Module A6

DEALING WITH DATA

Table of Contents

Introduction 6.1

6.1 Collecting data 6.2

6.2 Organising and displaying data from tables 6.5

6.3 Organising and displaying raw data 6.12
6.3.1 Frequency distribution table 6.12
6.3.2 Frequency histograms 6.16
6.3.3 Stem-and-leaf plot 6.21

6.4 Analysing data 6.26
6.4.1 Where is the centre of these data? 6.27
The mean 6.27
The mode 6.31
The median 6.32
A comparison of mean, median and mode 6.34
6.4.2 How spread out are these data? 6.36

6.5 Data with two variables 6.37

6.6 A taste of things to come 6.44

6.7 Post-test 6.47

6.8 Solutions 6.49

Introduction

It is often said that today we live in an information society. We just have to look at a newspaper or watch the TV to see that that is the case. Reporters often use expressions like:

'The statistics indicate.....'

'The data show....'

'Results of university tests show...'

We are surrounded with advertising trying to persuade us that one product is better than another. In our society today if we are to be informed citizens and not be taken advantage of we need to know how to interpret such information for ourselves. Also in your future tertiary studies you might be required to collect large amounts of information which has to be summarised and presented in a report or essay. Statistics is the science of gaining information from numerical facts. These numerical facts are called data. These words were first coined in the eighteenth century by political scientists in Germany.

In this module we aim to help you in the process of collecting, organising, presenting and analysing data. More formally, on successful completion of this module you should be able to:

- categorise data into different types;
- define and give examples of population, sample, variables and randomness;
- summarise and interpret data in bar charts;
- summarise and interpret data in frequency distributions and histograms;
- use and construct stem-and-leaf plots to help interpret data;
- demonstrate an understanding of the centre and dispersion of data;
- calculate and interpret mean, median, mode and range of a sample; and
- summarise and interpret data with two variables using scatterplots and lines of best fit.

(Reproduced with permission from G. Clark)

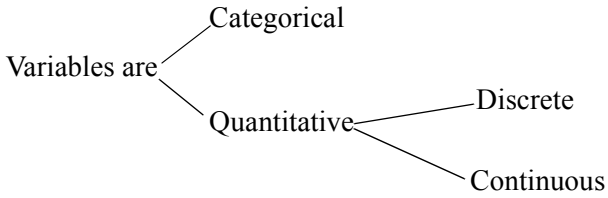
6.1 Collecting data

We see data everywhere but where does it come from and what exactly is it? Data is collected for individuals.....how many people live in Australia, are these people male or female, how tall are most Australians. All these figures are data (data is the plural of datum from the Latin). However, data are not only collected about people we also might want to collect data from things or animals. **Individuals** are the objects described by a set of data. **Variables** are the characteristics of those individuals, for example, number of people, the gender or the height of the people. Note that a variable might thus take on different values for different individuals.

Data collected as variables can be of different types. For examples when we ask whether a person is male or female only a certain number of answers are possible (usually only male and female). The same goes for blood types, eye colour, country of birth. All these variables would be **categorical variables**. That is the values for the variables are just labels for the different categories. The categories should be non-overlapping and should cover all possibilities.

On the other hand the value of the variable could be found by counting or measuring a quantity. Such variables would be called **quantitative variables**. Some of these data would be **continuous** (taking any value) such as height, temperature, money, while other data would be **discrete** (able to take only whole number values) such as number of people or number of trees in a paddock.

In summary



Example

Classify the following variables as either categorical, discrete or continuous

Effect of new drug (successful or not successful)

- Rainfall in mm
- Number of words in a paragraph
- Australian State
- Blood Pressure in mmHg
- Number of faults at tennis

Variable	Type	Reason
Effect of new drug (successful or not successful)	categorical	only two non-overlapping alternatives present
Rainfall in mm	continuous	an unlimited number of smaller and smaller measurements are possible
Number of words in a paragraph	discrete	can take only values which are whole numbers; this is a result of counting
Australian State	categorical	only six non-overlapping alternatives present
Blood Pressure in mmHg	continuous	an unlimited number of smaller and smaller measurements are possible
Number of faults at tennis	discrete	can take only values which are whole numbers; this is a result of counting

Now that we know what data we might want to collect, how do we go about collecting it? Let's look at the following example.

Suppose an entertainment park wishes to provide a family pass to their attraction. They would like their family pass to cater for families with the 'average' number of children. How do they go about finding the 'average' number of children in a family?

They must firstly find the number of children in every family with children in Australia. What a task!

All the children in Australia would be called a statistical **population**. That is, all possible values that could be collected: in our case the number of children in every family with children in Australia. In this and many other statistical investigations it is not practical to survey the entire population. In this case we might investigate a **sample** of the population. If this sample is to be later used to make conclusions about the entire population then it is important to make it representative of the entire population. It could include different types of families; country families, city families, or families from different nationalities, blended families and families with only one parent.

We call a sample that represents the entire population a **random sample**. This means that every member of the population has an equal chance of being selected for the sample. We have not gone out on purpose and selected every city family (for example) to survey and thus obtained a sample biased towards that group. In the remainder of this module we will be dealing only with samples collected randomly from a larger population.

Example

The following are some examples of populations, samples and variables that might have been of interest.

Population	Possible sample	Possible variable of interest
All words in a book	20 words from each page	length of words function of word (verb, noun) number of syllables in word
All trees in Australia	100 trees from each National Park	height of tree species of tree presence of flowers presence of possums
All Australian women who smoke	2000 women who smoke	age of women weight of women presence of cancer

Activity 6.1

1. Data from a study of local primary schools involving teachers, pupils and parents contained values for a range of different variables. Which of the variables were categorical, discrete or continuous?

   (a) Gender (male or female)

   (b) Age (years)

   (c) Income of parents (dollars and cents)

   (d) Siblings (number for each pupil)

   (e) Temperature of classrooms in winter (degrees Celsius)

   (f) Smoker (yes or no)

   (g) Class size (number of children)

2. You are interested in the relationship between English proficiency and 1st year university results of students whose second language is English. Answer the following questions.

   (a) Define the population.

   (b) Define the sample(s) of the population you might decide to use.

   (c) Define some variables which may be on some interest in this study.

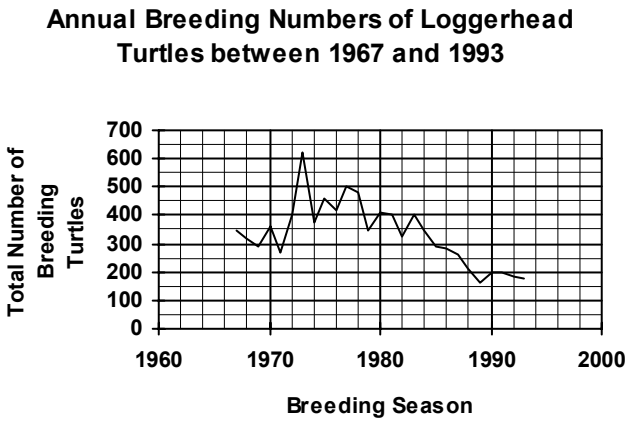
6.2 Organising and displaying data from tables

Previously in these modules we have examined vast amounts of data. These data were collected from a range of different samples for a multitude of reasons, and often were summarised into tables. Tables are useful so that we can see exactly the details of the data but to gain the attention of the reader or audience and to show the main points in a set of data then tables are often turned into pictures. These pictures can take the form of graphs. The following are some types of graphs we have seen before.

Line graphs

Line graphs are those where data have been plotted as a series of points that are joined by a line.

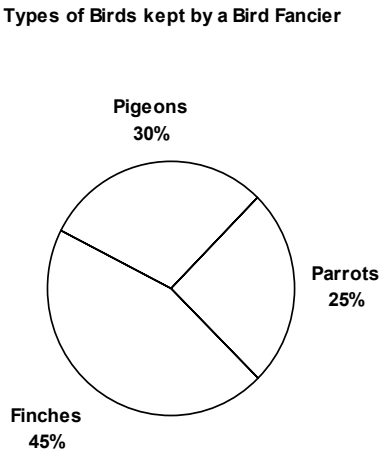
See module 5 for details on how to draw and interpret these graphs.



Pie Charts

Pie charts are pictures where data are represented as a circle in which each sector represents each category under study. They are used primarily for categorical data.

See module 3 for details on how to draw and interpret these graphs.



Bar Charts

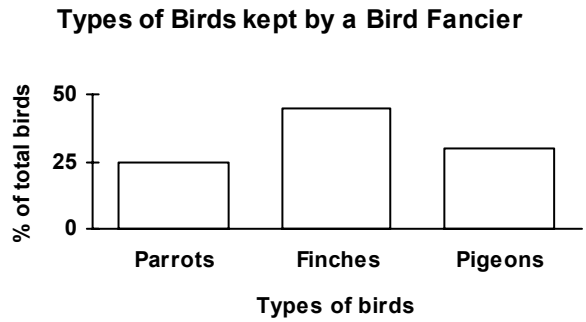
An alternative to the pie chart is a bar chart.

Let's examine the data from the pie chart above. These data were originally represented in a table.

Type of Bird	% of Total Birds
Parrots	25
Finches	45
Pigeons	30

In a bar chart the data are shown as a series of rectangles which represent each of the categories of the data. The height of the rectangle gives the value of the variable.

To draw this graph you must first consider what will be put on each axis. On the horizontal axis each category is given a certain length which will depend on the number of categories you have to fit across the page. The vertical axis shows the variable of interest this time the percentage of total birds. The graph should be high enough to cater for the maximum percentage shown in the table. The bar chart then looks like this.



From this graph we can quickly interpret that finches are the most common bird in the bird fancier's collection, with parrots being the least common.

Don't forget the following when drawing bar charts:

- label both horizontal and vertical axes;
- scale is correct and appropriate;
- make sure the rectangles are of equal width;
- make sure that the rectangles are the same distance apart;
- give the chart a title which details what the chart is about.

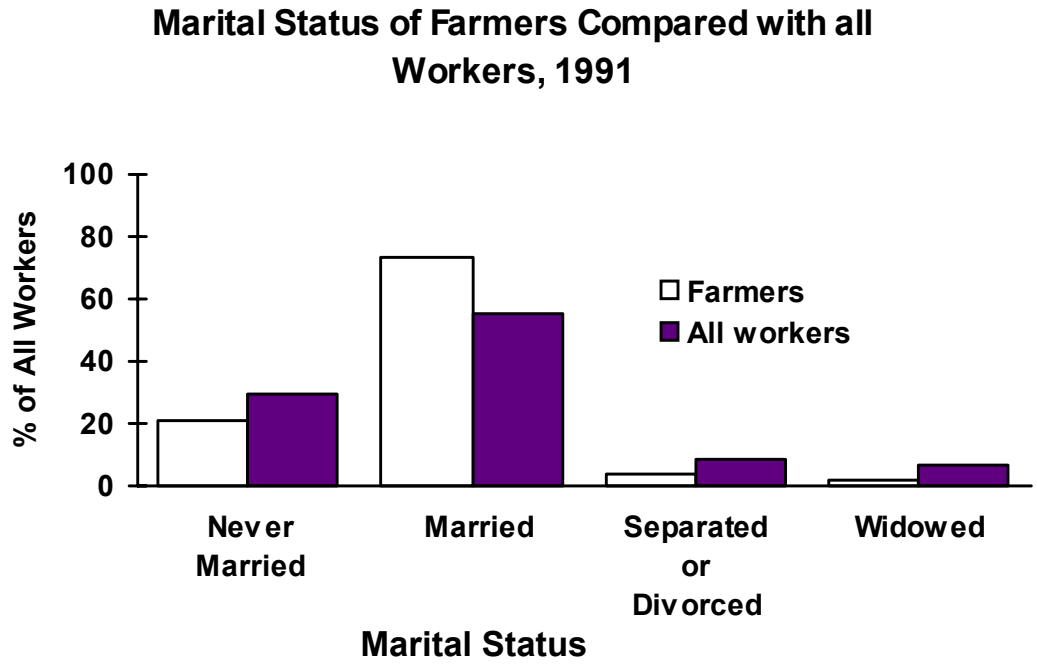
Often data contain information that has been broken into many different categories. This is where bar charts are more useful than pie charts. Consider the following set of data.

Marital Status of Farmers Compared with All Workers, 1991

	Never Married %	Married %	Separated or Divorced %	Widowed %
Farmers	20.8	73.2	4.0	2.0
All workers	29.6	55.7	8.3	6.4

(Source: *Australian Bureau of Statistics 1991.*)

These data could be included in a **paired bar chart** as shown below. It is easy to see from this chart the differences between farmers and other workers.



Activity 6.2

1. A survey of 200 people asking what cereal they ate for breakfast found the following results:

Cereal	Number of people	Percentage of people
Corn Flakes	50	
Rice Bubbles	42	
Nutri Grain	39	
Rolled Oats	23	
Muesli	11	
Coco Pops	10	
Other Cereals	25	

You have completed this table in module 4, add the figures to the above table from your previous working (or recalculate for practice) and draw a bar chart representing the cereal preferences of this group of people.

2. The following table shows the different items that have been found around the necks of 75 seals in the wild. This is a very serious problem in the wild and often leads to the death of the seal.

Item	Number	% of Total
Trawler nets	26	
Packaging bands	15	
Gillnet	8	
Rope	5	
Other	21	
Total	**75**	

Complete the above table and present the data pictorially in a bar chart.

3. 500 females and 500 males in a city were asked which political party they would support if there was an election on the following day. The results are given below.

Political Party	Males	Females	Total
National	140	100	240
Liberal	65	135	200
Labor	260	180	440
Democrats	20	60	80
Other	15	25	40

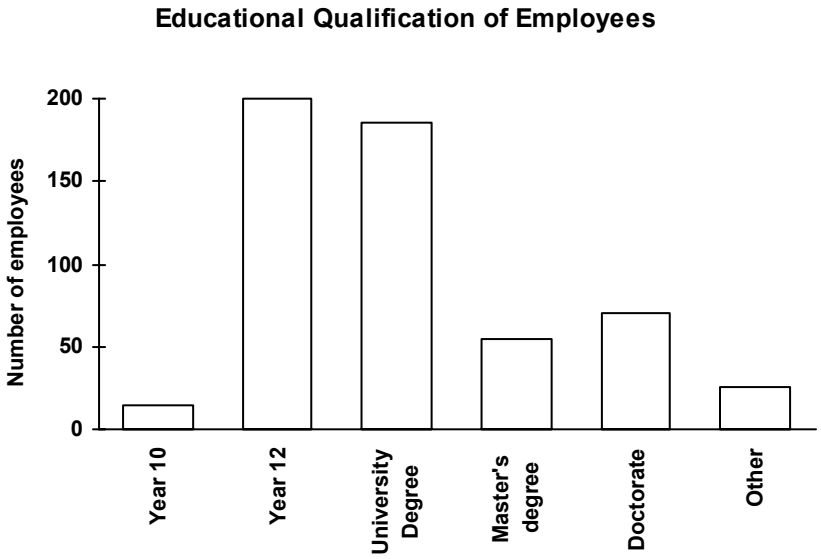
   (a) Display these data as a paired bar chart from which you could compare male and female political preferences.

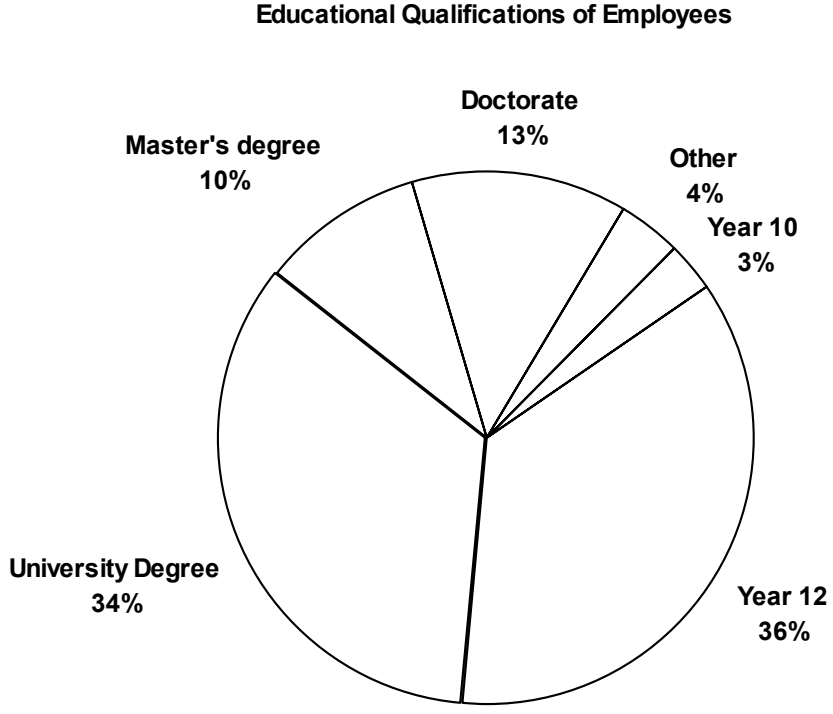
   (b) Write a few sentences comparing the different preferences of males and females.

4. The following are the qualifications of a group of workers at a local company. The local newspaper collects these figures from the company census with the aim of writing an article on the educational background of a typical company in that town.

Qualification	Number of Employees	% of Total
Year 10 Secondary School	15	2.7
Year 12 Secondary School	200	36.4
University Degree	185	33.6
Master's Degree	55	10.0
Doctorate Degree	70	12.7
Other	25	4.5
Total	550	99.9

The newspaper is trying to decide how to best present these data pictorially. The reporter has suggested that they use one of the following graphs.

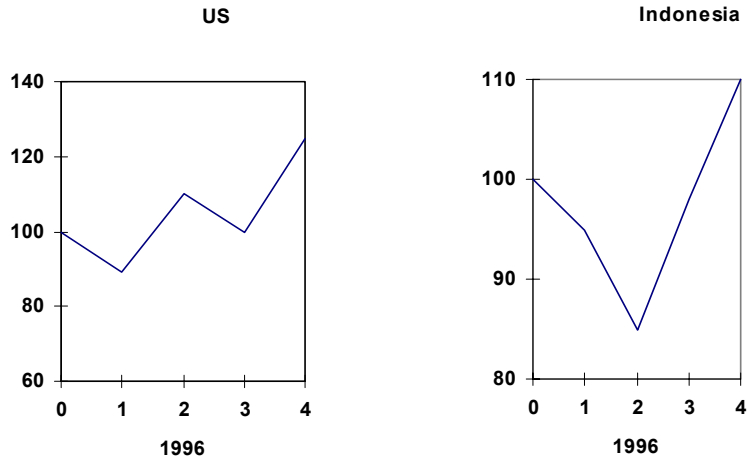




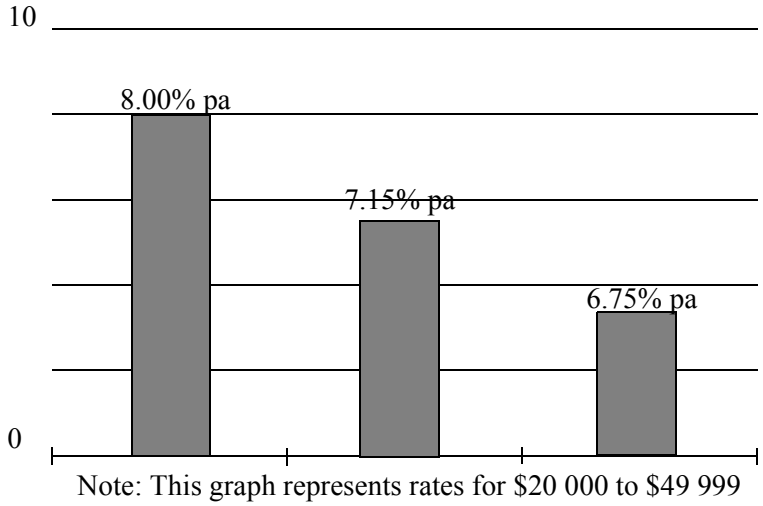
(a) Which graph would best describe the educational backgrounds for this set of workers. Explain the reason for your choice.

(b) Write a paragraph using one of the above graphs to describe the educational backgrounds of the workers. The paragraph should be suitable for use by the local newspaper.

In this section of work we have looked at a three different types of graph. This is only the tip of the iceberg when it comes to graphs that you might see in the daily news. The way that different types of graphs can be presented is only limited by our creativity. However, **beware** many graphs that we see daily are designed to deceive the unwary. It is up to us to be cautious in our interpretation of statistics. Here are a couple of examples of things to look for....



Graphs similar to these were seen in a financial newspaper. Note the different scales on the vertical axes making them difficult to compare. Also the titles and axes are lacking in details making them difficult to interpret.



This graph was seen in an advertisement for a popular credit union. Note that there is no detail on either axis and that the percentage written on the columns do not correspond with the scale on the left.

6.3 Organising and displaying raw data

So far we have looked at data that has already been collected and displayed in a table, but what if the data has just been collected and you are beginning with a large group of figures. For example:

You are a social worker who has just completed collecting data on the number of children in all the families within your jurisdiction. What you have as a result is a table which contains counts of the number of children in thirty families.

1	4	2	3	2	4	1	5	2	3
3	1	4	5	5	2	8	3	6	4
2	3	2	3	1	5	4	7	2	6

This type of table is very difficult to interpret, so we need a way to summarise these data into a form that is easier to read and interpret. The frequency distribution table is a way of doing this.

6.3.1 Frequency distribution table

The **frequency distribution table** is a common way of organising data so that it is very readable.

The first column of the table shows the variable being measured (in our case this is the number of children in the family). We often represent the variable as x. The second column is a tally column (if it is needed) and the third column is the **frequency** column. This column shows the number of times each score has been observed and is often represented by f. The table should always have a title. If we convert the table above into a frequency distribution table we get.

Frequency distribution table showing the number of children in each of 30 families surveyed.

Number of Children per family (x)	**Tally**	**Frequency (f)**
1	𝍸	4
2	𝍸 ||	7
3	𝍸 |	6
4	𝍸	5
5	𝍸	4
6	||	2
7	|	1
8	|	1
		$\Sigma f = 30$
This column shows all the possible values of the variable from smallest to largest	This column is the tally in groups of five. It is not really part of the formal table but is a way of keeping count.	This column is the frequency column and represents how often each value of the variable occurs.

Don't be put off by the use of Σf in the last column, it is just another way of saying the total or sum of the frequency column. It is the total frequency. Σ (sigma) is a Greek letter which means 'sum of'. It is always a good idea to find the sum of the frequency column as a check that you have not missed out any values when tallying. In our example the sum of the last column should be equal to 30 (because we started with 30 families).

It is much easier now to read off information from our survey. We can see that 7 families have 2 children but only 1 family had 8 children. We could also say that 8 of the 30 families have more than 4 children in their families (this means 5, 6, 7 or 8 children).

We could express some of this information as a percentage.

For example, what percentage of families have less than 4 children. Less than 4 children means 1, 2 or 3 children. From our table of values 17 families have less than 4 children. This represents $\frac{17}{30}$ or 57% of the families surveyed.

We will return to our survey later. For now try to construct some frequency distribution tables for the following data.

Activity 6.3

1. A survey of weights (to the nearest kilogram) of students in a maths tutorial produced the following data.

56	70	58	62	59	61	70	58
64	64	62	68	63	64	61	60
60	66	63	67	65	58	66	63
68	63	69	63	67	67	63	66

   (a) What was the lowest weight? What was the greatest weight?

   (b) Construct a frequency distribution table for the above data.

   (c) How many students weighed 70 kg?

   (d) How many students weighed less than 62 kg?

2. A golfer returned the following scores on 20 successive games on the same course.

85	81	81	83	84	78	87	79	82	86
83	80	82	81	79	84	85	81	79	80

   (a) Construct a frequency distribution table for the above data

   (b) On how many rounds did the golfer score 84?

   (c) On how many rounds did the golfer score less the 80?

3. When babies are born in hospital their blood group is determined as part of the routine testing performed on new-borns. Blood groups fall into one of four categories, A, B, AB or O.

   In a particular hospital the 24 babies born in one week had the following blood groups.

   O O A O AB B O A B O O A

   AOB O A A O O A O A A

   (a) Construct a frequency distribution table for the above data.

   (b) What was the most common blood group among these babies?

   (c) Did you know that approximately 49% of the Australian population has type O blood. What percentage of these babies had type O blood?

4. The following figures represent readings taken by a radar trap on a section of road with a speed limit of 80 km/h

81	80	70	76	90	82	73	78	81	80
81	83	73	75	78	75	79	76	84	102
93	81	100	85	73	81	76	77	76	73
74	83	71	71	85	96	86	79	81	83

   (a) Construct a frequency distribution table for the above data.

   (b) What number of the motorists were travelling above the speed limit?

5. A health clinic is having a free health check stand at the local shopping centre. The heights to the nearest centimetre of the first 50 clients are recorded below.

150	185	92	155	165	178	150	92	155	176
95	115	150	165	135	160	180	95	116	158
88	106	182	156	143	180	115	155	176	92
94	125	128	136	116	148	122	164	165	178
145	148	122	164	182	160	128	132	176	158

   (a) Construct a frequency distribution table for the above data.

   (b) Do you think that there would be a better way to present this data?

Grouped frequency distribution table

Let's consider the last question you did in the previous activity. These data ranged from the shortest person at 88 cm to the tallest at 185 cm. We would have had a very large table if we had included all the possible values for the heights between these two values.

A much more convenient way of representing these data is to use a **grouped frequency distribution table**. We could group the height measurements into **classes**.

In our example on heights we could have classes from 80 cm to 90 cm and from 90 cm to 100 cm and so on until we get to the class 180 cm to 190 cm. Looking at the first two of these classes you will see that 90 cm occurs in both. We must be very clear about which class 90 cm will fall into. In the case of this example we will put 90 cm into the class 80 to 90. It is very important to be clear about which class each observation falls into as you do not wish to count one observation twice.

Here is the grouped frequency distribution table for the data in activity 6.3, question 5.

Grouped frequency distribution table showing the heights of 50 people in a health survey.

Heights (cm)	Tally	Frequency (f)
80 up to and including 90	𝍸	1
90 up to and including 100	𝍸 𝍸	6
100 up to and including 110	𝍸	1
110 up to and including 120	𝍸 𝍸	4
120 up to and including 130	𝍸 𝍸	5
130 up to and including 140	𝍸	3
140 up to and including 150	𝍸 𝍸 𝍸	7
150 up to and including 160	𝍸 𝍸 𝍸	8
160 up to and including 170	𝍸 𝍸	5
170 up to and including 180	𝍸 𝍸 𝍸	7
180 up to and including 190	𝍸	3
		$\Sigma f = 50$

You should check that you get the same frequency as above for each of the classes.

Activity 6.4

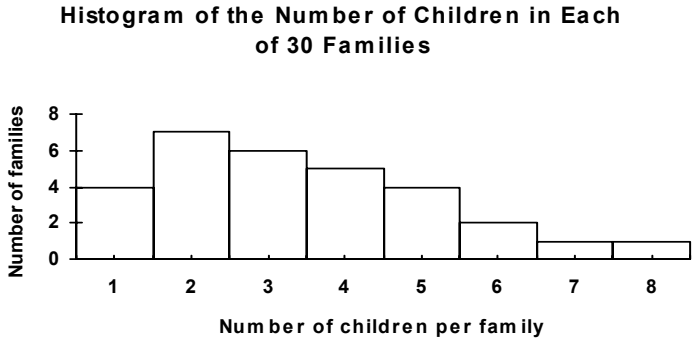
1. Using the data displayed in activity 6.3 question 1, construct a grouped frequency distribution, using class widths starting at 55 up to and including 57.

6.3.2 Frequency histograms

As before to gain a better understanding and to get a clearer picture of what is happening with the data it is often necessary to show the data pictorially. If the variable involved is quantitative (either discrete or continuous) then we can represent the frequency distribution as a **histogram**. Histograms are like bar charts.

For histograms showing discrete data, we construct a series of rectangles, usually of width 1 cm. Each rectangle is centred on an observed value of the variable and the height of each is equal to the frequency of that observed value. For example, to construct a histogram that displays the data on the number of children per family in the example in the section 6.3.1 follow the steps below.

- The number of children ranged from 1 to 8. Place these numbers along the horizontal axis. These should be evenly spaced, not missing out any values even if they have a frequency of zero.
- The frequencies ranged from 1 to 7, so use this to select a suitable scale for the vertical axis.
- Notice that all rectangles are of equal width and are centred on the number of children per family.



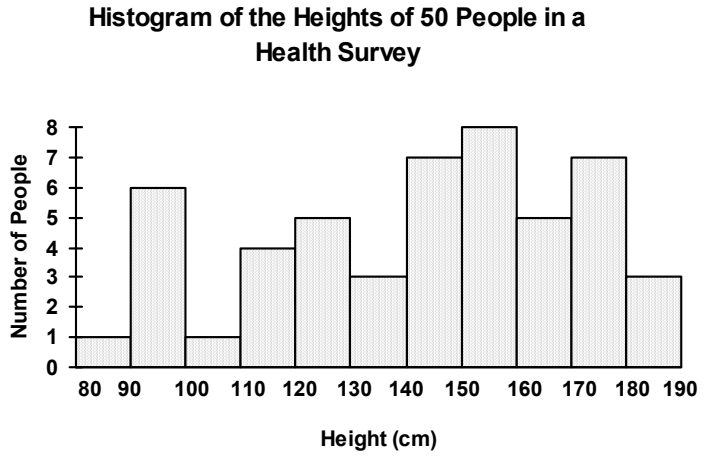
What can we say about these data when we have viewed the histogram. Firstly notice that the horizontal axis does not begin at zero so we have to be careful how we interpret the values on this axis. However bearing this in mind it still seems that the histogram is higher on the left than on the right, we say it is **skewed** to the right. That is, more families have less than 4 children than have more than 4. Only 2 families have 7 or 8 children compared with 11 families having 1 or 2.

To construct a histogram for continuous data we once again construct a series of rectangles of equal width. The height of each corresponds to the class frequency, as before. The width of each rectangle however will depend on the scale on the horizontal axis and classes that we decide to break the values of the variable up into. Refer back to the discussion of grouped frequency distributions. The principles developed there are used in constructing a histogram for continuous data. Consider the grouped frequency distribution we constructed from activity 6.3 question 5.

Grouped frequency distribution table showing the heights of 50 people in a health survey.

Heights (cm)	Frequency (f)
80 up to and including 90	1
90 up to and including 100	6
100 up to and including 110	1
110 up to and including 120	4
120 up to and including 130	5
130 up to and including 140	3
140 up to and including 150	7
150 up to and including 160	8
160 up to and including 170	5
170 up to and including 180	7
180 up to and including 190	3
	$\Sigma f = 50$

Using these data we could construct the following histogram.



What can we say about these data when we have viewed the histogram. It looks very different from the previous example with no definite patterns occurring either to the left or right. The peak at 90 to 100 cm is interesting and stands out from the other patterns. The researcher conducting this survey might go back to the original data and find out why this group occurred so frequently (e.g. were there a lot of children in the shopping centre on this day).

In summary to draw a histogram you should complete the following steps.

- Find the range of the data, that is the smallest to largest measurement.
- Use the range as a guide for choosing a suitable class width. This is especially important for continuous data. There is no correct answer in selecting class width, we choose those values which best suit our needs. If too few classes are used then the histogram will look like a skyscraper with all the values squeezed into a few classes. If too many classes are chosen then the graph will look squashed with most classes having only one or no observations. Both of these extremes will not help us to interpret the data easily. If you are not sure where to start try to use between 5 and 10 classes.
- Set up and complete a frequency distribution table with the chosen class widths if necessary.
- Use the values in the frequency distribution table to complete the histogram. If data are continuous use the end points of the class widths as labels on the horizontal axis as shown above.
- Give your graph a title and be sure to label the axes.

Activity 6.5

1. A golfer returned the following scores on 20 successive games on the same course.

85	81	81	83	84	78	87	79	82	86
83	80	82	81	79	84	85	81	79	80

   (a) Construct a frequency distribution table for the above data (see activity 6.3 question 2)

   (b) Use the above frequency distribution to construct a histogram representing the data.

2. Thirty 100 kg bags of sugar were selected at random at a sugar mill. Their weights (in kg) were:

98.2	102.7	96.1	105.3	95.1	97.5
92.5	96.4	98.3	101.2	97.3	96.3
105.0	100.5	93.8	96.1	98.9	95.3
102.0	105.0	102.0	96.8	97.2	95.6
93.4	99.9	93.9	103.7	96.5	102.1

   (a) Construct a grouped frequency distribution table for these data. Choose the class widths that you think appropriate.

   (b) Use the above frequency distribution table to construct a histogram to represent the data.

3. The following are the blood pressure reading in mmHg of a group outpatients at a local hospital.

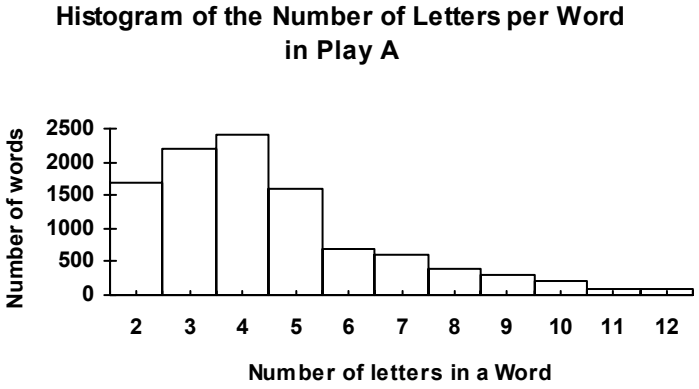
94	94	82	100	112	110	84	78	92	112
94	92	86	84	90	72	88	92	88	84
98	98	84	90	90	70	80	90	80	106
74	95	100	94	84	70	102	92	84	80
84	86	98	82	80	88	80	84	100	86

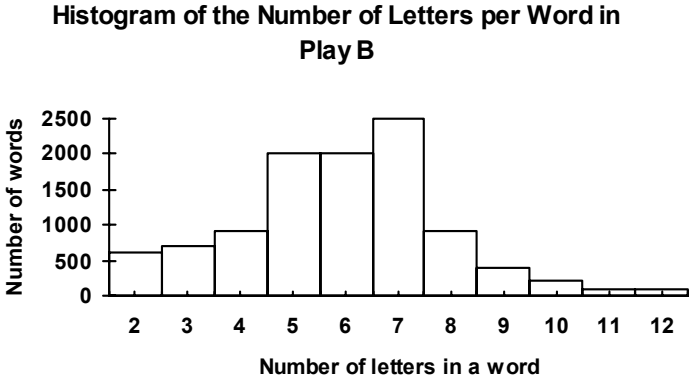
   (a) Construct a grouped frequency distribution table for the above data.

Choose the class widths that you think appropriate.

(b) Use the above frequency distribution table to construct a histogram which represents the data.

4. Examine the next two frequency distributions which display the frequency of words of different length in plays by two different playwrights.





(a) Describe in your own words the distribution of words of different length in Play A.

(b) In your own words compare the length of words in Play A with Play B

6.3.3 Stem-and-leaf plot

Histograms are not the only way to display frequency distributions so that they are easier to read. For small sets of data **stem-and-leaf plots** have also been found useful as they are often quicker to construct and present more detailed information than histograms. They are a cross between a table and a graph and are used to sort the data and display it at the same time. Consider the following set of data of the scores of 18 1st year students in a psychological test designed to measure the attitude of university students to study. Their scores were

124	100	121	115	153	167
103	165	167	129	200	148
140	152	137	141	101	126

For the stem-and-leaf plot you need to rewrite the data so that part of each number is identified as the stem and the rest of the number will be the leaf. In the example above the stem of 124 would be 12 and the leaf would be 4, the stem of 100 would be 10 and the leaf would be 0. Note that usually the leaf will consist of only one digit.

To construct the stem-and-leaf plot you need to:

- write the stem of each value vertically in increasing order from top to bottom. Draw a vertical line along side the list of stems.

10
11
12
13
14
15
16
17
18
19
20

- go through the data writing each leaf to the right of the vertical line next to its matching stem

10	0 3 1
11	5
12	4 1 9 6
13	7
14	8 0 1
15	3 2
16	7 5 7
17	
18	
19	
20	0

- write the stem again and rearrange the leaves in increasing order

10	0 1 3
11	5
12	1 4 6 9
13	7
14	0 1 8
15	2 3
16	5 7 7
17	
18	
19	
20	0

Check that you have 18 scores, the number you first started with.

There are two further things to add to complete the plot:

- A title

- An example of what the stem and leaf represent. For example, 11/6 represents 116. This is included as part of the heading and does not represent part of the data.

The completed stem-and-leaf plot is thus:

Stem-and-leaf plot of scores in attitude to study of 18 students

11	6 represents 116
10	0 1 3
11	5
12	1 4 6 9
13	7
14	0 1 8
15	2 3
16	5 7 7
17	
18	
19	
20	0

Notice that when you look more closely, the stem-and-leaf plot it is very similar to a histogram except that it is turned on its side. Stem-and-leaf plots are used in the same way as a histogram to help you see the pattern in the data or if there are any unusual points. The advantage of this type of graph over the histogram is that it preserves the actual values of each observation. However, for large data sets this type of plot is time consuming and we would not recommend it unless you were using a computer package to do all the work for you.

Activity 6.6

1. A manager of a ferry company records the following numbers of passengers on 25 runs of one of her ferries.

52	84	40	57	61
65	77	64	62	35
82	58	50	78	83
71	75	41	53	66
60	95	58	49	89

   (a) Construct a stem-and-leaf plot to represent the frequency distribution of these data.

   (b) From the stem-and-leaf plot determine what number of passengers the ferry most frequently carried.

2. A teacher of a small primary school gave a mathematics test to 20 Grade 7 students. They scored the following percentages.

67	85	46	64	73	52	79	61	76	58
58	84	32	57	68	42	76	52	74	47

   (a) Construct a stem-and-leaf plot of the data.

   (b) Use the stem-and-leaf plot to help describe in your own words the distribution of marks in the class.

   (c) If the passing mark was 50% how many students passed the test. Can you use the stem-and-leaf plot to find this out? Explain.

3. The following stem-and-leaf plots represent the rate of heart disease per 100 000 adults in different regions of the United Kingdom in 1990.

Heart disease in males

per 100 000

35	6 represents 356 per 100 000
32	0 3
33	
34	0
35	
36	1
37	
38	3
39	8
40	
41	1 8 8
42	
43	
44	2
45	2

Heart disease in females

per 100 000

35	6 represents 356 per 100 000
27	9
28	3 6
29	
30	
31	4
32	7
33	1 1 9
34	
35	
36	2
37	6
38	
39	0

(a) What was the highest and lowest rate of heart disease in males?

(b) What was the most frequent rate of heart disease in females?

(c) Use the stem-and-leaf plots above to make a comparison of the rate of heart disease of men and women in the UK.

6.4 Analysing data

So far we have collected, organised and displayed data. However, this is often not enough. In many cases we need to abbreviate the data even further so that we have measures such as the 'average' or the most typical value. These types of measures are referring to the centre of the

data and as such are called **measures of central tendency**. In the following section we will look at 3 measures of the centre of a distribution of data.

6.4.1 Where is the centre of these data?

The mean

The mean is the most commonly used measure of the centre of a group of data. You may have heard it referred to as the arithmetic average. In words we would say that the mean is equal to the sum of all the observations divided by the total number of observations.

$$\text{Mean} = \frac{\text{sum of all observations}}{\text{total number of all observations}}$$

This is sometimes abbreviated to the formula

$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ where n is the total sample size and $\sum_{i=1}^{n} x_i$ is the sum of all the observations.

Don't confuse mean with the English mean in the sense of 'nasty' although it might be considered mean to ask somebody all the things that mean can mean.

Example

If a nurse measured a patient's temperature (in °C) 4 times, what was the patient's mean temperature?

37.2 37.5 37.1 37.3

Add all the numbers given to get 149.1 then divide this total by 4 to get 37.275.

or

$$\text{Mean} = \frac{\text{sum of all observations}}{\text{total number of all observations}}$$

$$= \frac{37.2 + 37.5 + 37.1 + 37.3}{4}$$

$$= 37.275$$

$$\approx 37.3 \text{ (rounded to 1 decimal place)}$$

The mean temperature would be approximately 37.3 °C.

This method, however, becomes very tedious if we are working from a very large data set. To overcome this for large sets of data it is better to firstly arrange the data into a frequency distribution table and then calculate the mean from that table.

Recall this frequency distribution table from earlier. We can multiply the first two columns of this table together to produce a third column which we will call the *fx*.

Frequency distribution table showing the number of children in each of 30 families surveyed.

Number of Children per family (x)	**Frequency (f)**	$f \times x = fx$
1	4	1×4=4
2	7	2×7=14
3	6	3×6=18
4	5	4×5=20
5	4	5×4=20
6	2	6×2=12
7	1	7×1=7
8	1	8×1=8
	$\Sigma f = 30$	$\Sigma fx = 103$
	The total of this column is equal to the total number of observations	The total of this column is equivalent to the sum of all the observations.

This means that if we have a large number of observations that have been represented in a frequency distribution table then we can use this table to help us find the mean of the data.

$$\text{Mean} = \frac{\text{sum of all observations}}{\text{total number of all observations}}$$

This is sometimes abbreviated to the formula

$$x = \frac{\sum_{i=1}^{n} fx_i}{\sum_{i=1}^{n} f_i}$$

where $\sum_{i=1}^{n} fx_i$ is the sum of all the observations and $\sum_{i=1}^{n} f_i$ is the total number of observations.

Example

Use the frequency distribution above to find the mean number of children per family.

$$\text{Mean} = \frac{\text{sum of all observations}}{\text{total number of all observations}}$$

$$= \frac{\sum fx}{\sum f}$$

$$= \frac{103}{30}$$

$$\approx 3.433$$

The mean number of children would be approximately 3.4

Using the calculator to find the mean

Calculators can be a very useful tool in the calculation of statistics especially where data sets are large. It is important that if you are going to use your calculator to find the mean that you put your calculator in the statistics mode. As there are many different calculators it is impossible to list all the steps for each calculator. Check your calculator manual to find out the steps for your calculator. Alternatively, if you are having difficulties contact your tutor.

Example

Use your calculator find the mean $(\bar{x})$ if a nurse measured a patient's temperature (in °C) 4 times what was the patient's mean temperature?

37.2 37.5 37.1 37.3

Now try this on your calculator.

Write down your calculator steps in the space below.

Using your calculator the display should read 37.275.

Example

Use the frequency distribution table to find the mean number of children per family.

Frequency distribution table showing the number of children in each of 30 families surveyed.

Number of Children per family (x)	**Frequency (f)**	$f \times x = fx$
1	4	1×4=4
2	7	2×7=14
3	6	3×6=18
4	5	4×5=20
5	4	5×4=20
6	2	6×2=12
7	1	7×1=7
8	1	8×1=8
	$\Sigma f = 30$	$\Sigma fx = 103$

Note that when we enter these values into the calculator it is not necessary to enter each individually. The calculator is programmed so that it will allow you to enter multiples of the one value.

Write your calculator steps in the space below.

Now try this on your calculator.

Write down your calculator steps in the space below.

The display should read 3.43333333 (as before).

Activity 6.7

1. What is the mean rainfall for an 8 day period if the daily rainfall was

   7.2, 13.5, 2.1, 25, 4.6, 0, 7.2, 15.7 (rainfall is measured in mm).

2. Find the mean daily temperature if the temperature measured over a 21 day period are presented below.

Temperature in degrees Celsius (x)	**Frequency of each temperature (f)**
25	5
27	7
29	3
31	4
33	2

3. In a fishing competition there were 40 competitors. If the number of fish caught by each competitor was

8	7	7	4	7	3	7	7	6	7
3	2	7	9	7	3	8	2	6	7
5	10	8	7	8	9	7	7	8	10
7	6	5	7	3	7	7	5	4	10

   What was the mean number of fish caught?

The mode

The mode is another measure of the centre of a set of data. It is the most common observation made in a set of observations derived from the French word for fashionable. For example if we have a set of examination scores 55 55 65 62 55 78 99 55, then 55 would be the mode of these scores as it is the most common score.

When sets of data are summarised into frequency distributions it is easy to read off the mode by looking for the observation with the highest frequency. Look again at the distribution of number of children per family. In that distribution the mode was 2. This means that 7 families had 2 children each and that this was the most common number of children to have. In this case there was only one mode but in other examples there may be more than one, or the mode may not exist as all scores had the same frequency.

Activity 6.8

Return to activity 6.7 and calculate the mode for each of the distributions presented.

The median

The third measure of central tendency we shall discuss is the median. The median is the middle value in a set of observations, after they have been ranked in order (usually from smallest to largest). The median observation should therefore have the same number of observations on either side of it. If there are an odd number of observations the median is the middle observation but if there is an even number of observations the median will be the average of the two middle scores.

To find the median of a distribution.

- Arrange all observation in order of size, usually from smallest to largest.
- If the number of observations is odd, the median will be at the centre of the ordered list. You can find the middle score by adding one to the total number of observations and dividing by two. Then count up through the ordered set until you reach that observation.
- If the number of observations is even, the median will be midway between the two centre observations.

Example

The amount 5 people earned per hour is shown below.

$12, $18, $14, $11, $15

Find the median amount earned.

11 12 14 15 18 Rearrange data into ascending order

As there are 5 observations the median position will be $\frac{5+1}{2} = 3$

If we count along the data set the 3rd value is $14

The median is therefore $14.

Note that there are the same number of observations below 14 as there are above 14.

Example

Find the median of this maximum temperature data collected over 6 days during winter.

6.5 7.1 3.2 5.3 9.2 4.5 (all temperatures are in degrees Celsius)

3.2 4.5 5.3 6.5 7.1 9.2 Rearrange data into ascending order

As there are 6 observations the median position will be $\frac{6+1}{2} = 3.5$

This means that the median lies between the 3rd and 4th observation. To find the exact value of the median we take the midpoint between the 3rd and 4th value.

$\frac{5.3+6.5}{2} = 5.9$

The median is thus 5.9 degrees Celsius

Example

In a previous activity we looked at the distribution table which displayed daily temperature over a 21 day period. What would be the median of these data?

Temperature in degrees Celsius	Frequency of each temperature
25	5
27	7
29	3
31	4
33	2

The data are already arranged in ascending order, so the first step is to find the position of the median. The median will be the $\frac{21+1}{2}$ term, i.e. the 11th term.

To find the 11th term count down the 11 terms in the frequency distribution table. For example the first 5 terms are all 25, the next 7 are all 27 so that means that the 11th term is 27.

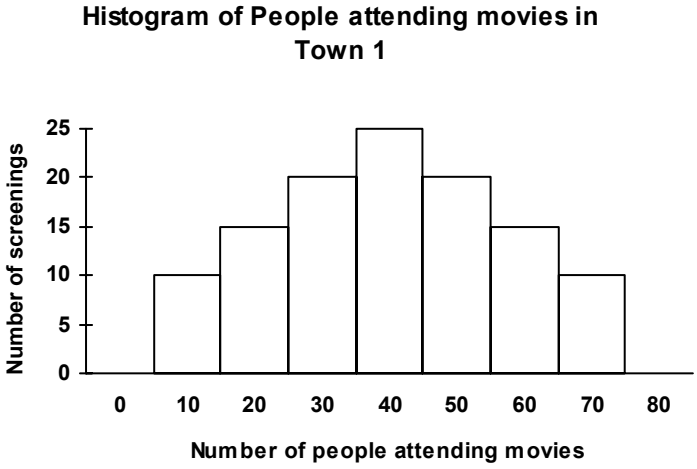
The median is 27 degrees Celsius.

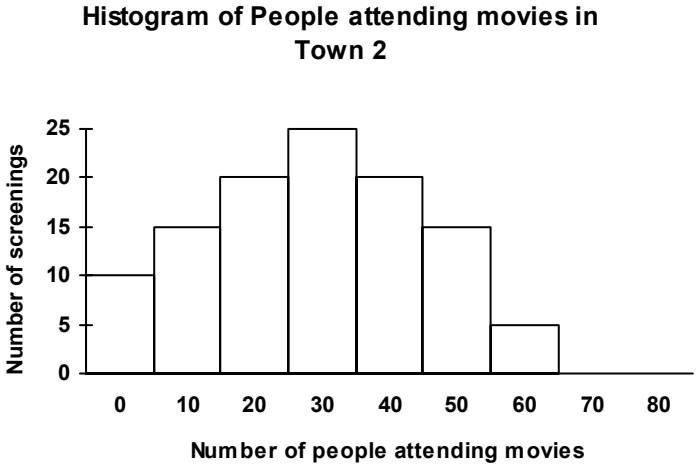
Activity 6.9

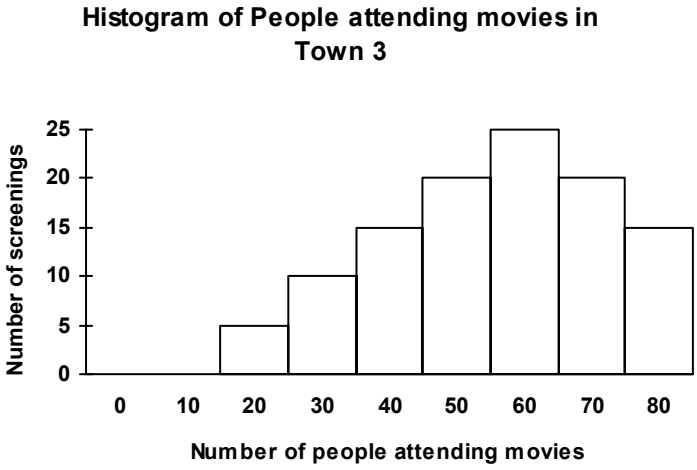
Find the median observation in the distributions detailed in activity 6.7.

A comparison of mean, median and mode

Now that we know how to calculate the mean, median and mode it is important to see when each of these measures is most useful. Let's look at the following three histograms of frequency distributions. The distributions represent the number of people attending three small cinemas around country Queensland.







Notice that although the histograms for these three distribution are about the same thing they have some distinct differences. Their general shapes are different. The first histogram is symmetrical (when it is cut in half both sides are the same). In the second histogram the rectangles are higher on the left hand side while on the third histogram the rectangles are higher on the right hand side. Also compare the values of the mean, median and mode for the three distributions. In the symmetrical histogram the mean, median and mode are the same, while in the unbalanced (or skewed) histograms the mean is different from the median and mode. This type of result will occur in many other distributions that are skewed to one side. In some cases the median and mode will also be different. It is often said that the mean is more liable to be influenced by extreme values in a distribution. However, it should not be concluded that because of this the mean is not useful. All three measures of central tendency measure slightly different aspects of a distribution and as such all are useful. In particular,

The mode is most useful when:

- categorical variables are being considered; and
- qualities like sizes of products are being considered e.g. a manager of a supermarket will always be interested in the most frequently bought size rather than the mean size.

The median is most useful when:

- the distribution has values which are either very small or very large (outliers) e.g. reports about incomes, house prices or other very skewed distributions usually use the median rather than the mean.

The mean is most useful because:

- it uses all the numbers in the calculation and is sensitive to small changes;
- many people intuitively understand the meaning of this measure; and
- many rigorous statistical theories have been developed around this measure.

In conclusion all three measures of central tendency have their advantages and disadvantages. When selecting the measure to summarise data it is always necessary to first consider what the statistics will be used for.

Activity 6.10

1. Below is the number of people in the same family undergoing counselling after a bus crash.

   4 2 2 4 6 3 2 3

   As the psychologist in charge you have to present a report on this event. Calculate the mean, median and mode for these data and comment on the most appropriate measure to use.

2. Levels of a drug in the blood of ten patients admitted to the outpatients clinic of a major hospital were

   75 103 66 61 91 61 74 103 99 93 (microunits per mL).

When reporting back to your superior about the group of admissions you would use either the mean, mean or mode to summarise these data. Calculate all three measures and give reasons for your decision to use the measure of your choice.

3. Eight houses all in the same street of a Toowoomba suburb had the following values

   $85 000 $115 000 $66 000 $77 000 $82 000 $89 000

   $79 000 $91 000

   Which measure of central tendency would best summarise these data. Explain your answer showing calculations for mean, median and mode.

6.4.2 How spread out are these data?

You will have noticed by now that when data is collected there is a considerable amount of variation. For example you would not expect every man of the same age to have the same weight. Take another look at the many histograms and frequency distributions we have constructed in this module so far. Many of them show considerable variation. As a consequence of this we have to be very cautious when using the mean, median or mode. We need to know something about the variation of the original data set before these measures are really useful. Consider the following.

You apply for two jobs where the amount of money you earn depends on how much you sell. You are told by both employers that the average weekly income of the other employees is $600. It sounds good so far, until you find out that at one place many of the employees earn just $50 while a few high fliers are earning $1200. Luckily you find out that at the second place salaries are more balanced with the weekly earnings always being between $400 and $700.

In this example the means for each set of data were the same but the variation in the data set was not.

It is always useful to have at least one measure of spread to accompany a measure of the centre of the data.

One simple way to measure spread is called the **range**, which is the difference between the smallest and largest values. For example in the situation above it would have been good to know that the range of weekly income for one group was $1150 ($1200 – $50) while the other was $300 ($700 – $400).

Remember that drawing a frequency distribution as a histogram or a stem-and-leaf plot would give you a good idea of how spread out the data were.

There are many other measures of spread that you will come across in other books or in your future studies but they are beyond the scope of this unit.

Activity 6.11

1. Calculate the range of the data given in activity 6.10. For each question write a sentence summarising the data including a measure of central tendency and the range.
2. The following are the marks out of 100 for two students (Joe and Chris) who sat for their final examination in 6 subject areas:

   Joe: 66 65 45 66 65 90

   Chris: 66 66 65 68 64 67

   (a) Calculate the mean and range of these data for each child.

   (b) Use this information to compare the average mark of the two students.

6.5 Data with two variables

So far we have looked at data sets where only one variable has been measured, for example height of men or temperature of water. But what about the situation that often occurs where we measure two variables and want to know something about the relationship that occurs between them.

Consider the situation of four men:

Adam is tall and light. He is 180 cm tall and weighs 60 kg.

Bill is the same height as Adam (180 cm) but weighs 110 kg. He is the heaviest of the four men.

Charles is short and heavy. He is 140 cm tall and weighs 80 kg.

Don is the shortest and the lightest of the four men, he is only 120 cm tall and weighs 50 kg.

Two variables are presented in these data, height and weight. We could summarise this type of data by representing the two variables together on a graph. It may well be that there is some relationship between the height and weight of a man.

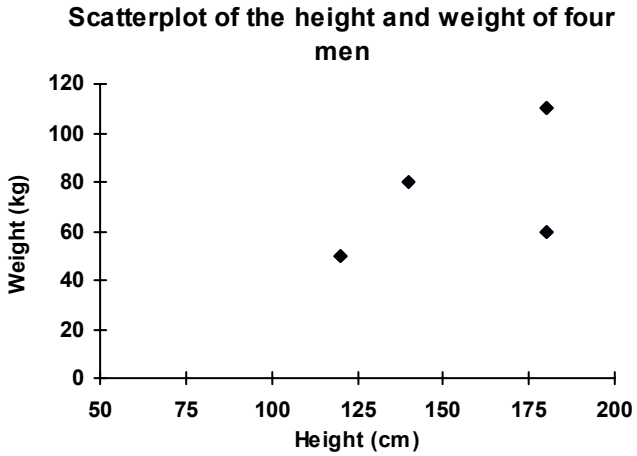
The best way to begin the investigation involving the relationship between these two variables is to draw a **scatterplot**. A scatterplot is a graph in which the values of one variable appear on the horizontal axis while those of the other variable appear on the vertical axis. The points drawn on the graph represent the values of each variable for that individual. From the data above we could construct a table that would help us plot the points on our scatterplot.

Name	Height (cm)	Weight (kg)
Adam	180	60
Bill	180	110
Charles	140	80
Don	120	50

To construct the scatterplot for these data you should:

- Draw a Cartesian plane, using only the necessary quadrant (best to use graph paper).
- Label the horizontal axis as height (cm) and the vertical axis as weight (kg). It normally doesn't matter which variable goes on which axis.
- Mark a scale on each axis that is appropriate for the range of each variable. In this case height could go from 50 to 200 and weight from 0 to 120.
- Plot a point on the graph for each pair of observations.
- Give the graph a title.

The scatterplot from the above data could look like this:

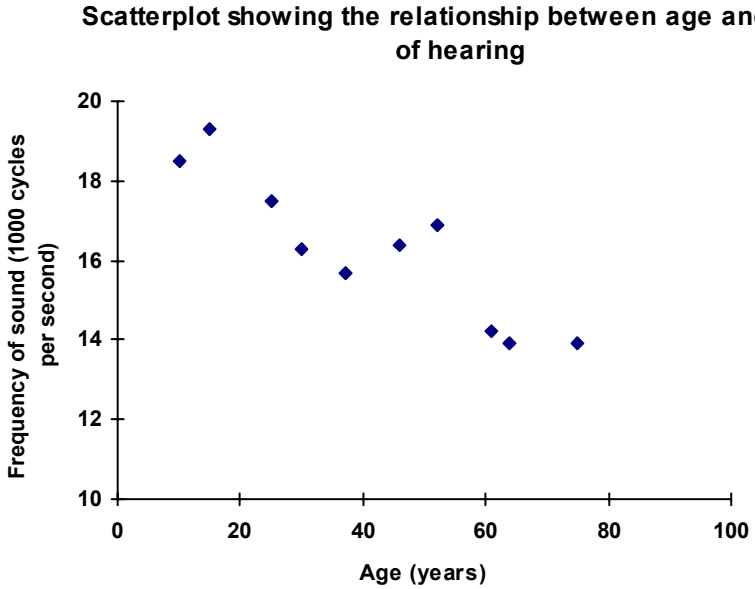


Let's consider another example.

Example

Ten people of different ages had their hearing checked to determine the relationship between age and hearing loss. Hearing was measured as the maximum frequency of the sound they could hear in 1000 cycles per second.

Person	Age (years)	Frequency (1000 cycles per second)
1	10	18.5
2	15	19.3
3	25	17.5
4	30	16.3
5	37	15.7
6	46	16.4
7	52	16.9
8	61	14.2
9	64	13.9
10	75	13.9



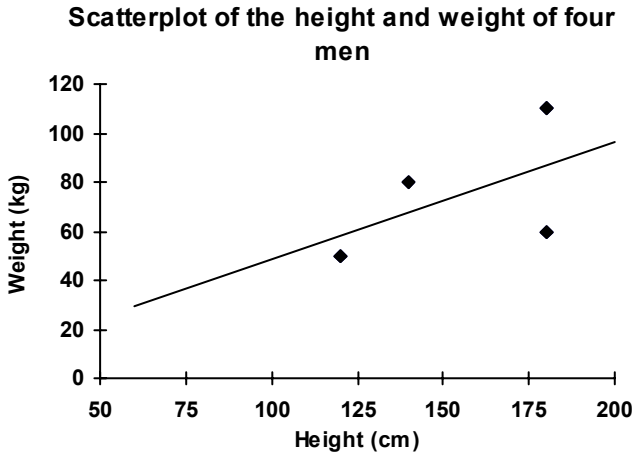
Important note: only the data for age and frequency are included in the plot. Do not include the person number. This is only included as a label for each observation and should not be a part of the scatterplot.

Before we go any further and draw some scatterplots, let's have a look at what these graphs tell us about the relationship between the two variables graphed.

In the first graph it looks like there is a relationship between the height and weight of a man. In particular as the height increases so does the weight. We can also see a relationship in the second graph. In this case as the age of the person increases their ability to hear decreases. Is it possible to make predictions about weight or hearing from these graphs?

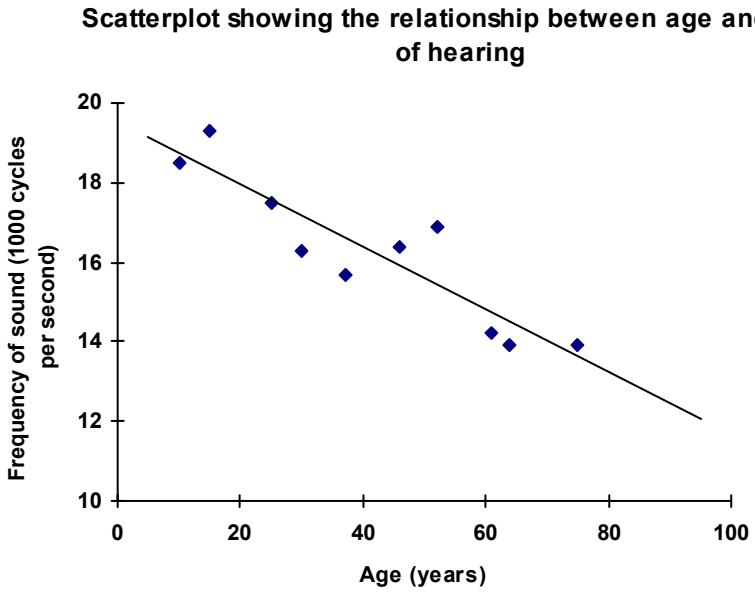
One thing that will help us understand what is happening in the graph is to draw a line through the points. This line is called the **line of best fit** and will give us a good idea of the **trend** the data follow. Ideally the line should be drawn through the middle of the points, dividing them into two balanced groups (one group each side of the line). This should be done by 'sighting' a line – taking your ruler and drawing a line that appears balanced. This method is only an estimation of where the line could be and different people will produced slightly different lines. If you study statistics further the guess work is taken out of this method by a more mathematical approach to fitting a line called 'Regression Analysis'. However, for our purposes a line of best fit, fitted by eye will give us a good idea of the trend in the data.

Let's fit some lines to the two examples above.



When we examine the trend in this scatterplot it appears that there is a relationship between height and weight of men, that is, the taller a man is then the heavier he might be predicted to be. We say that this is a positive relationship between these two variables.

We could now use the line of best fit to make some predictions about the weight of men of heights that we did not measure. For example reading from the line of best fit when a man is 160 cm tall he might weigh about 75 kg. We could also predict that if a man was 200 cm tall (a basket ball player?) then from the line of best fit we would predict that he would weight 90 kg.

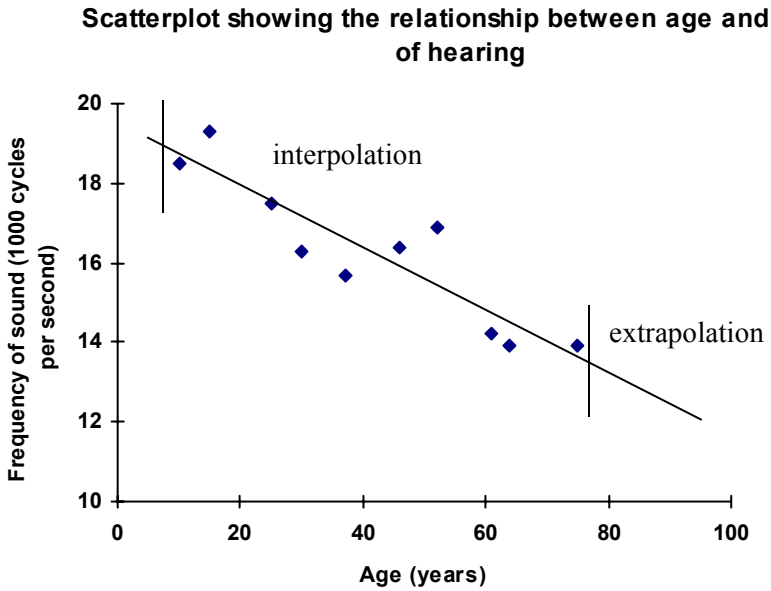


Similarly for the scatterplot of age versus level of hearing we could draw a line of best fit. Notice that this line is sloping downwards indicating that as people age the frequency of sound that they can hear is reduced (the older you are the worse your hearing is). As we did before we could use this graph to predict some values of level of hearing for people we did not measure. For example using the line of best fit we could say that if you were 42 years old you would be able to hear 17 000 cycles per second while if you were 90 years old you would be able to hear only 12 000 cycles per second.

The process of guessing values based on a known trend in some data is called:

interpolation if it is within the known range of values, and

extrapolation if it is outside the known range of values.



Watch out, however, for unrealistic predictions. For example could we predict the weight of a man who is 300 cm tall or the level of hearing of somebody 150 years old?

Activity 6.12

Draw scatterplots for the following data. Draw a line of best fit through the data points to illustrate the trend and then answer the questions.

1. The temperatures in a dam were measured at different depths producing the following data.

Depth (m)	Temperature (degrees)
0	24
1	20
2	18
3	15
4	10
5	9

   (a) Is there a trend? If so describe it.

   (b) What do you predict that the temperature would be at:

      (i) 2.5 m

      (ii) 6.0 m

2. Two students A and B sat for 5 tests in different subject areas. The teacher suspected these students of cheating so wanted to see if there was a relationship between the two results. The teacher examined the following data.

Subject	Person A's Results (%)	Person B's Results (%)
English	85	85
Mathematics	90	45
Art	40	95
History	55	80
Chemistry	80	50

   (a) Is there a trend? If so describe it.

   (b) Would this confirm the teacher's suspicions?

3. The percentage of the day that was sunshine was measured in winter and summer in a number of different towns to see if there was a relationship between percentage of sunshine in summer and winter. The following data resulted.

	Percentage Sunshine	
Town	**Summer (%)**	**Winter (%)**
A	56	62
B	61	74
C	59	68
D	58	67
E	73	79

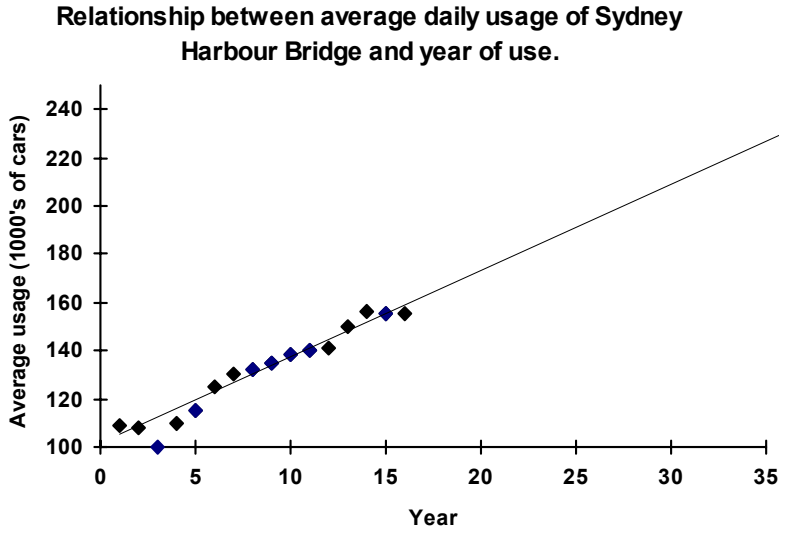
(a) Is there a trend? If so describe it.

(b) Predict the percentage summer sunshine in a town, if the percentage winter sunshine is

(i) 70

(ii) 78

4. In 1965 the NSW State Government began monitoring the annual average daily usage of the Sydney Harbour Bridge. The results of the survey are presented in the scatterplot below. The first year of the survey has been called Year 1 of the survey with the last year of the survey called Year 16.



(a) Describe in your own words the relationship between average daily usage and years.

(b) What year would be 35 years from the beginning of the survey?

(c) What would the average annual number of cars be predicted to be in this year?

(d) Will the prediction in part (c) be realistic? Explain your answer.

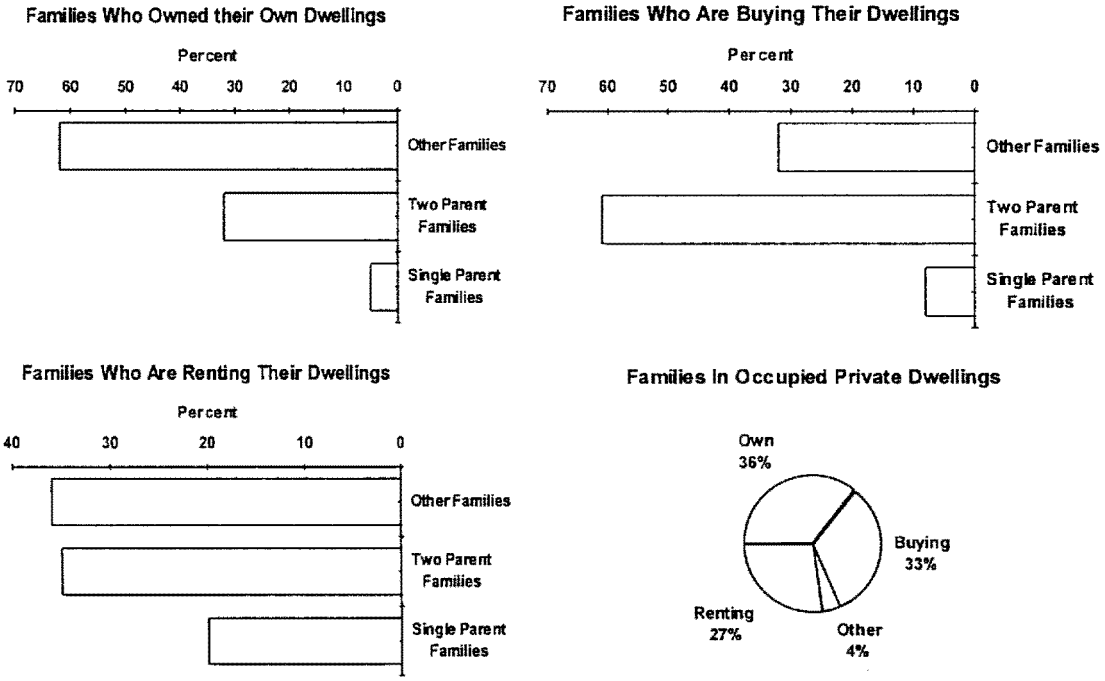
6.6 A taste of things to come

1. In a recent book about Queensland families it states:

Home ownership is one of the traditional goals of many Australians. The majority of Queensland families and households in 1991 owned their home or were paying off a mortgage. However, home ownership varies by family type, income level and life cycle stage.

Below are figures that display home ownership by family type:

In your own words summarise what you see as the type of home ownership patterns displayed in the figures below.



(Source: Australian Bureau of Statistics 1996, *Queensland Families: Facts and Figures.*)

2. An advertisement in a popular newspaper announced that great deals were to be made from purchasing a house with a view to increasing your income. You could own your own property in Brisbane and be earning substantial rents by spending as little as $100 000. The following are some examples they presented for our evaluation.

Some Examples of Market Value of House and Income per Week from Rental in Brisbane in 1997

Suburb	Price ($)	Rent per Week ($)
Albany Creek	186 000	215
Bald Hills	149 000	180
Carindale	170 000	200
Carindale	179 900	210
Carrara	192 000	230
Ferny Grove	245 000	270
Forest Lake	153 000	180
Middle Park	156 000	190
Nerang	195 000	230
Newmarket	199 000	210
Riverhills	155 000	185
Wynnum West	148 500	180

Are the claims of this advertiser backed up by the statistics they present? Answer the following questions to see what you think.

(a) What is the true relationship between house price and rent received per week? (Hint: Draw a scatterplot of the data presented with a line of best fit)

(b) If a house was worth $100 000 what rent would you receive for this property?

(c) What is wrong with the advertiser's claims based on the data they presented in the advertisement?

(d) What else would you do before you rushed to be a part of this deal?

3. In the study of communication different text books discuss the subject of communication in different ways. Here is an example from one such book.

   The telephone is extremely important in business. Workers in a research and development laboratory spend 35% of the work day talking on the telephone. Data collected on 4 086 communication episodes in a single Canadian manufacturing plant indicate the following: 526 (12.87%) of the episodes were conducted by telephone rather than face-to-face (2700; 66.08%) or in writing (860; 21.05%) .....

   The telephone is also an important part of everyday life. The average household initiates about 4 calls per day (the median is less than three calls) and between 40 and 50 percent of these calls go to other households within a 2 mile radius. The average length of a phone call is just over 4.5 minutes (the median is slightly over 1 minute). The most important single factor in boosting the number of calls in a household is the presence of a woman between the ages of 19 and 64.

   (Source: Lewis, P.V. 1987, *Organisational Communication,* Willey & Sons, New York.)

   (a) Considering the first paragraph. Can you understand what this author is saying? What would be a simple pictorial way of presenting these data for students that preferred visual learning?

   (b) Consider the second paragraph. In this section the author gives us some statistics on the mean and median number of calls and length of calls. Recall the differences between the mean and median and explain why these measures give different values.

You should now be ready to attempt questions 5–7 of Assignment 3A (see your Introductory Book for details). If you have any questions, please refer them to your course tutor.

6.7 Post-test

1. One hundred students, surveyed about their methods of reaching school in Brisbane, gave the following results:

Method	Number of students
Bus	40
Car	10
Cycle	30
Train	5
Walking	15

   (a) What type of variable are you working with in this study?

   (b) In this study there is no information on how the students were sampled. Do you think that this group of students represented a random sample? Give your reasons.

   (c) You need to represent these data graphically for a report on how students travel to school. Represent the data using a bar chart.

2. The following represents the number of accidents that 50 drivers had been involved in over a 5 year period.

1	2	4	1	6	2	1	1	1	2
0	2	0	4	1	3	2	4	0	0
2	1	3	1	5	2	3	2	0	2
0	3	4	2	2	0	4	3	2	3
2	1	0	5	1	2	2	5	6	1

   (a) Organise the data above into a frequency distribution table.

   (b) Construct a histogram to represent these data.

   (c) Find the mean, median and mode of these data

   (d) Using any of the information above do you think this sample is skewed in any way? Which measure of central tendency would tell you the most about these data?

3. Twenty nurses were asked to measure out 250 mL of a solution. The measuring was then checked and results recorded (in mL):

248.5	250.6	252.0	248.3	251.2
246.9	248.0	250.0	252.4	250.7
254.2	249.1	250.1	253.9	248.7
253.4	249.0	246.5	250.6	249.1

   (a) Using classes 246 up to and including 248 organise the data into a grouped frequency distribution table.

   (b) Use that table to draw a histogram.

   (c) Find the range of these data.

   (d) The amount of solution to be measured was 250 mL, use the histogram above and the range to comment on the accuracy of the measurements made by this group of nurses.

4. The following represent the Home Loan Mortgage Interest Rates in several countries in 1988 and 1989.

Country	Interest Rate (%)	
	August 1988	August 1989
Australia	14.5	17.0
Great Britain	12.0	14.0
United States	10.6	10.5
West Germany	6.6	7.6
Japan	5.7	6.0
Switzerland	4.9	4.9

   (a) Construct a scatterplot of these data to show the relationship between Mortgage Interest Rates in 1988 and 1989.

   (b) Draw a trend line through the data points to illustrate the trend.

   (c) If the interest rate was 9% in August 1988, in a certain country, what do you predict that it was for August 1989?

   (d) Can you think of another way to present these data graphically that would not involve a scatterplot?

6.8 Solutions

Solutions to activities

Activity 6.1

1.

(a) Gender is categorical.

(b) Age (years). If we consider whole years i.e. 1, 2 or 3 years then the variable could be discrete. If on the other hand we consider years as being able to take all parts of a year i.e. 1.27 years or 102.638 years then it will be continuous. In most instances age is considered to be a continuous variable.

(c) Income is continuous variable.

(d) Number of siblings for each pupil is a discrete variable.

(e) Temperature is a continuous variable.

(f) Smoker (yes or no) is a categorical variable.

(g) Class size is discrete because it is considering numbers of students.

2.

(a) Population could be all second language students

(b) Sample is all 1st year university second language students

(c) Variables of interest could be:

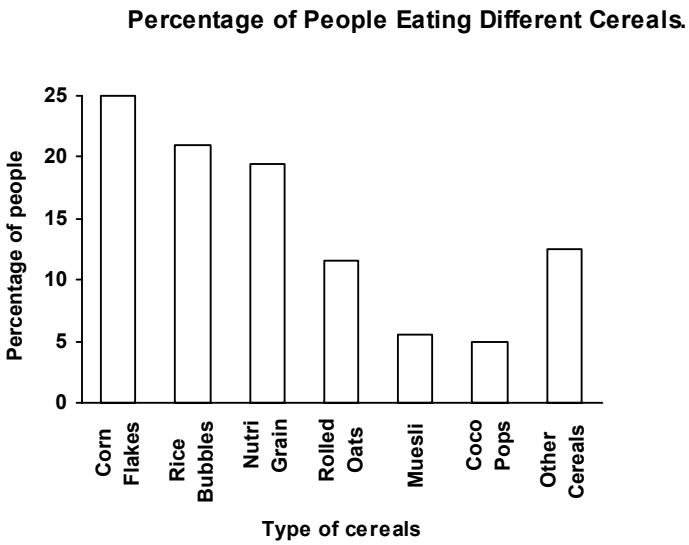
- number of words in vocabulary;
- number of complex sentences in written expression;
- number of questions correct in a test.

There are numerous other variables that could be considered. These will depend on the interests of the researcher.

Activity 6.2

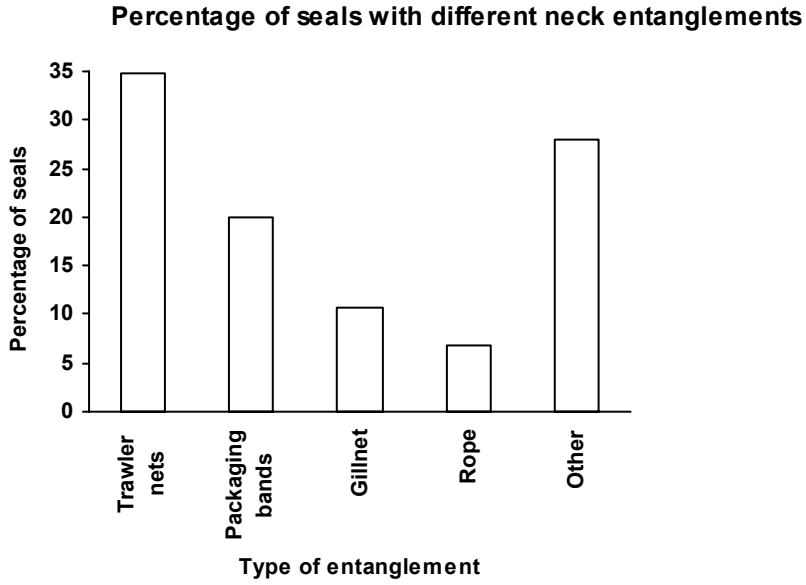
1.

Cereal	Number of people	Percentage of people
Corn Flakes	50	25
Rice Bubbles	42	21
Nutri Grain	39	19.5
Rolled Oats	23	11.5
Muesli	11	5.5
Coco Pops	10	5
Other Cereals	25	12.5

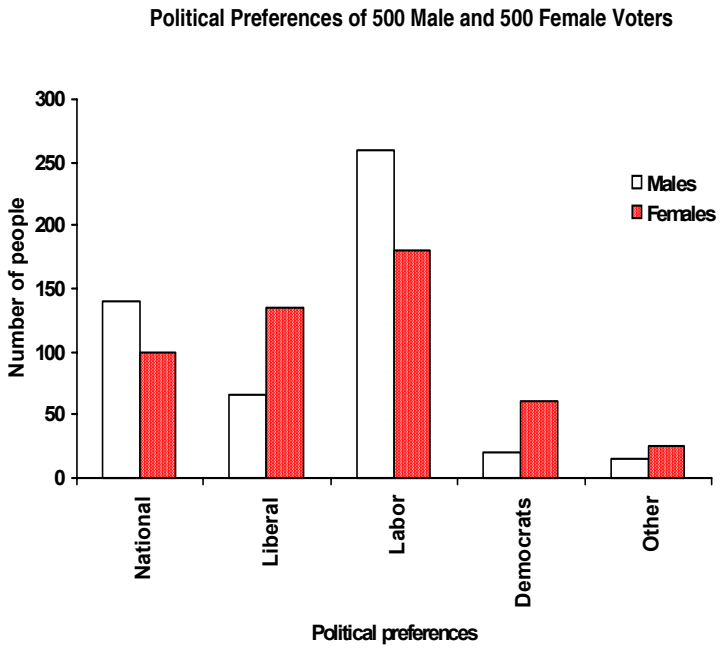


2.

Item	Number	% of Total
Trawler nets	26	34.7
Packaging bands	15	20.0
Gillnet	8	10.7
Rope	5	6.7
Other	21	28.0
Total	**75**	**100.1**



3. (a)



(b)The most popular party in this city was the Labor party. More males than females voted for this group. The National and the Liberal parties were the next most popular. The National party was the most popular with men while the Liberal party was the most popular with women. The remaining parties (Democrat and others) were the least popular, with more women voting for these in both cases.

4. (a) Both graphs present the same data but in different ways. The bar chart presents the raw data. Many find these types of figures hard to compare but the bar chart which has rectangles next to each other makes this type of comparison easier. The pie chart does not present the raw data it uses percentages. The problem with this type of graph is that we do not know how many workers the graph is based on. For example if the chart was only based on 6 workers then it is not based on a very good sample. We cannot tell this from the graph as it is presented here. Note that some readers will prefer different types of graphs depending on their past experiences e.g. it could depend on whether they like to read from rectangles or circles?

(b) There are a range of answers you could provide to this question. To be acceptable your answer should cover the following points:

- educational backgrounds divided into 6 categories;
- most workers had either year 12 or a university degree (385 out of 550) with year 12 being the most common;
- the least number of workers had year 10 (15 out of 550);
- postgraduate qualifications achieved by 125 out of 550 workers with doctorate being more common than master's degree;
- the educational backgrounds of one group of workers could not be classified (25 out of 550).

Activity 6.3

1. (a) Lowest weight 56 kg, greatest weight 70 kg.

   (b)

Frequency distribution table of weights (kg) of students in a mathematics tutorial

Weight (kg)	Frequency
56	1
57	0
58	3
59	1
60	2
61	2
62	2
63	6
64	3
65	1
66	3
67	3
68	2
69	1
70	2
	$\sum f = 32$

   (c) 2 students weighed 70 kg.

   (d) Totalling the frequencies of categories less than 62 kg we get, 1+0+3+1+2+2=9, 9 students weighed less than 62 kg.

2. (a)

Frequency distribution table of golf scores in 20 games

Golf score	Frequency
78	1
79	3
80	2
81	4
82	2
83	2
84	2
85	2
86	1
87	1
	$\sum f=20$

(b) 2 rounds

(c) Totalling the frequencies in groups less than 80 we get 1 + 3 = 4.

Golfer scored less than 80 on 4 rounds.

3. (a)

Frequency distribution table of blood groups of 24 babies.

Blood groups	Frequency
A	9
B	3
AB	1
O	11
	$\sum f=24$

(b) Most common blood group was O, with 11 babies.

(c) Percentage of O group in this sample was $\frac{11}{24} \times 100\ \% = 45.8\ \%$

4. (a)

Frequency distribution table of speed of cars caught by radar trap

Speed (km/h)	Frequency	Speed (km/h)	Frequency	Speed (km/h)	Frequency
70	1	81	6	92	
71	2	82	1	93	1
72		83	3	94	
73	4	84	1	95	
74	1	85	2	96	1
75	2	86	1	97	
76	4	87		98	
77	1	88		99	
78	2	89		100	1
79	2	90	1	101	
80	2	91		102	1
					Σf=40

Note: To save space this frequency distribution table has columns placed side by side.

(b) 19 motorists travelled above the speed limit of 80 km/h.

5. (a)

Frequency distribution table of heights of 50 clients at a shopping centre

Height (cm)	f	Height (cm)	f	Height (cm)	f	Height (cm)	f
88	1	111		135	1	159	
89		112		136	1	160	2
90		113		137		161	
91		114		138		162	
92	3	115	2	139		163	
93		116	2	140		164	2
94	1	117		141		165	3
95	2	118		142		166	
96		119		143	1	167	
97		120		144		168	
98		121		145	1	169	
99		122	2	146		170	
100		123		147		171	
101		124		148	2	172	
102		125	1	149		173	
103		126		150	3	174	
104		127		151		175	
105		128	2	152		176	3
106	1	129		153		177	
107		130		154		178	2
108		131		155	3	179	
109		132	1	156	1	180	2
110		133		157		181	
				158	2	182	2
						183	
						184	
						185	1
							$\Sigma f=50$

Note: To save space this frequency distribution table has columns placed side by side.

(b) Definitely, we could either use a computer to do this time consuming activity or present the data in a grouped distribution. See the next section of work.

Activity 6.4

1.

Frequency distribution table of weights (kg) of students in a mathematics tutorial

Weight (kg)	Frequency
55 up to and including 57	1
57 up to and including 59	4
59 up to and including 61	4
61 up to and including 63	8
63 up to and including 65	4
65 up to and including 67	6
67 up to and including 69	3
69 up to and including 71	2
	$\sum f=32$

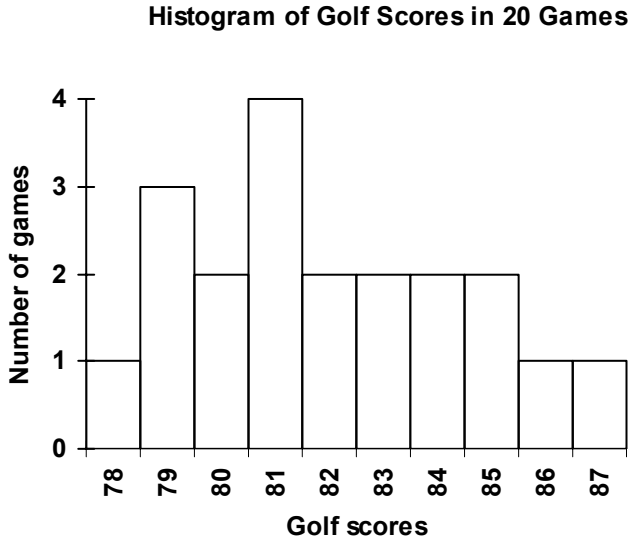
Activity 6.5

1. (a)

Frequency distribution table golf scores in 20 games

Golf score	Frequency
78	1
79	3
80	2
81	4
82	2
83	2
84	2
85	2
86	1
87	1
	$\sum f=20$

(b)

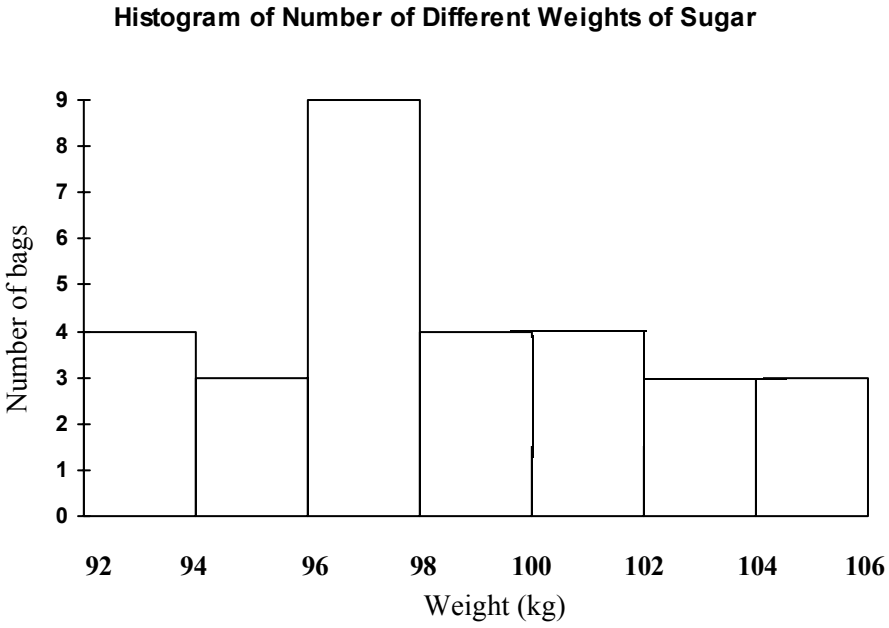


2. (a)

Frequency distribution table of weights of sugar

Weight of sugar (kg)	Frequency
92 up to and including 94	4
94 up to and including 96	3
96 up to and including 98	9
98 up to and including 100	4
100 up to and including 102	4
102 up to and including 104	3
104 up to and including 106	3
	$\sum f=30$

(b)

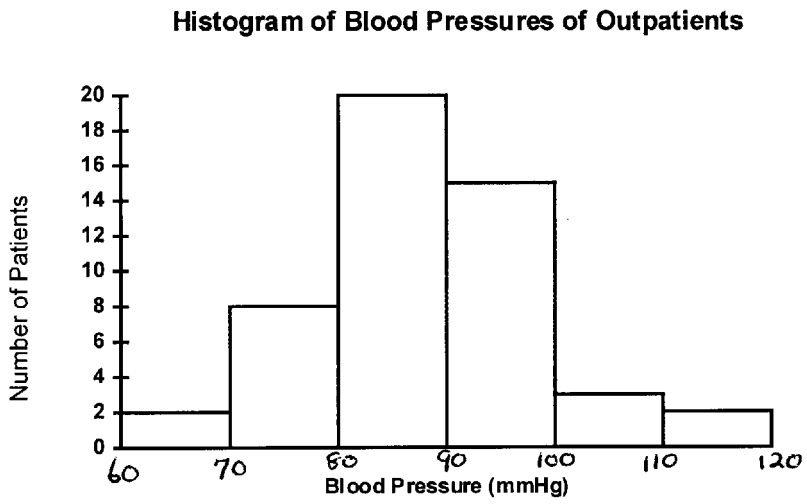


3. (a)

Frequency distribution table for blood pressures (mmhg) for outpatients at a local hospital

Blood Pressure (mmHg)	Frequency
60 up to and including 70	2
70 up to and including 80	8
80 up to and including 90	20
90 up to and including 100	15
100 up to and including 110	3
110 up to and including 120	2
	$\sum f=50$

(b)



4. (a) This answer could be expressed in a number of different ways but each should include:

- most frequent number of letters per word is 4 (approximately 2400 words);
- majority of words are less than 5 letters long;
- long words (greater than 10 letters) are rare;
- most of the distribution lies to the left of the graph;
- sample only includes words between 2 and 12 letters long.

(b) A number of different answers are possible but should all include the following points:

- both plays contain words between 2 and 12 letters;
- in Play A most common word length is 4 letters (approx. 2400 words);
- in Play B most common word length is 7 letters (approx. 2400 words);
- In Play A majority of words are between 2 and 5 letters long;
- in Play B the majority of the words are between 5 and 7 letters long;
- distribution from Play A lies to the left of the graph and in Play B lies to the middle of the graph.

Activity 6.6

1. (a)

Stem-and-leaf plot of number of passengers on 25 ferry runs

6	3 represents 63 people
3	5
4	0 1 9
5	0 2 3 7 8 8
6	0 1 2 4 5 6
7	1 5 7 8
8	2 3 4 9
9	5

(b) Most frequently carried is 58 passengers.

2. (a)

Stem-and-leaf plot of percentage scores in maths test of 20 students

2	5 represents 25 students
3	2
4	2 6 7
5	2 2 7 8 8
6	1 4 7 8
7	3 4 6 6 9
8	4 5

(b) A number of different answers are possible but all should included the following points

- most frequent scores were 76 and 52 and 58, all occurring twice
- most scores occurred between 50 and 60 and between 70 and 80 (5 each)
- scores ranged from 32 to 85

(c) 16 students passed the test. Stem-and-leaf plots make this easy to see as we can just count the scores above 50 from the plot.

3. (a) Lowest rate of heart disease is 320 men per 100 000; highest rate of heart disease is 452 men per 100 000.

(b) Most frequent rate of heart disease in females is 331 per 100 000.

(c) A number of answers are possible here but all should include:

- male rates extend between 320 and 452 while females are between 279 and 390 per 100 000 (i.e. lower than the males);
- three regions had male rates between 410 and 420 while 3 regions had female rates between 330 and 340.

Activity 6.7

1. Mean $= \dfrac{7.2 + 13.5 + 2.1 + 25 + 4.6 + 0 + 7.2 + 15.7}{8} \approx 9.4$

   Mean is approximately 9.4 mm.

2. To calculate the mean use this frequency distribution table and construct a further column for *fx*. Note you could also use the calculator to do all of this if you desired.

Temperature in degrees Celsius (x)	Frequency of each temperature (f)	fx
25	5	125
27	7	189
29	3	87
31	4	124
33	2	66
	Σf=21	Σfx=591

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{591}{21} \approx 28.14$$

   Mean temperature is approximately 28.14 degrees Celsius.

3. Mean is the sum of all the observations divided by the total number of observations. In this case it is 257 divided by 40. The mean is thus approximately 6.425 fish. You may have found this answer by constructing a frequency distribution table.

Activity 6.8

1. Mode is 7.2 mm

2. Mode is 27 degrees Celsius

3. First it is best to construct a frequency distribution table.

Frequency distribution table of number of fish caught at a fishing competition

No. of fish	Frequency
1	0
2	2
3	4
4	2
5	3
6	3
7	16
8	5
9	2
10	3
Total	$\Sigma f = 40$

Mode is 7 fish.

Activity 6.9

1. Place the measurements in order 0 2.1 4.6 7.2 7.2 13.5 15.7 25

   Median position will be $\frac{8+1}{2} = 4.5$.

   The median lies between 4th and 5th value $\frac{7.2+7.2}{2} = 7.2$

   The median is the 7.2 mm.

2. These readings are already in order, so we can read straight from the frequency distribution table. Median position will be $\frac{21+1}{2} = 11$. The 11th reading is 27.

   Median is 27 degrees Celsius.

3. Using the frequency distribution we constructed for the previous question we can again read straight from the frequency distribution table. Median position will be $\frac{40+1}{2} = 20.5$. Median lies between the 20 th and 21 st number. The 20 th number is 7 and the 21 st number is also 7. Median is thus 7 fish.

Activity 6.10

1. First arrange these data in order.2 2 2 3 3 4 4 6

   Mean = $\frac{2+2+2+3+3+4+4+6}{8} = 3.25$ people

   Mode is 2 people.

   Median: median position will be $\frac{8+1}{2} = 4.5$, the median will lie between the 4th and 5th terms. By counting along the row of numbers we can see that the median is 3.

   Median is 3 people.

   Mean, median and mode are all different. In this instance it may be that the median is the best measure of central tendancy because of the effect of the 6.

2. First arrange these data in order. 61 61 66 74 75 91 93 99 103 103

   Mean will be 82.6 microunits per mL.

   Modes are 61 and 103 microunits per mL (note that in this instance there are two modes, it is bimodal)

   Median lies between 5th and 6th terms i.e. between 75 and 91. We can find the midway point between these two numbers by adding them and then dividing by 2, $\frac{75+91}{2} = 83$.

   The median is 83 microunits per mL.

   In this instance either the mean or median would have been useful. Because the distribution has two modes at either end of the distribution the mode is not useful as a measure of central tendancy.

3. First put these data into order

   66 000 77 000 79 000 82 000 85 000 89 000 91 000 115 000

   Mean is $85 500.

   Mode: there is no distinct mode as there are no repeated values.

   Median: the position of the median will be the $\frac{8+1}{2} = 4.5$ term. That is, the median will be between the 4th and 5th terms. The values of these terms are 82 000 and 85 000. Midway between these terms is $\frac{82\,000 + 85\,000}{2} = 83\,500$.

   The median is $83 500.

   In this instance because of the extreme value of $115 000 the median would be the best measure of central tendancy. Note that this is often the case with house and property prices.

Activity 6.11

1. From data in activity 6.10.

   (i) mean = 3.25, range = 2 to 6.

   The mean number of people in a family undergoing counselling is 3.25, while between 2 and 6 in each family underwent counselling.

   (ii) mean = 82.6, range = 61 to 103

   The levels of drug in the blood ranged between 61 and 103 microunits per mL, with the mean value being 82.6 microunits per mL.

   (iii) median = \$83 5000, range = \$66 000 to \$115 000

   House prices ranged between \$66 000 and \$115 000 with the median being \$83 500.

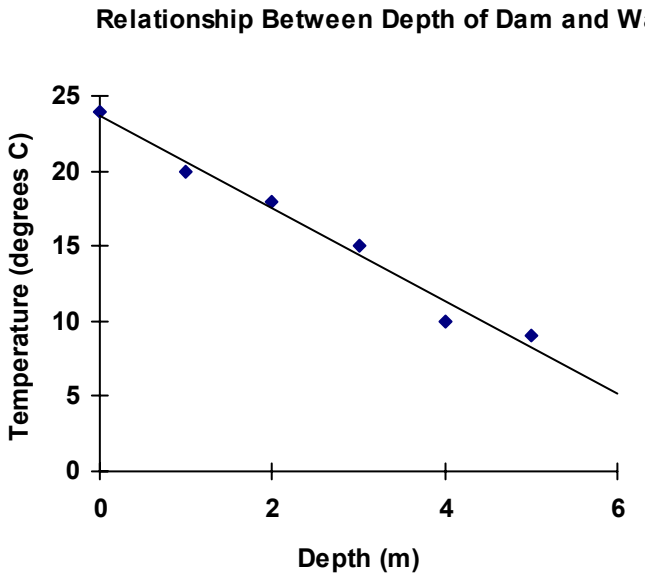
2. (a) Joe: mean ≈ 66.2 range = 45 to 90

   Chris: mean = 66.0 range = 64 to 68

   Although the mean scores for Chris and Joe are similar, the ranges are quite different. Joe scored between 45 and 90 while Chris scored between 64 and 68. It appears that Chris's exam scores are very consistent, while Joe has scored very high and very low in two subjects. In this instance it may not be desirable to compare the means alone, without taking into account the ranges from which they were developed.

Activity 6.12

1.

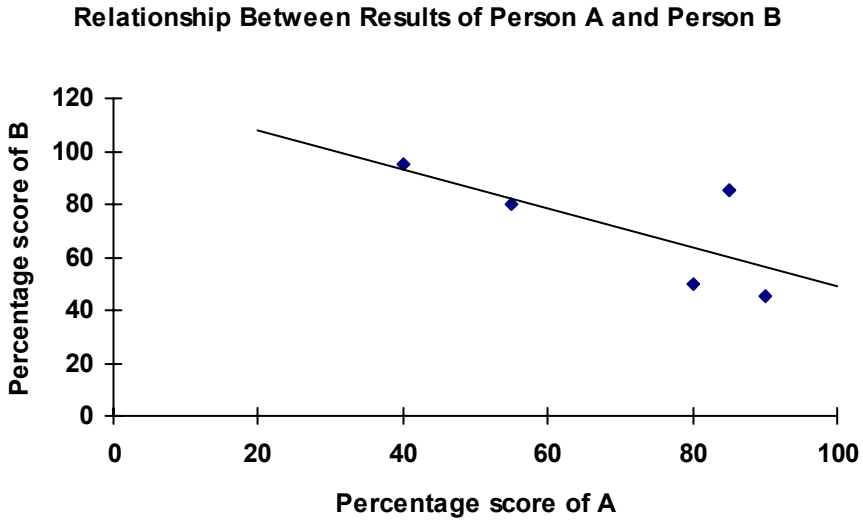


(a) Yes there is a trend. From the trend line we can see that as the depth of water in the dam increased the water temperature decreased.

(b) (i) When depth is 2.5 m, temperature would be approximately 15 degrees C.

(ii) When depth is 6.0 m, temperature would be approximately 6 degrees C.

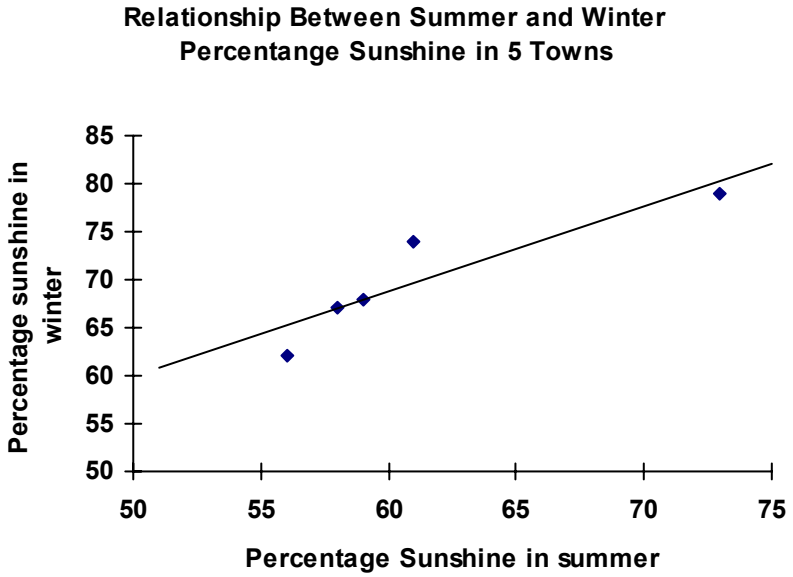
2.



(a) Yes there is a trend. As the marks for person A increased the marks for person B decreased.

(b) The above trend would suggest that cheating was unlikely.

3.



(a) Yes there is a trend. Percentage sunshine in summer and winter is related if the percentage sunshine in summer increases then the percentage sunshine in winter will also increase.

(b) (i) If winter percentage is 70 then the summer percentage should be approximately 60.

(ii) If the winter percentage is 78 then the summer percentage should be approximately 70.

4. (a) A number of answers would be possible, here is an example.

   The relationship between years of survey and average annual usage indicates that as years increased the number of cars surveyed increased steadily. When the survey commenced cars totalled 110 000 while 16 years later they totalled 150 000 cars per day.

   (b) 2 000

   (c) Using the trend line, when year 2 000 the number of cars would be 222 000 cars per day (approximately).

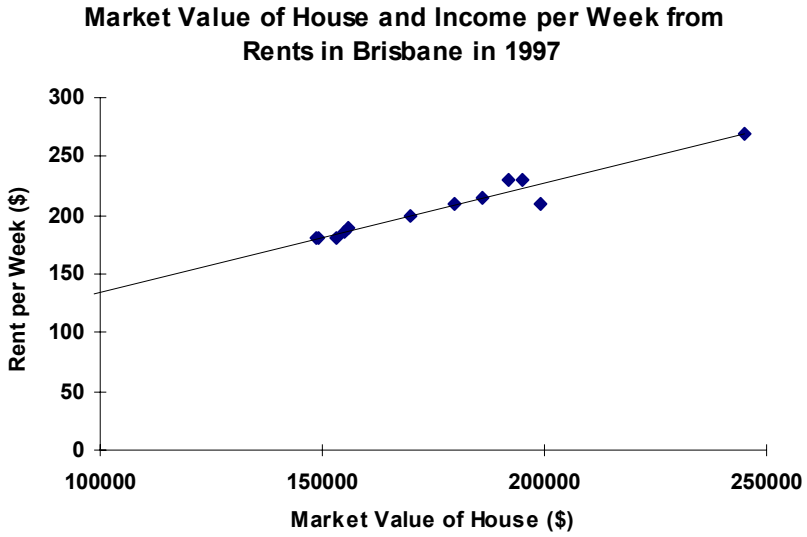
   (d) One should be cautious when making predictions well into the future away from that last available data point. The early data suggests that the relationship increases steadily however, with a gap of 17 years between the last measurement and the prediction, one doesn't know what other variable could have now come into play. For example an unexpected surge in population could cause car number to rise greatly, or changes in public transport or availability of parking could cause car numbers to stagnate or drop.

Solutions to a taste of things to come

1. A number of answers are possible for this question. Your answer should be something like the following.

   In 1991 the majority of Queensland families owned their own home (36%). Single parent families were the least likely to own a home (5%) with other types of family being more likely. Families who were buying their own home made up 33% of the sample and this group was made up predominately by two parent families. The third most frequent type of home ownership was renting. This group was made up predominately by two parent and other families (35% and 36% respectively). Single parent families represented a higher percentage in this type of occupation than in the other categories (20% compared with 10% and 5%). To summarise the patterns of home occupation and ownership were very different between different family types. Two parent families were mainly buying their dwelling, single parent families were renting and other family types owned their own homes.

2. (a) The relationship is represented by the scatter plot. The line of best fit indicates that as market value of the house increased so did the rent per week.



   (b) Approximately $135 per week.

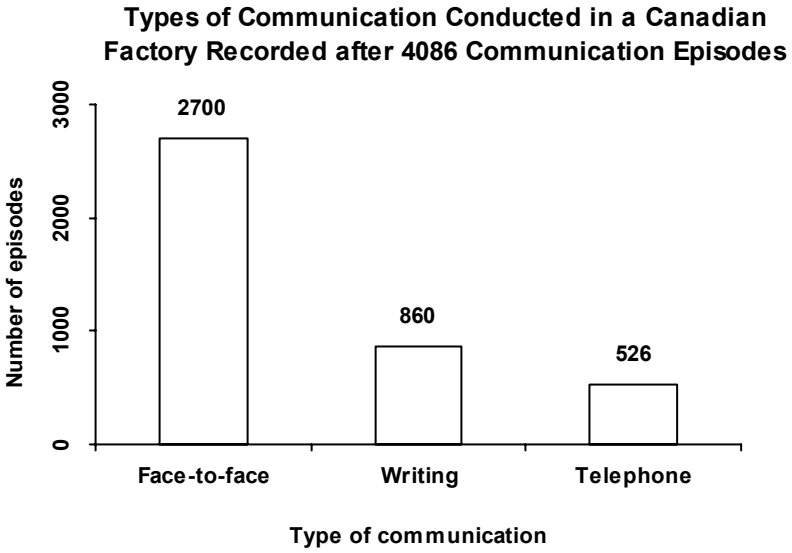
   (c) Advertiser claims that you would be earning substantial rents. You might look quickly at the figures and think that you could be earning rents equivalent to those in the table. This may not be the case. Also the $135 per week figure is based on a prediction well outside the extent of the available data. Caution should always be used in such cases as other assumptions should be considered.

   (d) There are a number of questions that you might care to ask. For example:

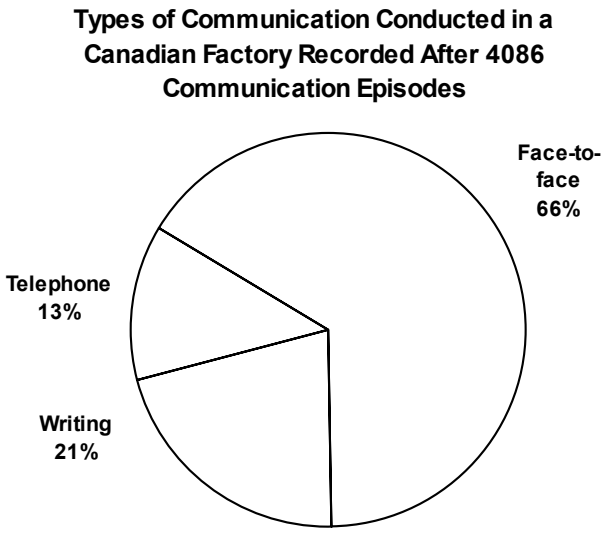
   - are there any hidden costs apart from the cost of the house;
   - what are other rents in the area; and
   - where is the house, perhaps rent is affected if it is located next to the local garbage dump;

The trend line only represents two variables, other factors not measured could come into play when predictions are made outside the extent of the measurements.

3. (a) Here are a couple of suggestions



OR



(b) When number of calls are considered the mean and median are similar indicating that there are no extreme values in the number of calls i.e. in the average household nobody is making 50 or 100 calls per day.

	Mean	Median
Number of calls per day	4	3
Length of calls (min)	4.5	1

On the other hand when you consider length of call the mean and median are quite different indicating that there are some instances when calls of very long length are made pushing the mean up to 4.5 minutes but leaving the median at 1 minute.

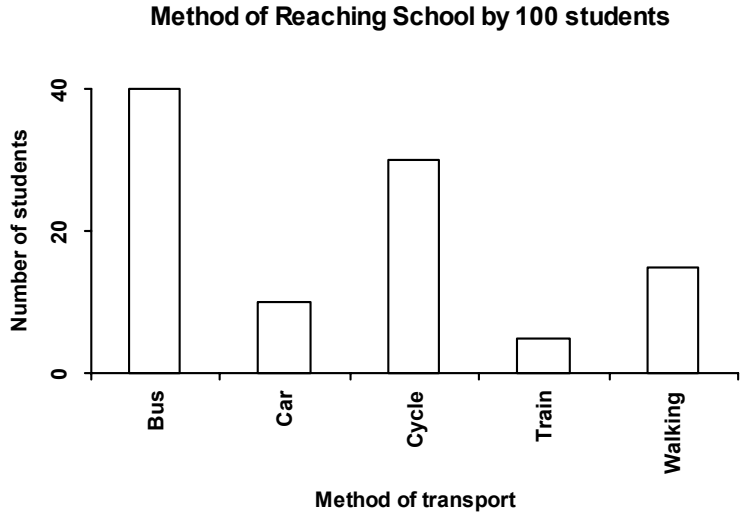
(Remember that the mean is said to be affected by extreme values while the median is not.)

Solutions to post-test

1. (a) Discrete

   (b) One would hope that the sample was randommly selected from all the schools in Brisbane otherwise it would be very biased towards certain locality. Questions should be put to the researcher to confirm this.

   (c)

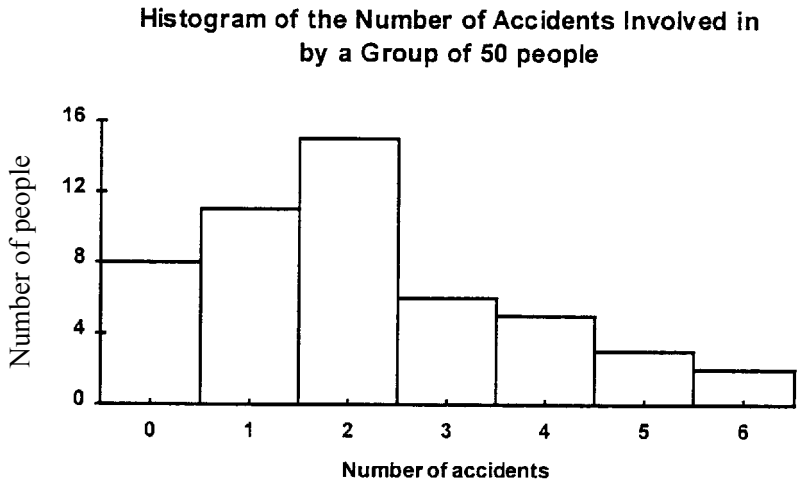


2. (a)

Frequency distribution for number of accidents in a group of fifty people

Number of Accidents	***f***	***fx***
0	8	0
1	11	11
2	15	30
3	6	18
4	5	20
5	3	15
6	2	12
	Σf=50	Σfx=106

(b)



(c) Mean is 2.12, Mode is 2 and Median is 2.

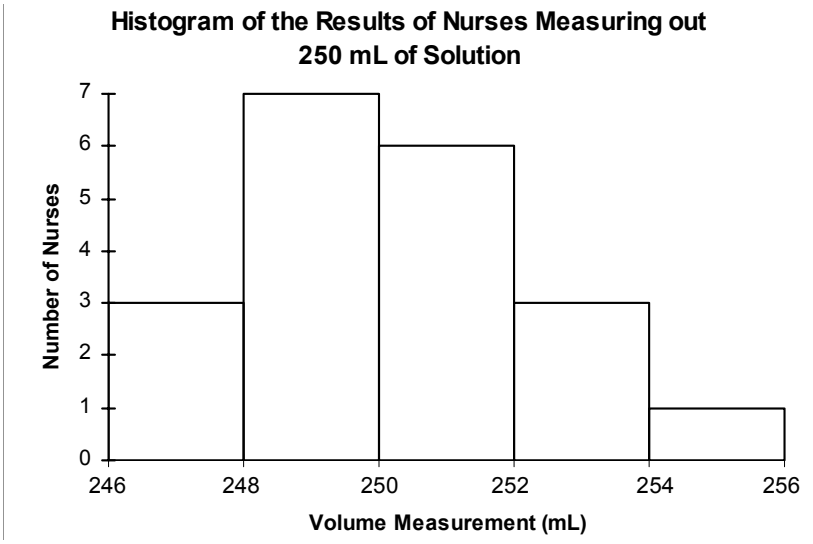
(d) All three measures of central tendancy tells us about the centre of the data and all three measures are similar. The only differences is the mean which is slightly different from the median reflecting the slightly skewed nature of the data to the right.

3. (a)

Frequency distribution of the results of nurses measuring out 250 ml of solution

Class	f
246 up to and including 248	3
248 up to and including 250	7
250 up to and including 252	6
252 up to and including 254	3
254 up to and including 256	1
	$\Sigma f=20$

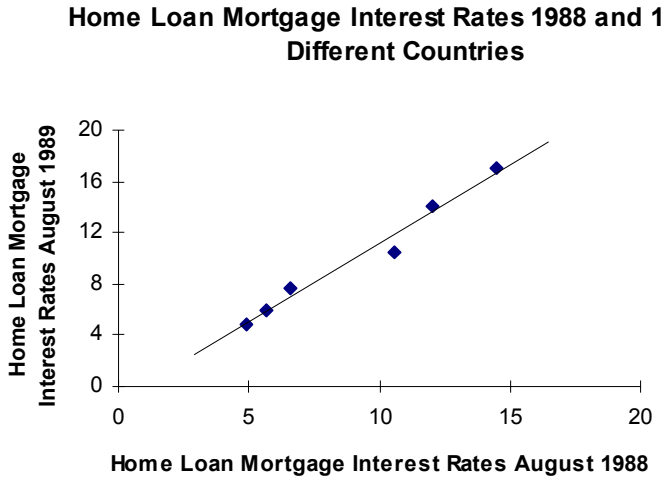
(b)



(c) Range is 254.2 – 246.5 = 7.7 mL

(d) The extent of the range of the data and the patterns in the histogram suggest that a number of nurses were quite inaccurate in their measurements of the sample. Seven of the 20 students made measurements over 2 mL outside the required standard.

4. (a) and (b)



(c) Approximately 10%

(d) Another way to present these data would be in a paired bar chart.

A Comparison of Home Loan Mortgage Interest Rates Between August 1988 and August 1989 for Different Countries