# MATHEMATICS

## Grade 9
## CAPS Learner Book
## TERM 4

ract

# CONTENTS

- Grade 9
- CAPS Learner Book
- TERM 4

## Term 4

# Chapter 22
# Collect, organise and summarise data

## 22.1 Collecting data

### Avoiding bias when selecting a sample

The methods that we use to collect data must help us to make sure that the data is reliable. This means that it is data that we can trust.

Data cannot be trusted unless it has been collected in a way that makes sure that every member of the population had the same chance of being selected in the sample.

It is not practical to taste all the oranges on a tree to know whether the oranges are sweet. Only a small number of oranges can be tested, otherwise the farmer would have too few oranges to sell. The oranges that are tested are called a **sample**, and all the oranges harvested from the tree are called the **population**.

Sample bias occurs when the particular section of the population from which the sample is drawn does not represent that population. The way to avoid sample bias is to take a **random** sample. A sample is random if **every member of the population has the same chance** of being selected. A random sample of the orange trees means that every tree should have a chance of being selected for the sample. Every person in the country should have a chance of being selected for the housing survey in a random sample.

An example of sample bias would be to survey only the people in Limpopo about their views on housing provision when you want to know the views of the whole country. For the sample to provide information on the population as a whole, each person in the country should have the same chance of being part of the survey.

Data can be collected through questionnaires, through observation and through access to databases.

### How to develop a good questionnaire

The questionnaire also has an important role in making sure that the information you collect is reliable. You should aim to get a high number of respondents and accurate information. If not enough people fill in the questionnaire, then you won't know whether the information you get reflects the real situation. Sampling techniques and rules developed by statisticians determine the numbers needed.

There are some important points to consider when designing a questionnaire. Two of the most important points are that the questions are **clear and accurate** and that people find the questionnaire relatively **easy to complete**.

1. Keep in mind the length of the questionnaire and the time that it takes to complete. Your participants are more likely to complete a short questionnaire that is quick and easy to complete. Exclude unnecessary information.

2. Write down a selection of questions that you think will provide the information that you want.

3. Check the wording for each question.

4. Order the items so that they are in a logical sequence. It might make sense to have the easiest questions first, but in some cases the more general questions should come first and the more specific questions towards the end of the questionnaire.

5. Then try the questionnaire out on a partner. Ask the following questions:
   - Is this question necessary? What information will be provided by the answer?
   - How easy will it be for the respondent to answer this question? How much time will it take to answer the question?
   - Do the questions ask for sensitive information? Will people want to answer the question? Will the respondent answer the question honestly?
   - Can the question be answered quickly?

6. Decide how the answers should be provided. Questions may require **open-ended** responses or **closed-ended** responses, as described below.

In an **open-ended** question, the person responds in his or her own words. Through his or her own words important information can be gained; the person is therefore free to write what he or she likes. A disadvantage is that you might not get the information you want and that it might take a long time to answer.

In a **closed-ended** question, the respondents are given some options to choose from. They tick the box which most closely represents their response. These options can be constructed in categories. For example, age may be categorised as follows:

Under 10 ☐ From 10 to 14 ☐ From 15 to 19 ☐ 20 and older ☐

### THINK ABOUT DATA COLLECTION AND DEVELOP A QUESTIONNAIRE

1. Which method for collecting data would be most appropriate for each of the cases below? Give reasons for your choice.
   (a) The number of learners who bring lunch to school. What are the contents of the school lunch?
   (b) Whether or not the tellers at a supermarket chain are happy with their conditions of work.
   (c) Whether or not the clients of a clinic are satisfied with the professional conduct of the medical staff.
   (d) The types of activities preschool children choose during their free time.
   (e) The number of Grade 9 learners in the Gauteng North district.
2. You are doing some market research for a new fast-food shop near the high school. The owners of the shop want to find out what kind of food and music the target market likes. The target market is learners from the high school. Develop a questionnaire to collect this information.

## 22.2 Organising data

There is a difference between **data** and **information**. Data is unorganised facts. When data is organised and analysed so that people can make decisions, it may be called information. Data can be organised in many different ways. Some methods are described below.

Data can be organised by making a **tally table**. Here is an example of a tally table showing the numbers of learners in a class that participate in different sports:

| Sport | Tally marks |
|---|---|
| Soccer | 𝍸 𝍸 𝍸 𝍸 𝍸 |
| Athletics | 𝍸 /// |
| Netball | 𝍸 𝍸 𝍸 𝍸 / |
| Chess | 𝍸 / |

The above data can also be organised in a **frequency table**:

| Sport | Frequency |
|---|---|
| Soccer | 25 |
| Athletics | 8 |
| Netball | 21 |
| Chess | 6 |

Numerical data sets with many items are often grouped into equal **class intervals** and represented in a table of frequencies for the different class intervals. This is very useful since it makes it easy to see how the data is spread out.

Here is an example of grouped data showing the heights of all the learners in a school. To make a frequency table for numerical data, the data has to be arranged from smallest to biggest first.

| Height in m | Number of learners (frequency) |
|---|---|
| < 1,20 m | 13 |
| 1,20 m – 1,30 m | 28 |
| 1,30 m – 1,40 m | 57 |
| 1,40 m – 1,50 m | 164 |
| 1,50 m – 1,60 m | 274 |
| 1,60 m – 1,70 m | 198 |
| 1,70 m – 1,80 m | 73 |
| > 1,80 m | 13 |

A value equal to the **lower boundary** of a class interval is counted in that interval.
For example, a length of 1,60 m is counted in the interval 1,60 – 1,70, and not in the interval 1,50 – 1,60 m.
However, 1,599 m is less than 1,60 m, so it belongs in the interval 1,50 m – 1,60 m.

A **stem-and-leaf display** is a useful way to organise numerical data. It also shows you what the "shape" of the data is like. Here is an example of a stem-and-leaf display:

Key: 35 | 4 means 354

| | |
|---|---|
| 34 | 0 4 |
| 35 | 4 8 8 |
| 36 | 0 1 6 8 |
| 37 | 1 3 5 8 8 8 9 |
| 38 | 2 4 9 |
| 39 | 0 3 4 4 5 6 9 |
| 40 | 0 3 7 |
| 41 | 1 |

The above stem-and-leaf display represents the following data about the masses in grams of the chickens in a sample of six-week-old chickens on a chicken farm:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 399 | 378 | 382 | 360 | 396 | 389 | 344 | 411 | 378 | 394 |
| 394 | 354 | 375 | 378 | 400 | 371 | 379 | 358 | 366 | 403 |
| 358 | 395 | 390 | 340 | 393 | 384 | 361 | 407 | 373 | 368 |

To make a stem-and-leaf display, it helps to first arrange the data from smallest to largest, as shown on the next page, for the above data set.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 340 | 344 | 354 | 358 | 358 | 360 | 361 | 366 | 368 | 371 |
| 373 | 375 | 378 | 378 | 378 | 379 | 382 | 384 | 389 | 390 |
| 393 | 394 | 394 | 395 | 396 | 399 | 400 | 403 | 407 | 411 |

The same data set is displayed in two slightly different ways below:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | 379 | | 399 | | |
| | | | 378 | | 396 | | |
| | | | 378 | | 395 | | |
| | | 368 | 378 | | 394 | | |
| | 358 | 366 | 375 | 389 | 394 | 407 | |
| 344 | 358 | 361 | 373 | 384 | 393 | 403 | |
| 340 | 354 | 360 | 371 | 382 | 390 | 400 | 411 |

In this display, the width of each class interval is 10, as in the stem-and-leaf display above.

| | | | | |
|---|---|---|---|---|
| | | 384 | | |
| | | 382 | 399 | |
| | | 379 | 396 | |
| | 368 | 378 | 395 | |
| | 366 | 378 | 394 | |
| | 361 | 378 | 394 | 411 |
| 354 | 360 | 375 | 393 | 407 |
| 344 | 358 | 373 | 390 | 403 |
| 340 | 358 | 371 | 389 | 400 |

In this display, the width of each class interval is 15.

## WORKING WITH GROUPED DATA

1. An organisation called Auto Rescue recorded the following numbers of calls from motorists each day for roadside service during March 2014:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 28 | 122 | 217 | 130 | 120 | 86 | 80 | 90 | 120 | 140 |
| 70 | 40 | 145 | 187 | 113 | 90 | 68 | 174 | 194 | 170 |
| l00 | 75 | 104 | 97 | 75 | 123 | 100 | 82 | 109 | 120 |
| 81 | | | | | | | | | |

   Set up a tally and frequency table for this set of data values, in intervals of width 40.

2. When geologists go out into the field they make sure they have their rulers and measurement instruments in their bags. They also have their "inbuilt rulers", for example their handspans. A handspan is the distance from the tip of the thumb to the tip of the fifth finger on an outstretched hand. Measure your handspan against the ruler! This frequency table shows the handspans of different Grade 9 learners, in cm.

| Handspan of Grade 9 learners (cm) | Frequency |
|---|---|
| 15–18 | 7 |
| 18–21 | 9 |
| 21–24 | 10 |
| 24 and greater | 4 |

(a) How many learner handspans were measured altogether?
(b) How many learner handspans are less than 21cm wide?
(c) How many handspans are 18 cm or wider?
(d) In which interval would you place a handspan of 18 cm?

## 22.3 Summarising data

The mean, median, mode and range are single numbers that provide some information about a data set, without listing all the data values.

The **mode** is the value that occurs most frequently. To find the mode, look for the number or category that is listed in the data set most often. Some data sets have more than one mode, and some may have none.

The **median** is the number that separates the set of values into an upper half and a lower half. The median can be found by arranging the values from small to big or big to small. If the data set consists of an even number of items, the median is the sum of the two middle values divided by 2.

The **mean** (average) of a set of numerical data is the sum of the values divided by the number of values in the data set.

Mean = the sum of the values ÷ the number of values.

The **range** is a number that tells us how spread out the data values are. It is the difference between the largest and smallest values.

The mean, median and mode do not work equally well for all sets of data. It depends on the kind of data, and also on whether the data is evenly spread out or not.

### ORGANISE, SUMMARISE AND COMPARE SOME DATA

1. A researcher analyses data about the people who are suffering from three different types of the flu virus: A, B and C. The ages of the people in the different groups are:

   Type A: 60, 80, 75, 87, 88, 49, 94, 84, 59, 86, 82, 62, 79, 89 and 78.

   Type B: 27, 39, 43, 29, 36, 70, 56, 25, 54, 36, 66, 45, 33, 46, 14 and 41.

   Type C: 33, 48, 64, 15, 31, 20, 70, 21, 18, 49, 21, 19, 57, 23, 29 and 20.

For each group:

- Draw a stem-and-leaf plot.
- Calculate the range, mean and median of the ages.
- Look at the shape of the stem-and-leaf displays as well as the summary measures. Discuss the spread of the data in each case, and compare the three different groups.

Work and report on your work.

2. Copy the table and fill in the statistic (mode, mean or median) that would best summarise each data set, and indicate the central tendency of the data:

| Data set | Best measure of central tendency |
|---|---|
| The shoe sizes of boys in Grade 9 | |
| An evenly spread set of measurement values, such as the heights of learners in a class | |
| A set of measurement values with a few very low values and mostly high values | |
| The number of siblings each person in your class has | |
| The sizes of properties in a town, where most people live in small apartments or RDP houses, and a few live on large properties | |

## EXTREME VALUES AND OUTLIERS

An **extreme value** or **outlier** is a data value that lies an abnormal distance from other values in a random sample from a population. Sometimes there are reasons why this data value is so different to the others. It is important to comment on the possible reasons.

When you are summarising data (and also when you analyse data), you need to decide whether or not an outlier makes sense in the context you are looking at.

It is possible that an outlier does not make sense, as it lies too far away and is an unreasonable measurement. Then you need to think about the fact that this data value may be an error. For example:

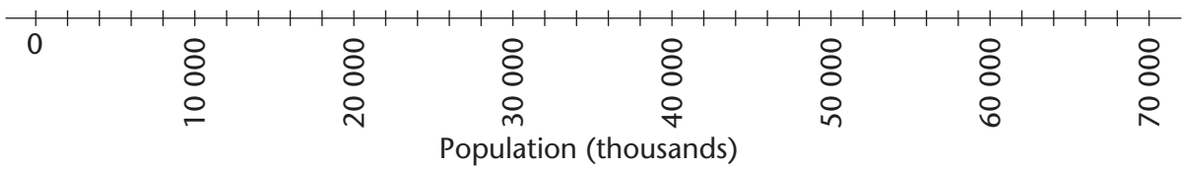
In this case, the value of 24 years old could be an unreasonable value. This depends on the context of the survey.

You will learn more about extreme values and outliers in Chapter 24.

Use this information about 14 countries to answer the questions that follow:

| Country | Total population (in 1 000s) | Total annual national income per person (US$) | Percentage of income spent on health |
|---|---|---|---|
| Angola | 18 498 | 4 830 | 4,6 |
| Botswana | 1 950 | 13 310 | 10,3 |
| DRC | 66 020 | 280 | 2,0 |
| Lesotho | 2 067 | 1 970 | 8,2 |
| Malawi | 15 263 | 810 | 6,2 |
| Mauritius | 1 288 | 12 580 | 5,7 |
| Mozambique | 22 894 | 770 | 5,7 |
| Namibia | 2 171 | 6 250 | 5,9 |
| Seychelles | 84 | 19 650 | 4,0 |
| South Africa | 50 110 | 9 790 | 8,5 |
| Swaziland | 1 185 | 5 000 | 6,3 |
| Tanzania | 43 739 | 1 260 | 5,1 |
| Zambia | 12 935 | 1 230 | 4,8 |
| Zimbabwe | 12 523 | 170 | Not available |

1. Look at the total population for each country.
   (a) Calculate the mean of the data.
   (b) Copy the number line below and draw a dot plot on the number line to show the data.




   (c) Find the median of the data.
   (d) What is the range of the data?
   (e) Which measure of central tendency do you think represents the data more accurately? Explain.

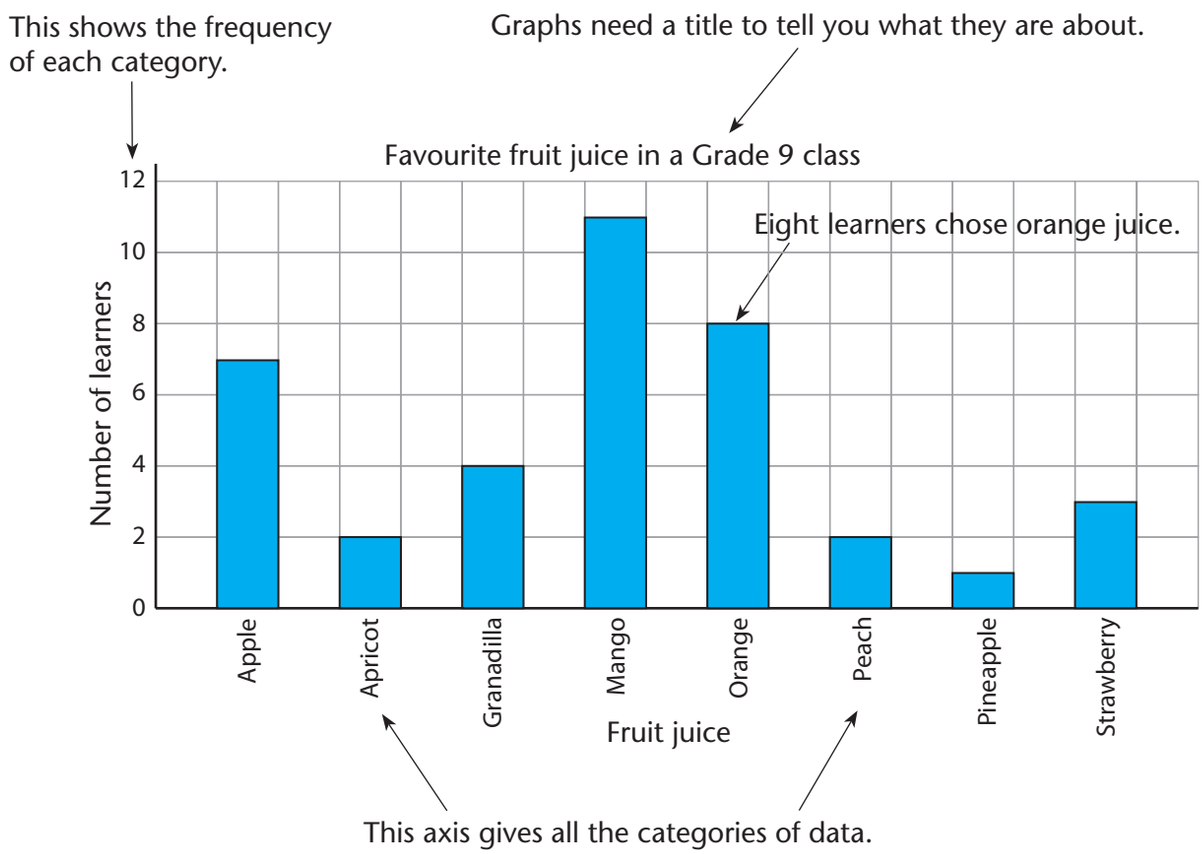
2. Look at the *Total annual national income per person in US dollars* column. Comment on the spread of the data.

# Chapter 23
# Representing data

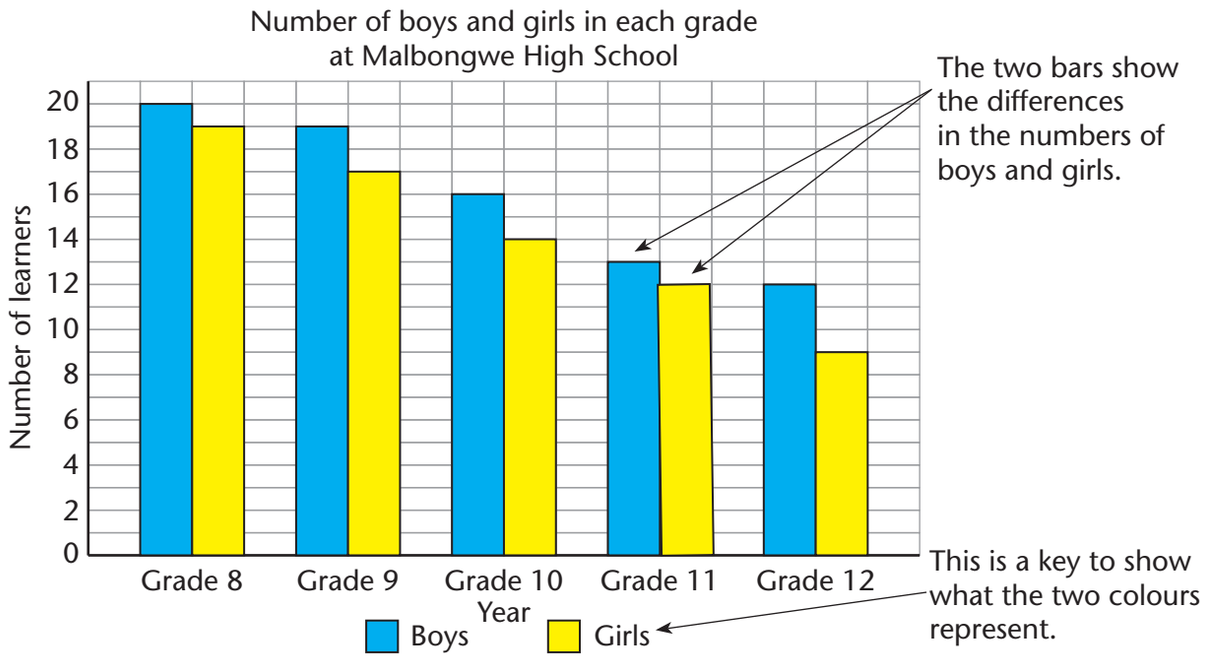
## 23.1 Bar graphs and double bar graphs

### REVISING BAR GRAPHS AND DOUBLE BAR GRAPHS

A **bar graph** shows categories of data along the horizontal axis, and the frequency of each category along the vertical axis. An example is given below.




A **double bar graph** shows two sets of data in the same categories on the same set of axes. This is useful when we need to show two groups within each category.




## DRAWING BAR GRAPHS AND DOUBLE BAR GRAPHS

1. Obese (very overweight) people have many health problems. It is a concern all around the world. Health researchers analysed the change over 28 years in the numbers of people who are overweight and obese in different areas of the world. The following table summarises some of the data:

   Percentage of population that is overweight and obese

| | 1980 | 2008 |
|---|---|---|
| Sub-Saharan Africa | 12% | 23% |
| North Africa and Middle East | 33% | 58% |
| Latin America | 30% | 57% |
| East Asia (low-income countries) | 13% | 25% |
| Europe | 45% | 58% |
| North America (high-income countries) | 43% | 70% |

   (a) The table summarises "some" of the data. What would some other important data be? Think of as many things as you can.
   (b) Which data stands out the most for you in the table above? Give your personal opinion.
   (c) On grid paper, plot a double bar graph to compare the data for the areas, and for the two years. Remember to give your graph a key.
   (d) Look carefully at the comparisons that the graph makes. Has your opinion of the most interesting differences changed, now that you see the double bar graph? Explain.

(e) In some countries, the obesity problem has been labelled "Obesity in the face of poverty". Write a short report on the data and your double bar graph to support this argument.

## 23.2 Histograms

### REVISING HISTOGRAMS

A histogram is a graph of the frequencies of data in different **class intervals**, as demonstrated in the example on the following page. Each class interval is used for a range of values. The different class intervals are consecutive and cannot have values that overlap. The data may result from counting or from measurement.

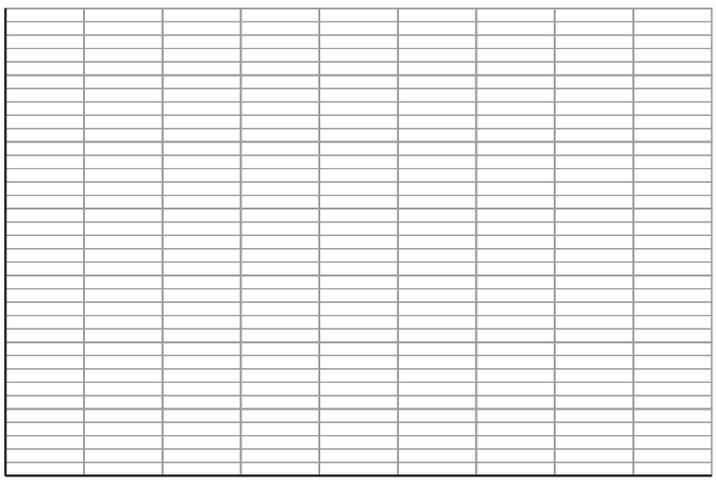
A histogram looks somewhat like a bar graph, but is normally drawn without gaps between the bars.

### REPRESENTING DATA IN HISTOGRAMS

1. (a) A fruit farmer wants to know which of his trees are producing good plums and which trees need to be replaced. He collects 100 plums each from two trees and measures their masses. The data below gives the mass of plums from the first tree:

| Mass of plums (g) | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 |
|---|---|---|---|---|---|
| Frequency | 6 | 18 | 34 | 30 | 12 |

Use the example below to represent the data in a histogram.



(b) Now draw another histogram to represent the following data giving the mass of the same type of plums from another tree in the orchard:

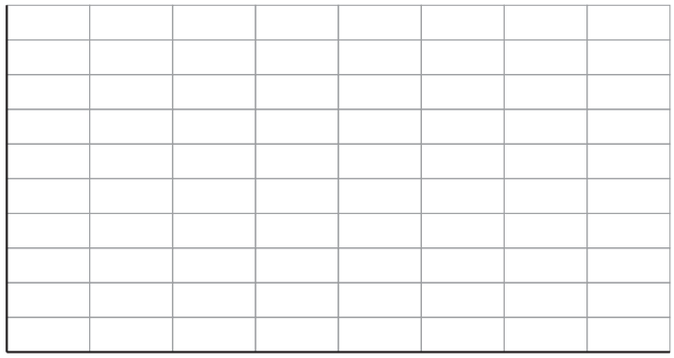
| Mass of plums (g) | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 |
|---|---|---|---|---|---|
| Frequency | 3 | 14 | 26 | 36 | 21 |

(c) Study the two histograms and then comment on the number of plums produced by the two trees.

2. (a) Use the example below to draw a histogram to represent the data in the table below. Group the data in intervals of 0,5 kg.

Birth weights (kg) of 28 babies at a clinic

| | | | | | | |
|---|---|---|---|---|---|---|
| 3,3 | 1,34 | 2,88 | 2,54 | 1,87 | 2,06 | 2,72 |
| 1,89 | 0,85 | 1,99 | 2,43 | 1,66 | 2,45 | 1,62 |
| 1,91 | 1,20 | 2,45 | 1,38 | 0,9 | 2,65 | 2,88 |
| 1,75 | 2,11 | 3,2 | 1,74 | 0,6 | 3,1 | 1,86 |



(b) Calculate the mean and median of the data.

(c) Records from the whole country show that the birth weight of babies ranges from 0,5 kg to 4,5 kg, and the mean birth weight is 3,18 kg. Use the graph and the mean and median to write a short report on the data from the clinic.

## 23.3 Pie charts

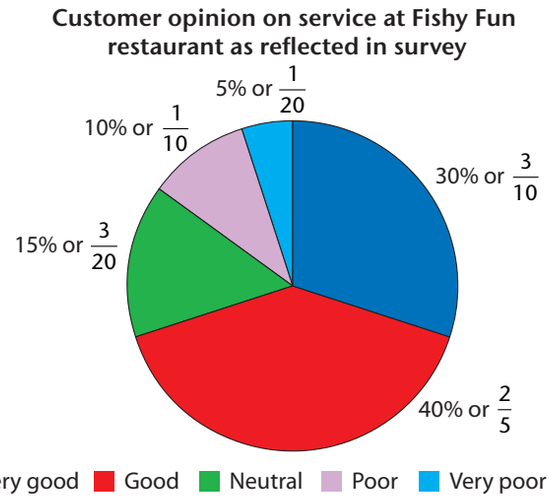
A **pie chart** consists of a circle divided into sectors (slices). Each sector shows one category of data. Bigger categories of data have bigger slices of the circle.

Here is an example of a pie chart:

This pie chart shows five categories of data.

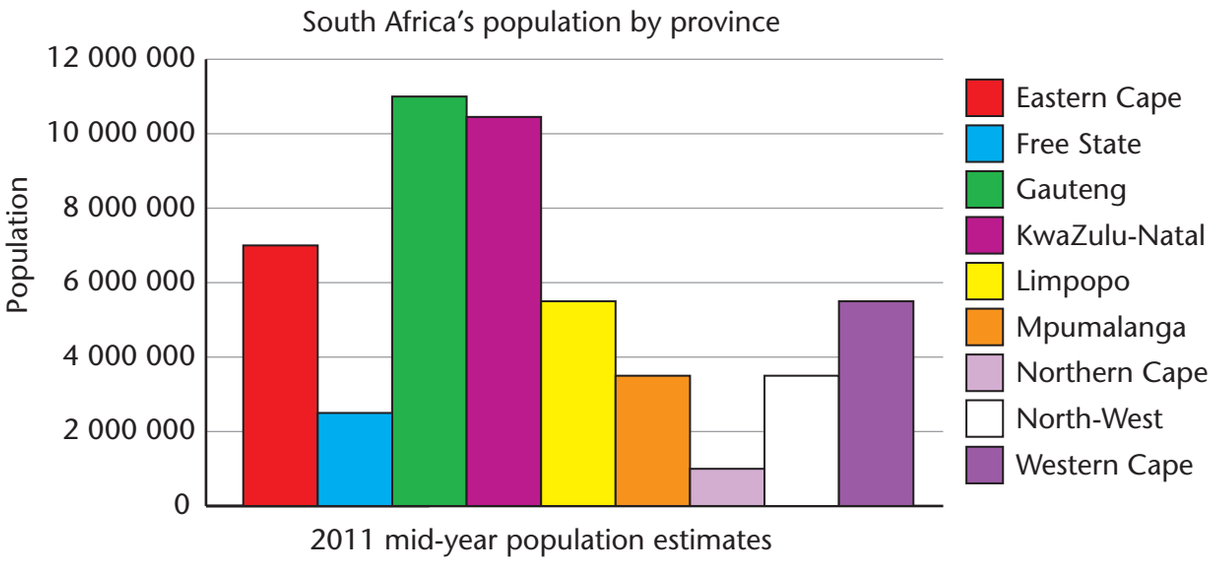
The size of each slice is the fraction or percentage of the whole that the category forms.

The key (or legend) shows the category that each colour stands for.



## DRAWING PIE CHARTS

1. The following bar graph shows the population of South Africa by province.




(a) Copy the table and write down the figures in the graph correct to the nearest 500 000.

| **Province** | E Cape | FS | Gau | KZN | Lim | Mpum | NC | NW | WC |
|---|---|---|---|---|---|---|---|---|---|
| **Population (× 1 000)** | | | | | | | | | |

(b) What is the total of the rounded off numbers?

(c) Work out the percentage of the whole for each province.

| **Province** | E Cape | FS | Gau | KZN | Lim | Mpum | NC | NW | WC |
|---|---|---|---|---|---|---|---|---|---|
| **Percentage of total** | | | | | | | | | |

(d) Draw a pie chart showing the data in the completed table. (Estimate the sizes of the slices.)

(e) Write a short report explaining the difference in the way the data is represented in the pie chart and the bar graph. Which do you think is a better method to show this data?

## 23.4 Broken-line graphs

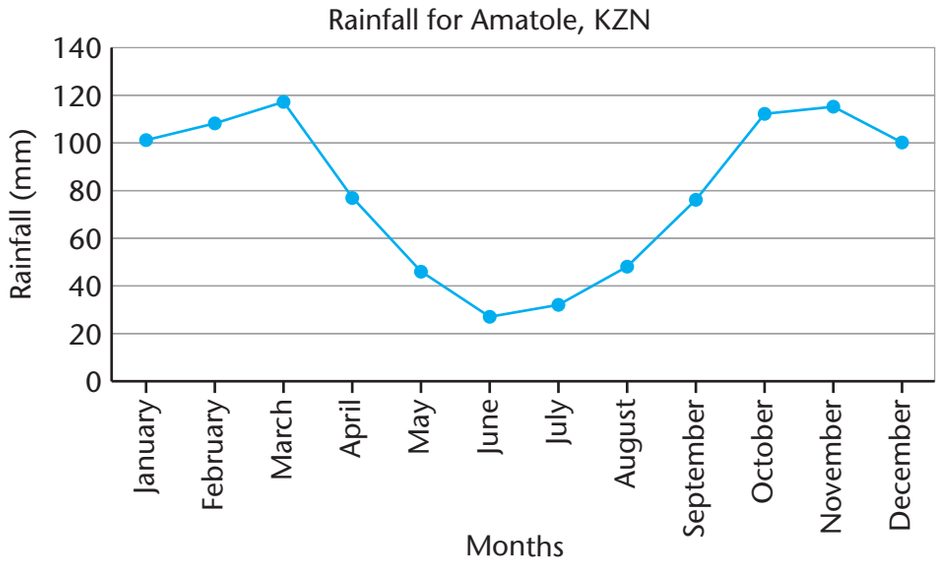
### BROKEN-LINE GRAPHS

**Broken-line graphs** are used to represent data that changes continuously over time. For example, the rainfall for a whole month is captured as one data point, even though the rain is spread out over the month, and it rains on some days and not on others. Broken line graphs are useful to identify and display trends.

Here is some data that can be represented with broken-line graphs:

| Rainfall at three locations in South Africa in 2012 | | | |
|---|---|---|---|
| | Amatole, KZN | Mahikeng, NW | Ceres, WC |
| | Rainfall (mm) | Rainfall (mm) | Rainfall (mm) |
| January | 101 | 118 | 27 |
| February | 108 | 90 | 23 |
| March | 117 | 86 | 41 |
| April | 77 | 61 | 60 |
| May | 46 | 14 | 130 |
| June | 27 | 6 | 168 |
| July | 32 | 3 | 152 |
| August | 48 | 7 | 162 |
| September | 76 | 18 | 88 |
| October | 112 | 46 | 60 |
| November | 115 | 75 | 41 |
| December | 100 | 86 | 36 |

Here is a broken line graph for the Amatole rainfall data:



1. During which four months does Amatole have the least rain?
2. During which six months does Amatole have the most rain?
3. During which months would you plan a hike if you were only considering the rainfall patterns?
4. What other factors should you consider when planning a hike in this region?
5. Make a broken-line graph for the Mahikeng rainfall data.
6. Make a broken-line graph for the Ceres rainfall data.
7. Write a few lines on the difference in rainfall patterns between Ceres and Mahikeng.
8. Draw a combined broken-line graph with the information from all three regions on one graph.

## 23.5 Scatter plots

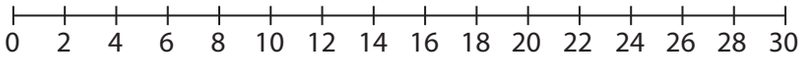
### UNDERSTANDING AND CONSTRUCTING SCATTER PLOTS

**Scatter plots** show how two sets of numerical data are related. Matching pairs of numbers are treated as coordinates and are plotted as a single point. All the points, made up of two data items each, show a scattering across the graph.

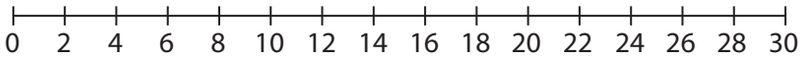
1. This table shows a data set with **two** variables. Study the information in the table.
2. Copy the number lines on the next page and make a dot for each learner's mark for each subject.

| Learners | Mathematics marks | Natural Sciences marks |
|---|---|---|
| **Zinzi** | **25** | **26** |
| John | 23 | 25 |
| **Palesa** | **22** | **25** |
| Siza | 21 | 23 |
| Eric | 20 | 23 |
| Chokocha | 19 | 21 |
| Gabriel | 17 | 20 |
| Simon | 16 | 19 |
| Miriam | 15 | 18 |
| Frederik | 15 | 16 |
| **Sibusiso** | **12** | **15** |
| Meshack | 11 | 13 |
| Duma | 11 | 12 |
| Samuel | 10 | 12 |
| Lola | 10 | 11 |
| Thandile | 9 | 10 |
| **Jabulani** | **8** | **10** |
| Manare | 7 | 9 |
| Marlene | 7 | 7 |
| Mary | 5 | 7 |

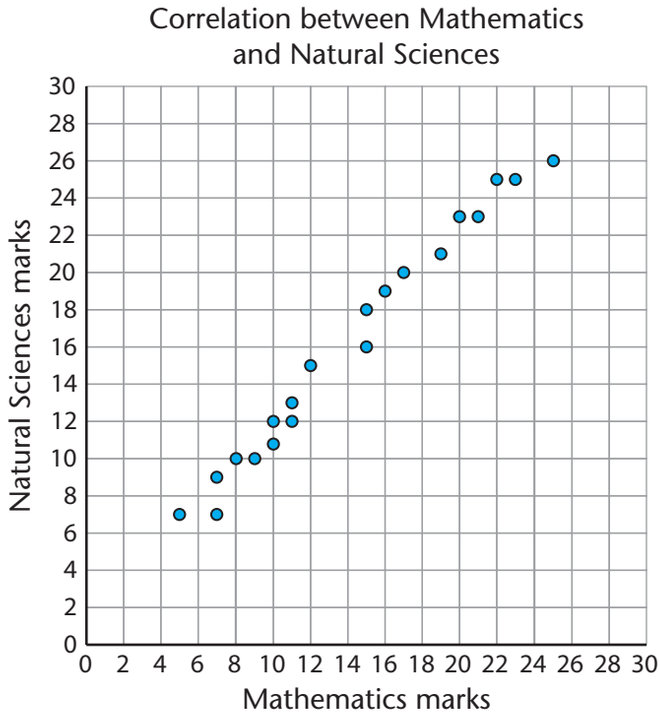
**Natural Sciences marks**




**Mathematics marks**




3. What if you were to show both sets of marks on the same graph, instead of a separate number line for each set? The graph below shows a scatter plot that represents both sets of data. Each dot represents one learner. Copy the scatter plot.




The scatter plot shows the **relationship** between the Natural Sciences mark and the Mathematics mark.

4. Find the dot for Sibusiso in the data set. He obtained a mark of 12 for the Mathematics test and a mark of 15 for Natural Sciences. Find 12 on the horizontal axis. Follow the vertical line up until you reach a blue dot. Find 15 on the vertical axis. Follow the line horizontally until you reach the same blue dot. This blue dot represents the two marks that belong to Sibusiso. On your scatter plot, circle the blue dot and label it "S".

5. Find the data points for Zinzi, Palesa, Jabulani and Mary. On your scatter plot, circle them and label them Z, P, J and M.

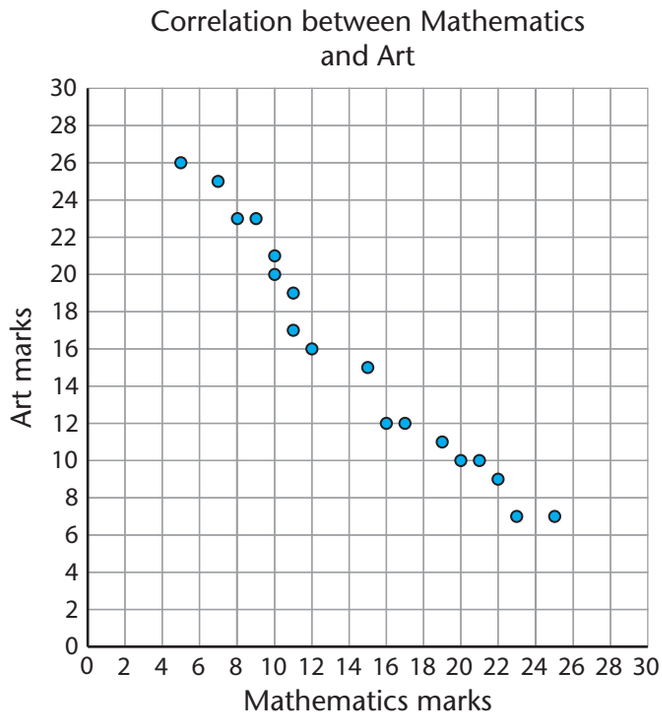
In the example on page 264, a higher Mathematics mark corresponds to a higher Natural Sciences mark. We say there is a **positive correlation** between the Mathematics marks and the Natural Sciences marks.

6. Study this data set and the scatter plot of the data given on the next page. Copy the scatter plot.

| Learner | Mathematics marks | Art marks |
|---|---|---|
| **Zinzi** | **25** | **7** |
| John | 23 | 7 |
| Jabulani | 22 | 9 |
| Siza | 21 | 10 |
| **Eric** | **20** | **10** |
| Chokocha | 19 | 11 |
| Gabriel | 17 | 12 |
| Simon | 16 | 12 |
| **Miriam** | **15** | **15** |
| **Frederik** | **15** | **15** |
| Sibusiso | 12 | 16 |
| Mishack | 11 | 17 |
| Duma | 11 | 19 |
| **Samuel** | **10** | **20** |
| Lola | 10 | 21 |
| Thandile | 9 | 23 |
| Palesa | 8 | 23 |
| Manare | 7 | 25 |
| Marlene | 7 | 25 |
| **Mary** | **5** | **26** |

7. Find Eric in the table. Note his marks for Mathematics and Art. Find the dot that represents his marks on the scatter plot. Encircle it and label it E.
8. Find Samuel in the table. Note his marks for Mathematics and Art. Find the dot that represents his marks. Encircle it and label it S.
9. Compare the two sets of marks for Eric and for Samuel. What do you notice about the marks?
10. On your scatter plot, find the data points on the scatter plot for Zinzi, Eric, Miriam, Frederik, Samuel and Mary. Circle the points and label them Z, E, M, F, S and Ma.

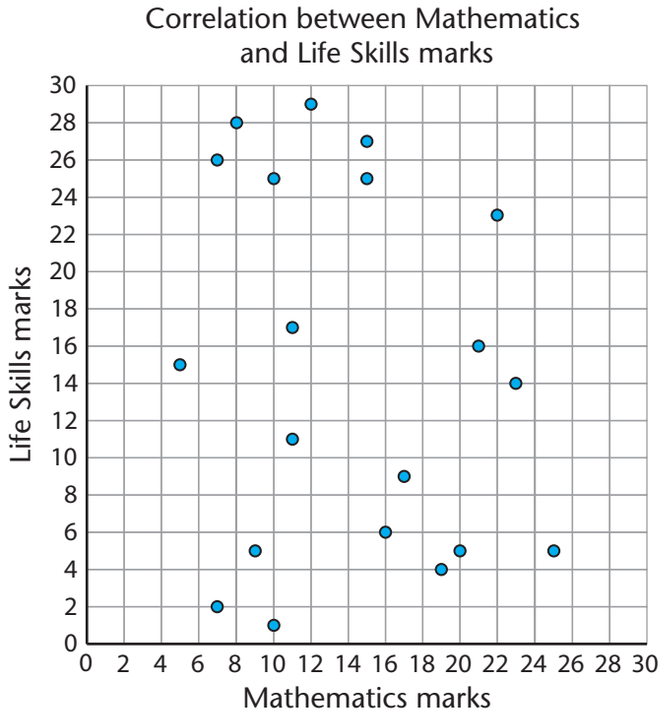
11. What do you notice about the pattern of marks in Mathematics and Art for this data set?




A **negative correlation** is a correlation in which an increase in the value of one piece of data tends to be matched by the decrease in the other set of data. Learners who obtain a high mark for Mathematics appear to obtain a low mark for Art. We say there is a negative correlation between the Mathematics and Art scores for this data set.

A correlation is an assessment of how strongly two sets of data appear to be connected. Two sets of data may be correlated or may show **no correlation**.

Here is the scatter plot for the Mathematics and Life Skills marks of the same group of learners. The table for this data is given on the next page.



12. Study the scatter plot on the previous page and the data table below. Copy the scatter plot.

13. On your scatter plot, find the data points on the scatter plot for Zinzi, Eric, Miriam, Lola and Mary. Circle the points and label them Z, E, M, L and Ma.

14. What do you notice about the pattern of marks in Mathematics and Life Skills for this data set?

| Learner | Mathematics | Life Skills |
|---|---|---|
| Zinzi | 25 | 5 |
| John | 23 | 14 |
| Jabulani | 22 | 23 |
| Siza | 21 | 16 |
| Eric | 20 | 5 |
| Chokocha | 19 | 4 |
| Gabriel | 17 | 9 |
| Simon | 16 | 6 |
| Miriam | 15 | 25 |
| Frederik | 15 | 27 |
| Sibusiso | 12 | 29 |
| Meshack | 11 | 17 |
| Duma | 11 | 11 |
| Samuel | 10 | 1 |
| Lola | 10 | 25 |
| Thandile | 9 | 5 |
| Palesa | 8 | 28 |
| Manare | 7 | 26 |
| Marlene | 7 | 2 |
| Mary | 5 | 15 |

## THE RELATIONSHIP BETWEEN ARM SPAN AND HEIGHT

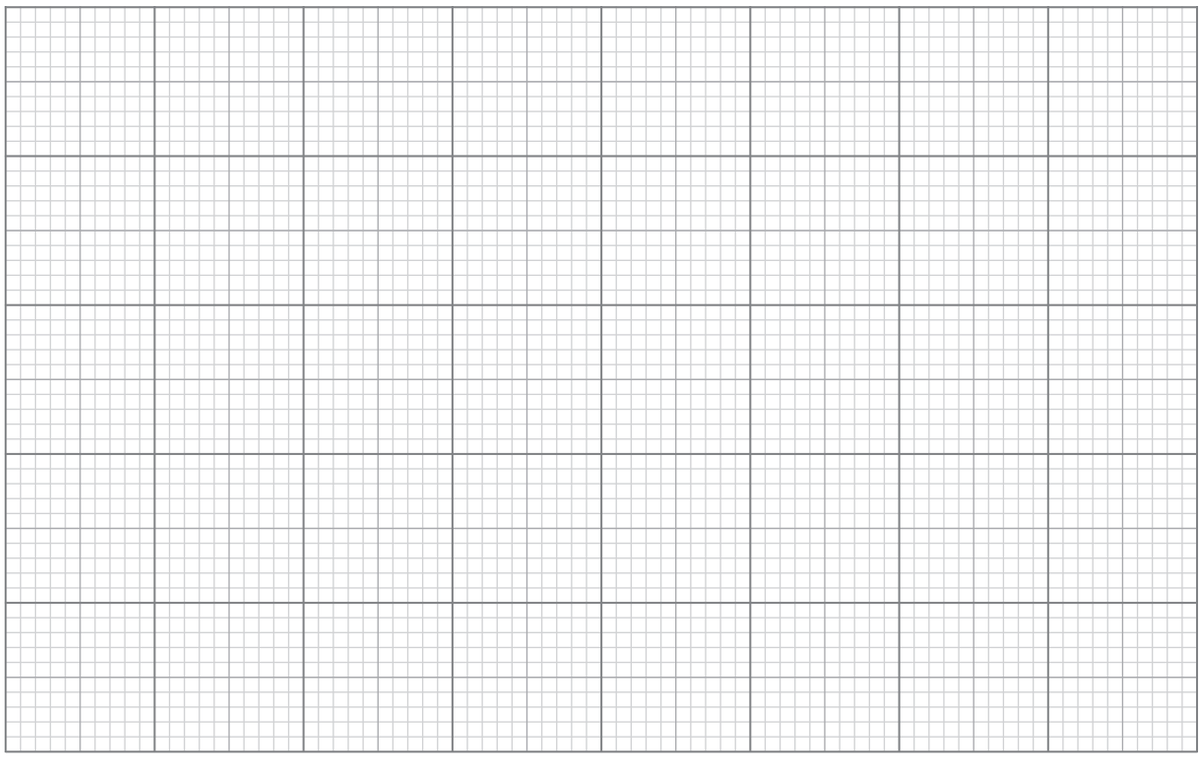
The idea that a person's arm span (the distance from the tip of the middle finger on one hand to the tip of the middle finger on the other hand when the arms are stretched out sideways), is the same as one's height has been explored many times.

A data set for 13 people is given on the next page.

1. Make a scatter plot of this data on a grid like the one below.

   For example, take Cilla's arm span. Find 156 on the horizontal axis. Follow a vertical line up. Then on the vertical axis find 162. Follow a horizontal line across. Where the two points meet, draw a dot.

| Person | Arm span | Height |
|---|---|---|
| Cilla | 156 | 162 |
| Meshack | 159 | 162 |
| Tony | 161 | 160 |
| Ellen | 162 | 170 |
| Karin | 170 | 170 |
| Sibongile | 173 | 185 |
| Gabriel | 177 | 173 |
| Alpheus | 178 | 178 |
| Mfiki | 188 | 188 |
| Nathi | 188 | 182 |
| Manare | 188 | 192 |
| Khanyi | 196 | 184 |



2. What would you say about the correlation between the arm span and the height?

# Chapter 24
# Interpret, analyse and report on data

## 24.1 Which graph is best?

You have learnt that certain types of graphs are best for displaying certain kinds of information. The type of graph depends mostly on the type of data that needs to be represented. Here is a summary of the advantages of different types of graphs:

**Tables** show more information than graphs but the patterns are not as easy to see. They do not give a visual impression of particular trends.

**Pie charts** show a whole divided into parts. They show how the parts relate to each other and how the parts relate to a whole. They do not show the quantities involved.

**Bar graphs** show the amounts or quantities involved but do not show the relationship as effectively as pie charts. They are useful for showing **quantitative** data. Bar charts allow us to compare the quantities of different categories, for example, the sales of different items.

A **double-bar graph** is used to compare two or more things for each category. For example, we could use a double-bar graph to compare the differences between males and females.

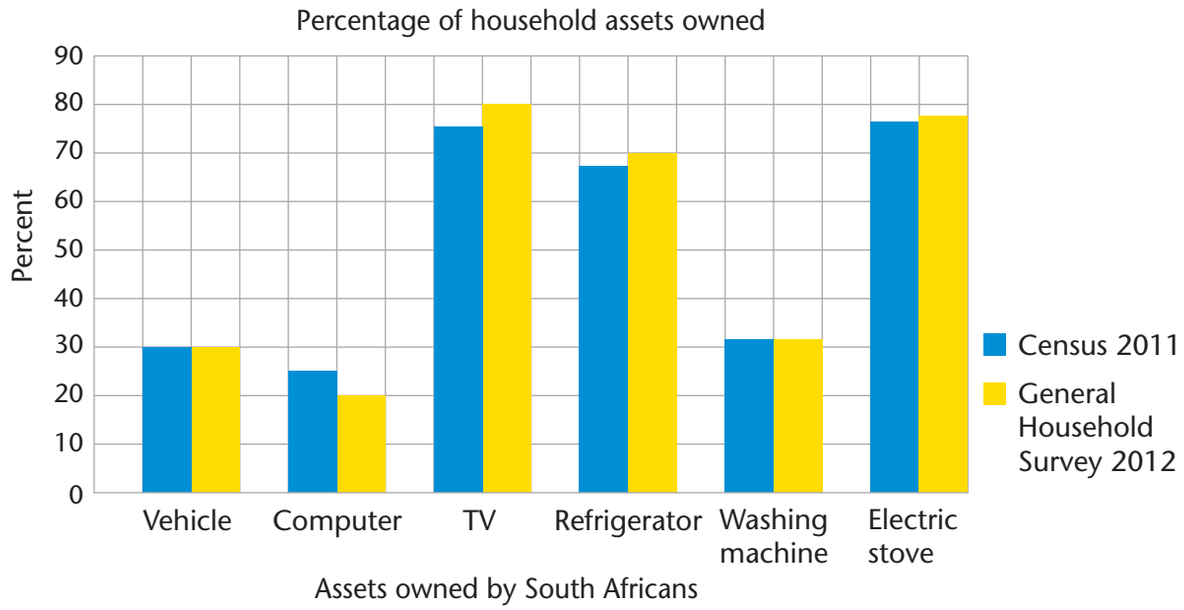
**Histograms** are used to represent numerical data that is grouped into equal class intervals. Histograms are useful to show the way the data is spread out.

**Broken-line graphs** show trends or changes in quantities over time.

## CHOOSE THE BEST REPRESENTATION

1. Which kind of graph is best to represent each of the following? Explain your answers.
   (a) Showing the value of the rand against the US dollar over several years
   (b) Comparing the monthly sales of six different makes of car in 2014 and 2015
   (c) The proportion of people of different age groups in a town
   (d) The quantities of different crops produced on a farm
   (e) The percentages of different goods sold to make up the total sales for a shop
   (f) The change in HIV infection rates over time

2. This graph was published by Statistics South Africa to show the assets owned by South Africans. The blue bar shows the Census 2011 results and the yellow bar shows the General Household Survey 2012 results.




Give reasons for your answers to the following questions:
   (a) Is it useful to show the differences in the results of Census 2011 and the General Household Survey 2012?
   (b) Is it useful to collect data on assets that people own?
   (c) Is it useful to show that lower percentages of people own certain assets?
   (d) The different coloured bars represent the two different surveys. Draw up a table to show the data in table form. (Read the percentages as accurately as you can from the graph and round off the data to the nearest whole number for the table.)
   (e) Does the table show the data as effectively as the double bar chart? Give your own opinion.

3. The table below shows the employment status of people ages 15–64 years in South Africa. Discuss some ways of representing the data (e.g. graphs). Justify your answers.

| | Jul–Sept 2012 | Apr–June 2013 | Jul–Sep 2013 |
|---|---|---|---|
| | Number of people (thousands) | | |
| **Population 15–64 years old** | 33 017 | 33 352 | 33 464 |
| **Labour force** | 18 313 | 18 444 | 18 638 |
| **Employed** | **13 645** | **13 720** | **14 028** |
| Formal sector (non-agricultural) | 9 663 | 9 694 | 10 008 |
| Informal sector (non-agricultural) | 2 197 | 2 221 | 2 182 |
| Agriculture | 661 | 712 | 706 |
| Private households | 1 124 | 1 093 | 1 132 |
| **Unemployed** | **4 668** | **4 723** | **4 609** |
| **Not economically active** | **14 705** | **14 908** | **14 826** |
| Discouraged work-seekers | 2 170 | 2 365 | 2 240 |
| Other (not economically active) | 12 535 | 12 543 | 12 586 |
| **Unemployment rate (%)** | 25,5 | 25,6 | 24,7 |

(a) The percentages of the employed, unemployed, and not economically active people in July–September 2013
(b) The change in the employment rates over three time periods
(c) The proportions of employed people who work in the formal sector, informal sector, agriculture and private households
(d) The numbers of the employed and unemployed over the three time periods

## 24.2 The effects of summary statistics on how data is reported

Information articles often use averages to report information. The articles might not use the exact terms for average that you have learnt about: the mean, median and mode. Instead, they may use terms such as "most". However, it is important to be sure about the kind of average to which a report refers, because an average gives us different information.

- Remember that the **mean** is useful for describing a set of measurement values, but can also be used for other numerical data sets. The word "average" usually refers to the "mean" if it is not explained further. The mean is not reliable if a data set is too spread out.
- The **median** is the value in the middle of a data set when it is arranged in order. Half the values in the data set are lower than the median and half of them are higher than the median. The median is often the average used when data values are not uniformly distributed, because the mean is affected by extreme values in the data set, while the median is not. For example, house prices vary widely, so the median would be a better description of the data than the mean. When the median is given in a report, the writer should state that he or she is using the median or middle value.
- The **mode** is the number that occurs most often in a set of data. For example, if we collect data about people's favourite colours, the data set would be a list of colours, and the mode would be the colour that comes up most often. The mode can also be used for numbers. Not all data sets have a mode, because sometimes none of the numbers occurs more than once.

**Example:** The standard way of reporting house prices in South Africa and internationally is the median house price, which is used by economists in financial reports. The median is regarded as more useful than the mean house price because the sale of a few expensive houses would increase the mean, but would not affect the median.

If a bank gives bonds for eight houses to the value of R100 000, and for two houses to the value of R1 million, then the mean would be R280 000. This does not seem to be an accurate reflection of the value of the houses, because it is distorted by the higher values. The median house price would be R100 000, which is an accurate reflection of the prices.

Remember that the median is the middle point, and half of the values fall below the median, and half above. If the median is lower than the mean, this shows us that there are high values that are distorting the mean.

### USING DIFFERENT SUMMARY STATISTICS

1. What kind of average is used in each of these statements?
   (a) The average family has 2,6 children.
   (b) Most families have three children.
   (c) Most people prefer red cars.
   (d) The average height for women is 1,62 m.
   (e) More people shop after work than at any other time during the day.

2. The mean monthly salary of all the staff at company ABC is R8 000 per month, but the median salary is R5 000.
   (a) Explain why the two summary statistics are so different.
   (b) Which summary statistic gives a better idea of the salaries at the company? Give reasons for your answer.

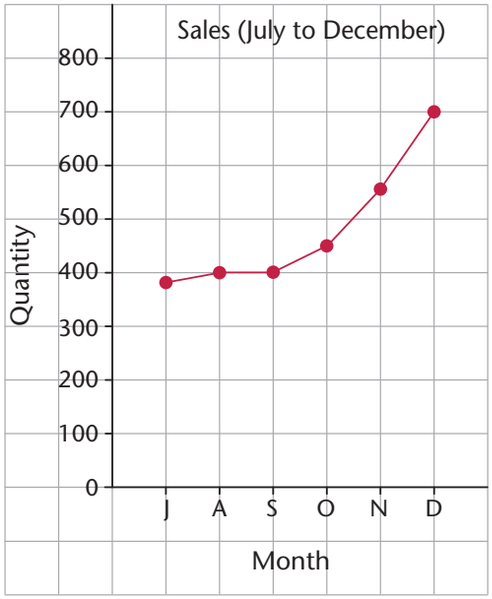
## 24.3 Misleading graphs

The media (i.e. newspapers, magazines and television), regularly use graphs to show information. Unfortunately, the information is often manipulated to emphasise a particular result. This may be because the writer simply wants to make his or her argument more obvious to the reader.

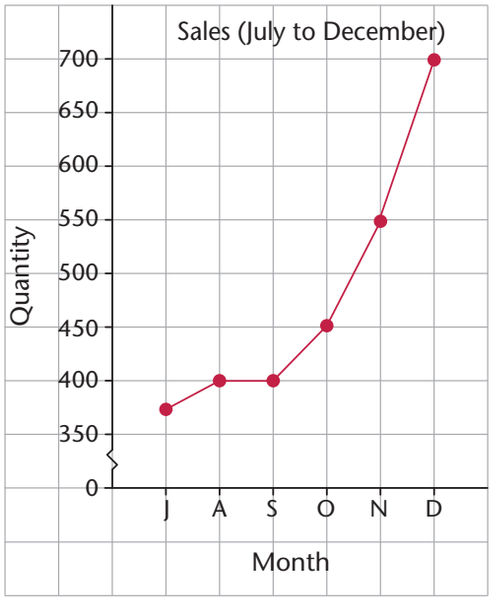
### Changing the scale of the axis

If you change the scale of the vertical axis on bar graphs and line graphs, you will change the way the graphs look. For a bar graph, the larger the spaces between the numbers on the vertical axis, the bigger the difference between the bars. The smaller the spaces between the numbers on the axis, the smaller the difference in the height of the bars. The same is true for a line graph which will either have sharp points or be much flatter, depending on how you have changed the scale.

**Example:** The two broken-line graphs on the next page show the same sales data for a business over a period of six months. Which graph gives the more accurate impression?




*Graph A*



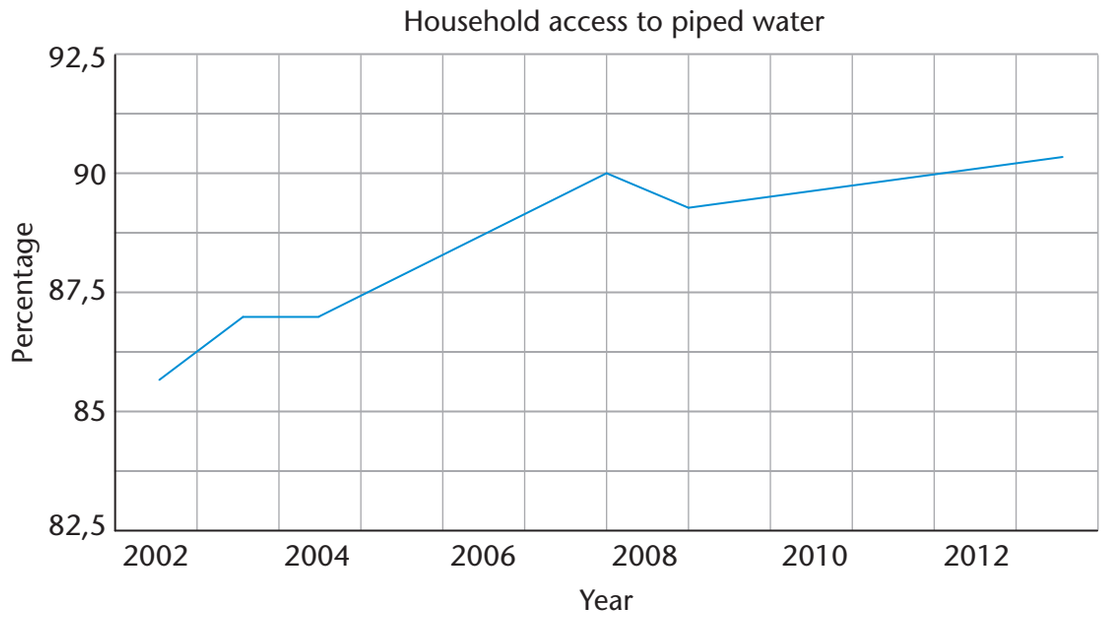

*Graph B*

Graph B has a different scale on the vertical axis. The vertical axis does not start at 0 and so **two** blocks on the vertical axis represent 100 items instead of only **one** block, as in Graph A. This makes it look as if the sales increased rapidly over the six months.

Note that it is not necessarily wrong to change the scale on the axes or not to start at 0. For example, graphs showing stock exchange fluctuations rarely show the origin on the graph and stockbrokers are taught to interpret the graphs in that form. Sometimes small changes in data values have important effects and in these cases, it may be valid to change the scale to show these.

## ANALYSING GRAPHS

1. This graph from Statistics South Africa shows the increase in the percentage of households that had access to piped water over a ten-year period.

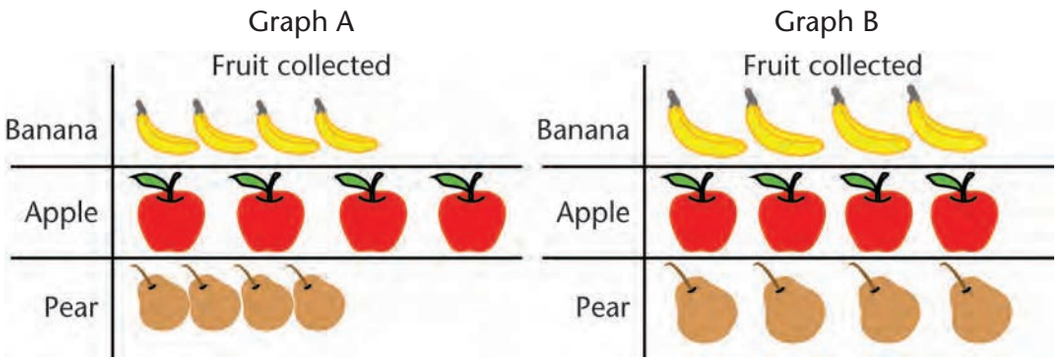



   (a) Comment on the scale used on the vertical axis. Is this a misleading graph?
   (b) How could you redraw the graph so that the differences on the graph are more noticeable?
   (c) How could you draw the graph so that the differences are less noticeable?

2. In this graph the height of the houses represents the number of sales.


   Do you think that this graph is misleading? Give reason(s) for your answer.

3. Look at the two graphs below:

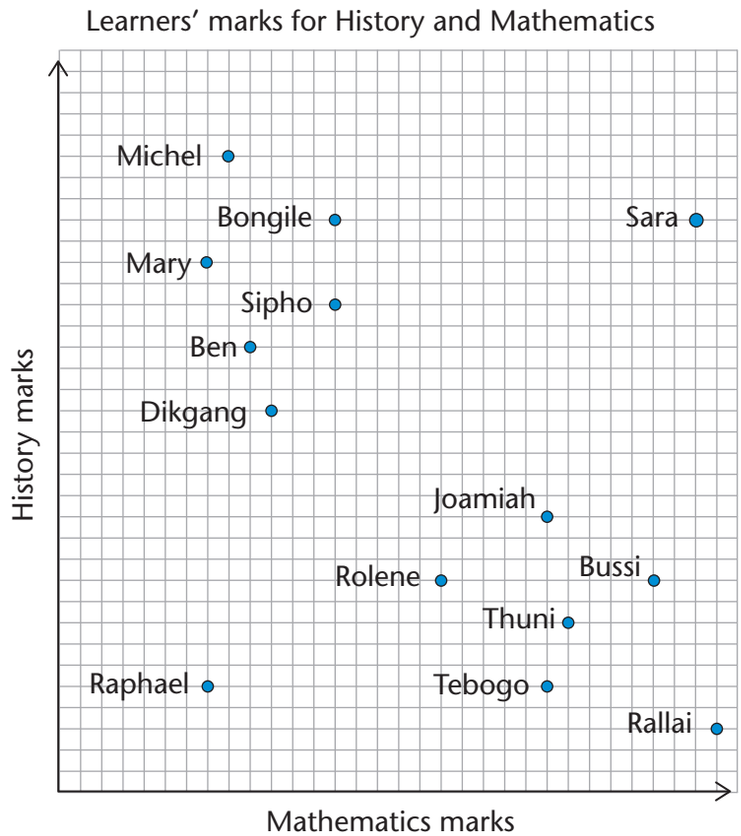



Which graph do you think is drawn correctly? Explain your answer.

## 24.4 Analysing extreme values and outliers

A data item that is very different from all (or most) of the other items in a data set is called an **outlier**.

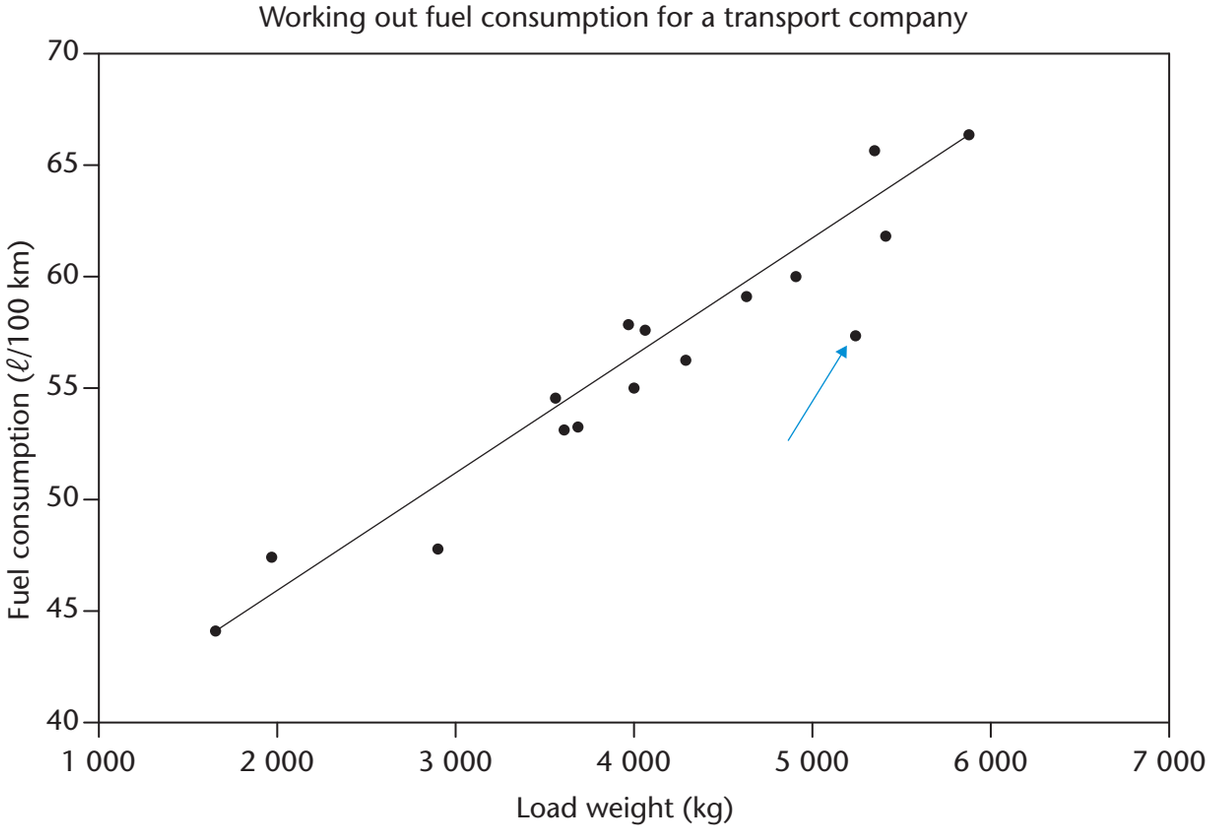
It is sometimes difficult to notice outliers in numerical data. However, outliers often become clearly noticeable when data is displayed with graphs.



1. The scatter plot on the previous page shows the performance of a group of learners in Mathematics and History. Which of the points on the scatter plot can be regarded as outliers? Give reasons for your answer.

Outliers in data sets can be very important. We need to decide if there is a particular reason for the value being so different to the others. Sometimes it gives us important information. In some cases, the data collected for that point could be wrong.

The scatter plot below is for data collected by a transport company.




The company uses just one type of truck. Before each transport job, the company has to specify the price for the job. In order to specify a price before a job, the company needs to estimate how much their costs will be for doing the job. One of the main costs is the cost of fuel, and the main factor influencing the amount of fuel used is the distance. The load weight also plays a role: the greater the load weight, the higher the fuel consumption (litres/100 km).

The table on the next page gives information that was recorded for previous transport jobs. The jobs are numbered from 1 to 16 and for each job the values of the four variables *distance*, *load weight*, *amount of fuel used* and *fuel consumption rate* are given.

2. (a) Which of the four variables are represented on the scatter plot given above?
   (b) What are the values of these two variables for the point indicated by the blue arrow on the scatter plot?

| Job number | Distance (km) | Load weight (kg) | Fuel used (ℓ) | Fuel consumption (ℓ/100 km) |
|---|---|---|---|---|
| 1 | 1 304 | 5 445 | 879 | 67,4 |
| 2 | 1 320 | 2 954 | 639 | 48,4 |
| 3 | 1 151 | 4 705 | 698 | 60,6 |
| 4 | 1 371 | 4 378 | 787 | 57,4 |
| 5 | 325 | 3 673 | 176 | 54,2 |
| 6 | 1 630 | 5 995 | 1 113 | 68,3 |
| 7 | 1 023 | 5 357 | 600 | 58,7 |
| 8 | 620 | 4 988 | 382 | 61,6 |
| 9 | 73 | 1 992 | 35 | 47,9 |
| 10 | 1 071 | 5 529 | 680 | 63,5 |
| 11 | 370 | 4 140 | 218 | 58,9 |
| 12 | 1 423 | 4 062 | 843 | 59,2 |
| 13 | 394 | 4 068 | 221 | 56,1 |
| 14 | 1 536 | 1 678 | 682 | 44,4 |
| 15 | 1 633 | 3 736 | 887 | 54,3 |
| 16 | 435 | 3 644 | 241 | 55,4 |

3. (a) Consider the scatter plot and the data set. What is the effect of load weight on fuel consumption?
   (b) Is job 7 an exception in this respect? Explain your answer.

4. Further investigations revealed that the driver for jobs 2 and 7 was the same person, and that he was not the driver for any other jobs. What may this indicate?

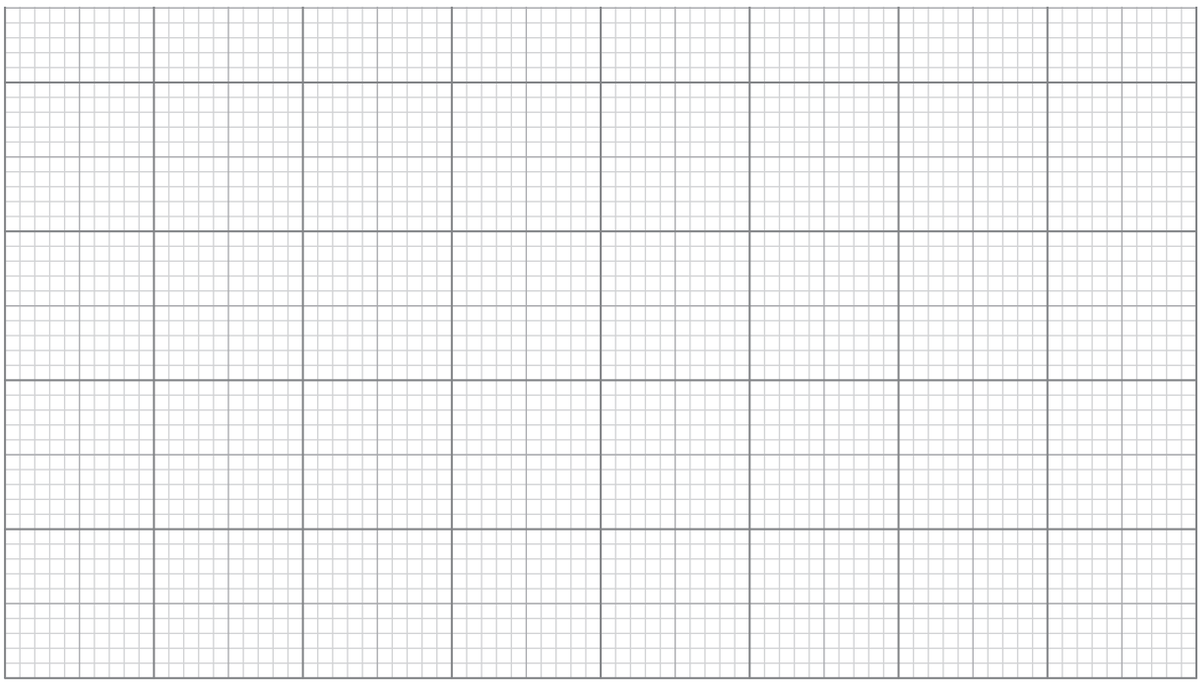
## FIND OUTLIERS

Researchers collected data on the population of some African countries (including the Seychelles), which included the income per person and the percentage of the income spent on health.

| Country | Total population (in 1 000s) | Total annual national income per person (US$) | Percentage of income spent on health |
|---|---|---|---|
| Angola | 18 498 | 4 830 | 4,6 |
| Botswana | 1 950 | 13 310 | 10,3 |
| DRC | 66 020 | 280 | 2,0 |

| Country | Total population (in 1 000s) | Total annual national income per person (US$) | Percentage of income spent on health |
|---|---|---|---|
| Lesotho | 2 067 | 1 970 | 8,2 |
| Malawi | 15 263 | 810 | 6,2 |
| Mauritius | 1 288 | 12 580 | 5,7 |
| Mozambique | 22 894 | 770 | 5,7 |
| Namibia | 2 171 | 6 250 | 5,9 |
| Seychelles | 84 | 19 650 | 4,0 |
| South Africa | 50 110 | 9 790 | 8,5 |
| Swaziland | 1 185 | 5 000 | 6,3 |
| Tanzania | 43 739 | 1 260 | 5,1 |
| Zambia | 12 935 | 1 230 | 4,8 |

1. What are the three variables in this table?
2. Why do you think it is important to look at income per person in this case, rather than the total income?
3. On graph paper, plot the points for the national income per person and the percentage spent on health care for each country.



4. Write a short report on the data in the table and what the scatter plot shows you about the data. Comment on the general trend and any outliers.

# Chapter 25
# Probability

## 25.1 Simple events

**REVISION**

| yellow | green | pink | blue | red | brown | grey | black |
|---|---|---|---|---|---|---|---|

1. (a) Suppose the eight coloured buttons above are in a bag and you draw one button from the bag without looking. Can you tell what colour you will draw?
   (b) Suppose you repeatedly draw a button from the bag, note its colour, then put it back. Can you tell in approximately what fraction of all the trials the button will be yellow?

Archie has a theory. Because the eight possible outcomes are equally likely, he believes that if you perform eight trials in a situation like the above you will draw each colour once.

2. If Archie's theory is correct, how many times will each colour be drawn if 40 trials are performed?

3. If Archie's theory is correct, in what fraction of the total number of trials will each colour be drawn?

4. If Archie's theory is correct, how many times will each of the colours be drawn if a total of 40 trials is performed? Copy the table on the following page and write your answers in the second row of the table. Write the predicted relative frequencies in row 3 as fortieths, and in row 4 as two hundredths.

Each time you draw a button from the bag without looking, you perform a **trial**. If you do this and put the button back, and repeat the same actions eight times, you have performed eight trials.

The number of times an event occurs during a set of trials is called the **frequency** of the event.

When the frequency of an event is expressed as a fraction of the total number of trials, it is called the **relative frequency**.

| Colour | Yellow | Green | Pink | Blue | Red | Brown | Grey | Black |
|---|---|---|---|---|---|---|---|---|
| Frequencies predicted by Archie | | | | | | | | |
| Relative frequencies predicted by Archie expressed in fortieths | | | | | | | | |
| Relative frequencies predicted by Archie expressed in two hundredths | | | | | | | | |

The relative frequency for each colour that Archie predicted is called the **probability** of drawing that colour. If all the outcomes are equally likely, then:

$$\text{probability of an outcome} = \frac{1}{\text{the total number of equally-likely outcomes}}$$

You will now investigate whether or not Archie's theory is correct.

5. (a) Make eight small cards and write the name of one of the above colours on each card, so that you have cards with the eight colour names. Perform eight trials to check whether or not Archie's theory is correct. Copy the table below and record your results (your tally marks 1 and your frequencies 1) in the relevant row of the table.
   (b) Find out what any four of your classmates found when they did the experiment. Enter their results in your table too (Friend 1, 2, 3 and 4 frequencies).

Table for the results of the experiments

| Colour | Yellow | Green | Pink | Blue | Red | Brown | Grey | Black |
|---|---|---|---|---|---|---|---|---|
| Your tally marks (1) | | | | | | | | |
| Your frequencies (1) | | | | | | | | |
| Friend 1 frequencies | | | | | | | | |
| Friend 2 frequencies | | | | | | | | |
| Friend 3 frequencies | | | | | | | | |
| Friend 4 frequencies | | | | | | | | |
| Total frequencies for 5 experiments | | | | | | | | |

6. (a) What was the total number of trials in the five experiments you recorded in the table?
   (b) What is the total of the frequencies for the different colours, in the last row of your table?

7. Is Archie's theory correct?

Bettina has a different theory to Archie's. She believes that if one does many trials with the eight buttons in a bag, each colour will be drawn in **approximately** one-eighth of the cases. In other words, Bettina believes that the relative frequency of each outcome will be close to the probability of that outcome, but may not be equal to it.

8. (a) You and your four classmates performed 40 trials in total. Copy the table below and enter the results in the second row of the table. Also express each frequency as a fraction of 40, in fortieths and in two hundredths.

| Colour | Yellow | Green | Pink | Blue | Red | Brown | Grey | Black |
|---|---|---|---|---|---|---|---|---|
| Actual frequencies obtained in your experiments (40 trials) | | | | | | | | |
| Relative frequencies as fortieths | | | | | | | | |
| Relative frequencies as two hundredths | | | | | | | | |
| Probability as two hundredths | | | | | | | | |

(b) Do your experiments show that Bettina's theory is correct or not?

Jayden believes that when more trials are performed, the relative frequencies will get closer to the probabilities.

You will now do an investigation to find out whether Jayden's theory is true.

## INVESTIGATE WHAT HAPPENS WHEN MORE TRIALS ARE DONE

1. Perform 40 trials by drawing one card at a time from eight small cards with the names of the colours written on them, and enter your results in the second and third rows of a table like the one shown below.

| Colour | Yellow | Green | Pink | Blue | Red | Brown | Grey | Black |
|---|---|---|---|---|---|---|---|---|
| Tally marks | | | | | | | | |
| Frequencies | | | | | | | | |
| Relative frequencies as fortieths | | | | | | | | |
| Relative frequencies as two hundredths | | | | | | | | |
| Probabilities as two hundredths | | | | | | | | |

2. Make a copy of the table on page 282, but leave out the row for tally marks, the row for the relative frequencies as fortieths and the row for the probabilities, on a loose sheet of paper. Exchange it with a classmate. Copy the following Tables 1 and 2 and enter the results of your classmate on Table 1 and 2. Also enter your own results for question 1 on the tables.
3. Get hold of the data reports of three other classmates, and enter these on the tables as well.
4. Add the frequencies of the various colours in the five sets of data for 40 trials each, and calculate the relative frequencies expressed as two hundredths.
5. Is the range of relative frequencies for 200 trials smaller than the ranges for the five different sets of 40 trials each? What does this indicate with respect to Jayden's theory?

When only a small number of trials are done, the actual relative frequencies for different outcomes may differ a lot from the probabilities of the outcomes.

When many trials are done, the actual relative frequencies of the different outcomes are quite close to the probabilities of the outcomes.

Table 1: Frequencies for five sets of 40 trials each

| Colour | Yellow | Green | Pink | Blue | Red | Brown | Grey | Black |
|---|---|---|---|---|---|---|---|---|
| Frequencies for your own 40 trials in question 1 | | | | | | | | |
| Frequencies for 40 trials by classmate 1 | | | | | | | | |
| Frequencies for 40 trials by classmate 2 | | | | | | | | |
| Frequencies for 40 trials by classmate 3 | | | | | | | | |
| Frequencies for 40 trials by classmate 4 | | | | | | | | |
| Total frequencies for 200 trials | | | | | | | | |
| Relative frequencies for 200 trials as two hundredths | | | | | | | | |

Table 2: Relative frequencies for each of the five sets of 40 trials each (expressed as two hundredths)

| Colour | Yellow | Green | Pink | Blue | Red | Brown | Grey | Black |
|---|---|---|---|---|---|---|---|---|
| Relative frequencies for your own 40 trials | | | | | | | | |
| Relative frequencies for 40 trials by classmate 1 | | | | | | | | |
| Relative frequencies for 40 trials by classmate 2 | | | | | | | | |
| Relative frequencies for 40 trials by classmate 3 | | | | | | | | |
| Relative frequencies for 40 trials by classmate 4 | | | | | | | | |

6. How many different three-digit numbers can be formed with the symbols 3 and 5, if no other symbols are used? You may use one, two or three of the symbols in each number, and you may repeat the same symbol.

## 25.2 Compound events

### TOSSING A COIN AND GIVING BIRTH

1. Simon threw a coin and the outcome was heads. He will now throw the coin again.
   (a) What are the possible outcomes?
   (b) What is the probability of each of the possible outcomes?
   (c) What are the possible outcomes if Simon throws the coin for a third time?
   (d) What is the probability of each of the possible outcomes for the third throw?

What happens when a coin is thrown for a second time has nothing to do with what happened when it was thrown the first time.

The first throw and the second throws are called **independent events**, i.e. what happened on the first throw cannot influence what will happen on the second throw.

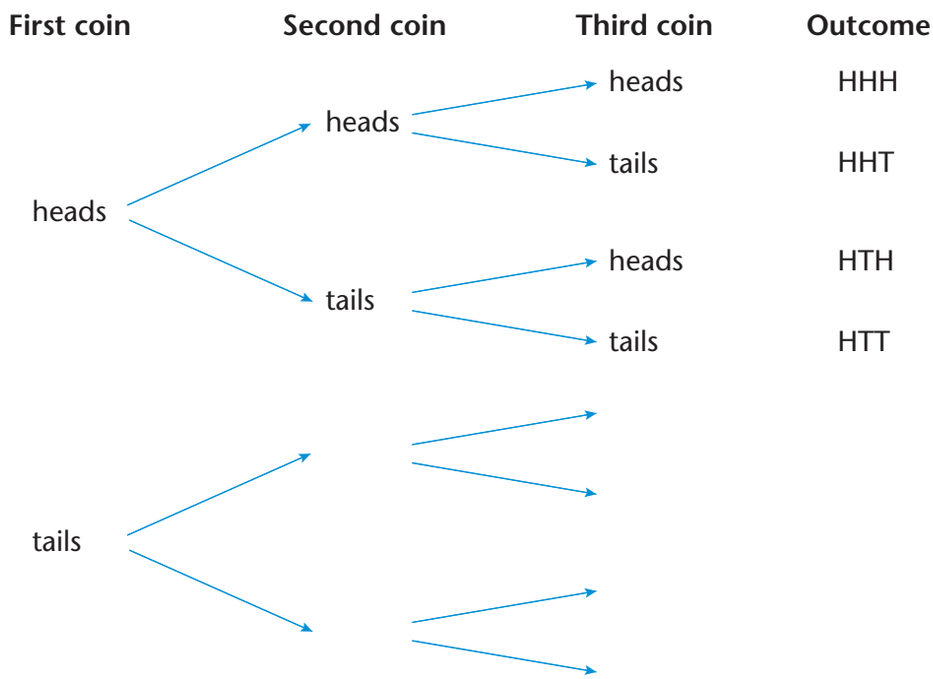
2. (a) If an event has four different equally-likely outcomes, what is the probability of each of the four outcomes?
   (b) Does that mean that if the event is repeated four times, each of the four outcomes will happen once?
   (c) Does your answer in (a) mean that if the event is repeated 100 times, each of the four outcomes will happen 25 times?

3. (a) What are the possible outcomes when two coins are thrown? Copy and complete the **two-way table** below to answer this question. One possible outcome is already given.

| | **Heads** | **Tails** |
|---|---|---|
| **Heads** | | H T |
| **Tails** | | |

(b) Do you think these four outcomes are equally likely?
(c) What is the probability of each of the four outcomes?
(d) What is the probability of getting a head and a tail?

4. Let us consider the possible outcomes if three coins are thrown.
Below is a tree diagram that can help you figure out what the different possible outcomes are. Complete the diagram by filling in the missing information.




5. (a) Do you think the eight different outcomes in question 4 are equally likely?
(b) What is the probability of each of the eight outcomes?
(c) What is the probability of throwing two heads and one tail?

6. In question 6 on page 284 you were asked to write down the various numbers that can be formed by using symbols 3 and 5. Think of all the four-letter codes that you can form by using only two letters, P and Q. Any letter can be used more than once in one code. First think about how you will go about finding all the possibilities in a systematic way, and then try to set up a tree diagram to help you.

(a) Draw a tree diagram to help you to solve this problem. List all the outcomes.
(b) If the codes are formed by randomly choosing the letters, what is the probability that the code will consist of the same letter being used four times?
(c) What is the probability that the code will consist of two Ps and two Qs?

When a woman is pregnant, the baby can be a boy or a girl. Suppose we make the assumption that the two possibilities are equally likely, so the probability of a boy is $\frac{1}{2}$ and the probability of a girl is $\frac{1}{2}$.

7. (a) Copy and complete this two-way table to show the possible outcomes of the gender of the two children in a family.

| | **Boy** | **Girl** |
|---|---|---|
| | | |
| | | |

(b) List the possible outcomes.
(c) What is the probability that the two children in the family will be of the same gender?
(d) What is the probability that the eldest child will be a boy and that they will then have a girl?

8. A certain woman already has a boy. She now expects a second child. What is the probability of it being a boy again, if we make the assumption that a baby being a boy or a girl are equally likely events?

The assumption that a boy or a girl being born are equally likely events may not actually be true. However, probabilities can only be calculated and used to make predictions if it is assumed that outcomes are equally likely.

9. (a) A woman gets married and plans to have a baby in one year and another baby in the next year. What is the probability that both babies will be girls?
(b) A woman gets married and plans to have a baby in each of the first three years of the marriage. What is the probability that she will have a boy in the first year, and girls in the second and third years?