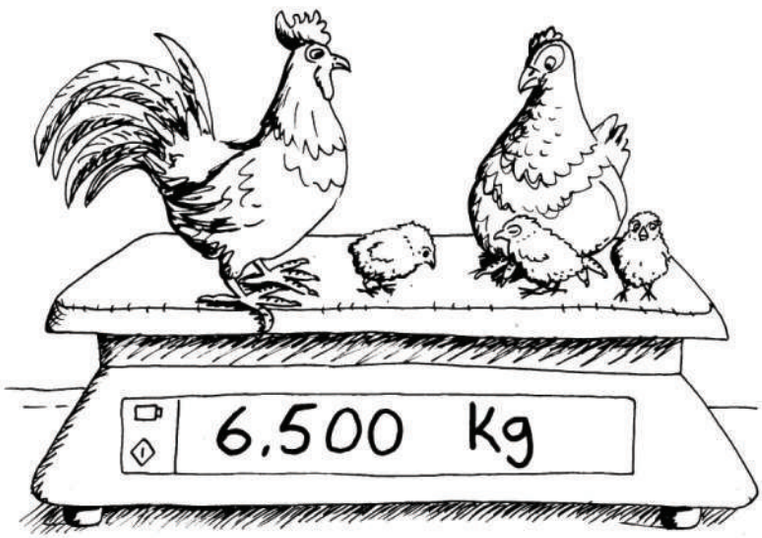
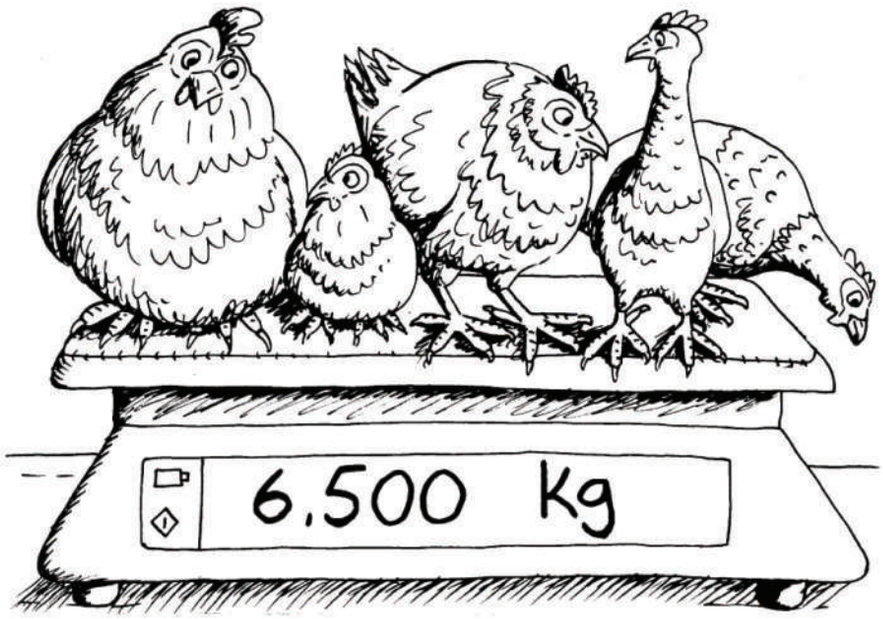
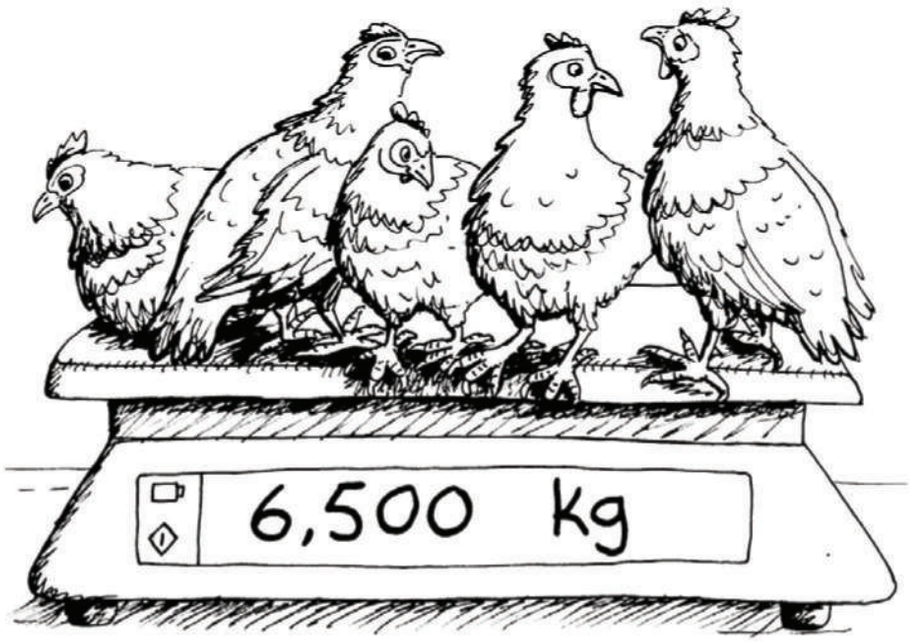
Chapter 8
Collect, organise and summarise data

You have learnt how to collect, organise and summarise data in previous grades. In Grade 9, you need to decide which methods are best in certain situations and you need to be able to justify your choices.







8 Collect, organise and summarise data

8.1 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 don'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 will more likely complete a short questionnaire that is quick and easy to complete. Exclude information that is not needed.

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 blow.

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 free to write what they like. 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 schools. What are the contents of the school lunch?

   .....

   .....

   .....

   (b) Whether the tellers at a supermarket chain are happy with their conditions of work.

   .....

   .....

   .....

   (c) Whether 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, on the next page.

8.2 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 organized 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 6-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 here 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
100	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 in 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?

.....

8.3 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 don't 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 below and on the next page.

.....

.....

Type A:

.....

.....

.....

.....

Type B:

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Type C:

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

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

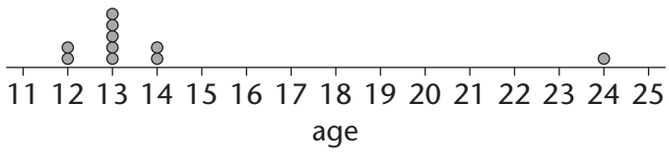
Data set	Best measure of central tendency
The shoe sizes of boys in Grade 9	
An evenly-spread set of measurement values, such as the heights of learners in a class	
A set of measurement values with a few very low values and mostly high values	
The number of siblings each person in your class has	
The 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 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:



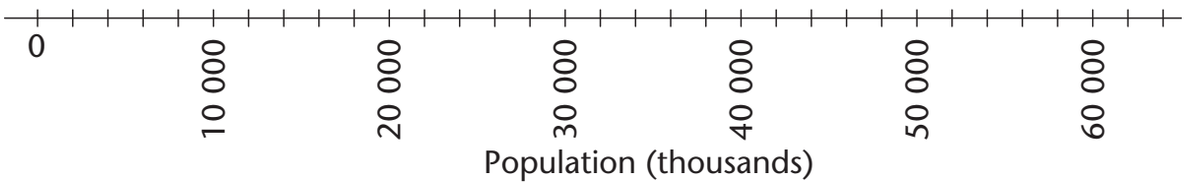
In this case, the value of 24 years old could be an unreasonable value. This depends on the context of the survey.

You will learn more about extreme values and outliers in Chapter 10.

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) Draw a dot plot on the number line below 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*. Comment on the spread of the data. .....

.....