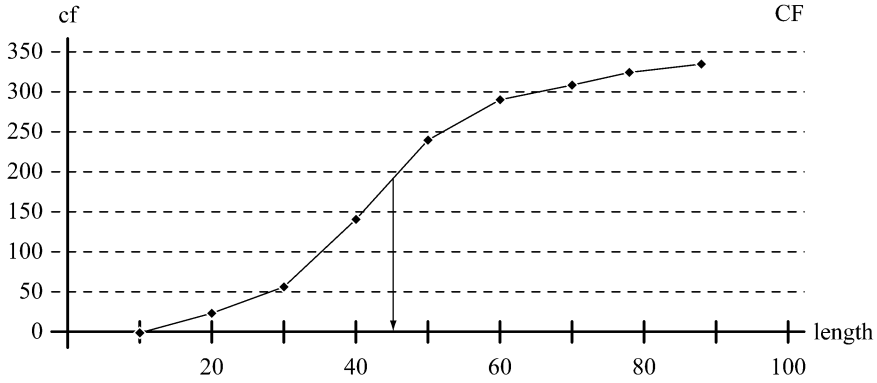
   Draw accurate diagram in your histogram as illustrated above. The median should be close to the calculated value of 13.53.

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

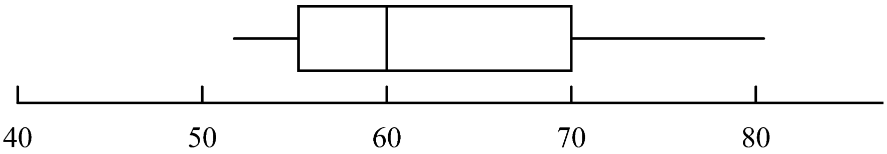
1. a)

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.

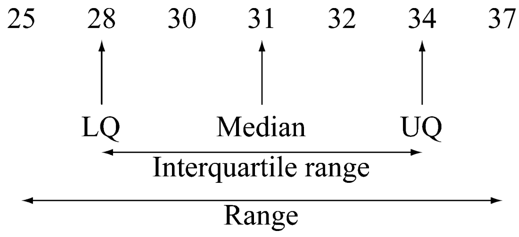
The **median** is the middle value, $\frac{7+1}{2}$ = 4th value, which is 31 cm.

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.

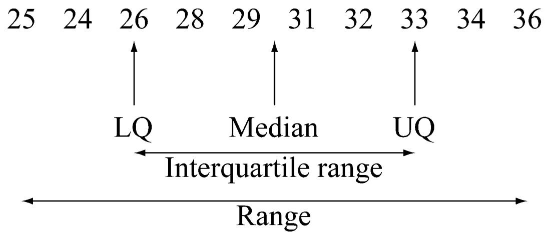
The **range** is 36 – 24 = 12 cm.

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	x^2	f	fx	fx^2
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
$\sum x=$	$\sum x^2=$	$\sum f=$	$\sum fx=$	$\sum fx^2=$
105	1603	60	887	13259

The mean is $\frac{\sum fx}{\sum f} = \bar{x} = \frac{887}{60} = 14.78$ h (2 dp)

The standard deviation s.d. = $\sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2} = \sqrt{\frac{13259}{60} - (14.78)^2} = 1.6$ 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:

$$\frac{\sum fm}{\sum f} = \overline{m} \text{ and s.d. } = \sqrt{\frac{\sum fm^2}{\sum f} - \overline{m}^2}$$

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.

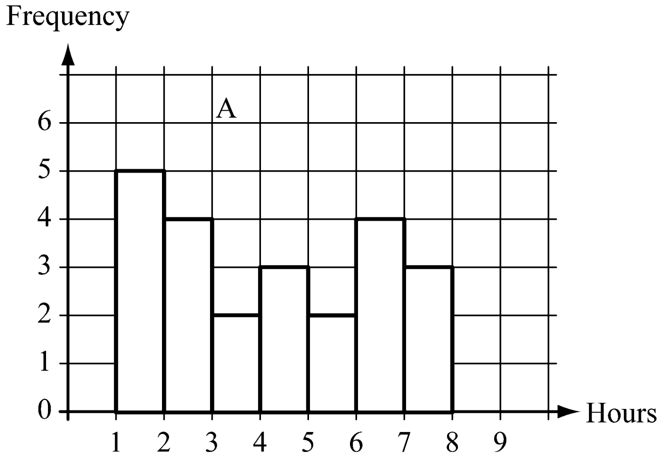
Number of words	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39
Novel A	12	15	10	26	14	8	4
Novel B	8	18	12	31	18	4	2

a) Calculate for each novel an estimate for the mean number of words in a sentence and the standard deviation.

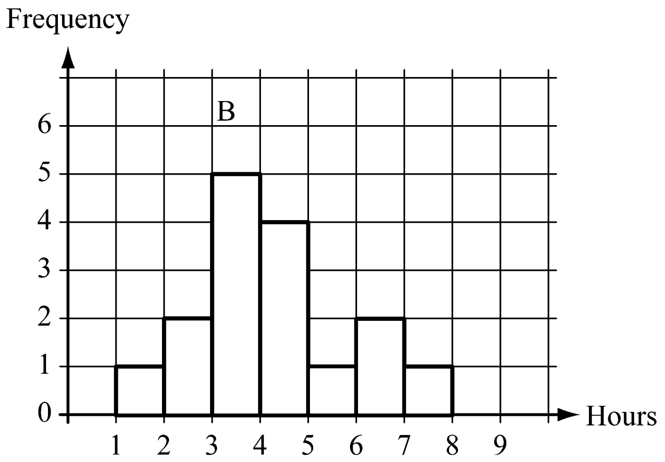
b) Compare the results of a. Which novel do you think is easier to read? Justify your answer.

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

G

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