Introduction to Descriptive Statistics

Jackie Nicholas

Mathematics Learning Centre
University of Sydney
NSW 2006

Acknowledgements

Contents

1 Measures of Central Tendency

1.1 The Mean, Median and Mode

When given a set of raw data one of the most useful ways of summarising that data is to find an average of that set of data. An average is a measure of the centre of the data set. There are three common ways of describing the centre of a set of numbers. They are the mean, the median and the mode and are calculated as follows.

The mean	−	add up all the numbers and divide by how many numbers there are.
The median	−	is the middle number. It is found by putting the numbers in order and taking the actual middle number if there is one, or the average of the two middle numbers if not.
The mode	−	is the most commonly occurring number.

Let's illustrate these by calculating the mean, median and mode for the following data.

Weight of luggage presented by airline passengers at the check-in (measured to the nearest kg).

$$18 \qquad 23 \qquad 20 \qquad 21 \qquad 24 \qquad 23 \qquad 20 \qquad 20 \qquad 15 \qquad 19 \qquad 24$$

$$\text{Mean} = \frac{18 + 23 + 20 + 21 + 24 + 23 + 20 + 20 + 15 + 19 + 24}{11} = 20.64.$$

Median = 20.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20 \quad 21 \quad 23 \quad 23 \quad 24 \quad 24$$

$$\uparrow$$

middle value

Mode = 20. The number 20 occurs here 3 times.

Here the mean, median and mode are all appropriate measures of central tendency.

Central tendency describes the tendency of the observations to bunch around a particular value, or category. The mean, median and mode are all measures of central tendency. They are all measures of the 'average' of the distribution. The best one to use in a given situation depends on the type of variable given.

For example, suppose a class of 20 students own among them a total of 17 pets as shown in the following table. Which measure of central tendency should we use here?

Type of Pet	Number
Cat	5
Dog	4
Goldfish	3
Rabbit	1
Bird	4

If our focus of interest were on the *type* of pet owned, we would use the mode as our average. 'Cat' would be described as the 'modal category', as this is the category that occurs most often.

If, on the other hand, we were not interested in the type of pet kept but the average *number* of pets owned then the mean would be an appropriate measure of central tendency. Here the mean is $\frac{17}{20} = 0.85$.

Also, if we are interested in the average number of pets per student then our data might be presented quite differently as in the table below.

Number of Pets	Tally	Frequency
0	⫫⫫ ⫫⫫ |	11
1	||||	4
2	|||	3
3	|	1
4	|	1

Now we are concerned only with a quantity variable and the average used most commonly with quantity variables is the mean. Here, again, the mean is 0.85.

$$\text{Mean} = \frac{(11 \times 0) + (4 \times 1) + (3 \times 2) + (1 \times 3) + (1 \times 4)}{20} = 0.85.$$

Note that (4×1) is really $1 + 1 + 1 + 1$, since 4 students have 1 pet each, and (3×2) is really $2 + 2 + 2$, since 3 students have 2 pets each. Since there are 20 scores the median score will occur between the tenth and the eleventh score. The median is 0, since the tenth and the eleventh scores are both 0, and the mode is 0.

The mean has some advantages over the median as a measure of central tendency of quantity variables. One of them is that all the observed values are used to calculate the mean. However, to calculate the median, while all the observed values are used in the ranking, only the middle or middle two values are used in the calculation. Another is that the mean is fairly stable from sample to sample. This means that if we take several samples from the same population their means are less likely to vary than their medians.

However, the median is used as a measure of central tendency if there are a few extreme values observed. The mean is very sensitive to extreme values and it may not be an appropriate measure of central tendency in these cases. This is illustrated in the next example.

Let's look again at our pets example and suppose that one of the students kept 18 goldfish.

Number of Pets	Tally	Frequency
0	卌 卌 |	11
1	||||	4
2	||	2
3	|	1
4	|	1
18	|	1

The mean is now 1.8, but the median and the mode are still 0. The effect of the outlier was to significantly increase the mean and now the median is a more accurate measure of the centre of the distribution.

With the exception of cases where there are obvious extreme values, the mean is the value usually used to indicate the centre of a distribution. We can also think of the mean as the balance point of a distribution.

For example, consider the distribution of students' marks on a test given in Figure 1. Without doing any calculation, we would guess the balance point of the distribution to be approximately 58. (Think of it as the centre of a see-saw.)
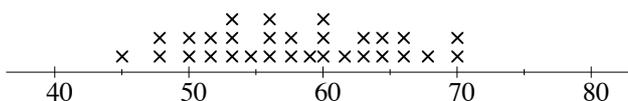
Figure 1: Students' marks on a test.

1.2 Exercises

1. Ten patients at a doctor's surgery wait for the following lengths of times to see their doctor.

   5 mins   17 mins   8 mins   2 mins   55 mins
   9 mins   22 mins   11mins   16 mins   5 mins

   What are the mean, median and mode for these data? What measure of central tendency would you use here?

2. What is the appropriate measure of central tendency to use with these data?

Method of Transport	Number of Students
Walk	5
Car	4
Train	15
Bicycle	10
Motorbike	6
Bus	10
Total	50

3. Which measure of central tendency is best used to measure the average house price in Sydney?

4. Without doing any calculation, estimate the mean of the distribution in Figure 2.

Figure 2: Students' marks on a test.

2 Measures of Dispersion

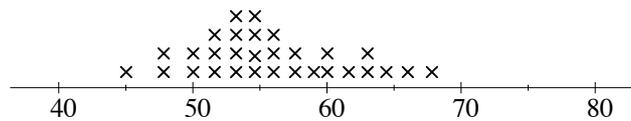The mean is the value usually used to indicate the centre of a distribution. If we are dealing with quantity variables our description of the data will not be complete without a measure of the extent to which the observed values are spread out from the average.

We will consider several measures of dispersion and discuss the merits and pitfalls of each.

2.1 The Range

One very simple measure of dispersion is the range. Lets consider the two distributions given in Figures 3 and 4. They represent the marks of a group of thirty students on two tests.
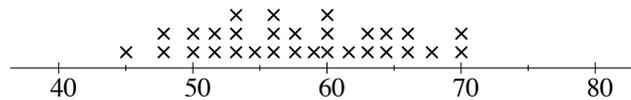
Figure 3: Marks on test A.
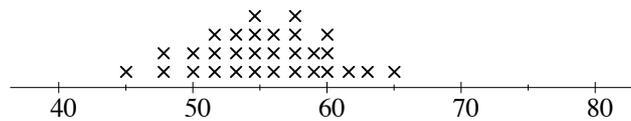
Figure 4: Marks on test B.

Here it is clear that the marks on test A are more spread out than the marks on test B, and we need a measure of dispersion that will accurately indicate this.

On test A, the range of marks is $70 - 45 = 25$.

On test B, the range of marks is $65 - 45 = 20$.

Here the range gives us an accurate picture of the dispersion of the two distributions.

However, as a measure of dispersion the range is severely limited. Since it depends only on two observations, the lowest and the highest, we will get a misleading idea of dispersion if these values are outliers. This is illustrated very well if the students' marks are distributed as in Figures 5 and 6.

Figure 5: Marks on test A.

Figure 6: Marks on test B.

On test A, the range is still $70 - 45 = 25$.

On test B, the range is now $72 - 40 = 32$, but apart from the outliers, the distribution of marks on test B is clearly less spread out than that of A.

We want a measure of dispersion that will accurately give a measure of the variability of the observations. We will concentrate now on the measure of dispersion most commonly used, the standard deviation.

2.2 Standard Deviation

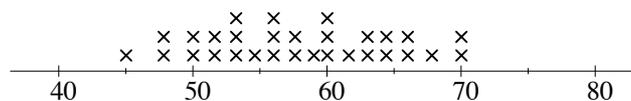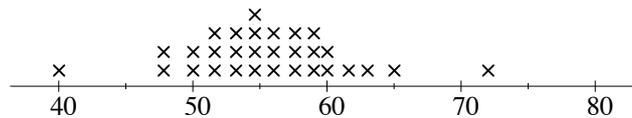Suppose we have a set of data where there is no variability in the observed values. Each observation would have the same value, say 3, 3, 3, 3 and the mean would be that same value, 3. Each observation would not be different or *deviate* from the mean.

Now suppose we have a set of observations where there is variability. The observed values would deviate from the mean by varying amounts.

The standard deviation is a kind of average of these deviations from the mean.

This is best explained by considering the following example.

Take, for example, the following grades of 6 students:

$$56 \quad 48 \quad 63 \quad 60 \quad 51 \quad 52.$$

Mean $= 55$.

To find how much our observed values deviate from the mean, we subtract the mean from each.

Observed values	56	48	63	60	51	52
Deviations from Mean	+1	−7	+8	+5	−4	−3

We cannot, at this stage, simply take the average of the deviations as their sum is zero.

$$(+1) + (-7) + (+8) + (+5) + (-4) + (-3) = 0$$

We get around this difficulty by taking the square of the deviations. This gets rid of the minus signs. (Remember $(-7) \times (-7) = 49$.)

Deviations	+1	−7	+8	+5	−4	−3
Squared deviations	1	49	64	25	16	9

We can now take the mean of these squared deviations. This is called the variance.

$$\text{Variance} = \frac{1 + 49 + 64 + 25 + 16 + 9}{6} = 27.33.$$

The variance is a very useful measure of dispersion for statistical inference, but for our purposes it has a major disadvantage. Because we squared the deviations, we now have a quantity in square units. So to get the measure of dispersion back into the same units as the observed values, we define standard deviation as the square root of the variance.

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{27.33} = 5.228.$$

The standard deviation may be thought of as the 'give or take' number. That is, on average, the student's grade will be 55, give or take 5 marks. The standard deviation is a very good measure of dispersion and is the one to use when the mean is used as the measure of central tendency.

Example: Calculate the mean and standard deviation of the following set of data.

Birthweight of ten babies (in kilograms)

2.977   3.155   3.920   3.412   4.236   2.593   3.270   3.813   4.042   3.387

Solution:

Birthweight in kilograms	Deviations from Mean score $-$ mean	Squared Deviations (score $-$ mean)^2
2.977	-0.5035	0.2535
3.155	-0.3255	0.1060
3.920	0.4395	0.1932
3.412	-0.0685	0.0047
4.236	0.7555	0.5708
2.593	-0.8875	0.7877
3.270	-0.2105	0.0443
3.813	0.3325	0.1106
4.042	0.5615	0.3153
3.387	-0.0935	0.0087
Sum = 34.805	Sum = 0	Sum = 2.3948

$$\text{Mean} = \frac{\text{sum of observations}}{\text{number of observations}} = \frac{34.805}{10} = 3.4805 = \mu.$$

$$\text{Variance} = \frac{\text{sum of squared deviations}}{\text{number of observations}} = \frac{2.3948}{10} = 0.2395 = \sigma^2.$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{0.2395} = 0.4894 = \sigma.$$

2.2.1 Exercises

1. Which of the following lists has the greatest standard deviation?

   a. 98   99   100   101   102

   b. 2   4   6   8   10

   c. 2   10

2. Two distributions are given in Figures 7 and 8. Which distribution has the greater standard deviation?
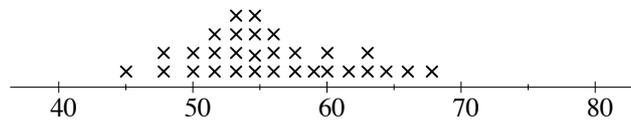
Figure 7:
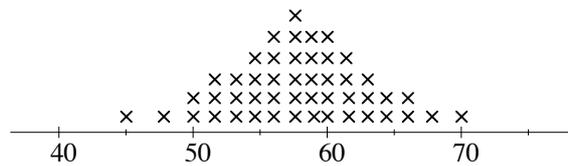
Figure 8:

3. The following values are the number of customers a restaurant served for lunch on ten consecutive days.

   46   50   51   60   62   64   72   41   53   55

   Find the mean and standard deviation of these data.

4. The ages of seven geography students who went on a field trip were

   20   19   19   25   20   18   19

   and the age of the instructor who accompanied them was 52.

   Find the mean and standard deviation of ages of these eight people.

5. The raw scores that eight students got on a history test were:

   69   84   93   61   79   88   57   67

   Find the mean and standard deviation of these scores.

2.3 The Interquartile Range

The interquartile range is another useful measure of dispersion or spread. It is used when the median is used as the measure of central tendency. It gives the range in which the middle 50% of the distribution lies. In order to describe this in detail, we first need to discuss what we mean by quartiles.

2.3.1 Quartiles

Suppose we start with a large set of data, say the heights of all adult males in Sydney. We can represent these data in a graph, which if smoothed out a bit, may look like Figure 9.
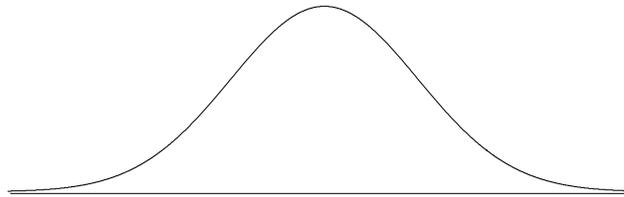
Figure 9: Graph representing heights of adult males.

As the name 'quartile' suggests, we want to divide the data into four equal parts. In the above example, we want to divide the area under our curve into four equal areas.

The second quartile or median

It is easy to see how to divide the area in Figure 9 into two equal parts, since the graph is symmetric. The point which gives us 50% of the area to the left of it and 50% to the right of it is called the second quartile or median. This is illustrated in Figure 10.

50% of the area   50% of the area

median or second quartile
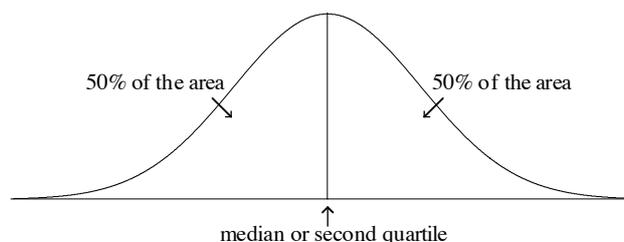
Figure 10: Graph showing the median or second quartile.

This exactly corresponds to our previous idea of median as the middle value.

The first quartile

The first quartile is the point which gives us 25% of the area to the left of it and 75% to the right of it. This means that 25% of the observations are less than or equal to the first quartile and 75% of the observations greater than or equal to the first quartile. The first quartile is also called the 25th percentile. This is illustrated in Figure 11.
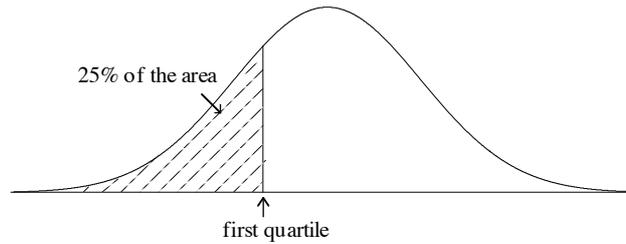
Figure 11: Graph showing the first quartile.

The third quartile

The third quartile is the point which gives us 75% of the area to the left of it and 25% of the area to the right of it. This means that 75% of the observations are less than or equal to the third quartile and 25% of the observation are greater than or equal to the third quartile. The third quartile is also called the 75th percentile. This is illustrated in Figure 12.

Figure 12: Graph showing the third quartile.

Summary

The first (Q_1), second (Q_2) and third (Q_3) quartiles divide the distribution into four equal parts. This is illustrated in Figure 13.

Figure 13: Graph showing all quartiles.

2.3.2   Quartiles for small data sets

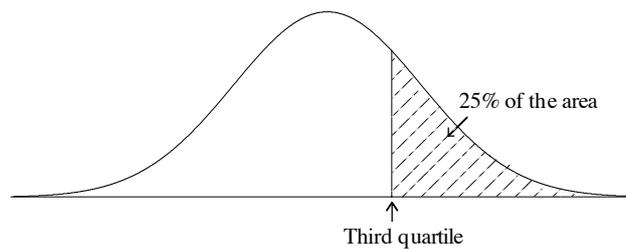Suppose now we have a small data set of twelve observations which we can write in ascending order as follows. (A data set, where the number of observations is a multiple of four, has been chosen to avoid some technical difficulties.)

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20 \quad 21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$

In this case, we want to divide the data into four equal sets, so that there are 25% of the observations in each.

First, we find the median just as we did in Chapter 1.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20 \uparrow 21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$
$$\text{median}$$

The **median** is 20.5 (half way between the 6th and 7th observations), and divides the data into two equal sets with exactly 50% of the observations in each: the 1st to the 6th observations in the first set and the 7th to 12th observations in the other.

To find the first quartile we consider the observations less than the median.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20$$

The **first quartile** is the median of these data. In this case, the first quartile is half way between the 3rd and 4th observations and is equal to 19.5.

Now, we consider the observations which are greater than the median.

$$21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$

The **third quartile** is the median of these data and is equal to 23.5.

So, for our small data set of 12 observations, the quartiles divide the set into four subsets.

$$15 \quad 18 \quad 19 \uparrow 20 \quad 20 \quad 20 \uparrow 21 \quad 23 \quad 23 \uparrow 24 \quad 24 \quad 25$$
$$\qquad\qquad Q_1 \qquad\qquad\qquad Q_2 \qquad\qquad\qquad Q_3$$

We will now use the quartiles to define a measure of spread called interquartile range.

2.3.3 The interquartile range

The interquartile range quantifies the difference between the third and first quartiles. If we were to remove the median (Q_2) from Figure 13 we would have a graph like that in Figure 14.
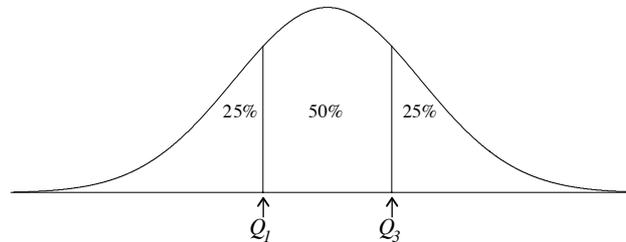

Figure 14: Graph showing the first and third quartiles.

From Figure 14, we see that 50% of the area is between the first and third quartiles. This means that 50% of the observations lie between the first and third quartiles.

We define the interquartile range as:

$$\text{The interquartile range} = \text{Third quartile} - \text{First quartile.}$$

For our small data set, the first quartile was 19.5 and our third quartile was 23.5. So, the interquartile range is $23.5 - 19.5 = 4$.

We will use the interquartile range later to draw a box-plot. For now we are interested in it as a measure of spread.

The interquartile range is particularly useful to describe data sets where there are a few extreme values. Unlike the range, and to a lesser extent the standard deviation, it is not sensitive to extreme values as it relies on the spread of the middle 50% of the distribution. So, if there are data sets which have extreme values, it can be more appropriate to use the median to describe central tendency and the interquartile range to describe the spread.

In the following exercises, data sets, where the number of observations is a multiple of four, have been given. Of course, the quartiles can be found for all other sized data sets, but we will restrict ourselves to these simple cases to avoid any technical difficulties. It is not necessary that you know how to calculate quartiles for all cases, but it is important that you understand the concept.

2.3.4 Exercises

1. For the following data sets, calculate the quartiles and find the interquartile range.

   a. The following numbers represent the time in minutes that twelve employees took to get to work on a particular day.

   | 18 | 34 | 68 | 22 | 10 | 92 | 46 | 52 | 38 | 29 | 45 | 37 |

b. The number of people killed in road traffic accidents in New South Wales from 1989 to 1996 is given below.

| 960 | 797 | 663 | 652 | 560 | 619 | 623 | 583 |

Source: Statistics–A Powerful Edge, Australian Bureau of Statistics, 1998.

c. The following data are the final marks of 40 students for the University Preparation Course, 'Preparatory Mathematics' in 1998.

$$61 \quad 77 \quad 51 \quad 85 \quad 55 \quad 77 \quad 70 \quad 56 \quad 41 \quad 61$$

$$28 \quad 87 \quad 23 \quad 22 \quad 86 \quad 63 \quad 99 \quad 94 \quad 38 \quad 25$$

$$90 \quad 59 \quad 87 \quad 53 \quad 29 \quad 86 \quad 33 \quad 87 \quad 75 \quad 50$$

$$59 \quad 77 \quad 77 \quad 71 \quad 99 \quad 78 \quad 70 \quad 93 \quad 78 \quad 93$$

Source: Mathematics Learning Centre, 1998.

2. The curve in Figure 15 represents the marks of a large number of students on an English exam. Estimate the quartiles and calculate the interquartile range.
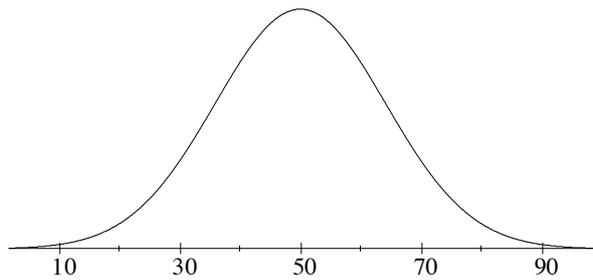

Figure 15: Graph showing marks on English exam.

3   Formulae for the Mean and Standard Deviation

So far we have avoided giving the formulae for mean or standard deviation but no discussion would be complete without them. If you are not familiar with sigma notation, do not attempt this section. An explanation of sigma notation can be found in the Mathematics Learning Centre booklet: *Introductory Algebra for Social Scientists*.

3.1   Formulae for Mean and Standard Deviation of a Population

The formula for the mean (average) of N observations is given by:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

where x_1 is the value of the first observation, x_2 is the value of the second observation, etc.

Example: The weights of five children in a family are:

$$x_1 = 3.5 \text{kg} \qquad x_2 = 12.3 \text{kg} \qquad x_3 = 17.7 \text{kg} \qquad x_4 = 20.9 \text{kg} \qquad x_5 = 23.1 \text{kg}.$$

Find the mean and standard deviation of the weights of these children.

Solution:

$$\begin{aligned}
\mu = \frac{1}{N} \sum_{i=1}^{N} x_i &= \frac{1}{N}(x_1 + x_2 + x_3 + x_4 + x_5) \\
&= \frac{1}{5}(3.5 + 12.3 + 17.7 + 20.9 + 23.1) \\
&= \frac{1}{5}(77.5) \\
&= 15.5.
\end{aligned}$$

A measure of how spread out the scores are, called the variance, has the following formula:

$$\begin{aligned}
\sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \\
&= \frac{1}{5}((-12)^2 + (-3.2)^2 + (2.2)^2 + (5.4)^2 + (7.6)^2) \\
&= \frac{1}{5}(246) \\
&= 49.2.
\end{aligned}$$

The standard deviation is the square root of the variance so,

$$\sigma = \sqrt{\sigma^2}$$

$$= \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$= 7.0 \qquad \text{to one decimal place.}$$

3.2 Estimates of the Mean and Variance

We have, so far, concerned ourselves with the mean, variance, and standard deviation of a population. These have been written using the Greek letters μ, σ^2, and σ respectively.

However, in statistics we are mainly concerned with analysing data from a sample taken from a population, in order to make inferences about that population. Our data sets are usually random samples drawn from the population.

When we have a random sample of size n, we use the sample information to estimate the population mean and population variance in the following way.

The mean of a sample of size n is written as $\overline{x}$ (read x bar).

To find the sample mean we add up all the sample scores and divide by the number of sample scores. This can be written using sigma notation as:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The sample mean is used to estimate the population mean. If we took many samples of size n from the population, calculated their sample means, and then averaged them, we would get a value very close to the population mean. We say that the sample mean is an unbiased estimator of the population mean.

An estimate of the population variance of a sample of size n is given by s^2 where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Notice that we are dividing by $n - 1$ instead of n as we did to find the population variance. We need to do this because the value obtained if we divide by n, tends to underestimate the population variance. Calculated in this way, s^2 is an unbiased estimator of population variance. In fact, s^2 can be described as the *estimated population variance.* (It is sometimes called the 'sample variance' but this is strictly speaking not accurate.)

3.2.1 Exercises

1. Suppose that the data set in Exercises 2.2.1 number 3 is a random sample of 10 days taken from a restaurant's records. Calculate the estimated population variance, s^2 for these data.

2. Suppose that the data in Exercise 2.2.1 number 5 are a random sample of scores on a history test. Calculate the mean, $\overline{x}$, and the estimated population standard deviation, s, of these data.

4 Presenting Data Using Histograms and Bar Graphs

We can now calculate two very important characteristics of a distribution, namely its 'average value' and a measure of its spread. In this section we will discuss one way of organising our data to give a visual representation of our data set. One of the most effective ways of presenting data is by a histogram.

Before we discuss histograms, we need to revise some facts about area.

4.1 Areas

Do you remember how to find the area of a rectangle? A rectangle with length l and breadth b is given in Figure 16.
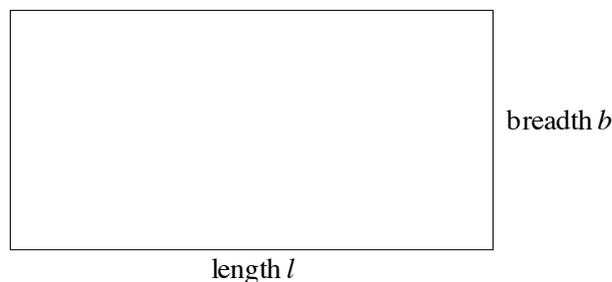
breadth b

length l

Figure 16: Rectangle of length l and breadth b.

The area of a restangle = length × breadth.

For example, a rectangle of length 4 units and breadth 2 units has an area of 8 square units. This is illustrated in Figure 17.
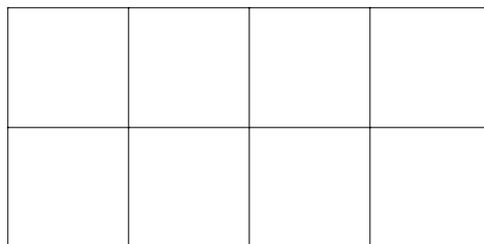
Figure 17: Rectangle of length 4 units and breadth 2 units.

When lengths are measured in centimetres (cm), the unit of area is the square centimetre (cm^2). When lengths are measured in metres (m), the unit of area is the square metre (m^2).

Figure 18 is a shape constructed from two rectangles. Its area is the sum of the area of rectangle A and the area of rectangle B.
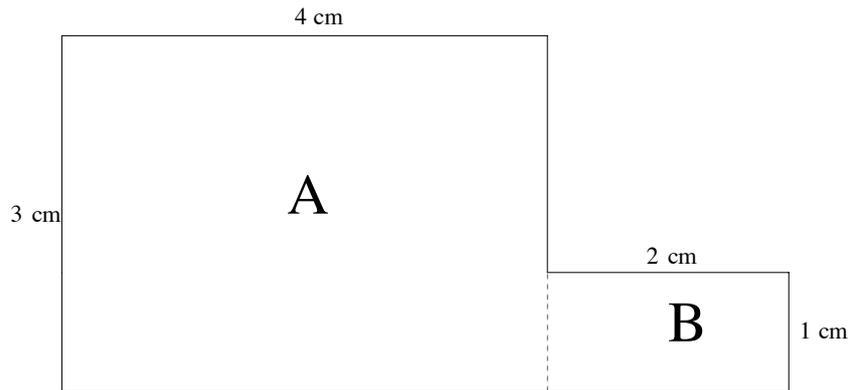
Figure 18: Shape constructed from two rectangles.

Area of rectangle A $= 3 \times 4 = 12\text{cm}^2$.

Area of rectangle B $= 2 \times 1 = 2\text{cm}^2$.

So, Total area $= 12 + 2 = 14\text{cm}^2$.

4.2 Histograms

In statistics, data is often represented using a histogram. A histogram is constructed by dividing the data into a number of *classes* and then number in each class or *frequency* is represented by a vertical rectangle. The area of the rectangle represents the frequency of each class.

The table below gives the marks of 80 students on an exam. The data has already been *grouped* for us into 10 classes. The exam scores are given in whole marks.

Range of marks	Frequency
1–10	2
11–20	2
21–30	4
31–40	6
41–50	7
51–60	8
61–70	15
71–80	22
81–90	10
91–100	4
Total	80

Each of the intervals from 1–10 marks, 11–20 marks and so on is called a class interval. In this example, each class interval is an interval of 10 marks, namely the marks 1 to 10 including both 1 and 10, or 11 to 20 including both 11 and 20, etc.

The table tells us that, for example, the class interval 21–30 has a frequency of 4. This means that 4 students scored marks between 21 and 30 inclusive but we don't know their exact marks.

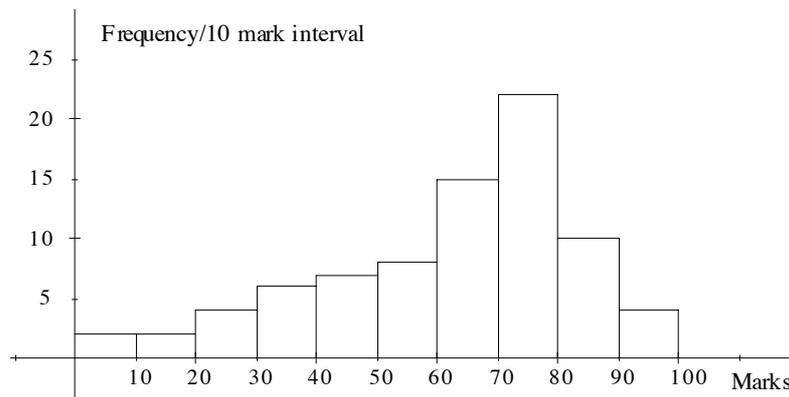A histogram of these data has been drawn in Figure 19.

Figure 19: Histogram of students' exam marks.

Here we have used the right hand endpoint of the class intervals to indicate our horizontal scale. All the class intervals have the same width, 10 marks.

The height of each column represents the frequency per 10 mark interval.

The area of each column represents the number of members in each class interval, or frequency.

For the interval 21– 30,

Area of the rectangle = no. of 10 mark intervals×frequency/10 mark interval = $1 \times 4 = 4$.

Since each column has the same width, i.e. one, its height is equal to its area. The total area enclosed represents the total number in the sample.

If we are given a histogram, we can use it to get information about the sample. For example, we can use the histogram in Figure 20 to estimate the number of people with marks between 26 and 40. We want to find the area of the histogram between the two dotted lines in Figure 20. The area is shaded to help you.

This area is $(\frac{1}{2} \times 4) + (1 \times 6) = 8$. That is, we take half the area of the rectangle 21–30 and add the area of the rectangle 31–40. So we estimate 8 people have marks between 26 and 40.
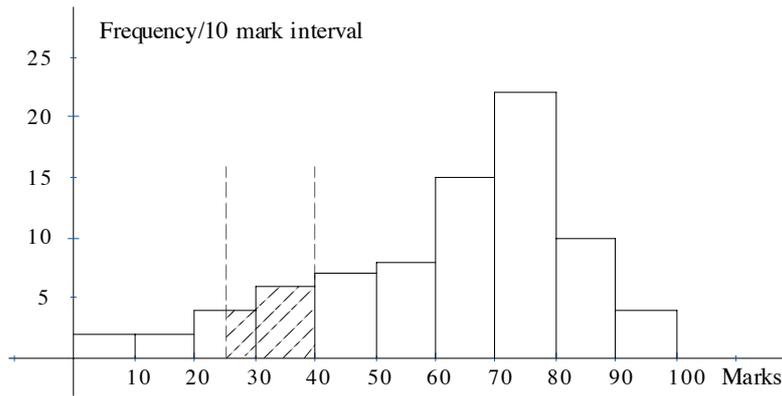
Figure 20: Histogram of students' exam marks.

Now suppose the information is grouped differently.

Range of marks	Frequency	Frequency per 10 marks
1–50	21	4.2
51–60	8	8
61–70	15	15
71–80	22	22
81–100	14	7
Total	80	

Here all marks of 50 and below are grouped in one class interval, and marks above 80 are also grouped together. In drawing this histogram it is extremely important that the *area* of each column, rather than its height, represents the frequency. The correct units for the vertical axis is again frequency/10 mark interval. The histogram for these data is drawn in Figure 21.
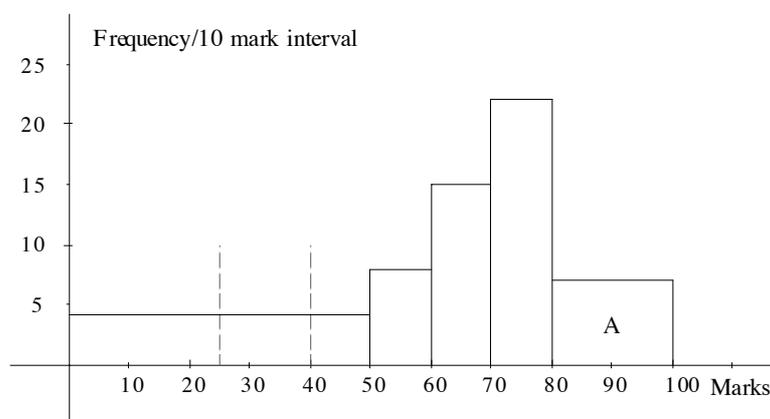

Figure 21: Histogram of students' exam marks with some groups combined.

For example, the number of people who obtained more than 80 marks is the area of rectangle A.

Area of restangle A = no. of 10 mark intervals $\times$ frequency/10 mark interval = $2\times7 = 14$.

This histogram can also be used to estimate the number of people with marks between 26 and 40. Again we find the area enclosed by the dotted lines drawn at 25 and 40. This time the estimate is $1.5 \times 4.2 = 6.3$, so our estimate would be 6.

4.2.1 Exercises

1. In the previous example, which estimate for the number of people with marks between 26 and 40 do you think is closer to the true value: 6 or 8? Why?

2. Histograms of marks (out of 100) on three different exams for a group of 150 students are given in Figures 22, 23 and 24. The pass mark for each exam is 50. For each exam, was the percentage who passed about 50%, well over 50% or well under 50%?
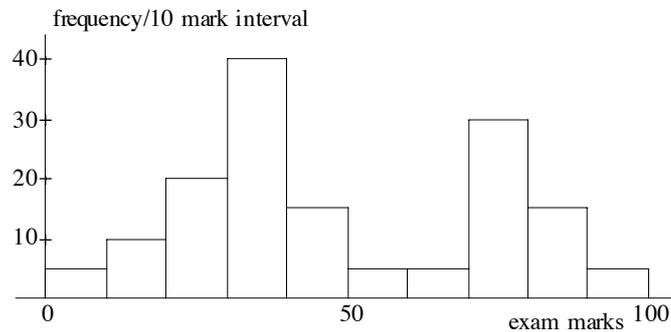

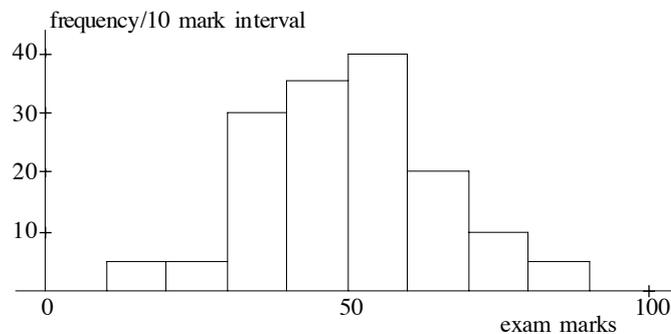Figure 22: Histogram of marks of students on exam 1.

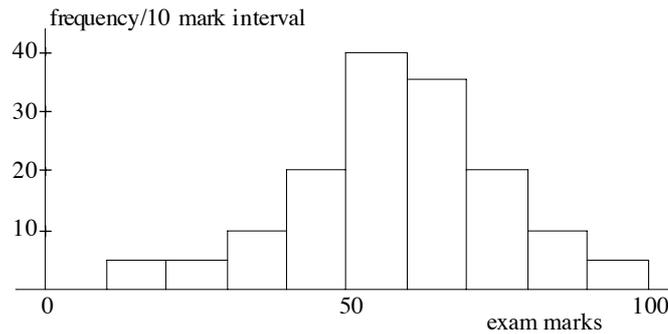Figure 23: Histogram of marks of students on exam 2.

frequency/10 mark interval

Figure 24: Histogram of marks of students on exam 3.

3. **a.** Calculate the percentage of students who passed the exam in each of Figures 22, 23 and 24.

b. What percentage of students scored 65% or more in each of these exams?

4. The gross incomes (in whole dollars only) of 100 employees, including part-time, of a medium size company is given in the following table.

Income Level	Frequency
1–4,999	2
5,000–9,999	7
10,000–14,999	12
15,000–19,999	16
20,000–24,999	18
25,000–34,999	24
35,000–49,999	11
50,000–99,999	8
100,000–149,999	2

Draw a histogram to represent these data.

5. The table below gives the number of students in a given age range who commensed a Bachelor's degree at the University of Sydney in 1997. The age is given in whole years and so, the range 20–24 includes those students who have had their twentieth birthday, and those who are 24 but have not yet turned 25. Thus, the class interval 20–24 is a **5** year interval.

Note also that the 'Under 20' and '45 and over' classes are open ended. Decide on a reasonable lower end point and upper end point respectively and use them to draw a histogram for these data.

Age in years	Frequency
Under 20	5267
20–24	1758
25–29	528
30–34	262
35–39	192
40–44	102
45 and over	130

Source: Statistics 1997, University of Sydney

4.3   Constructing Histograms and Bar Graphs from Raw Data

So far we have drawn our histogram from data which has already been grouped; that is, it has been divided into a number of class intervals for us. However it is likely that we will be given raw data or indeed collect it ourselves. How we decide to group our raw data will depend on the data itself, but a useful rule of thumb to use is to aim for between seven and twelve class intervals.

Suppose we have the following data.

Heights of 40 children in cms (to the nearest cm) in a Sydney day care centre.

109 92 60 77 103   88 91 93 57 73

65 68 72 79   83   86 79 98 62 69

71 74 82 84   90 100 96 80 84 93

69 75 80 77   82   78 84 68 90 79

Here the measurements given are to the nearest cm, although it is theoretically possible for height to take on any value. This is an example of continuous data.

First of all, we decide on the class intervals we are going to use, and group our data into these class intervals. We must take care when doing this not to put an observation into more than one class.

Class interval	Tally	Frequency
55–60	||	2
61–65	||	2
66–70	||||	4
71–75	⊬|	5
76–80	⊬||||	8
81–85	⊬||	6
86–90	||||	4
91–95	||||	4
96–100	|||	3
101–105	|	1
106–110	|	1

Here we have grouped the data into 11 class intervals. We could draw our histogram using these 11 class intervals as in Figure 25.
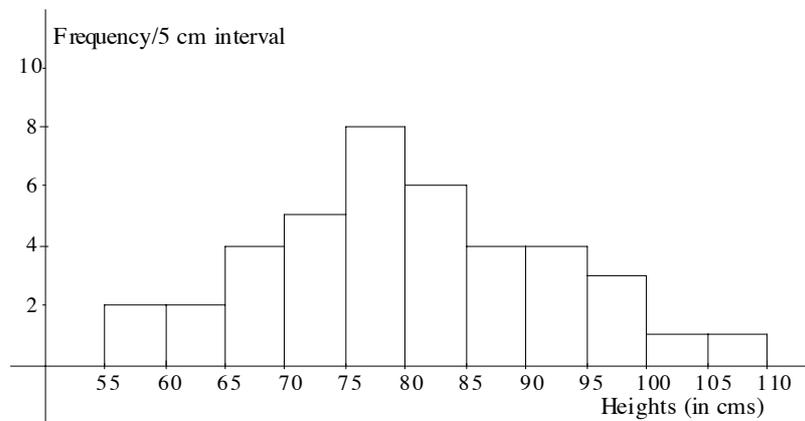

Figure 25: Histogram of childrens' heights.

However, we might choose to combine the last two intervals into one, since there is only one observation in each class. If we do this, the frequency of the combined interval, 101–110, is now 2, so the height of the interval must be one to maintain its area as 2.

Area of rectangle = no. of 5cm intervals $\times$ frequency/ 5cm interval = $2 \times 1 = 2$.
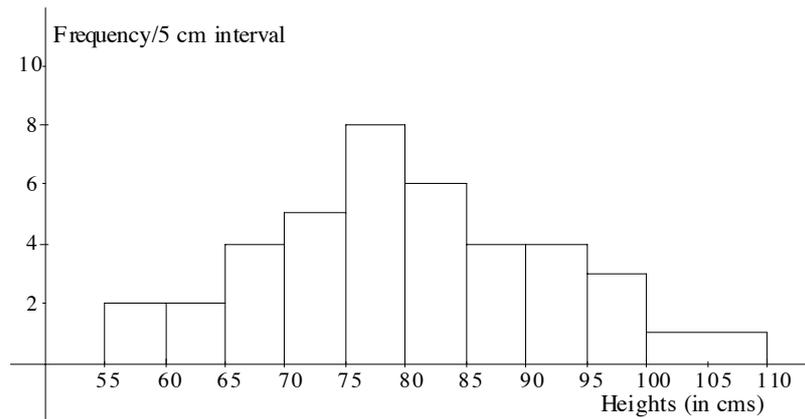
This is illustrated in Figure 26.

Figure 26: Histogram of childrens' heights with combined classes.

Example:

The following list gives the number of rooms, excluding the bathroom and kitchen, in 50 dwellings.

$$
\begin{array}{cccccccccc}
2 & 6 & 4 & 3 & 3 & 4 & 4 & 7 & 5 & 4 \\
5 & 3 & 7 & 5 & 5 & 4 & 4 & 5 & 6 & 2 \\
6 & 3 & 4 & 4 & 5 & 8 & 6 & 5 & 5 & 3 \\
3 & 3 & 7 & 5 & 4 & 4 & 5 & 4 & 1 & 6 \\
5 & 4 & 4 & 8 & 6 & 2 & 3 & 3 & 6 & 4
\end{array}
$$

Draw up a frequency table and draw a bar graph to represent these data.

Solution:

This is an example of discrete data, that is you can have 1 room, 2 rooms, 3 rooms etc, but nothing in between is possible. We use a bar graph to represent discrete data. In a bar graph, the height of the rectangle gives the frequency of the class.

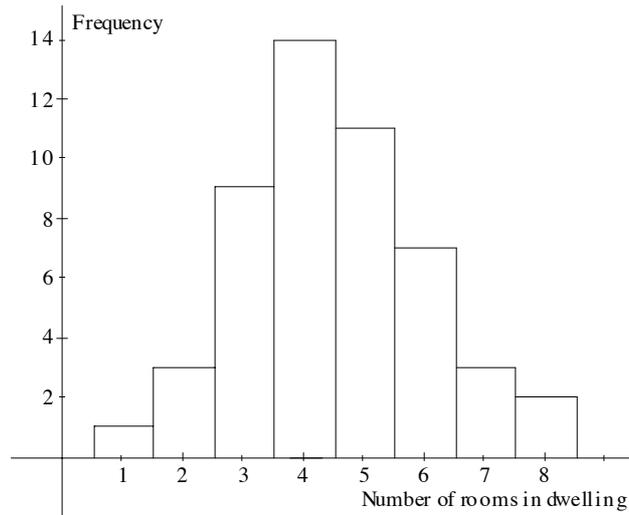| Number of rooms	Tally	Frequency
1	|	1
2	|||	3
3	卌 ||||	9
4	卌 卌 ||||	14
5	卌 卌 |	11
6	卌 ||	7
7	|||	3
8	||	2

Figure 27: Bar graph of number of rooms in 50 dwellings.

In the bar graph drawn in Figure 27, we have centred each class interval on the actual number of rooms as this is discrete data and so there is nothing between, say, 3 rooms and 4 rooms. We could also leave a gap between the rectangles to make this clear.

4.3.1 Exercises

1. Draw a histogram for the following data set. Weights (in whole kilograms) of a class of 30 female students in year 10.

57	46	61	66	48	59	55	56	60	49
44	53	68	57	55	54	49	50	52	54
62	59	51	52	63	54	47	53	56	60

2. Draw a bar graph for the following data set.

   Number of pets owned by a class of 20 students. This is the raw data we used in Chapter 1.

2	0	0	1	1	0	0	0	0	0
0	0	4	1	1	2	18	0	0	3

3. The number of people killed in road traffic accidents in New South Wales from 1987 to 1996 is given in the following table.

Year	Number of deaths
1987	959
1988	1037
1989	960
1990	797
1991	663
1992	652
1993	560
1994	619
1995	623
1996	583

Source: Statistics—A Powerful Edge, Australian Bureau of Statistics, 1998.

Draw a bar graph to represent these data.

5   The Box-plot

The box-plot is another way of representing a data set graphically. It is constructed using the quartiles, and gives a good indication of the spread of the data set and its symmetry (or lack of symmetry). It is a very useful method for comparing two or more data sets.

The box-plot consists of a scale, a box drawn between the first and third quartile, the median placed within the box, whiskers on both sides of the box and outliers (if any). This is best illustrated using a diagram such as Figure 28.
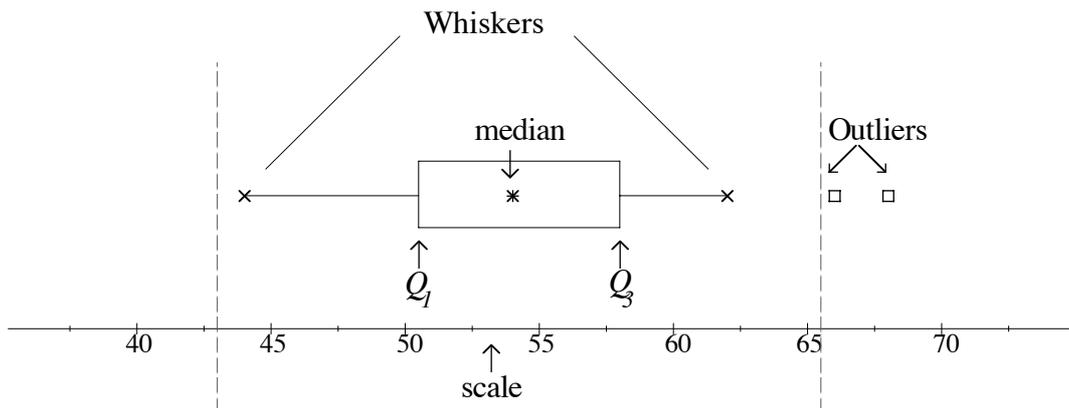

Figure 28: Annotated example of a box-plot.

The two dashed vertical lines in Figure 28 are the lower and upper outlier thresholds and are not normally included in a box-plot.

The following data set was used to construct the box-plot in Figure 28.

$$57 \quad 46 \quad 61 \quad 66 \quad 48 \quad 59 \quad 55$$

$$56 \quad 60 \quad 49 \quad 44 \quad 53 \quad 68 \quad 57$$

$$55 \quad 54 \quad 49 \quad 50 \quad 52 \quad 54 \quad 62$$

$$59 \quad 51 \quad 52 \quad 53 \quad 54 \quad 47 \quad 53$$

5.1   Constructing a Box-plot

Step 1: Order the data and calculate the quartiles.

$$44 \quad 46 \quad 47 \quad 48 \quad 49 \quad 49 \quad 50$$

$$51 \quad 52 \quad 52 \quad 53 \quad 53 \quad 53 \quad 54$$

$$54 \quad 54 \quad 55 \quad 55 \quad 56 \quad 57 \quad 57$$

$$59 \quad 59 \quad 60 \quad 61 \quad 62 \quad 66 \quad 68$$

Now we calculate the median, the first quartile and the third quartile.

For these data, median $= 54$, the first quartile $= 50.5$ and the third quartile $= 58$.

With this information we can begin to construct the box-plot.

Step 2: Draw the scale and mark on the quartiles.

Mark the median at the correct place above the scale with a asterix, draw a box around this asterix with the left hand side of the box at the first quartile, 50.5, and the right hand side of the box at the third quartile, 58. This is illustrated in Figure 29.
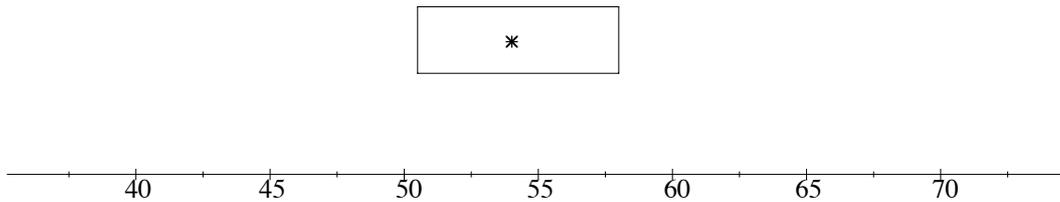
Figure 29: Box constructed using first and third quartiles.

Step 3: Calculate the interquartile range and determine the position of the outlier thresholds.

$$\text{Interquartile range} = \text{third quartile} - \text{first quartile} = 58 - 50.5 = 7.5.$$

The position of the lower outlier threshold is found by subtracting the interquartile range from the first quartile, $50.5 - 7.5 = 43$.

The position of the upper outlier threshold is found by adding the interquartile range to the third quartile, $58 + 7.5 = 65.5$.

(Some texts add or subtract $1.5 \times$ interquartile range.)

We now add the outlier thresholds to our diagram. This is illustrated in Figure 30.
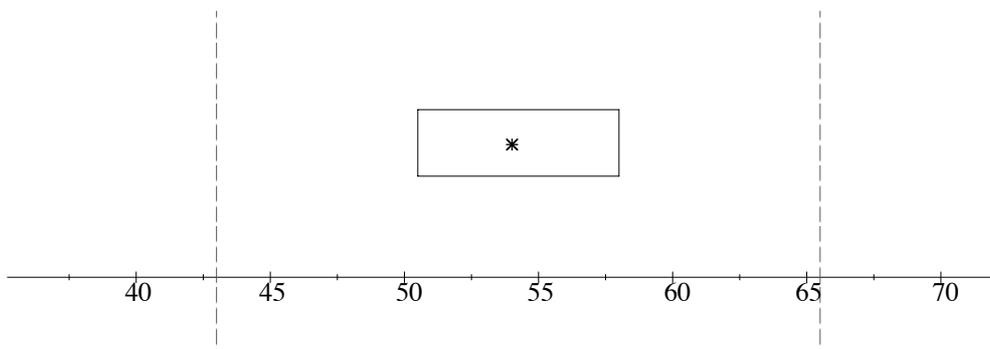

Figure 30: Diagram illustrating the outlier thresholds.

Step 4: Use the outlier thresholds to draw the whiskers.

To draw the left hand whisker, we need the smallest data value that lies inside the outlier thresholds. In this example, it is the value 44. This is drawn on our diagram with a small cross level with the asterix. A horizontal line is now drawn to the left hand side of the box.

To draw the right hand whisker, we find the largest data value that lies inside the outlier thresholds. In this example, the value is 62. This is drawn on the right hand side of the box with a small cross and connected to the box by a horizontal line.
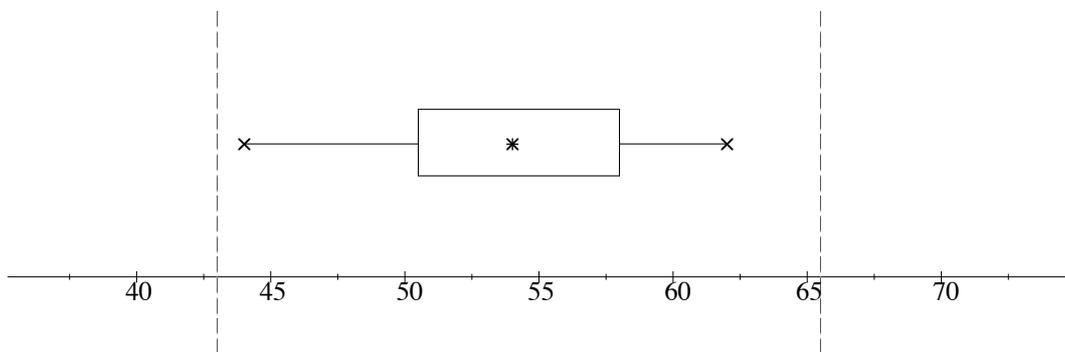
This is illustrated in Figure 31.

Figure 31: Diagram illustrating the whiskers.

Step 5: Determine the outliers and remove the outlier thresholds.

Values (if any) that lie *outside* the outlier thresholds are called outliers. In this example, 66 and 68 are outliers. These are placed on the diagram using a small square or circle.

Finally, the outlier thresholds are removed.

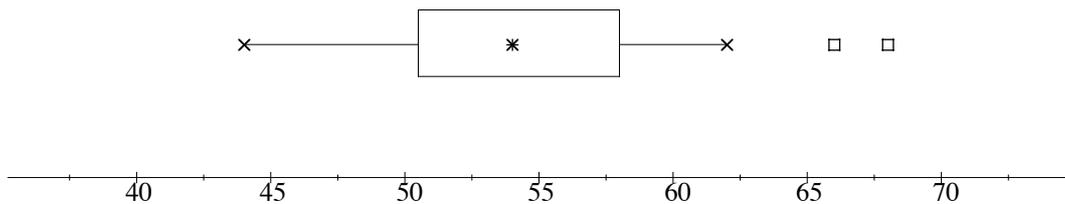The completed box-plot is illustrated in Figure 32.

Figure 32: Completed box-plot.

5.2 Using Box-plots to Compare Data Sets

Box-plots are frequently used to compare data sets as the differences in shape, spread and location are easily seen.

For example, Figure 33 gives box-plots for the final marks of the University of Sydney university preparation course, *Preparatory Mathematics* for the years 1996, 1997 and 1998.
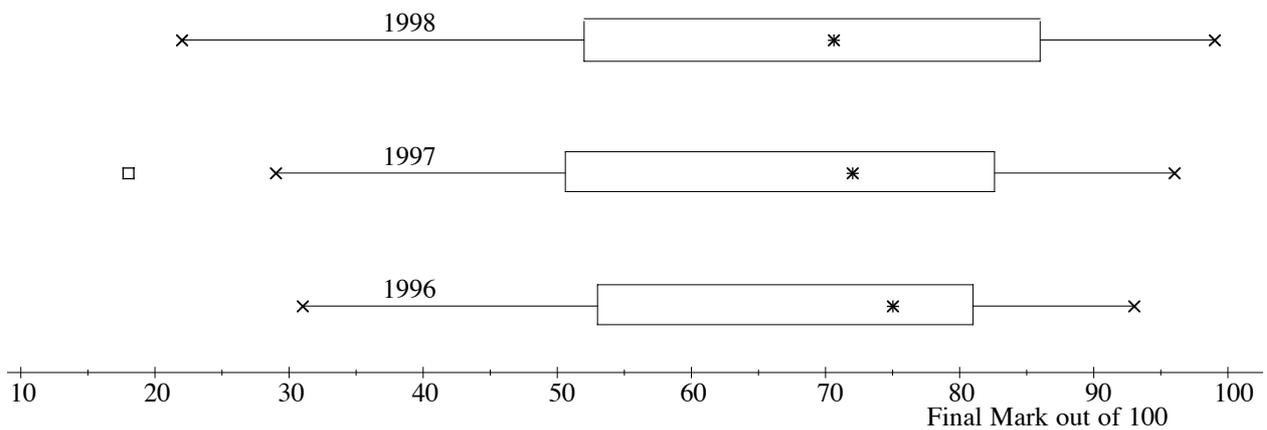

Figure 33: Box-plots of final marks for the years 1996, 1997 and 1998.

The marks from all years are left-skewed, but those from 1996 and 1997 quite markedly so. 1996 had the highest median score but the least spread. The marks from 1998 vary more than those in 1996 and 1997. In all years, over 75% of students passed the course.

5.3 Exercises

1. The following four data sets give the daytime temperature in Brisbane, Hobart, Melbourne and Sydney for February 1998. Draw a box-plot for each data set.

Brisbane						
31	30	30	30	30	29	31
30	31	29	29	30	31	29
29	30	28	29	29	29	29
28	29	27	29	28	29	29

Hobart						
20	29	22	25	20	19	28
24	22	23	25	26	21	22
20	16	23	26	22	18	21
20	19	24	24	22	18	20

Melbourne						
26	35	23	27	26	25	33
26	25	29	30	28	24	28
23	21	24	32	27	23	24
24	23	31	32	35	23	23

Sydney						
28	28	29	29	30	27	30
25	24	25	24	29	26	28
29	31	23	26	29	31	26
27	26	27	25	29	37	25

2. Comment on the differences in the shape, spread, and location of the box-plots in 1.

6   Solutions to Exercises

6.1   Solutions to Exercises from Chapter 1

1.2 Exercises

1. Mean $=15$ mins,   Median $= \frac{9+11}{2} = 10$ mins,   Mode $= 5$ mins.

    The median would be the preferred measure of central tendency to use here and not the mean, since there is an outlier of 55 mins. This is making the assumption that the outlier is a freak value and should be disregarded. The mode would not be suitable, because it is just chance that two people waited for the same period of time, and all the others waited for different time periods.

2. The mode is the only possible measure of central tendency to use here, since we are dealing with category data. The modal category is 'train'.

3. The median is used to indicate average house prices in Sydney. The inclusion of the very expensive houses (those worth millions of dollars) in the calculation of the mean would make the 'average' house price too high to be representative of the general market. Nor is the mode suitable because it could happen by chance that a very large number of houses all had the same non-representative value.

4. The actual value for the mean is 56. How close to this value did you get with your guess?
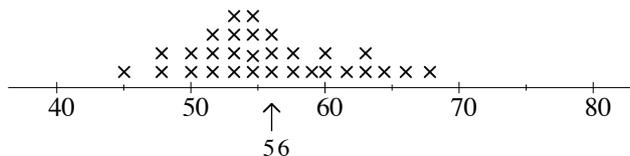

Figure 34: Students'marks on a test.

6.2   Solutions to Exercises from Chapter 2

2.2.1 Exercises

1. c) has the greatest standard deviation. The standard deviation of c) is 4, the standard deviation of b) is 2.83, and the standard deviation of a) is 1.41.

2. a) has the greater standard deviation. It is more spread out than b).

3. Number of customers served in a restaurant on ten consecutive days.

Number of Customers	Deviations from Mean $(x - \mu)$	Squared Deviations $(x - \mu)^2$
46	-9.4	88.36
50	-5.4	29.16
51	-4.4	19.36
60	4.6	21.16
62	6.6	43.56
64	8.6	73.96
72	16.6	275.56
41	-14.4	207.36
53	-2.4	5.76
55	-0.4	0.16
Sum $= 554$	Sum $= 0$	Sum $= 764.4$

Mean $= 55.4$, Variance $= \frac{764.4}{10} = 76.44$, Standard deviation $= \sqrt{76.44} = 8.743$.

4. Mean $= 24$, Standard deviation $= 10.77$.

5. Mean $= 74.75$, Standard deviation $= 12.296$.

2.3.4 Exercises

1. a. First quartile $= 25.5$, Median $= 37.5$, Third quartile $= 49$, IQR $= 23.5$.

b. First quartile $= 601$, Median $= 637.5$, Third quartile $= 730$, IQR $= 129$.

c. First quartile $= 52$, Median $= 70.5$, Third quartile $= 86$, IQR $= 34$.

2. Our estimate puts the first quartile at 40, the median at 50 and the third quartile at 60. This gives an interquartile range of 20. This means that the middle 50% of marks lie within 20 marks of each other.

6.3 Solutions to Exercises from Chapter 3

3.2.1 Exercises

1. $s^2 = 84.93$.

2. $\bar{x} = 74.75$, $s = 13.14$.

6.4   Solutions to Exercises from Chapter 4

4.2.1 Exercises

1. The estimate 6 obtained from grouped data is less likely to be reliable than the estimate 8 from the original histogram.

2. Figure 22: well under 50%, Figure 23: about 50%, Figure 24: well over 50%.

3. **a.** In Figure 22, 60 students passed the exam, and 90 failed. So, 40% of the students passed. In Figure 23, 75 students passed and 75 failed, so 50% passed. In Figure 24, 110 students passed and 40 failed, so about 73% of students passed the exam.

    b. In Figure 22, about 53 students scored 65% or more, so about 35% of students scored 65% or more. In Figure 23, about 25 students scored 65% or more, so about 17% of students scored 65% or more. In Figure 24, about 53 students scored 65% or more, so about 35% of students scored 65% or more on the exam.

4. The first five class intervals have a range of $5000. So, in the following table we have calculated the frequency per $5000 for each class.

Income Level	Frequency per $5000
0–4999	2
5000–9,999	7
10,000–14,999	12
15,000–19,999	16
20,000–24,999	18
25,000–34,999	12
35,000–49,999	3.67
50,000–99,999	0.8
100,000–149,999	0.2

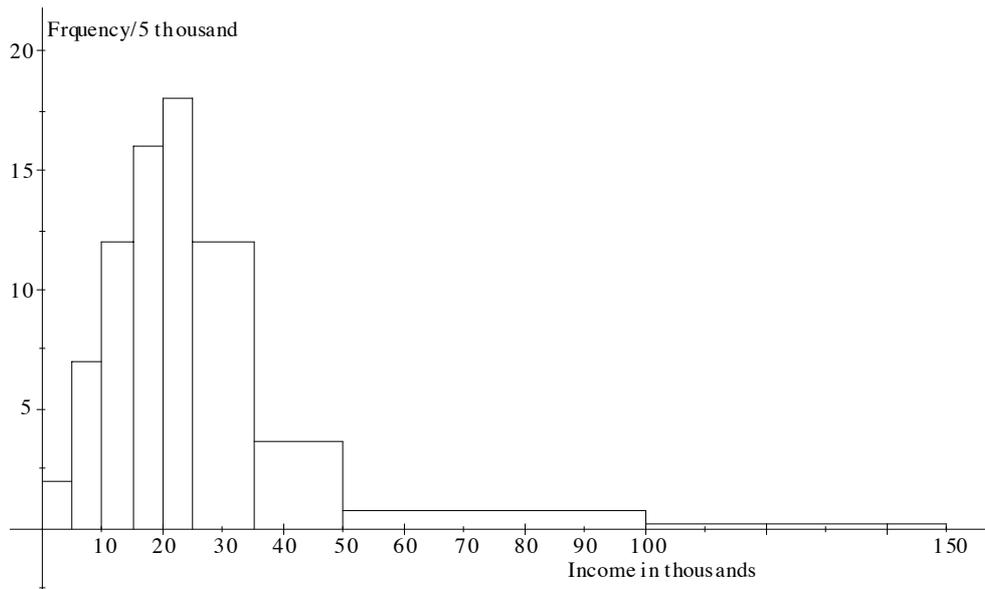The histogram is drawn in Figure 35.

Figure 35: Histogram of gross incomes of the employees of a medium size company.

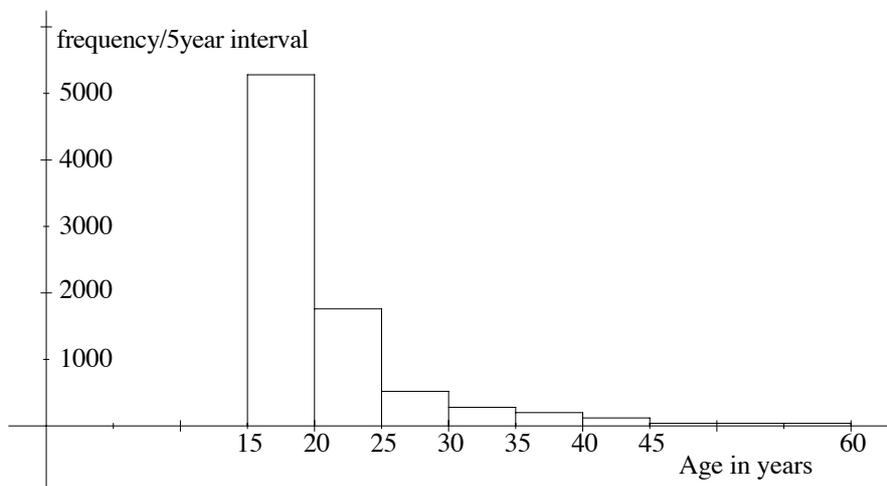**5.** The histogram is is given in Figure 36.

Figure 36: Histogram of ages of the students who commensed a Bachelor's degree in 1997.

We have closed off the first class interval at 15 years (even though it would be very unlikely for a student to be that young). The last class interval has been closed off at age 59. So, the last class interval has a range of 15 years. You may have made different choices and so your histogram will be slightly different.

4.3.1 Exercises

1. The data on the weights of 30 female students in year 10 has been grouped into 6 class intervals, each with range 5 kilograms, and given in the following table.

Weight in Kilograms	Frequency
40–44	1
45–49	5
50–54	9
55–59	8
60–64	5
65–69	2

A histogram representing these data has been drawn is Figure 37.
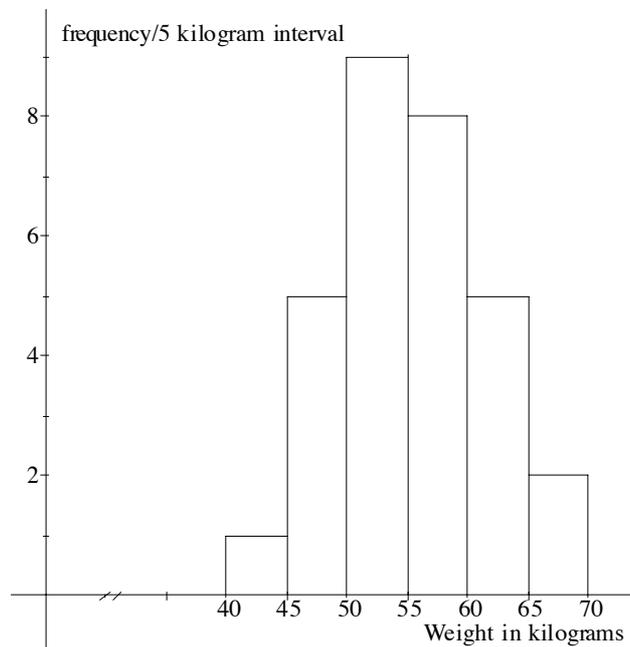
Figure 37: Histogram of weights of 30 female year 10 students.

2. A bar graph has been drawn to represent the number of pets owned by a group of 20 students. Note that since this is discrete data, the height of the rectangle gives the frequency of the class.
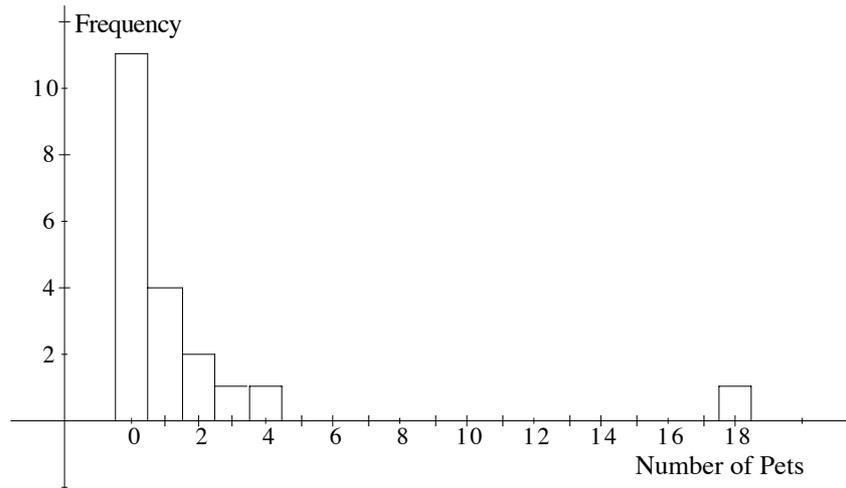
Figure 38: Bar graph of the number of pets owned by 20 students.

3. The data giving road fatalities in NSW form 1987 to 1996 are discrete data. A bar graph is given in Figure 39.
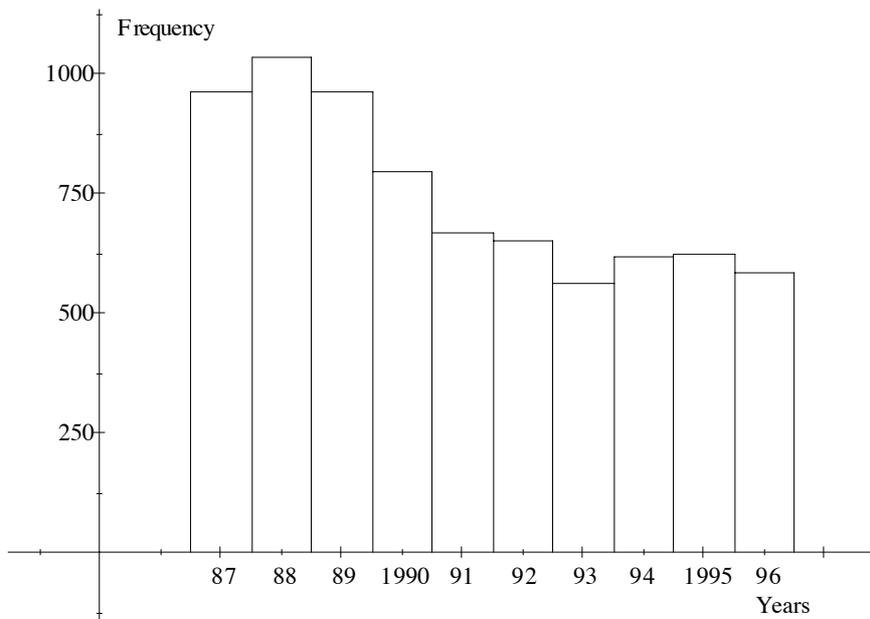

Figure 39: Bar graph of number of road fatalities in NSW from 1987 to 1996.

6.5 Solutions to Exercises from Chapter 5

5.3 Exercises

1. For Brisbane,

First quartile = Median = 29,   Third quartile = 30,   Interquartile range = 1.

So,

Lower threshold = 28,   Upper threshold = 31.

The left hand whisker is positioned at 28; the right hand whisker at 31. 27 is the only outlier.

For Hobart,

First quartile = 20,   Median = 22,   Third quartile = 24,   Interquartile range = 4.

So,

Lower threshold = 16,   Upper threshold = 28.

The left hand whisker is positioned at 16; the right hand whisker at 28. 29 is the only outlier.

For Melbourne,

First quartile = 23.5,   Median = 26,   Third quartile = 29.5,   Interquartile range = 6.

So,

Lower threshold = 17.5,   Upper threshold = 35.5.

The left hand whisker is positioned at 21; the right hand whisker at 35. There are no outliers.

For Sydney,

First quartile = 25.5,   Median = 27.5,   Third quartile = 29,   Interquartile range = 3.5.

So,

Lower threshold = 22,   Upper threshold = 32.5.

The left hand whisker is positioned at 23; the right hand whisker at 31. 37 is the only outlier.

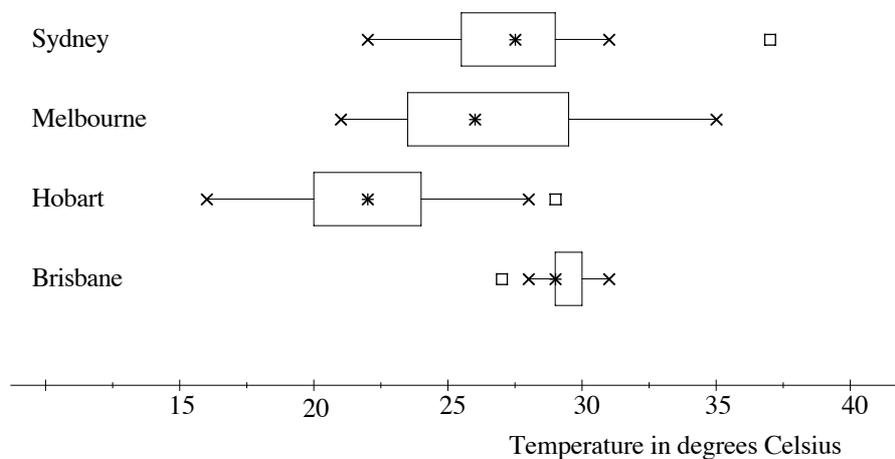Box-plots for these data sets are given in Figure 40.

Figure 40: Box-plots for temperatures of four Australian cities for February 1998.

2. Brisbane was the city with the highest median temperature for February 1998. The daytime temperature for the month was very consistent. The first quartile and the median were both 29°, so at least a quarter of the temperatures were that value. There was very little variation in temperature for the whole month.

Hobart had the lowest median temperature. The distribution of temperatures for the month is symmetric if the outlier is ignored.

Melbourne had the second lowest median temperature but was the city with the most temperature variation. The distribution of temperatures is slightly right skewed (the tail is on the right).

Sydney was the city with the second highest median temperature with a smaller variation in temperatures than that of Hobart and Melbourne. Ignoring the outlier (the highest temperature for the month in all four cities), the distribution of temperatures for the month is left skewed.

MATHEMATICS
LEARNING CENTRE

THE UNIVERSITY OF
SYDNEY