

Presentation of data

9

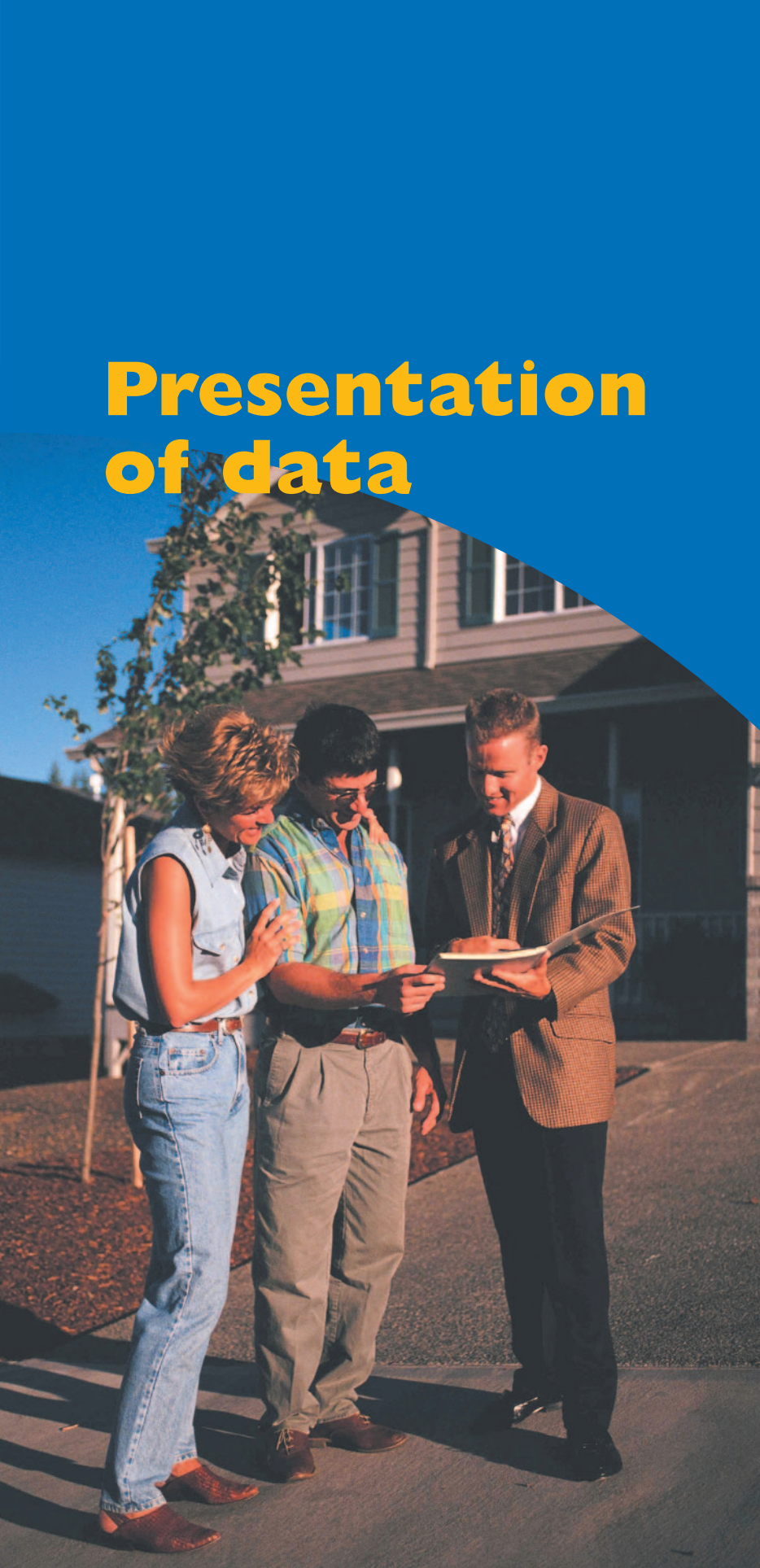
syllabus reference

Topic:

- Applied statistical analysis

In this chapter

- 9A Types of variables and data
- 9B Collection of data
- 9C Stem plots
- 9D Frequency histograms and bar charts
- 9E Describing the shape of stem plots and histograms
- 9F Cumulative data



Introduction

Karen is a real estate agent. At the end of each year it is part of her job to perform a statistical analysis of house prices in the local area. In her real estate agency, currently there are 60 houses for sale and Karen has summarised their prices in the table below.

Price range	Number of houses
\$75 000–\$100 000	1
\$100 000–\$125 000	5
\$125 000–\$150 000	7
\$150 000–\$175 000	6
\$175 000–\$200 000	11
\$200 000–\$225 000	14
\$225 000–\$250 000	9
\$250 000–\$275 000	4
\$275 000–\$300 000	0
\$300 000–\$325 000	1
\$325 000–\$350 000	0
\$350 000–\$375 000	0
\$375 000–\$400 000	2

Karen must make a presentation on her findings to the Real Estate Institute. What are the different ways in which she can present this information?

In this chapter we will look at different types of data and consider statistics in terms of their presentation.

Defining statistics

People have always been interested in the collection of information about themselves and their environment. The collection of such information in a systematic fashion is called *statistics*.

Types of variables and data

A statistical investigation usually involves looking at a characteristic of a population. Because this characteristic varies for different members of the population this characteristic is called the *variable*.

Once a known piece of information is assigned to a variable this then becomes a piece of data. For example, if we are studying the maximum daily temperature, we refer to the temperature as the variable as it can be different every day. If we say that it was 25°C on Thursday, a value has been assigned to the variable and this is now a piece of data. The variable that is being studied in an investigation, or a piece of data, can be described as either *categorical* or *qualitative*.

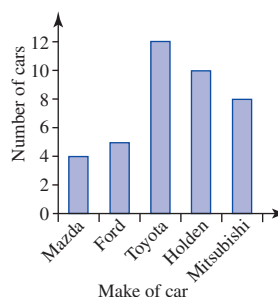
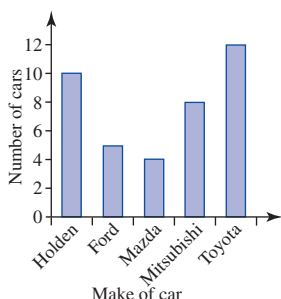
Categorical

Categorical data cannot be measured; they can only be put into categories.

An example of categorical data is makes of cars. The categories for the data would be all possible makes of cars such as Ford, Holden, Toyota, Mazda etc. Other questions that would lead to categorical data would be things such as:

- What is your hair colour?
- Who is your favourite musical performer?
- What method of transport do you use to get to school?

Categorical data cannot be placed in a specific order. Although the graphs show the same information, they look different.



With categorical variables, the order in which the columns appear is not important and can be changed without altering the meaning, as they are non-numerical. The frequency of each type of car, which is numerical, is unaffected by the order of the columns.

Quantitative

Quantitative data can be measured. They are data to which we can assign a numerical value.

Data concerning quantitative variables are collected by measurement or by counting. For example, the data collected by measuring the heights of students are quantitative in nature. The data collected by counting the ages of students in years are also quantitative.

WORKED Example 1

State whether the following variables are categorical or quantitative in nature.

- a** The value of sales recorded at each branch of a fast-food outlet
- b** The breeds of dog that appear at a dog show

THINK

- a** The value of sales at each branch can be measured.
- b** The breeds of dog at a show cannot be measured.

WRITE

- a** The value of sales is quantitative.
- b** The breeds of dog is categorical.

There are two types of quantitative data and variables.

1. **Discrete** — These take only exact values, most often integers. For example, the number of children in a family or the marks achieved on a test. (This is discrete even though you may have half marks.)
2. **Continuous** — These can take any value within certain limits. For example, a person's height or the daily temperature are continuous variables because they can be measured to any degree of accuracy.

WORKED Example 2

State whether each of the following pieces of quantitative data are discrete or continuous.

- a The number of people in each car that passes through a tollgate
- b The mass of a baby at birth



THINK

- a
 - 1 The number of people in the car must be a whole number.
 - 2 Give a written answer.
- b
 - 1 A baby's mass can be measured to various degrees of accuracy.
 - 2 Give a written answer.

WRITE

a
The data are quantitative and discrete.

b
The data are quantitative and continuous.

Types of data

Consider Karen's summary of house prices.

- 1 Are the data that Karen has collected categorical or quantitative?
- 2 Are house prices an example of discrete or continuous data?

remember

1. Variables and data can be classified as either:
 - (a) categorical — the data are in categories, or
 - (b) quantitative — the data can be either measured or counted.
2. Quantitative variables and data can be either:
 - (a) discrete — the data can take only certain values, usually whole numbers, or
 - (b) continuous — the data can take any value depending on the degree of accuracy.

EXERCISE 9A**Types of variables and data****WORKED Example 1**

- 1 State whether the variables in each of the following situations would be categorical or quantitative.

- a The number of matches in each box is counted for a large sample of boxes.
- b The sex of respondents to a questionnaire is recorded as either M or F.
- c A fisheries inspector records the lengths of 40 cod.
- d The occurrence of hot, warm, mild and cool weather for each day in January is recorded.
- e The actual temperature for each day in January is recorded.
- f Cinema critics are asked to judge a film by awarding it a rating from one to five stars.

WORKED Example 2

- 2 State whether the quantitative data considered in each of the following situations are discrete or continuous.

- a The heights of 60 tomato plants at a plant nursery
- b The number of jelly beans in each of 50 packets
- c The time taken for each student in a class of six-year-olds to tie their shoelaces
- d The petrol consumption rate of a large sample of cars
- e The IQ (intelligence quotient) of each student in a class

- 3 For each of the following, state if the data are categorical or quantitative. If quantitative, also state if the variables are discrete or continuous.

- a The number of students in each class at your school
- b The teams people support at a football match
- c The brands of peanut butter sold at a supermarket
- d The heights of people in your class
- e The interest rate charged by each bank
- f A person's pulse rate

- 4 An opinion poll was conducted. A thousand people were given the statement 'Euthanasia should be legalised'. Each person was offered five responses: strongly agree, agree, unsure, disagree and strongly disagree. Describe the data type in this example.

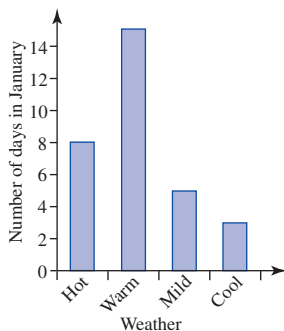
- 5 A teacher marks her students' work with a grade A, B, C, D, or E. Describe the data type used.
- 6 A teacher marks his students' work using a mark out of 100. Describe the data type used.

7 **multiple choice**

The number of people who are using a particular bus service are counted over a two week period. The data formed by this survey would be an example of:

- A categorical data
 B quantitative and discrete data
 C quantitative and continuous data
 D numerical data
 E insufficient information

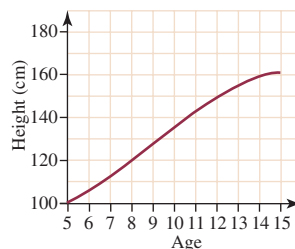
- 8 The following graph shows the number of days of each weather type for the Gold Coast in January.



Describe the data in this example.

- 9 The graph at right shows a girl's height each year for 10 years.

Describe the data in this example.



Collection of data

A common method of collecting data is through a *poll*. A poll is the recording of responses to a set of questions known as a *questionnaire*.

Gallup poll

The most famous poll is named after its founder, the American statistician, George Gallup, who was born in 1901.

Find out about Gallup and his work and how Gallup polls are used today.

The first step in gathering the relevant data for a statistical investigation is to target the population to be investigated. This means identifying the sections of the population for whom the statistical investigation will have relevance.

For example, if investigating the medical needs of a community, we would not conduct our survey at the local fitness club! For such a survey we would survey doctors and other medical personnel, as well as a selection of patients who use the existing facilities.

Identifying the target population

For each of the following statistical investigations, identify the population that you would target for a survey.

- 1 The school 'End of Year' Committee wants to find out the preferred venue, band and meals for the Year 12 farewell.
- 2 The local council wants to know what sporting facilities are needed in the local area.
- 3 A newspaper wants a survey to predict the winner of a forthcoming election.
- 4 A group of people planning to build a preschool would like to know what facilities attract people to a particular preschool.
- 5 A recording label wants to estimate the potential success of a 'grunge' band.

When starting an investigation, we must determine the quantity of data needed for the database. Consider the case of a company calculating the TV ratings. Does the company need to find out what every household is watching? Obviously they do not so they ask a selection of homes to record their TV viewing.

Now consider the case of selecting a band to play at the Year 12 farewell. In this case it is reasonable to ask every Year 12 student their opinion.

Data can be collected in one of two ways:

1. **Census.** In a census an entire population is counted. Australians complete 'The Census' every five years. This is a survey conducted by the Bureau of Statistics of every household in the nation. For the purposes of most statistical investigations, a census is where everyone in the target population is surveyed, such as the Year 12 example above.
2. **Sample.** A sample is a more practical method for conducting most surveys. Only a selection of the target population is surveyed with the results taken to be representative of the whole group. The TV ratings example is one where a sample is used.

WORKED Example 3

In each of the following, state if the information was obtained by census or sample.

- a A school uses the roll to count the number of students absent each day.
- b The television ratings, in which 2000 families complete a survey on what they watch over a one week period.
- c A light globe manufacturer tests every hundredth light globe off the production line.
- d A teacher records the examination marks of her class.

THINK

- a Every student is counted at roll call each morning.
- b Not every family is asked to complete a ratings survey.
- c Not every light globe is tested.
- d The marks of every student are recorded.

WRITE

- a Census
- b Sample
- c Sample
- d Census

Sampling methods

To ensure that the results of your sample are representative of the whole population, the method of sampling is important. There are three main methods of choosing a sample: random sample, stratified sample and systematic sample.

Method 1. Random sample

In a random sample, those to be surveyed are selected by chance. When a random sample is conducted, every person in the target population should have an equal chance of being selected. For example, the names of the people to complete your survey may be drawn from a hat. If this method is used, you should get a good mixture of people in your survey.

Suppose that we are going to survey students in a school. We want a mixture of students and could choose a fixed number of students from each year. Suppose we decide to survey 60 students. We could select 12 from each year, but if we did this the survey would not have the correct proportion of students from each year. For example, 22.5% of the students at this school are in Year 8, but only 20% of the survey participants are in Year 8.

Year	No. of students
Year 8	180
Year 9	190
Year 10	185
Year 11	135
Year 12	110
Total	800



Graphics Calculator **tip!**

Choosing a random sample

Suppose that we are to choose a random sample of 20 students from the population of 800. To choose a random sample each student would be allocated a number between 1 and 800 and the graphics calculator could then be used to make a random choice using the random number generator.

1. To find the random number generator press **MATH**.
2. Press **▶ ▶ ▶** to choose the **PRB** menu.
3. Choose option 5: **randInt**.
4. Select the lower limit, upper limit and number of values, separated by a comma, then close the brackets. To do this press **1, 800, 20**.
5. Use the scroll to see the entire list of numbers.

A scientific calculator will generate random numbers. Your calculator may generate a random integer as does the graphics calculator, or may generate a random decimal between 0 and 1. To generate random integers from this decimal we multiply the decimal by the number the sample is being chosen from (in the example above, 800) and round the result up to the next whole number.



WORKED Example 4

Three students from a school are to be selected to participate in a statewide survey of school students. There are 750 students at the school. To choose the participants, a random decimal generator is used with the results 0.983, 0.911, and 0.421. What are the roll numbers of the students who should be selected?

THINK

- 1 Multiply the results of the random number generator by the size of the population.
- 2 Round up to whole numbers.

WRITE

$$0.983 \times 750 = 737.25$$

$$0.911 \times 750 = 683.25$$

$$0.421 \times 750 = 315.75$$

The 738th, 684th and the 316th people on the roll would be surveyed.

Any other method may not give a truly representative sample. For example, if you survey people in the playground you may:

- have a tendency to ask people you know
- choose an area where a lot of students from a particular year tend to sit
- choose more of one gender than another.

Method 2. Stratified sample

In this type of sample you deliberately choose people to complete your survey who are representative of the whole population. In the school survey you would need to select five *strata* that had the correct proportion of students from each year. For example, if 20% of the school population are in Year 8 then 20% of your sample should be from Year 8.

WORKED Example 5

Adrian is conducting a survey of school students. The table at right shows the number of students in each year of school.

If Adrian decides to survey 50 students, how many should be chosen from each year?

Year	No. of students
Year 8	150
Year 9	180
Year 10	140
Year 11	90
Year 12	80
Total	640

THINK

- 1 Calculate the percentage of the school population in each year.

WRITE

$$\begin{aligned} \text{Year 8} &= \frac{150}{640} \times 100\% \\ &= 23.4\% \end{aligned}$$

$$\begin{aligned} \text{Year 9} &= \frac{180}{640} \times 100\% \\ &= 28.1\% \end{aligned}$$

$$\begin{aligned} \text{Year 10} &= \frac{140}{640} \times 100\% \\ &= 21.8\% \end{aligned}$$

$$\begin{aligned} \text{Year 11} &= \frac{90}{640} \times 100\% \\ &= 14.1\% \end{aligned}$$

$$\begin{aligned} \text{Year 12} &= \frac{80}{640} \times 100\% \\ &= 12.5\% \end{aligned}$$

Continued over page 

THINK

- 2 Calculate the number of students that should be chosen from each year to do the survey.

- 3 Give a written answer.

WRITE

$$\begin{array}{ll} \text{Year 8} = 23.4\% \text{ of } 50 & \text{Year 9} = 28.1\% \text{ of } 50 \\ = 11.7 & = 14.05 \\ = 12 \text{ (nearest whole no.)} & = 14 \text{ (nearest whole no.)} \end{array}$$

$$\begin{array}{ll} \text{Year 10} = 21.8\% \text{ of } 50 & \text{Year 11} = 14.1\% \text{ of } 50 \\ = 10.9 & = 7.05 \\ = 11 \text{ (nearest whole no.)} & = 7 \text{ (nearest whole no.)} \end{array}$$

$$\begin{array}{l} \text{Year 12} = 12.5\% \text{ of } 50 \\ = 6.25 \\ = 6 \text{ (nearest whole no.)} \end{array}$$

Adrian should choose twelve Year 8 students, fourteen Year 9 students, eleven Year 10 students, seven Year 11 students and six Year 12 students.

**Method 3. Systematic sample**

Systematic sampling is where those chosen for the sample are chosen in a systematic or organised way. This method is mostly used in quality control situations. For example, suppose that the quality and strength of sports shoes is being tested. The quality control department may test every 20th pair of shoes that comes off the production line. In doing a survey, every 20th person on the school roll may be surveyed.

remember

- Before beginning a statistical investigation it is important to identify the target population.
- The survey can be conducted either by:
 - Census — the entire target population is surveyed, or
 - Sample — a selection is surveyed such that those selected are representative of the entire target population.
- There are three methods for selecting a sample.

Method 1. Random sample — chance is the only factor in deciding who is surveyed. This is best done using a random number generator.

Method 2. Stratified sample — those sampled are chosen in proportion to the entire population.

Method 3. Systematic sample — a system is used to choose those who are to be in the sample.

EXERCISE 9B**Collection of data****WORKED Example****3**

- 1 A school conducts an election for a new school captain. Every teacher and student in the school votes. Is this an example of a census or a sample? Explain your answer.
- 2 A survey is conducted by a council to see what sporting facilities the community needs. If 500 people who live in the community are surveyed, is this an example of a census or a sample?
- 3 For each of the following surveys, state whether a census or a sample has been used.
 - a Two hundred people in a shopping centre are asked to nominate the supermarket where they do most of their grocery shopping.
 - b To find the most popular new car on the road, 500 new car buyers are asked what make and model car they purchased.
 - c To find the most popular new car on the road, the make and model of every new car registered are recorded.
 - d To find the average mark in the mathematics half-yearly exam, every student's mark is recorded.
 - e To test the quality of tyres on a production line, every 100th tyre is road tested.
- 4 For each of the following, recommend whether you would use a census or a sample to obtain the results.
 - a To find the most watched television program on Monday night at 7:30 pm
 - b To find the number of cars sold during a period of one year
 - c To find the number of cars that pass through the tollgates on the Sydney Harbour Bridge each day
 - d To find the percentage of computers produced by a company that are defective
- 5 An opinion poll is conducted to try to predict the outcome of an election. Two thousand people are telephoned and asked about their voting intention. Is this an example of a census or a sample?

WORKED Example**4**

- 6 A factory has 500 employees. Each employee has an employee number between 1 and 500. Five employees are selected to participate in an Occupational Health and Safety survey. To choose the participants, a random number generator is used. The results are 0.326, 0.352, 0.762, 0.989 and 0.018. What are the employee numbers of those to participate in the survey?
- 7 A school has 837 students. A survey of 10 students in the school is to be conducted. A random number generator is used to select the participants. If the random numbers chosen are:
0.988 0.251 0.498 0.661 0.247 0.031 0.967 0.932 0.229 0.443
what are the roll numbers of the students who should be selected?
- 8 A survey is to be conducted of 20 out of 50 000 people in a country town. Those selected are to be chosen using a random number generator.
 - a Use your calculator to generate 20 random numbers.
 - b Calculate the electoral roll numbers of the people who should be chosen for the survey.

- 9 For each of the following, state whether the sample used is an example of random, stratified or systematic sampling.
- a Every 10th tyre coming off a production line is tested for quality.
 - b A company employs 300 men and 450 women. The sample of employees chosen for a survey contains 20 men and 30 women.
 - c The police breathalyse the driver of every red car.
 - d The names of the participants in a survey are drawn from a hat.
 - e Fans at a football match fill in a questionnaire. The ground contains 8000 grandstand seats and 20 000 general admission seats. The questionnaire is then given to 40 people in the grandstand and 100 people who paid for a general admission seat.

10 multiple choice

Which of the following is an example of a systematic sample?

- A The first 20 students who arrive at school each day participate in the survey.
- B Twenty students to participate in the survey are chosen by a random number generator.
- C Twenty students to participate in the survey are selected in proportion to the number of students in each school year.
- D Ten boys and 10 girls are chosen to participate in the survey.

11 multiple choice

Which of the following statistical investigations would be practical to complete by census?

- A A newspaper wants to know public opinion on a political issue.
- B A local council wants to know if a skateboard ramp would be popular with young people in the area.
- C An author wants a cricket player's statistics for a book being written.
- D An advertising agency wants to know the most watched program on television.

WORKED
Example

5

- 12 The table below shows the number of students in each year at a school.

Year	No. of students
8	110
9	90
10	80
11	70
12	50
Total	400

If a survey is to be given to 40 students at the school, how many from each Year should be chosen if a stratified sample is used?

- 13 A company employs 300 men and 200 women. If a survey of 60 employees using a stratified sample is completed, how many people of each gender participated?



- 14** The table below shows the age and sex of the staff of a corporation.

Age	Male	Female
20–29	61	44
30–39	40	50
40–49	74	16
50–59	5	10

A survey of 50 employees is to be done. Using a stratified survey, suggest the breakdown of people to participate in terms of age and gender.



Census or sample?

For each of the following statistical investigations, state whether you would gather data using a census or sample. For those for which you would use a sample, state the best method for selecting the sample.

- 1** A company wants to test the life of its batteries.
- 2** A sporting club wants to elect a new club president.
- 3** A market research company wants to determine the most popular brand of toothpaste.
- 4** A theme park wants to know from which State and suburb its visitors come.
- 5** A Gallup poll is conducted to determine the preferred prime minister.

Bias

A high school is having a disco and the organisers expect that most students will attend. A new DJ is employed to run the disco. To gain information about what music should be played, he conducts a survey. The DJ does not have time to complete a school census and so he selects a sample of 10 students.

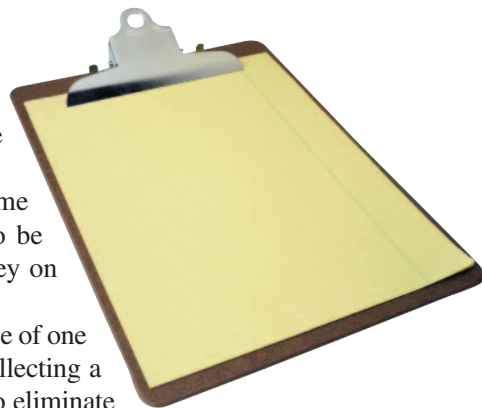
Now suppose that the DJ visited the school basketball courts at lunchtime and surveyed the ten Year 11 boys who were playing at that time. What sort of results would

you expect the DJ to get? Would these results be representative of the whole school's taste in music?

It is unlikely that the Year 11 boys would like the same music as the Year 7 girls. The DJ's results are said to be biased because the sample chosen to conduct the survey on was not representative of the whole school population.

Bias occurs when the sample chosen is more likely to be of one opinion than representing the total population. When collecting a sample, methods for selecting the sample are designed to eliminate bias. Consider the above example with regard to the following points.

- A random sample would most probably have selected a mixed group of students to survey. It is unlikely that a random number generator would have selected 10 students of the same gender in the same year.
- A stratified sample would ensure that a mix of boys and girls from all years was chosen.



Biased sampling

Discuss the problem caused by each of the following biased samples.

- 1 A survey is to be conducted to decide the most popular sport in a local community. A sample of 100 people was questioned at a local football match.
- 2 A music store situated in a shopping centre wants to know the type of music that it should stock. A sample of 100 people was surveyed. The sample was taken from people who passed by the store between 10 and 11 am on a Tuesday.
- 3 A newspaper conducting a Gallup poll on an election took a sample of 1000 people from one Brisbane suburb.

Women and work

A class has been given the task of conducting a statistical analysis that has the title 'Women and the Australian workforce'.

Julie approaches the assignment by surveying 1000 women to find out what percentage are engaged in full-time work.

Ricardo decides to survey 1000 full-time workers to find out what percentage are women.

- 1 Discuss the way in which each student should select his or her sample.
- 2 Is either approach more susceptible to bias?
- 3 Which of the following questions would it be easier to answer?
 - a What percentage of women are engaged in full-time work?
 - b What percentage of the full-time workforce are women?

Explain your answer with reference to your answer to questions 1 and 2.
- 4 Are the answers to questions a and b the same? Explain your answer.

Cost of a house

Remember Karen at the real estate agency? She collected information on the prices of houses for sale through the real estate agency where she works.

- 1 Are the data collected an example of a census or a sample? If they are a sample, describe the type of sample that has been taken.
- 2 Are Karen's data subject to any bias? Explain.

Displaying data

Once a data set has been collected it can be displayed in tabular and graphical form, for various purposes. The type of display chosen depends on the type of data that are being represented.

Stem plots

A *stem-and-leaf plot*, or *stem plot* for short, is a way of displaying a set of data. It is best suited to data which contain up to about 50 observations (or records).

The following stem plot shows the ages of people attending an advanced computer class.

The ages of the members of the class are
16, 22, 22, 23, 30, 32, 34, 36, 42, 43, 46, 47, 53, 57 and 61.

A stem plot is constructed by breaking the numerals of a record into two parts — the stem, which in this case is the first digit, and the leaf, which is always the last digit.

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

WORKED example 6

The number of cars sold in a week at a large car dealership over a 20-week period is given below.

16 12 8 7 26 32 15 51 29 45
19 11 6 15 32 18 43 31 23 23

Construct a stem plot to display the number of cars sold in a week at the dealership.

THINK

- In this example the observations are one- or two-digit numbers and so the stems will be the digits referring to the 'tens', and the leaf part will be the digits referring to the units.
Work out the lowest and highest numbers in the data in order to determine what the stems will be.
- Before we construct an ordered stem plot, construct an unordered stem plot by listing the leaf digits in the order they appear in the data.

- Now rearrange the leaf digits in numerical order to create an ordered stem plot.
Include a key so that the data can be understood by anyone viewing the stem plot.

WRITE

Lowest number = 6

Highest number = 51

Use stems from 0–5.

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

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

Key: 2|3 = 23 cars

WORKED Example 7

The masses (in kilograms) of the members of an Under-17 football squad are given below.

70.3 65.1 72.9 66.9 68.6 69.6 70.8
 72.4 74.1 75.3 75.6 69.7 66.2 71.2
 68.3 69.7 71.3 68.3 70.5 72.4 71.8

Display the data in a stem plot.

THINK

- ① In this case the observations contain 3 digits. The last digit always becomes the leaf and so in this case the digit referring to the tenths becomes the leaf and the two preceding digits become the stem.

Work out the lowest and highest numbers in the data in order to determine what the stems will be.

- ② Construct an unordered stem plot. Note that the decimal points are omitted since we are aiming to present a quick visual summary of data.

- ③ Construct an ordered stem plot. Provide a key.

WRITE

Lowest number = 65.1
 Highest number = 75.6
 Use stems from 65–75.

Stem	Leaf
65	1
66	9 2
67	
68	6 3 3
69	6 7 7
70	3 8 5
71	2 3 8
72	9 4 4
73	
74	1
75	3 6

Stem	Leaf
65	1
66	2 9
67	
68	3 3 6
69	6 7 7
70	3 5 8
71	2 3 8
72	4 4 9
73	
74	1
75	3 6

Key: 74|1 = 74.1 kg

Sometimes data which are very bunched make it difficult to get a clear idea about the data variation. To overcome the problem, we can split the stems. Stems can be split into halves or fifths.

WORKED Example 8

A set of golf scores for a group of professional golfers trialling a new 18-hole golf course is shown on the following stem plot.

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

Key: 6|1 = 61

Produce another stem plot for these data by splitting the stems into:

- a** halves **b** fifths.

THINK

- a** By splitting the stem 6 into halves, any leaf digits in the range 0–4 appear next to the first 6, and any leaf digits in the range 5–9 appear next to the second 6. Likewise for the stem 7.

- b** Alternatively, to split the stems into fifths, each stem would appear 5 times. Any 0s or 1s are recorded next to the first 6. Any 2s or 3s are recorded next to the second 6. Any 4s or 5s are recorded next to the third 6. Any 6s or 7s are recorded next to the fourth 6 and finally any 8s or 9s are recorded next to the fifth 6. This process would be repeated for those observations with a stem of 7.

WRITE

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

Key: 6|1 = 61

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

Key: 6|1 = 61

remember

1. A stem-and-leaf plot is a useful way of displaying data containing up to about 50 observations (or records).
2. A stem plot is constructed by breaking the numerals of a record into two parts, a 'stem' and a 'leaf'. The last digit is always the leaf and any preceding digits form the stem.
3. When asked to represent data using a stem-and-leaf plot, you should always assume that the plot will be drawn with the data ordered.
4. If data are bunched then it may be useful to break the stems into halves or even fifths.

EXERCISE 9C

Stem plots

- 1 In each of the following, write down all the pieces of data shown on the stem plot. The key used for each stem plot is $3|2 = 32$.

a

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

b

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

c

Stem	Leaf
10	1 2
11	5 8
12	2 3 3
13	6 6 7
14	1 3 4
15	5 5 6 7

d

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

e

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



- 2 The money (to the nearest dollar) earned each week by a busker over an 18-week period is shown below. Construct a stem plot for the busker's weekly earnings.

5 19 11 27 23 35 18 42 29
31 52 43 37 41 39 45 32 36

- 3 The ages of those attending an embroidery class are given below. Construct a stem plot for these data.

39 68 51 57 63 51 37 42
63 49 52 61 58 59 49 53

- 4 The number of dogs brought into a dog refuge each week over a 20-week period is given below. Construct a stem plot for these data.

28 18 9 16 8 30 26 41 43 54
32 26 29 30 21 35 45 23 19 27

5 multiple choice

The observations shown on the stem plot at right are:

- A** 4 10 27 28 29 31 34 36 41
B 14 10 27 28 29 29 31 34 36 41 41
C 4 22 27 28 29 29 30 31 34 36 41 41
D 14 22 27 28 29 30 30 31 34 36 41 41
E 4 2 27 28 29 29 30 31 34 36 41

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

Key: $2|5 = 25$

- 6 The ages of the mothers of a class of children attending an inner city kindergarten are given below. Construct a stem plot for these data.

32 37 29 23 29 32 25 38 39 32
28 30 33 34 28 35 29 33 30

- 7 The number of people attending a Neighbourhood Watch committee meeting each fortnight for a year is given below. Construct a stem plot to display these data.

14 17 19 21 18 23 16 11 25 23 20
13 15 19 21 23 29 22 18 21 19 18
19 22 20

- 8 The number of hit outs made by each of the principal ruckmen in each of the AFL teams for Round 11 is recorded below. Construct a stem plot to display these data.

Team	Number of hit outs
Collingwood	20
Bulldogs	34
Kangaroos	29
Port Adelaide	24
Geelong	21
Sydney	31
Melbourne	29
Brisbane	25

Team	Number of hit outs
Adelaide	32
St Kilda	34
Essendon	31
Carlton	26
West Coast	29
Fremantle	22
Hawthorn	33
Richmond	28

WORKED Example

7

- 9 The heights of members of a squad of basketballers are given at right in metres. Construct a stem plot for these data.

1.96 1.85 2.03 2.21 2.17 1.89
1.99 1.87 1.95 2.03 2.09
2.05 2.01 1.96 1.97 1.91

- 10 The 2001 median house price of a number of Brisbane suburbs is given below. Construct a stem plot for these data.

Suburb	\$(000)
Auchenflower	233
Bulimba	217
Balmoral	203
Cannon Hill	298
Carrara	246
Coorparoo	210
Doomben	205
Eagle Farm	202
Fairfield	212
Holland Park	242

Suburb	\$(000)
Indooroopilly	255
Rosalie	221
Spring Hill	252
Mt Gravatt	228
Nudgee	290
Paddington	285
Sandgate	244
Sth Brisbane	290
Woolloongabba	283

WORKED Example

8

- 11 The data at right give the head circumference (to the nearest cm) of 16 four-year-old girls.

48 49 47 52 51 50 49 48
50 50 53 52 43 47 49 50

Construct a stem plot for head circumference, using:

- a the stems 4 and 5 b the stems 4 and 5 split into halves
c the stems 4 and 5 split into fifths.

- 12 A random sample of 20 screws is taken and the length of each is recorded to the nearest millimetre (at right).

23 15 18 17 17 19 22
19 20 16 20 21 19 23
17 19 21 23 20 21

Construct a stem plot for screw length using:

- a the stems 1 and 2 b the stems 1 and 2 split into halves
c the stems 1 and 2 split into fifths.

- 13 The number of seconds for which 12 Grade-2 children can hold their breath under water is given below.

8.2 9.2 8.1 8.5 9.3 8.9 8.9 9.5 8.9 9.0 9.1 9.7

Construct a stem plot for holding breath using:

- a the stems 8 and 9 b the stems 8 and 9 split into halves
c the stems 8 and 9 split into fifths.

Frequency histograms and bar charts

Frequency histograms and bar charts display data in graphical form.

Frequency histograms

A histogram is a useful way of displaying large data sets (say, over 50 observations). The vertical axis on the histogram displays the frequency and the horizontal axis displays class intervals of the variable (for example height, income etc.).

When data are given in raw form — that is, just as a list of figures in no particular order — it is helpful to first construct a frequency table.

WORKED Example 9

The data below show the distribution of masses (in kilograms) of 60 students in Year 7 at Northwood State High School. Construct a frequency histogram to display the data more clearly.

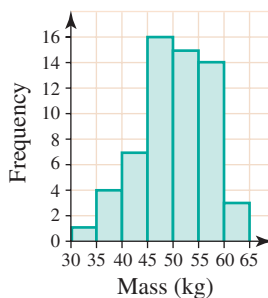
45.7	34.2	56.3	38.7	52.4	45.7	48.2	52.1	58.7	62.3
45.8	52.4	60.2	48.5	54.3	39.8	36.2	54.3	39.7	46.3
45.9	52.3	44.2	49.6	48.6	42.5	47.2	51.3	43.1	52.4
48.2	51.8	53.8	56.9	53.7	42.9	46.7	51.9	56.2	61.2
48.3	45.7	43.5	43.8	58.7	59.2	58.7	54.6	43.0	48.2
48.4	56.8	57.2	58.3	57.6	53.2	53.1	58.7	56.3	58.3

THINK

- First construct a frequency table.
The lowest data value is 34.2 and the highest is 62.3. Divide the data into class intervals. If we started the first class interval at, say, 30 kg and ended the last class interval at 65 kg, we would have a range of 35. If each interval was 5 kg, we would then have 7 intervals which is a reasonable number of class intervals. While there are no set rules about how many intervals there should be, somewhere between about 5 and 15 class intervals is usual. So, in this example, we would have class intervals of 30–34.9 kg, 35–39.9 kg, 40–44.9 kg and so on. Count how many observations fall into each of the intervals and record these in a table.
- Check that the frequency column totals 60. The data are in a much clearer form now.
- A histogram can be constructed.

WRITE

Class interval	Frequency
30–34.9	1
35–39.9	4
40–44.9	7
45–49.9	16
50–54.9	15
55–59.9	14
60–64.9	3
Total	60



WORKED Example 10

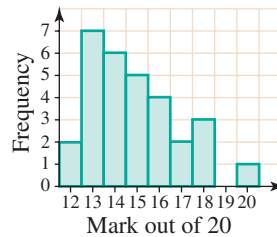
The marks out of 20 received by 30 students for a book-review assignment are given in the frequency table below.

Mark	12	13	14	15	16	17	18	19	20
Frequency	2	7	6	5	4	2	3	0	1

Display these data on a histogram.

THINK

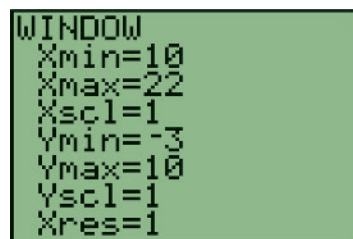
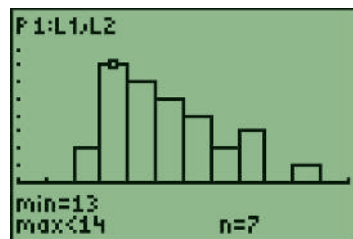
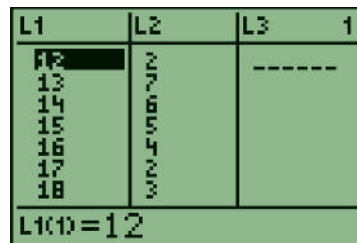
In this case we are dealing with integer values. Since the horizontal axis should show a class interval, we extend the base of each of the columns on the histogram halfway below each score and halfway above it.

WRITE**WORKED Example 11**

Construct a histogram using the data in worked example 10 and a graphics calculator.

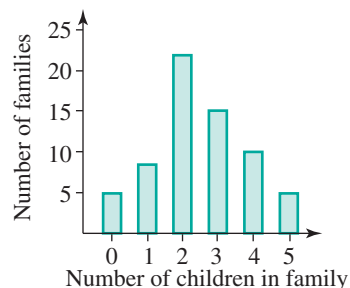
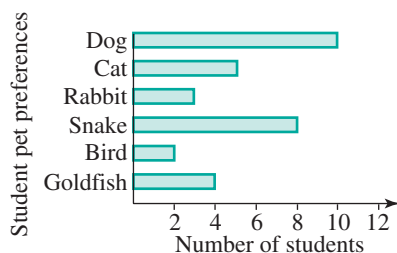
THINK

- 1 Enter the data.
 - (a) Clear any previous equations.
 - (b) Press **(Y=)** and clear any functions.
 - (c) Press **(STAT)**, select **1:Edit** and press **(ENTER)**.
 - (d) Enter the marks in **L1** and the frequency in **L2**.
- 2 Set up the calculator for graphing.
 - (a) Press **(2nd) [STAT PLOT]** and select **1:Plot1**. Press **(ENTER)**.
 - (b) Select **On** and press **(ENTER)**.
 - (c) Select the type of graph required. The histogram is the third along on the top row.
 - (d) At **Xlist** type in **L1** (press **(2nd) [L1]**).
 - (e) At **Freq** type in **L2** (press **(2nd) [L1]**).
 - (f) Press **(ZOOM)** and highlight **9:Zoom Stat**; press **(ENTER)**.
 - (g) If not all of the histogram is shown, press **(WINDOW)** and reset the *x*- and *y*-range and step values.

DISPLAY

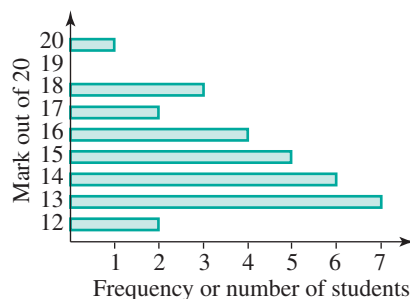
Bar charts

A *bar chart* is similar to a histogram. However, it consists of bars of equal width separated by small, equal spaces and may be arranged either *horizontally* or *vertically*.



In bar charts the frequency is graphed against a variable as shown in both figures above.

The variable may or may not be numerical. However, in this chapter we consider only numerical variables. The numerical variable should take discrete values; that is, it should take only certain values (such as whole hours or number of people) rather than being continuous (such as the height of people) which could take any value within a range. This is because the scale is broken by the gaps between the bars. The numerical values are generally close together and have little spread, like consecutive years.



The bar chart above right represents the data presented in worked example 5. Of course, it could have been drawn with vertical bars (columns).

Segmented bar charts

A segmented (divided) bar chart is a single bar which is used to represent all the data being studied. It is divided into segments, each segment representing a particular group of the data. Generally, the information is presented as percentages and so the total bar length represents 100% of the data.

Consider the following table, showing fatal road accidents in Australia.

ROAD TRAFFIC ACCIDENTS INVOLVING FATALITIES									
Accidents involving fatalities									
Year	NSW	Vic.	Qld	SA	WA	Tas.	NT	ACT	Aust.
1991	585	435	362	166	187	65	60	16	1876
1992	578	365	364	142	171	56	42	18	1734
1993	518	381	357	191	190	47	40	11	1735
1994	557	346	367	145	195	51	36	15	1712
1995	563	371	408	163	194	53	56	14	1822
1996	544	383	338	162	220	53	58	17	1775

ROAD TRAFFIC ACCIDENTS INVOLVING FATALITIES									
Persons killed									
Year	NSW	Vic.	Qld	SA	WA	Tas.	NT	ACT	Aust.
1991	663	503	395	184	207	75	67	17	2113
1992	649	396	416	165	200	74	54	20	1974
1993	581	435	396	218	209	58	44	12	1953
1994	647	378	422	159	211	59	41	17	1934
1995	620	418	456	181	209	57	61	15	2017
1996	587	418	385	181	247	64	72	23	1977

Source: Federal Office of Road Safety, Road Fatalities Australia, 1996. (From ABS Yearbook, 1998.)

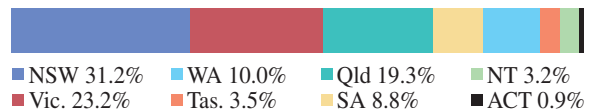
It is appropriate to represent the number of accidents involving fatalities in all States and territories during 1991 as a segmented bar chart.

Firstly, using the data on page 392, we convert each State's proportion of accidents out of the total to a percentage.

State	Number of accidents	Percentage
NSW	585	$585 \div 1876 \times 100\% = 31.2\%$
Vic.	435	$435 \div 1876 \times 100\% = 23.2\%$
Qld	362	$362 \div 1876 \times 100\% = 19.3\%$
SA	166	$166 \div 1876 \times 100\% = 8.8\%$
WA	187	$187 \div 1876 \times 100\% = 10.0\%$
Tas.	65	$65 \div 1876 \times 100\% = 3.5\%$
NT	60	$60 \div 1876 \times 100\% = 3.2\%$
ACT	16	$16 \div 1876 \times 100\% = 0.9\%$

The segmented bar chart is drawn to scale. An appropriate scale would be constructed by drawing the total bar 10 cm long, so that 1 mm represents 1%. That is, NSW's accidents would be represented by a segment of 31.2 mm, Victoria's by a segment of 23.2 mm and so on.

Each segment is then labelled directly, or a key may be used.



remember

1. On a frequency histogram the vertical axis displays the frequency and the horizontal axis displays the class intervals.
2. Data given in raw form should be summarised first in a frequency table.

EXERCISE 9D

Frequency histograms and bar charts

WORKED
Example 9

- 1 Construct a frequency table for each of the following sets of data.

- a 3 4 4 5 5 6 7 7 7 8 8 9 9 10 10 12
 b 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.6 5.2 3.6 2.5 4.3 2.5 3.7 4.5 6.3 1.3
 c 11 13 15 15 16 18 20 21 22 21 18 19 20 16 18 20 16 10 23 24 25
 27 28 30 35 28 27 26 29 30 31 24 28 29 20 30 32 33 29 30 31 33 34
 d 0.4 0.5 0.7 0.8 0.8 0.9 1.0 1.1 1.2 1.0 1.3 0.4 0.3 0.9 0.6

WORKED
Example 9, 10

- 2 Using the frequency tables from question 1, construct a histogram for each set of data.

WORKED
Example 11

- 3 Using a graphics calculator, construct a histogram for each of the sets of data given in question 1. Compare this histogram with the one drawn for question 2.

- 4 Using the frequency table from question 1a, construct a bar chart for the data.

- 5 The data at right represent the number of hours each week that 40 teenagers spent on household chores. Represent these data by a bar chart.

2	5	2	0	8	7	8	5	1	0
2	1	8	0	4	2	2	9	8	5
7	5	4	2	1	2	9	8	1	2
8	5	8	10	0	3	4	5	2	8

Segmented bar chart

Compare the proportion of fatal accidents in the States and territories during the period 1996 to 2001 by drawing a segmented bar chart similar to that on page 393 for each of the 5 years since 1991.

- 1 What conclusions can be made from your charts?
- 2 Would you say that this presentation is misleading in any way?
- 3 Give reasons for your answers.

Looking at cost

Let's return to house prices. Look at the table of data that Karen has collected.

Price range	Number of houses
\$75 000–\$100 000	1
\$100 000–\$125 000	5
\$125 000–\$150 000	7
\$150 000–\$175 000	6
\$175 000–\$200 000	11
\$200 000–\$225 000	14
\$225 000–\$250 000	9
\$250 000–\$275 000	4
\$275 000–\$300 000	0
\$300 000–\$325 000	1
\$325 000–\$350 000	0
\$350 000–\$375 000	0
\$375 000–\$400 000	2

Present these data in an appropriate type of graph.

Using a database

If you have access to a computer database, collect the following information on each property displayed in the real estate agents window in your local area.

Category: _____ (e.g. house, unit, vacant land, business)

Area: _____

Number of bedrooms: _____

Number of bathrooms: _____

Special features: _____

Price: _____

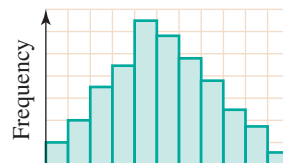
Enter your data into your database, and experiment to determine the different ways that you can sort, select and display the data.

Describing the shape of stem plots and histograms

Symmetric distributions

The data shown in the histogram at right can be described as *symmetric*.

There is a single peak and the data trail off on both sides of this peak in roughly the same fashion.



Similarly in the stem plot at right, the distribution of the data could be described as symmetric.

The single peak for these data occur at the stem 3. On either side of the peak, the number of observations reduces in approximately matching fashion.

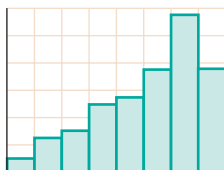
Stem	Leaf
0	7
1	2 3
2	2 4 5 7 9
3	0 2 3 6 8 8
4	4 7 8 9 9
5	2 7 8
6	1 3

Skewed distributions

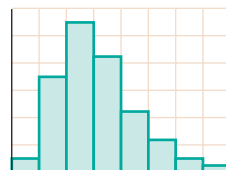
Each of the histograms below show examples of skewed distributions.

The figure below left shows data which are *negatively skewed*. The data in this case peak to the right and trail off to the left.

The figure below right shows *positively skewed* data. The data in this case peak to the left and trail off to the right.



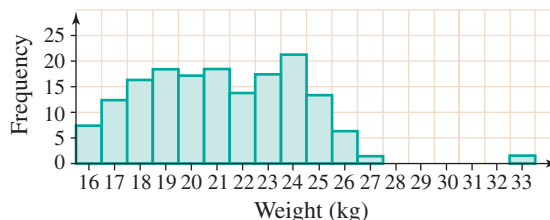
Negatively skewed distribution



Positively skewed distribution

Outliers

When one observation lies well away from other observations in a set, we call it an *outlier*. Sometimes an outlier occurs because data have been incorrectly obtained or misread. For example, below we see a histogram showing the weights of a group of 5-year-old boys.



The outlier, 33, may have occurred because a weight was incorrectly recorded as 33 rather than 23 or perhaps there was a little boy in this group who, for some medical reason, weighed a lot more than his counterparts. When an outlier occurs, the reasons for its existence should be checked.

WORKED Example 12

The ages of a group of people who were taking out their first home loan are shown below.

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

Key: 1|9 = 19 years

Describe the shape of the distribution of these data and comment on the existence of any outliers.

**THINK**

- 1 Check whether the distribution is symmetric or skewed. The peak of the data occurs at the stem 2. The data trail off as the stems increase in value. This seems reasonable since most people would take out a home loan early in life to give themselves time to pay it off.
- 2 Check whether there is an outlier.

Note: It is unusual to have a 77-year-old person taking out a first home loan. Maybe this observation was incorrectly recorded or maybe exceptional circumstances apply in this case.

WRITE

The data are positively skewed.

The observation 77 is an outlier.

remember

1. When data are displayed in a histogram or a stem plot, we say that their distribution is:
 - (a) symmetric if there is a single peak and the data trail off on either side of this peak in roughly the same fashion
 - (b) negatively skewed if the data peak to the right and trail off to the left.
 - (c) positively skewed if the data peak to the left and trail off to the right.
2. An outlier is an observation that lies well away from the rest of the data.

EXERCISE 9E

Describing the shape of stem plots and histograms

WORKED
Example

12

- 1 For each of the following stem plots, describe the shape of the distribution of the data and comment on the existence of any outliers.

a

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

Key: 1|2 = 12

b

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

Key: 2|6 = 2.6

c

Stem	Leaf
2	3 5 5 6 7 8 9 9
3	0 2 2 3 4 6 6 7 8 8
4	2 2 4 5 6 6 6 7 9
5	0 3 3 5 6
6	2 4
7	5 9
8	2
9	7
10	
11	
12	
13	5

Key: 10|4 = 104

d

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

Key: 2|7 = 27

e

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

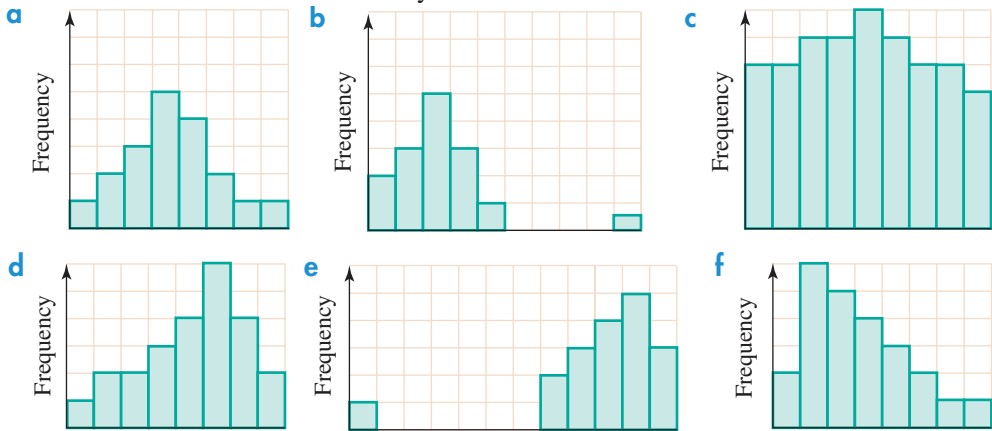
Key: 4|3 = 0.43

f

Stem	Leaf
60	2 5 8
61	1 3 3 6 7 8 9
62	0 1 2 4 6 7 8 8 9
63	2 2 4 5 7 8
64	3 6 7
65	4 5 8
66	3 5
67	4

Key: 62|3 = 623

- 2 For each of the following histograms, describe the shape of the distribution of the data and comment on the existence of any outliers.



3 **multiple choice**

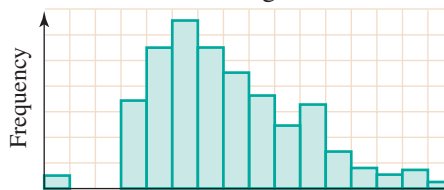
The distribution of the data shown in this stem plot could be described as:

- A** negatively skewed
- B** negatively skewed with one outlier
- C** positively skewed
- D** positively skewed with one outlier
- E** symmetric.

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

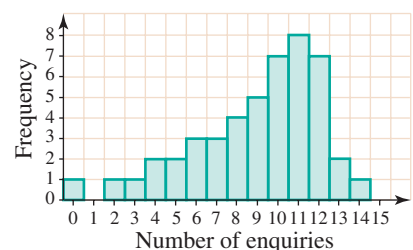
4 **multiple choice**

The distribution of the data shown in this histogram could be described as:



- A** negatively skewed
- B** negatively skewed with one outlier
- C** positively skewed
- D** positively skewed with one outlier
- E** symmetric.

- 5 The average number of product enquiries per day received by a group of small businesses who advertised in the *Yellow Pages* telephone directory is given at right. Describe the shape of the distribution of these data and comment on the existence of any outliers.



- 6 The number of nights per month spent interstate by a group of flight attendants is shown on the stem plot at right. Describe the shape of the distribution of these data and explain what this tells us about the number of nights per month spent interstate by this group of flight attendants.

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

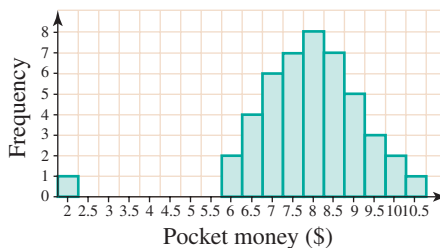
Key: 1|4 = 14 nights

- 7 The mass (to the nearest kilogram) of each dog at a dog obedience school is shown on the stem plot at right.
- Describe the shape of the distribution of these data and comment on the existence of any outliers.
 - What does this information tell us about this group of dogs?

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

Key: 0|4 = 4 kg

- 8 The amount of pocket money (to the nearest 50 cents) received each week by students in a Grade-6 class is illustrated in this histogram.



- Describe the shape of the distribution of these data and comment on the existence of any outliers.
 - What conclusions can you reach about the amount of pocket money received weekly by this group of students?
- 9 The number of hours of exercise completed each week by a group of employees at a company is shown on the stem plot at right.
- Describe the shape of the distribution of these data and comment on the existence of any outliers.
 - What does this tell us about the number of hours of exercise completed weekly by the employees in this company?

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

Key: 0|1 = 1 hr

Cumulative data

Cumulative frequency

It is often useful to consider the number of data points that are less than or equal to a particular score. In such cases it is helpful to include a *cumulative frequency* column on the frequency distribution table.

The *cumulative frequency* is the number of records equal to and less than a particular score. The cumulative frequency of a particular score is obtained by adding the frequency of that score to the sum of the frequencies of all preceding scores.

In other words, if all the data were sorted in order of size, the cumulative frequency would give a 'running total' of the number of observations up to each score. Consider the following data that show the heights (in cm) of 40 girls who are competing in trials to form a basketball squad.

181	191	185	175	192	186	188	182	179	172
186	188	193	198	182	175	176	188	180	191
185	186	183	180	179	175	180	188	190	193
172	179	188	183	186	188	182	193	194	181

A frequency distribution table including a cumulative column could be drawn as follows:

Height	Frequency	Cumulative frequency
170–	3	3
175–	6	9
180–	12	21
185–	10	31
190–	8	39
195–	1	40

The cumulative frequency column in this case records the number of girls who had a height as indicated by the particular group or those preceeding it. For example the figure 31 in the cumulative frequency column can be interpreted as: 'There were 31 girls who had a height of 189 cm or less'. The figure 31 in this example was found by totalling $3 + 6 + 12 + 10$. Note that the final number in the cumulative frequency column should always equal the total number of scores.

Ogives

An *ogive* (also called a *cumulative frequency polygon*) is a line graph of the cumulative frequency results.

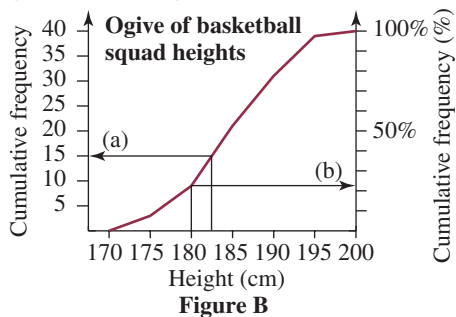
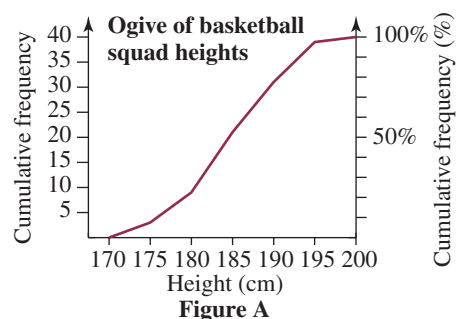
An ogive is appropriate only for displaying grouped data. The graph is started on the horizontal axis at a point corresponding to the lowest possible score in the smallest group. In the case of the basketball squad data the graph will start at 170. The ogive is then drawn by plotting the value of the cumulative frequency of each group against each group end point. For the basketball squad data the points which form the rest of the ogive will be (175, 3), (180, 9), (185, 21), (190, 31), (195, 39) and (200, 40).

The 'S' shape of this ogive is typical of most sets of data.

A percentage axis was added on the right-hand side of the ogive. A percentage axis can be added by ruling a vertical line from the end point of the ogive to the horizontal axis. The end point is labelled 100% and then the axis is scaled from 0–100% appropriately. A percentage axis is not an essential feature of an ogive but it will help to answer a lot of questions like some of the following:

1. How many girls had a height of less than 182 cm? (Find 182 on the height axis, rule a vertical line to the ogive then horizontally to the frequency axis. See line (a) in figure B.) Answer: About 15 girls would have a height of less than 182 cm.
2. What percentage of girls had a height of less than 180 cm? (Find 180 on the height axis then rule a vertical line to the ogive then horizontally to the percentage axis. See line (b) in figure B.) Answer: About 22% of girls had a height of less than 180 cm.
3. What percentage of girls had a height more than 180 cm? Answer: About 22% had a height of less than 180 cm so there must be 78% with a height of more than 180 cm.

Note that when you are interpreting ogives, less than ($<$) and less than or equal to ($\leq$) make no difference to the approach that we take to solving a question or its answer. Example 1 above would have had exactly the same answer if the question had been: ‘How many girls had a height of less than or equal to 182 cm?’



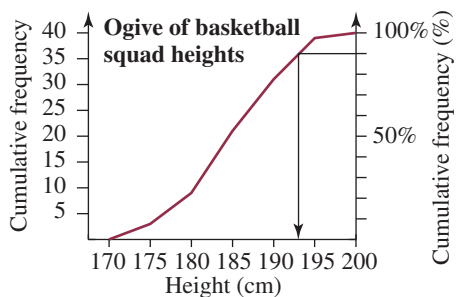
Percentiles



A percentile is the score below which a particular percentage of the distribution of data lies.

For example the 90th percentile is the score below which 90% of the data lies. In the case of the basketball squad data the 90th percentile could be found by finding 90% on the percentage axis, going horizontally to the ogive, then vertically down to the ‘height axis’.

The 90th percentile would be about 193 cm. This could be interpreted as: ‘90% of the girls would have a height of 193 cm or less’.



WORKED Example 13

Forty sample pieces of rope are tested in an effort to determine their breaking strain. The maximum load that could be attached to each was recorded.

Breaking strain (kg)	Frequency
40–	2
45–	6
50–	8
55–	10

Breaking strain (kg)	Frequency
60–	9
65–	4
70–	1

- Add a cumulative frequency column to the table.
- Represent the data using an ogive.
- What number of sample pieces broke under a strain of less than 52 kg?
- Find the 75th percentile and write a sentence to explain what it means.
- The manufacturer of the rope wishes to label the rope with an appropriate breaking strain. What should the rope be rated at if the manufacturer wants 90% of all ropes to be at least as strong as the labelled rate?

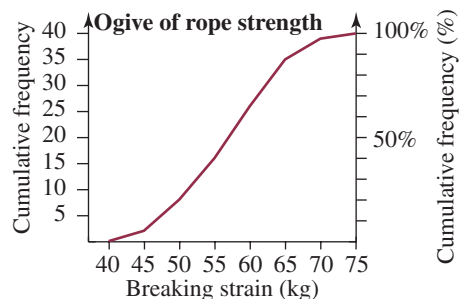
THINK

- The cumulative frequency column is a 'running total' of the amounts in the frequency column.
The final entry in the cumulative frequency column should match the number of observations.

WRITE

- | Breaking strain (kg) | Frequency | Cumulative frequency |
|----------------------|-----------|----------------------|
| 40– | 2 | 2 |
| 45– | 6 | 8 |
| 50– | 8 | 16 |
| 55– | 10 | 26 |
| 60– | 9 | 35 |
| 65– | 4 | 39 |
| 70– | 1 | 40 |

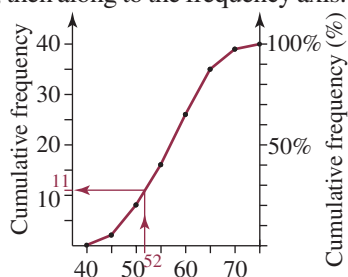
- The maximum height of the ogive will be 40 because there were 40 observations.
The ogive will start at 40 kg on the horizontal axis. The next point will represent the end of the first group. By the end of the first group (that is, 45 kg) there had been 2 observations entered. So (45, 2) is the next point on the ogive. The last point is (75, 40). Join the points to complete the ogive. Draw a vertical line from the end point of the ogive to the horizontal axis. Label the top end point as 100%, then scale the axis appropriately.



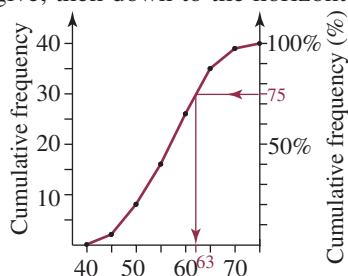
Continued over page

THINK

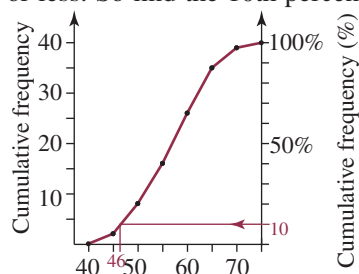
- c Find 52 kg on the horizontal axis, go up to the ogive, then along to the frequency axis. Arrive at 11.



- d Find 75% on the percentage axis, go across to the ogive, then down to the horizontal axis.



- e If 90% of the ropes are to withstand the strain then 10% of the ropes will break with this strain or less. So find the 10th percentile.



Find 10% on the percentage axis, go across to the ogive, then down to the horizontal axis.

WRITE

- c About 11 pieces of rope broke under a strain of less than 52 kg.

- d The 75th percentile is 63 kg. 75% of the sample pieces broke under a strain of 63 kg or less.

- e 10% of the ropes will break under a strain of less than 46 kg. So if the rope is marketed as 46 kg breaking strain then 90% of the ropes will withstand the strain.

A different display

Display Karen's real estate data in an ogive.

remember

1. The *cumulative frequency* is the number of times that a score plus all lower scores occur in the set of data. It is obtained by adding together all the preceding data in the frequency column.
2. An *ogive* is a line graph of the cumulative frequency results.
3. A percentile is the score below which a particular percentage of the data lies.

EXERCISE 9F Cumulative data

**WORKED
Example**
13a, b

- 1 The frequency table below shows the lengths of 77 flathead caught in a fishing competition.

- Copy the table and add a cumulative frequency column to it.
- Prepare an ogive of the data.

Length of fish (mm)	Frequency
300–310	9
310–320	15
320–330	20
330–340	12

Length of fish (mm)	Frequency
340–350	8
350–360	7
360–370	4
370–380	2

- 2 The following frequency table shows the times taken for 60 people involved in a psychology experiment to complete a simple manipulative puzzle.

Time taken (sec)	Frequency
6–8	1
8–10	4
10–12	15
12–14	18

Time taken (sec)	Frequency
14–16	12
16–18	8
18–20	2

- Copy the table and add a cumulative frequency column to it.
- Prepare an ogive of the data.

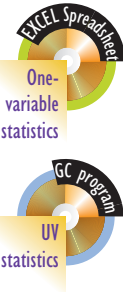
**WORKED
Example**
13

- 3 The salaries of the 40 employees of a small manufacturing company are represented by the accompanying frequency table.

Salary (\$1000)	Frequency
15–	6
20–	12
25–	8
30–	7

Salary (\$1000)	Frequency
35–	5
40–	1
45–	1

- Copy the table and add a cumulative frequency column to it.
- Prepare an ogive of the data.
- How many employees are earning less than \$22 000?
- How many employees are earning less than \$31 000?
- Find the 75th percentile of the data and write a sentence explaining what it means.
- Find the 50th percentile of the data and write a sentence explaining what it means.
- Find the 25th percentile of the data and write a sentence explaining what it means.
- The management decides to award pay rises to its highest earning employees. The top 10% of employees will all get a pay rise. How much salary would an employee need before qualifying for a pay rise?

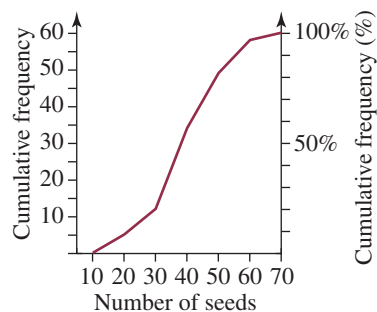


- 4 A manufacturer of surf clothing needs to know how many clothes of different sizes to produce. The manager organises a survey of young people which provides the following data:

Waist size (cm)	Frequency
70–	13
75–	28
80–	46
85–	30

Waist size (cm)	Frequency
90–	17
95–	8
100–	7
105–	1

- Copy the table and add a cumulative frequency column to it.
 - Prepare an ogive of the data.
 - How many young people had a waist size of less than 82 cm?
 - How many young people had a waist size of greater than 94 cm?
 - Find the 90th percentile of the data and write a sentence explaining what it means.
 - Find the 50th percentile of the data and write a sentence explaining what it means.
 - The manager decides that production costs can be minimised by only making garments fitting sizes between 78 cm and 100 cm. What percentage of the population will not be catered for by this manufacturer?
- 5 A biologist who counts the number of seeds in each of 60 pumpkins presents his findings on the ogive at right.
- How many pumpkins contained 30 or fewer seeds?
 - How many pumpkins contained more than 50 seeds?
 - What percentage of pumpkins had fewer than 45 seeds?
 - What percentage of pumpkins had fewer than 20 seeds?
 - Find the 90th percentile of the data and write a sentence explaining what it means.
 - Find the 75th percentile of the data and write a sentence explaining what it means.
 - Find the 50th percentile of the data and write a sentence explaining what it means.
 - The worst 20% of pumpkins (in terms of their seed numbers) are to be kept aside for further investigation. Find the maximum number of seeds for any pumpkin in this group.



- 6 A time trial is a race in which each competitor rides separately, racing 'against the clock'. The following are the times (in seconds) of 20 competitors in a 1 km cycling time trial.

75	72	68	78	75
68	77	80	85	82
73	75	82	90	92
75	73	72	70	83

- a Copy and complete the frequency distribution table at right.
 b Prepare an ogive of the data.
 c How many riders finished with a time of 82 seconds or better?
 d Find the 90th percentile and write a sentence which explains what it means.
 e The top 20% of riders are to be selected for a special training squad. What time would be needed to qualify for the squad?

Time	Tally	Frequency	Cumulative frequency
65–			
70–			
75–			
80–			
85–			
90–			

- 7 The following data, collected from a maternity hospital, give the birth weights (in kg) of 30 babies.

3.7	3.2	3.8	4.1	2.9	3.3	3.6	3.1	3.6	3.9
4.2	2.5	2.7	3.9	3.6	3.2	3.0	2.9	3.4	3.0
3.1	2.8	2.9	3.2	3.1	3.8	3.9	3.3	4.4	3.4

- a Copy and complete the following frequency distribution table.
 b Prepare an ogive of the data.
 c What percentage of babies had a birth weight of 3.8 kg or less?
 d Find the 50th percentile and write a sentence that explains what it means.
 e Babies of weight 2.6 kg or less are given special attention by medical staff. What percentage of babies will be given special attention?

Weight	Tally	Frequency	Cumulative frequency
2.4–			
2.8–			
3.2–			
3.6–			
4.0–			
4.4–			



- 8 (You might like to attempt this question using a graphics calculator.) The number of people who are using a particular plane flight is recorded on a weekly basis for 20 weeks. The results are as follows:

85	103	98	72	88
86	92	108	112	84
95	99	84	79	88
93	87	79	81	80

- Summarise the data using a frequency distribution table. Use a class size of 10. Include a cumulative frequency column.
- Prepare an ogive of the data.
- How many of the flights had less than 95 passengers?
- Find the 75th percentile and write a sentence which explains what it means.
- Find the 50th percentile and write a sentence which explains what it means.
- Find the 25th percentile and write a sentence which explains what it means.
- The management of the airline is considering using a smaller aircraft for the flight. One of the management team calculates that it would be more profitable to use a smaller plane if 20% of current flights have 80 or fewer passengers. Should a smaller aircraft be used?



- 9 (You might like to attempt this question using a graphics calculator.) Thirty samples of paint are exposed to high temperature until they begin to blister. The following data show the temperature ($^{\circ}\text{C}$) at which each began to blister.

75	88	76	92	77	88	82	81	93	96
81	88	75	79	80	81	77	88	92	79
81	83	88	82	85	79	93	84	84	77

- Represent the data on a frequency distribution table using a class size of 5° . Include a cumulative frequency column.
- Prepare a histogram of the data.
- Prepare an ogive of the data.
- How many paint samples blistered at a temperature of 85° or less?
- What percentage of the samples blistered at a temperature of 85° or less?
- What percentage of the samples could withstand a temperature of 88° ?
- Find the 90th percentile and write a sentence explaining what it means.
- The manufacturer wishes to guarantee the paint at high temperatures. What is the highest temperature at which the paint could be rated if the management wants at most 20% of the paint returned because it failed to meet its rating?

10 multiple choice

Which of the following is *untrue* of ogives?

- Ogives display cumulative data.
- Ogives are a type of line graph.
- Ogives tell the number of scores that are of a particular value or less.
- An ogive cannot be interpreted without having the original experimental data at hand.
- An ogive can also be called a cumulative frequency polygon.

11 multiple choice

The 90th percentile of some data is 25. This means that:

- 10% of the data are less than or equal to 25
- 10% of the data are greater than or equal to 25
- 90% of the data are less than or equal to 25
- 90% of the data are greater than or equal to 25
- 25% of the data are less than or equal to 90.

summary

Classification of variables and data

- Variables and data can be classified as being categorical or quantitative.
- Categorical data are non-numerical. For example, a survey of car types is not numerical.
- Quantitative data can be either counted or measured. For example, a survey of the daily temperature is quantitative.
- Quantitative data and variables can be either discrete or continuous.
- Discrete data can take only certain values, generally whole numbers, but not always.
- Continuous data can take any value within a certain range.

Data collection

- A statistical investigation can be done using either a census or a sample.
- A census is when an entire population is used in the investigation.
- A sample is when a small group is used in the investigation and the results are taken to be representative of the whole group. There are three types of sample.
 1. Random sample — chance is the only factor in deciding who participates.
 2. Stratified sample — the sample taken is chosen so that it has the same characteristics as the whole population.
 3. Systematic sample — there is a method for deciding who participates in the sample.

Bias

- If the sample is poorly chosen the results of the investigation will be biased. This means the results will be skewed towards one section of the population.

Stem plots

- A stem-and-leaf plot is a useful way of displaying data up to about 50 observations.
- A stem plot is constructed by breaking the numerals of a record into two parts: a *stem* and a *leaf*. The last digit is always the leaf and any preceding digits the stem.
- When asked to represent data using a stem-and-leaf plot, it is always assumed the stem-and-leaf plot will be ordered.
- If data are bunched then it may be useful to break the stems into halves or even fifths.

Frequency histograms and bar charts

- On a frequency histogram the vertical axis displays the frequency and the horizontal axis displays the class intervals.
- Data given in raw form should be summarised first in a frequency table.

Describing the shape of stem plots and histograms

- When data are displayed in a histogram or stem plot, we say that the distribution of those data is:
 1. *symmetric*, if there is a single peak and the data trail off on either side of the peak in roughly the same fashion

2. *negatively skewed*, if the data peak to the right and trail off to the left
 3. *positively skewed*, if the data peak to the left and trail off to the right.
- An *outlier* is an observation that lies well away from the rest of the data.

Cumulative data

- The cumulative frequency of a score is the number of times that score plus all lower scores occur in the set of data.
- An ogive is a line graph of the cumulative frequency results.
- A percentile is the score below which a particular percentage of the data lies.

CHAPTER

review

9A

9A

9B

9B

9B

9B

9B

- 1 State whether each of the following data types are categorical or quantitative.
 - a The television program that people watch at 7:00 pm
 - b The number of pets in each household
 - c The amount of water consumed by athletes in a marathon run
 - d The average distance that students live from school
 - e The mode of transport used between home and school
- 2 For each of the quantitative data types below, determine if the data are discrete or continuous.
 - a The dress sizes of Year 11 girls
 - b The volume of backyard swimming pools
 - c The amount of water used in households
 - d The number of viewers of a particular television program
 - e The amount of time Year 11 students spent studying
- 3 For each of the following statistical investigations, state whether a census or a sample has been used.
 - a The average price of petrol in Sydney was estimated by averaging the price at 40 petrol stations.
 - b The Australian Bureau of Statistics has every household in Australia complete an information form once every five years.
 - c The performance of a cricketer is measured by looking at his performance in every match he has played.
 - d Public opinion on an issue is sought by a telephone poll of 2000 homes.
- 4 Name and describe three different methods for selecting a sample.
- 5 Which method of sampling has been used for each of the following?
 - a The quality-control department of a tyre manufacturing company road tests every 50th tyre that comes off the production line.
 - b To select the students to participate in a survey, a spreadsheet random number generator selects the roll numbers of 50 students.
 - c An equal number of men and women are chosen to participate in a survey on fashion.
- 6 Use your random number generator to select 10 numbers between 1 and 1000.
- 7 The table at right shows the number of students in each year of school.
In a survey of the school population, how many students from each year should be chosen, if a sample of 60 is selected using a stratified sample?

Year	No. of students
8	200
9	189
10	175
11	133
12	124

9C

- 8 The money (rounded to the nearest whole dollar) raised by Year-12 students is shown below.
78 84 61 73 71 83 87 65 60 67 71 82 84 79 78
Construct a stem plot for the amount raised using:
- the stems 6, 7 and 8
 - the stems 6, 7 and 8 split into halves
 - the stems 6, 7 and 8 split into fifths.

9D

- 9 The frequency table below shows the speeds of cars recorded by police. The cars were travelling through a 60 km/h zone. Construct a histogram to display the data.

Class interval	Frequency
50–51.9	3
52–53.9	5
54–55.9	6
56–57.9	7
58–59.9	9
60–61.9	10
62–63.9	9

Class interval	Frequency
64–65.9	10
66–67.9	8
68–69.9	5
70–71.9	3
72–73.9	4
74–75.9	2

9E

- 10 The money raised (to the nearest whole dollar) by each student in a Grade-3 class on the school walkathon is shown in the stem plot at right.
Describe the shape of the distribution of these data and comment on the existence of any outliers.

Key: 0|8 = \$8

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



9F

- 11 A researcher investigates the amount of money spent by 50 people in a local Coles supermarket on Friday night between 6 pm and 8 pm. His findings are presented in the ogive shown at right.

- How many people spent \$50 or less?
- How many people spent more than \$100 on shopping?
- What percentage of people spent less than \$80?
- Find the 80th percentile of the data and write a short sentence explaining what it means.
- If Nathan was in the lowest 10% of buyers, what is the maximum amount of money he could have spent?

