Back to Menu >>>

CHAPTER 2

CORE

Summarising numerical data: the median, range, interquartile range and box plots

- How can we describe a distribution with just one or two statistics?
- What is the median, how is it calculated and what does it tell us?
- What are the range and the interquartile range (*IQR*), and how are they calculated?
- What is a five-number summary?
- What is a box plot and why is it useful?

2.1 Will less than the whole picture do?

Even when we have constructed a frequency table, histogram or stem plot to display a set of numerical data, we are still left with a large amount of information to digest. One way of overcoming the problem is to try to summarise the information. Just a few numbers obtained from the data can be used to describe the essential features of the distribution. We call these numbers **summary statistics**.

The two most commonly used types of summary statistics may be classified as:

- measures of centre (about which point is the distribution centred?)
- measures of spread (how are the scores in the distribution spread out?)

We have met the concept of centre and spread in Chapter 1 when we were using histograms and stem plots to describe the distribution of numerical variables. In this chapter and the next, we will aim to come up with more precise ways of defining and quantifying (giving values to) these concepts.

Firstly we will consider a set of summary statistics that are based on ordering the data.

Back to Menu >>>

2.2 The median, range and interquartile range (IQR)

The median

The median is the **midpoint** of a distribution: 50% of values in the data set are less than or equal to the median.

Calculating the median

The **median** is the **middle** value in a data set. Its value is found by listing all the data values in numerical order. We then find the value that divides the distribution into two equal parts. For small data sets, the median can be easily located by the eye. However, for larger data sets the following rule for locating the median is helpful.

A rule for determining the location of the median in a data set

For n ordered data values, the **median**, M, is located at the $\left(\frac{n+1}{2}\right)$th position.

Example 1 **Finding the median value in a data set**

Order each of the following data sets, locate the median, and then write down its value.

a 2 9 1 8 3 5 3 8 1 **b** 10 1 3 4 8 6 10 1 2 9

Solution

a 1 Write down the data set.

2 9 1 8 3 5 3 8 1

2 Order the data set.

1 1 2 3 3 5 8 8 9

3 Locate the position of the median in the data set. For n data points, the median is located at the $\left(\frac{n+1}{2}\right)$th position in the data set.

$n = 9$

median is $\left(\frac{9+1}{2}\right)$th or 5th term

4 Write down the value of the median.

$\therefore M = 3$

b 1 Write down the data set.

10 1 3 4 8 6 10 1 2 9

2 Order the data set.

1 1 2 3 4 6 8 9 10 10

3 For n data points, the median is located at the $\left(\frac{n+1}{2}\right)$th position in the data set.

$n = 10$

median is $\left(\frac{10+1}{2}\right)$th or 5.5th term

4 The 5.5th term lies mid-way between the 5th and 6th terms. The median value is taken to be the average value of these two terms. Determine this value and write it down.

$\therefore M = \left(\frac{4+6}{2}\right) = 5$

Note: There is a way to check that you are correct when calculating a median. Count the number of data values each side of the median. They should be equal.

Back to Menu >>>

Using a stem plot to help locate medians

The process of calculating a median is very simple in theory but can be tedious in practice, particularly if the data set is large. However, if an ordered stem plot of the data is available it is a quick and easy process.

Example 2 Finding the median value from an ordered stem plot

The ordered stem shows the distribution of life expectancies (in years) in 23 countries.

Identify the position of the median in the stem plot and write down its value.

Life expectancy (years)

5	2
5	5 6
6	4
6	6 6 7 9
7	1 2 2 3 3 4 4 4 4
7	5 5 6 6 7 7

Solution

1 For n data values, the median is located at the $\left(\frac{n+1}{2}\right)$th position in the data set.

$n = 23$

median is $\left(\frac{23+1}{2}\right)$th or 12th term

2 Count in 12 terms from either end of the stem plot to locate the median. Write down its value.

$\therefore$ median value $=$ 73 years

Note: Again you can check to see whether the value you have calculated for the median is correct by counting the number of data values each side of the median. They should be equal.

Having found a way of making the concept of centre more precise, we now look at ways of doing the same with the concept of spread.

The range

The range

The **range**, R, is the simplest measure of spread of a distribution. It is the difference between the largest and smallest values in the data set, so that:

$$R = \text{largest data value} - \text{smallest data value}$$

For example, for the life expectancies data used in Example 2, we can see that the highest life expectancy in the 23 countries was 77 years. The lowest (smallest) was 52 years. Therefore, the range of life expectancies is given by:

$$R = 77 - 52 = 25 \text{ years}$$

The range was the measure of spread we used in Chapter 1 when describing the spread of a histogram or stem plot. We did this because it was simple to use. However, the range as a measure of spread has its limitations. Because the range depends only on the two extreme values in a set of data it is not always an informative measure of spread. For example, the largest and smallest values in a data set might be outliers and not at all typical of the rest of the

Back to Menu >>>

values. Furthermore, any two sets of data with the same highest and lowest values will have the same range, irrespective of the way in which the data values are spread out in between. However, the range is useful to know because it gives us an indication of the **absolute spread** of the distribution.

The interquartile range (*IQR*)

Just as the median is the point that divides a distribution in half, **quartiles** are the points that divide a distribution into **quarters**. We will use the symbols, Q_1, Q_2 and Q_3, to represent the quartiles. Note that $Q_2 = M$, the median.

The interquartile range

The **interquartile range** (*IQR*) is defined to be the spread of the middle 50% of data values, so that:

$$IQR = Q_3 - Q_1$$

To calculate the *IQR*, it is necessary to first calculate the quartiles, Q_1 and Q_3. In principle, this is straight forward as:

- Q_1 is the midpoint of the lower half of the data values
- Q_3 is the midpoint of the upper half of the data values

Again, if the data has been ordered, the computation of the quartiles is relatively straightforward.

A comment on calculating quartiles

A practical problem arises when calculating quartiles if the median corresponds to an actual data value. This will happen whenever there is an odd number of data values. The question is what to do with the median value when calculating quartiles. One strategy is to omit it, which means that there will always be slightly less than 50% of the data values in each 'half' of the distribution. This is the approach we will take. It is also the approach taken by most commonly used graphics calculators and many statistical packages. The other approach, used by some statistical packages, is to put the median into both 'halves' before calculating the median. This ensures that there are exactly 50% of values in each half, but at the expense of creating another data value out of nowhere. More sophisticated methods do exist for calculating the quartiles, but they are necessarily more time consuming and generally only give results that are marginally different from those determined using either of these methods.

Example 3 Finding quartiles from an ordered stem plot

Use the stem plot to determine the quartiles Q_1 and Q_3, the *IQR* and the range, *R*, for life expectancies. The median life expectancy is 73.

Life expectancy (years)

5	2
5	5 6
6	4
6	6 6 7 9
7	1 2 2 3 3 4 4 4 4
7	5 5 6 6 7 7

Back to Menu >>>

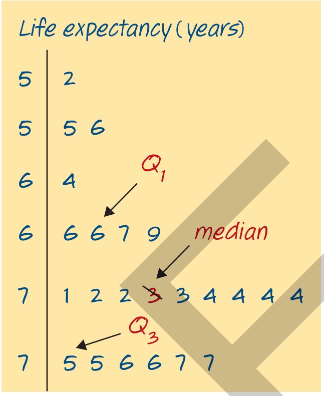
Solution

1 Mark the median value, 73, on the stem plot.

2 To find the quartiles, **the median value is excluded**. This leaves 11 values below the median and 11 values above the median. Then:

- Q_1 = midpoint of the bottom 11 data values
- Q_3 = midpoint of the top 11 data values

Mark Q_1 and Q_3, on the stem plot. Write these values down.



$Q_1 = 66$, $Q_3 = 75$

3 Determine the *IQR* using $IQR = Q_3 - Q_1$

$\therefore IQR = Q_3 - Q_1 = 75 - 66 = 9$

4 Determine the range using

R = largest data value − smallest data value

$R = 77 - 52 = 25$

Note: To check that these quartiles are correct, write the data values down in order, and mark in the median and the quartiles. If correct, the median divides the data set up into four equal groups.

$$\underbrace{52\ \ 55\ \ 56\ \ 64\ \ 66}_{5\text{ values}}\ \overset{Q_1}{66}\ \underbrace{67\ \ 69\ \ 71\ \ 72\ \ 72}_{5\text{ values}}\ \overset{Q_2\ (=M)}{73}\ \underbrace{73\ \ 74\ \ 74\ \ 74\ \ 74}_{5\text{ values}}\ \overset{Q_3}{75}\ \underbrace{75\ \ 76\ \ 76\ \ 77\ \ 77}_{5\text{ values}}$$

Why is the *IQR* a more useful measure of spread than the range?

The *IQR* is a measure of spread of a distribution that includes the middle 50% of observations. Since the upper 25% and lower 25% of observations are discarded, the interquartile range is generally not affected by the presence of outliers. This makes it a more useful measure of spread than the range.

Exercise 2A

1 Write down in a few words the meaning of the following terms:

a range **b** median **c** quartiles **d** interquartile range

2 Locate the medians of the following data sets. In each case, check that the median divides the ordered data set into two **equal** groups.

a 4 9 3 1 8 6 **b** 10 9 12 20 14

c 103 109 99 112 87 90 103 100

d 0.01 1.03 0.4 2.05 0.59 0.009 0.63

Back to Menu >>>

3 The prices of nine second-hand mountain bikes advertised for sale were as follows:

\$650 \$3500 \$750 \$500 \$1790 \$1200 \$2950 \$430 \$850

What is the median price of these bikes? Check that an equal number of bikes have prices above and below the median.

4 Find the median, M, the quartiles, Q_1 and Q_3, and the IQR for each of the following sets of numbers:

a 16 18 14 12 11 9 12 14 16 **b** 7 14 21 28 14 21 28 28 14 21 28

c 3 4 8 2 4 7 9 3 7 4 12 16 18 5

5 The stem plot shows the distribution of infant mortality rates (deaths per 1000 live births) in 14 countries.

a Determine the median M.

b Determine the quartiles Q_1 and Q_3.

c Calculate the IQR.

d Calculate the range R.

e By writing the data values out in a line, check that the quartiles and the median have divided the data set up into four equal groups.

Infant mortality rate (/1000)

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

6 The stem plot shows the distribution of test scores for 20 students.

a Determine the median M.

b Determine the quartiles Q_1 and Q_3.

c Calculate the IQR.

d Calculate the range R.

Test scores

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

7 The stem plot shows the distribution of university participation rates (%) in 23 countries.

a Determine the median M.

b Determine the quartiles Q_1 and Q_3.

c Calculate the IQR.

d Calculate the range R.

University participation rates (%)

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

2.3 The five-number summary and the box plot

The five-number summary

Knowing the median and quartiles of a distribution means we know quite a lot about the centre of the distribution. If we also knew something about the tails (ends) of the distributions then we would have a good picture of the whole distribution. This can be achieved by recording the smallest and largest values of the data set. Putting all this information together gives the **five-number summary**.

Five-number summary

A listing of the median M, the quartiles Q_1 and Q_3, and the smallest and largest data values of a distribution, written in the order:

Minimum, Q_1, M, Q_3, Maximum

is known as a **five-number summary**.

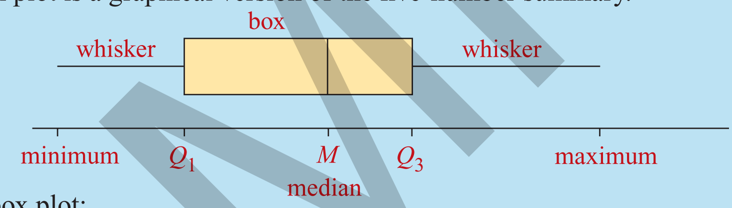
The five-number summary can be used to construct a new graph known as the **box plot**. The box plot is an extremely powerful tool for describing data distributions.

The box plot

In its simplest form, the **box plot** (or box-and-whisker plot as it is sometimes called) is a *graphical version of a five-number summary*. As we shall see, a box plot is a very compact way of displaying the location, spread and general shape of a distribution. It is also a very useful tool for comparing distributions of various related subgroups. Box plots can be drawn either vertically or horizontally.

The box plot

A box plot is a graphical version of the five-number summary.

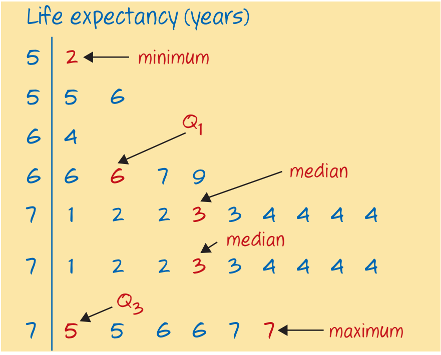


In a box plot:

- a box is used to represent the middle 50% of scores
- the median is shown by a vertical line drawn within the box
- lines (called whiskers) are extended out from the lower and upper ends of the box to the smallest and largest data values of the data set respectively

How to construct a box plot from a stem plot

The stem plot shows the distribution of life expectancies (in years) in 23 countries. Display the data in the form of a box plot.



Life expectancy (years)

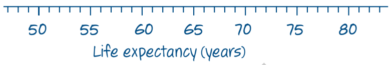
5	2
5	5 6
6	4
6	6 6 7 9
7	1 2 2 3 3 4 4 4 4
7	1 2 2 3 3 4 4 4 4
7	5 5 6 6 7 7

Back to Menu >>>

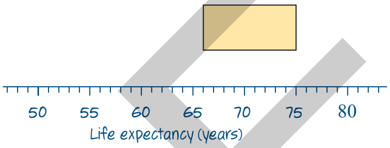
1 Use the stem plot to write down the five-number summary.

Min = 52, Q_1 = 66, M = 73, Q_3 = 75, Max = 77

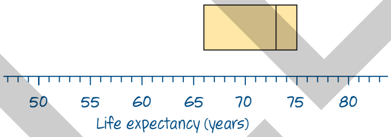
2 Draw in a labelled and scaled number line that covers the full range of values.



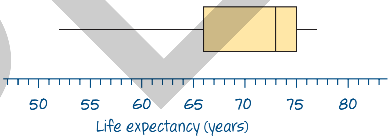
3 Draw in a box starting at $Q_1 = 66$ and ending at $Q_3 = 75$.



4 Mark in the median value with a vertical line segment at $M = 73$.



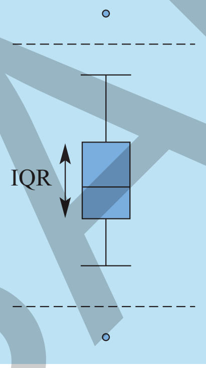
5 Draw in the whiskers: lines joining midpoint of the ends of the box to the minimum and maximum values, 52 and 77.



Box plots with outliers

The box plot with outliers is a more sophisticated form of the box plot and is designed to identify any outliers that may be present in the data. How this is done is illustrated below.

Anatomy of a box plot with outliers



Maximum value:	possible outlier
Upper fence:	$Q_3 + 1.5 \times IQR$ (not drawn in)
Upper adjacent value:	highest data value inside fence
Third quartile:	Q_3
Median:	M
First quartile:	Q_1
Lower adjacent value:	lowest data value inside fence
Lower fence:	$Q_1 - 1.5 \times IQR$ (not drawn in)
Minimum value:	possible outlier

Two new things to note in a box plot with outliers are that:

- any points more than 1.5 *IQR*s away from the end of the box are classified as possible outliers (possible, in that it may be that they are just part of a distribution with a very long tail and we do not have enough data to pick up other values in the tail)
- the whiskers end at the highest and lowest data values that lie within 1.5 *IQR*s from the ends of the box

Box plots with outliers take more time to construct than standard box plots. However, they are normally constructed with the aid of a graphics calculator. Your prime task is to be able to recognise and interpret them, not just construct them.

Back to Menu >>>

How to construct a box plot with outliers using the TI-Nspire CAS

Display the following set of 19 marks in the form of a box plot with outliers.

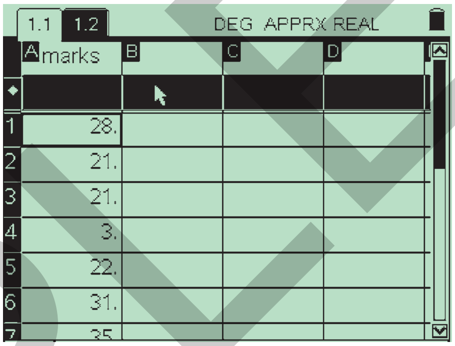
28 21 21 3 22 31 35 26 27 33 36 35 23 24
43 31 30 34 48

Steps

1 Start a new document: ctrl + N.

2 Select **3:Add Lists & Spreadsheet**.
Enter the data into a list called *marks*, as shown.

3 Statistical graphing is done through the **Data & Statistics** application.



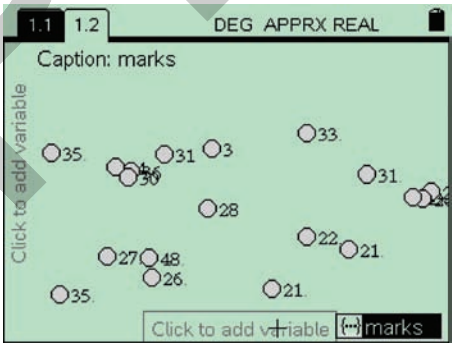
1.1 1.2 DEG APPRX REAL

	A marks	B	C	D
♦				
1	28.			
2	21.			
3	21.			
4	3.			
5	22.			
6	31.			
7	35.			

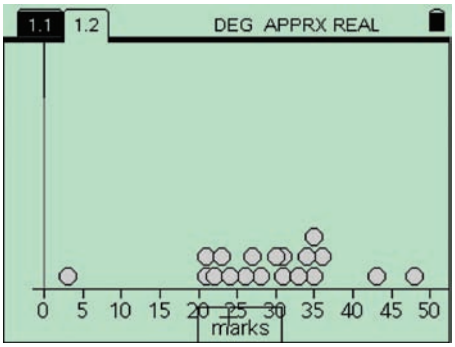
Press ⌂ and select **5:Data & Statistics**.

Note: A random display of dots will appear – this is to indicate list data are available for plotting. It is not a statistical plot.

a Move your cursor to the text box area below the horizontal axis. Press ☝ when prompted and select the variable *marks*. Press enter to paste the variable *marks* to that axis.

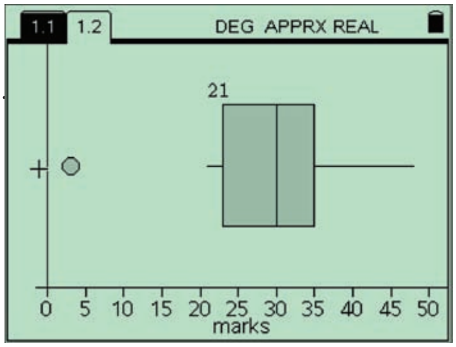


b A dot plot is displayed as the default plot. To change the plot to a box plot with outliers press menu/**1:Plot Type/2:Box Plot**.



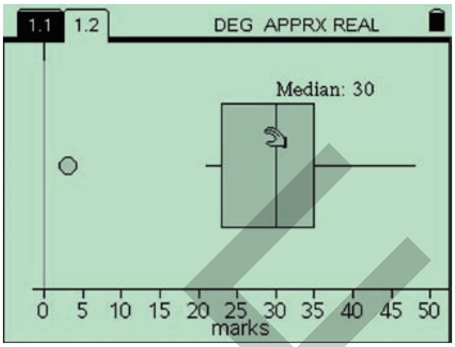
Your screen should now look like that shown opposite.

Hint: ctrl + menu will give you contextual menus. You can change the box plot properties such as extend the whiskers or alter Window Settings as required.



4 Data analysis

Key values can be read from the box plot by using the horizontal arrow keys (◀ and ▶) to move the cursor from point to point on the box plot. A ☝ will show each time a key point is reached. To see the outlier value, hold the centre mouse button ◉ until ✊ appears. Press ◉ to exit this point.



Starting at the far left of the plot, we see that the

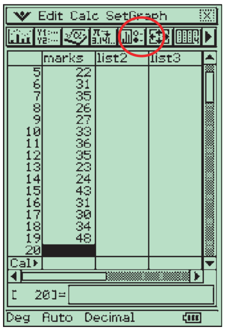
- minimum value is 3 (i.e. the outlier)
- lower adjacent value is 21
- first quartile is 23 (**Q_1 = 23**)
- median is 30 (**Median = 30**)
- third quartile is 35 (**Q_3 = 35**)
- maximum value is 48

How to construct a box plot with outliers using the ClassPad

Display the following set of 19 marks in the form of a box plot with outliers.

28 21 21 3 22 31 35 26 27 33 36 35 23 24
43 31 30 34 48

1 Open the **Statistics** application and enter the data into the column labelled **marks**. Your screen should look like the one shown.



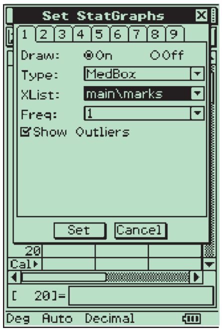
	marks	list2	list3
5	22		
6	31		
7	35		
8	26		
9	27		
10	33		
11	36		
12	35		
13	23		
14	24		
15	43		
16	31		
17	30		
18	34		
19	48		
20			

[20]=

2 Open the **Set StatGraphs** dialog box by tapping ▥ in the toolbar. Complete the dialog box as given below. For

- **Draw**: select **On**
- **Type**: select **MedBox** (▾)
- **XList**: select **main \ marks** (▾)
- **Freq**: leave as **1**

Tap the **Show Outliers** box to add a tick (☑).

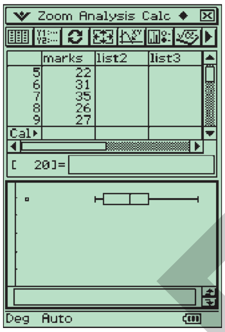


Draw: ● On ○ Off
Type: MedBox
XList: main\marks
Freq: 1
☑ Show Outliers

[20]=

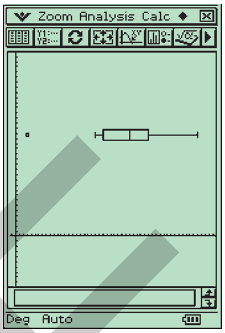
Back to Menu >>>

3 Tap SET to confirm your selections and plot the box plot. The graph is drawn in an automatically scaled window, as shown.



	marks	list2	list3
5	22		
6	31		
7	35		
8	26		
9	27		

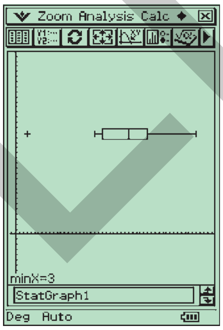
Cal▸
[20]=



4 Tap the Resize icon at the bottom of the screen for a full-screen graph.

Note: If you have more than one graph on your screen, tap the data screen, select **StatGraph** and turn off any unwanted graphs.

5 To read key values from the boxplot, tap in the toolbar. This places a marker on the box plot (+), as shown. The horizontal cursor arrows (◀ and ▶) can then be used to move from point to point on the box plot.

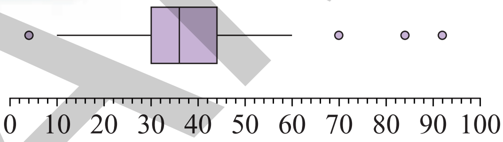


minX=3
StatGraph1

Starting at the far left of the plot, we see that the

- minimum value is 3 (**minX** = **3**; i.e. the outlier)
- lower adjacent value is 21 (***xc*** = **21**)
- first quartile is 23 (**Q$_1$** = **23**)
- median is 30 (**Med: 30**)
- third quartile is 35 (**Q$_3$** = **35**)
- maximum value is 48 (**maxX** = **48**)

Example 4 Reading values from a box plot



For the box plot above, write down the values of:

a the median

b the quartiles Q_1 and Q_3

c the interquartile range (IQR)

d the minimum and maximum values

e the values of any possible outliers

f the smallest value in the upper end of the data set that will be classified as an outlier

g the largest value in the lower end of the data set that will be classified as an outlier

Back to Menu >>>

Solution

a median (vertical line in the box) — $M = 36$

b quartiles Q_1andQ_3 (end points of box) — $Q_1 = 30,\ Q_3 = 44$

c interquartile range ($IQR = Q_3 - Q_1$) — $IQR = Q_3 - Q_1 = 44 - 30 = 14$

d minimum and maximum values (extremes) — $Min = 4,\ Max = 92$

e the values of any outliers (dots) — 4, 78, 84 and 92

f upper fence (given by $Q_3 + 1.5 \times IQR$)

$$\text{upper fence} = Q_3 + 1.5 \times IQR$$
$$= 44 + 1.5 \times 14 = 65$$

Any value above 65 is an outlier.

g lower fence (given by $Q_1 - 1.5 \times IQR$)

$$\text{lower fence} = Q_3 - 1.5 \times IQR$$
$$= 30 - 1.5 \times 14 = 9$$

Any value below 9 is an outlier.

Exercise 2B

1 The ordered stem plot shows the distribution of infant mortality rates for 14 countries. Use the stem plot to construct:

a a five-number summary

b a box plot

Infant mortality rates

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

2 The ordered stem plot shows the price (in \$000s) of 23 houses sold in a country town. Use the stem plot to construct:

a a five-number summary

b a standard box plot

House prices

13	6 7
14	3 6 8 8 9
15	2 5 8 8 8
16	4 5 5 6 7 9
17	8 8 9
18	2 9

3 University participation rates (%) in 21 countries are listed below:

3 3 7 8 9 12 13 15 17 20 21
22 25 26 26 26 27 30 36 37 55

a Use a calculator to construct a box plot with outliers for this data. Name variable, *unirate*.

b Use the box plot to write down a five-number summary for this data. Identify any outliers and their values.

4 The reaction times (in milliseconds) of 18 people are listed below:

38 36 35 35 43 46 42 64 40 48 35 34 40 44 30 25 39 31

a Use a calculator to construct a box plot with outliers for this data. Name variable, *rtime*.

b Use the box plot to write down a five-number summary for this data. Identify any outliers and their values.

Back to Menu >>>

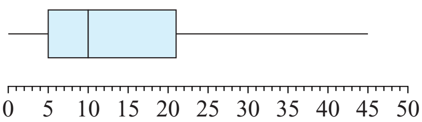
5 For each of the box plots below, estimate the values of:

i the median M **ii** the quartiles Q_1 and Q_3

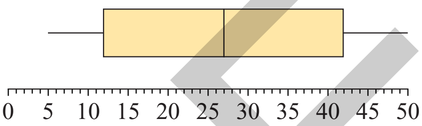
iii the interquartile range IQR **iv** the minimum and maximum values

v the values of any outliers

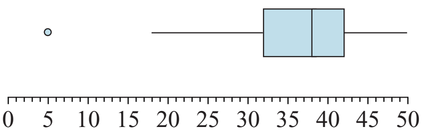
a



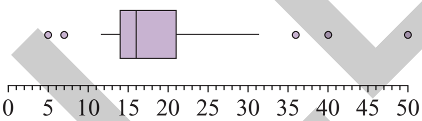
b



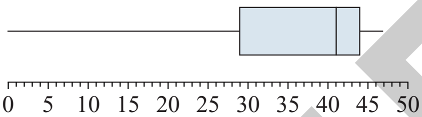
c



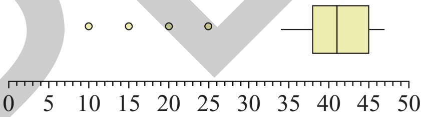
d



e



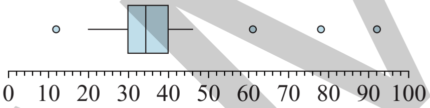
f



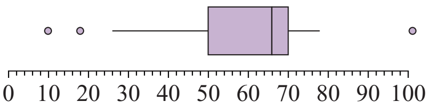
6 For the box plots below, determine the location of:

i the upper fence **ii** the lower fence

a



b



2.4 Relating a box plot to distribution shape

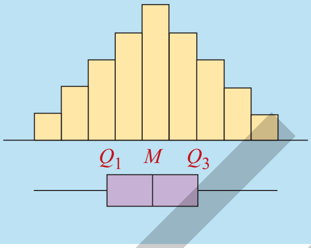
There are an almost infinite variety of quantities that can be subjected to statistical analysis. However, the types of distributions that arise tend to fall into a relatively small number of characteristic forms or shapes. Furthermore, each of these shapes tends to have quite distinct box plots.

A symmetric distribution

A **symmetric distribution** is evenly spread out around the median. There is also a strong tendency for data values to cluster around the centre of the distribution rather than at the extremes. Examples include the heights of a sample of 16-year-old girls or the scores obtained on an intelligence test.

Back to Menu >>>

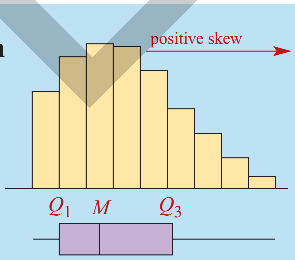
For a **symmetric** distribution, the **box plot** is also symmetric. The **median** is generally in the **middle** of the box and the **whiskers** are approximately **equal** in length.



Positively skewed distributions

Positively skewed distributions are characterised by a cluster of data values at the left-hand end of the distribution with a **gradual tailing off to the right**. An example of such a distribution would be the distribution of male road deaths with age. In this case there is a disproportionate number of deaths in the age group 18–25 years.

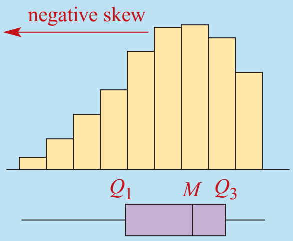
The box plot of a **positively skewed** distribution has the **median off-centre** and generally to the left. The **left-hand whisker** will be **short**, while the **right-hand whisker** will be **long**, reflecting the gradual tailing off of data values to the right.



Negatively skewed distributions

Negatively skewed distributions are characterised by a clustering of data values to the right-hand side of the distribution, with a **gradual tailing off to the left**. An example would be the age of home owners, as few young people own homes but many older people do.

The box plot of a **negatively skewed** distribution has the **median off-centre** and generally to the right. The **right-hand whisker** will be **short**, while the **left-hand whisker** will be **long**, reflecting the gradual tailing off of data values to the left.



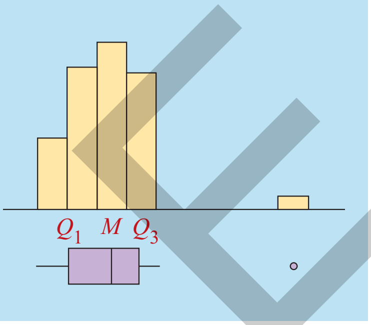
Distributions with outlier(s)

Sports data often contains outliers. For example, the heights of the players in a football side vary but do so within a limited range. One exception is the ‘knock’ ruckman, who may be exceptionally tall and well outside the normal range of variation. In statistical terms, the exceptionally tall ruckman is an outlier, because his height does not fit in the range of heights that might be regarded as typical for the team.

Back to Menu >>>

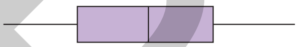
Less interesting but of practical importance, an outlier might signal an error in the data. For example, when studying the age distribution of residents in a large country town, a value of 165 would show up as an outlier in the box plot and signal a possible recording or data entry error.

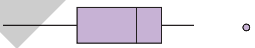
Distributions with outliers are characterized by **gaps** between the main body and data values in the tails. The histogram opposite, displays a distribution with an outlier. In the corresponding box plot, the **box and whiskers** represents the **main body of data** and the **dot** indicates the **outlier**.

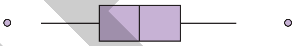


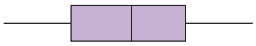
Exercise 2C

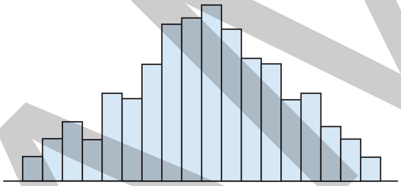
Match these box plots with their histograms.

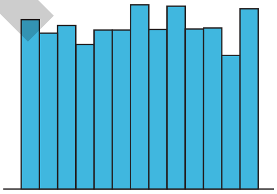

Box plot 1

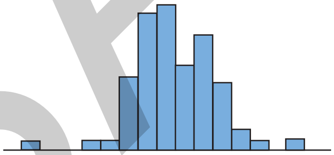

Box plot 2

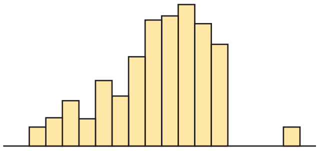

Box plot 3


Box plot 4


Histogram A


Histogram B


Histogram C


Histogram D

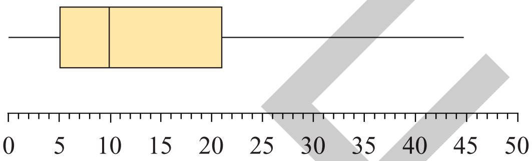
2.5 Interpreting box plots: describing and comparing distributions

Because of the wealth of information contained in a box plot, it is an extremely powerful tool for describing a distribution. At a glance, we can see the **shape** of the distribution. We can also see whether or not there are any **outliers**. Furthermore, the **centre** of the distribution is clearly identified and given a value by the **median**. Finally, the **spread** of the distribution can be seen in two ways. The first is given by the length of the box. This corresponds to the spread of the middle 50% of values, the ***IQR***. The second measure of spread given by a box plot is the

range. When there are no outliers, this is given by the length of the box plus the whiskers. If there are outliers, these are also included in determining the value of the range.

Example 5 Using a box plot to describe a distribution without outliers

Describe the distributions represented by the box plot in terms of shape, centre, spread. Give appropriate values.

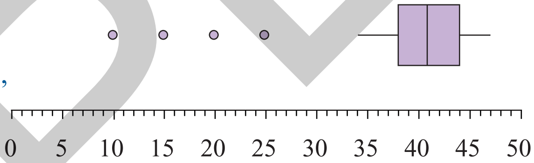


Solution

The distribution is positively skewed with no outliers. The distribution is centred at 10, the median value. The spread of the distribution, as measured by the IQR is 16 and, as measured by the range, 45.

Example 6 Using a box plot to describe a distribution with outliers

Describe the distributions represented by the box plot in terms of shape and outliers, centre, spread. Give appropriate values.

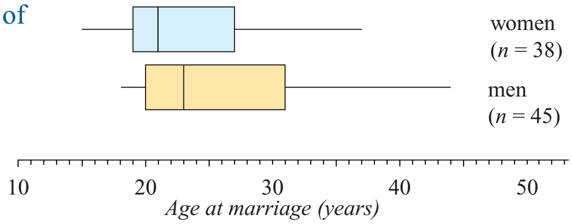


Solution

The distribution is symmetric but with outliers. The distribution is centred at 41, the median value. The spread of the distribution as measured by the IQR is 6 and, as measured by the range, 37. There are four outliers: 10, 15, 20 and 25.

Example 7 Using a box plot to compare distributions

The parallel boxplots show the distribution of age at marriage of 45 married men and 38 married women.



a Compare the two distributions in terms of shape (including outliers, if any), centre and spread. Give appropriate values at a level of accuracy that can be read from the plot.

b Comment on how the age at marriage of men compares to women for this data.

Solution

a The distributions of age at marriage are positively skewed for both men and women. There are no outliers. The median age at marriage is higher for men ($M = 23$ years) than women ($M = 21$ years). The IQR is also greater for men ($IQR = 11$ years) than women ($IQR = 8$ years). The range of age at marriage is also greater for men ($R = 26$ years) than women ($R = 22$ years).

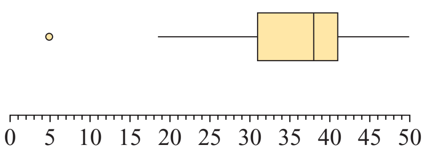
b For this group of men and women, the men on average married at an older age and the age at which they married is more variable.

Back to Menu >>>

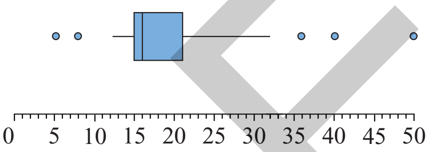
Exercise 2D

1 Describe the distributions represented by the following box plots in terms of shape, centre, spread and outliers (if any). Give appropriate values.

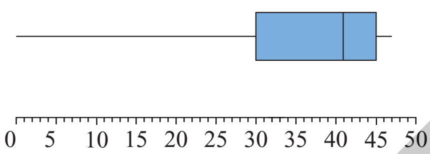
a



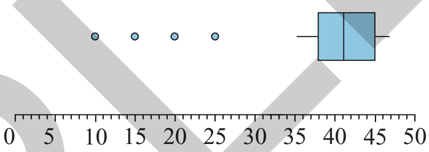
b



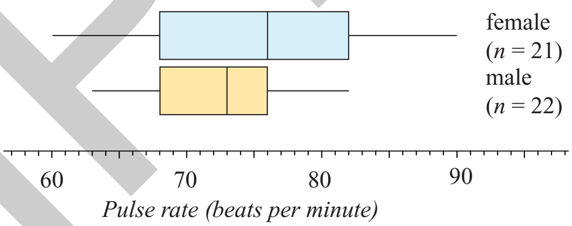
c



d



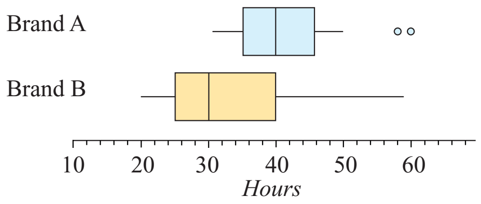
2 The parallel box plots show the distribution of pulse rate of 21 adult females and 22 adult males.



a Compare the two distributions in terms of shape (including outliers, if any), centre and spread. Give appropriate values at a level of accuracy that can be read from the plot.

b Comment on how the pulse rates of females compare to the pulse rates of men for this data.

3 The lifetimes of two different brands of batteries were measured and the results displayed in the form of parallel box plots.



a Compare the two distributions in terms of shape (including outliers, if any), centre and spread. Give appropriate values at a level of accuracy that can be read from the plot.

b Comment on how the lifetime of Brand A compares to the lifetime of Brand B batteries for this data.

MC TEST D&D TEST

Key ideas and chapter summary

Summary statistics — **Summary statistics** are used to give numerical values to special features of a data distribution such as centre and spread.

The median — The **median** is a summary statistic that can be used to locate the **centre** of a distribution. It is the midpoint of a distribution dividing an ordered data set into two equal parts.

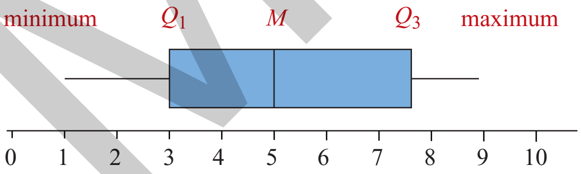
Quartiles — **Quartiles** are summary statistics that divide an ordered data set into four equal groups.

- The first quartile, Q_1, marks off the first 25% of values.
- The second quartile, Q_2 (which is also the median), marks off the first 50% of values.
- The third quartile, Q_3, marks off the first 75% of values.

The interquartile range — The **interquartile range** is defined as $IQR = Q_3 - Q_1$. The IQR gives the **spread** of the middle 50% of data values.

Five-number summary — The median, the first quartile, the third quartile, along with the minimum and the maximum values in a data set, are known as a **five-number summary**.

Box plots — A standard **box plot** is a graphical representation of a five-number summary.

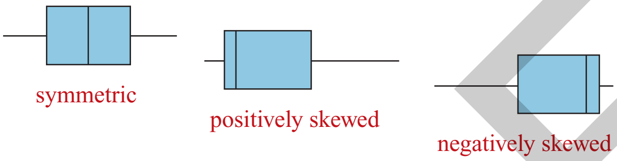


Back to Menu >>>

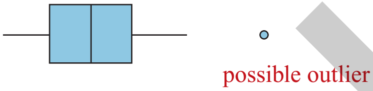
Interpreting box plots

Box plots are powerful tools for picturing and comparing data sets as they give both a visual view and a numerical summary of a distribution.

- **shape:** symmetric or skewed (positive or negative)?



- **outliers:** values that appear to stand out

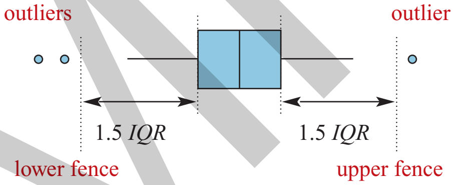


- **centre:** the midpoint of the distribution (the median)
- **spread:** the *IQR* and the range of values covered

Outliers

In a box plot, outliers are defined as being those values that are

- greater than $Q_3 + 1.5 \times IQR$ (upper fence)
- less than $Q_1 - 1.5 \times IQR$ (lower fence)



Skills check

Having completed this chapter you should be able to:

- locate the median and the quartiles of a data set and hence calculate the *IQR*
- produce a five-number summary from a set of data
- construct a box plot from a stem plot
- construct a box plot from raw data using a graphics calculator
- use a box plot to identify key features of a data set, such as shape (including outliers if any), centre and spread
- use the information in a box plot to describe and compare distributions

Back to Menu >>>

Multiple-choice questions

The following information relates to Questions 1 to 3

The following is a set of test marks: 11, 4, 13, 15, 16, 19, 8, 10, 12

1 The median value is:

A 10 **B** 11 **C** 12 **D** 12.5 **E** 13

2 The first quartile is:

A 9 **B** 10 **C** 11 **D** 12 **E** 12.5

3 The range is:

A 11 **B** 12 **C** 12.5 **D** 13 **E** 15

The following information relates to Questions 4 to 5

The following is a set of test marks: 11, 4, 13, 15, 16, 19, 8, 10, 12, 16

4 The median value is:

A 10 **B** 11 **C** 12 **D** 12.5 **E** 13

5 The interquartile range is:

A 5 **B** 6 **C** 7 **D** 8 **E** 9

6 The following is an ordered set of 10 daily maximum temperatures (in degrees Celsius):

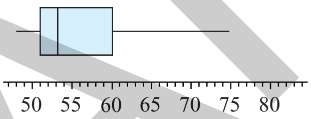
22 22 23 24 24 25 26 27 28 29

The five-number summary for these temperatures is:

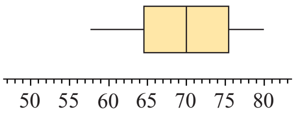
A 22, 23, 24, 27, 29 **B** 22, 23, 24.5, 27, 29 **C** 22, 24, 24.5, 27, 29
D 22, 23, 24.5, 27.5, 29 **E** 22, 24, 24.5, 27, 29

The following information relates to Questions 7 to 15

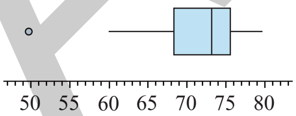
A



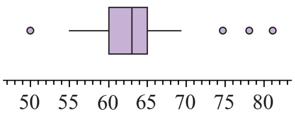
B



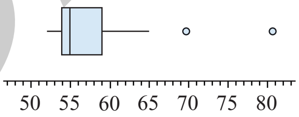
C



D



E



7 The median of box plot D is closest to:

A 50 **B** 60 **C** 63 **D** 65 **E** 70

8 The *IQR* of box plot B is closest to:

A 10 **B** 20 **C** 25 **D** 65 **E** 75

9 The range of box plot E is closest to:

A 4 **B** 13 **C** 20 **D** 30 **E** 80

10 The description that best matches box plot A is:

A symmetric **B** symmetric with outliers **C** negatively skewed
D positively skewed **E** positively skewed with outliers

11 The description that best matches box plot B is:

A symmetric **B** negatively skewed with an outlier
C negatively skewed **D** positively skewed **E** positively skewed with outliers

12 The description that best matches box plot C is:

A symmetric **B** negatively skewed with an outlier
C negatively skewed **D** positively skewed **E** positively skewed with outliers

13 The description that best matches box plot D is:

A symmetric **B** symmetric with outliers
C negatively skewed **D** positively skewed **E** positively skewed with outliers

14 The description that best matches box plot E is:

A symmetric **B** symmetric with outliers **C** negatively skewed
D positively skewed **E** positively skewed with an outlier

15 To be an outlier in box plot D, a score must be:

A either less than 52.5 or greater than 72.5 **B** greater than 72.5
C either less than 55 or greater than 70 **D** greater than 70
E less than 55

Extended-response questions

1 A group of 16 obese people attempted to lose weight by joining a regular exercise group. The following weight losses in kilograms were recorded.

26 14 7 38 23 21 17 4 18 34 24 29 2 13 33 15

a Use your calculator to construct a box plot for the data. Name variable, *wloss*.

b Use the box plot to locate the median and the quartiles Q_1 and Q_3.

c Complete the following statements:

'The middle 50% of the people who exercised had weight losses between ☐ kilograms and ☐ kilograms.'

'Twenty-five per cent of people lost less than ☐ kilograms.'

d Use the box plot to describe the distributions of weight loss in terms of shape, centre, spread and outliers (if any). Give appropriate values.

2 The weights (in kg) carried by the horses in a handicap race at a country meeting are given below.

60 57 57 55 54 53 53 53 52 52 51.5 51

a Use your calculator to construct a box plot. Name variable, *hweight*.

b Complete a five-number summary for the weights carried by the horses.

c What is the interquartile range?

d Use the box plot to describe the distributions of weight carried by the horses in terms of shape, centre, spread and outliers (if any). Give appropriate values.

3 The strike rates (runs/100 balls) of cricketers playing in a one-day cricket competition are given below.

31 70 63 59 85 54 61 60 69 61 54 56 63 95 81 67

a Use your calculator to construct a box plot for the data. Name variable, *srate*.

b Use the box plot to locate the median and the quartiles Q_1 and Q_3.

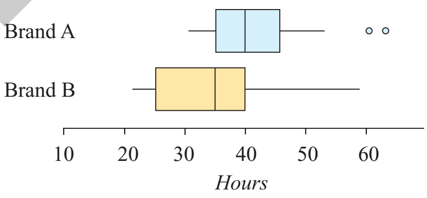
c Complete the following statements:

'The top 25% of the players had strike rates above □ runs/100 balls.'

'Fifty per cent of players had strike rates less than □ runs/100 balls.'

d Use the box plot to describe the distribution of strike rates in terms of shape, centre, spread and outliers (if any). Give appropriate values.

4 The life of two different brands of batteries was measured and the results displayed in the form of parallel box plots.



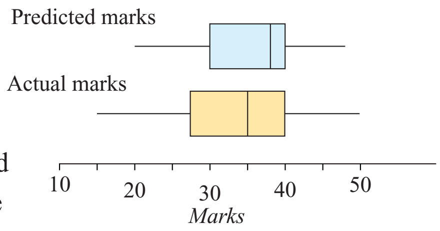
a On average, which brand of batteries had the longest life? Explain.

b Which brand of batteries had a more variable lifetime? Explain.

c What do the two outliers for Brand A represent?

d Both brands of batteries cost the same. On the basis of this information, which brand of battery would you buy and why?

5 To find out how well she could estimate her students' marks on a test, a statistics teacher set a test and then, before marking the test, predicted the mark she thought her students would get. After marking the test, she produced a parallel box plot to enable her to compare the two sets of marks. The box plots are shown opposite. The test was marked out of 50.



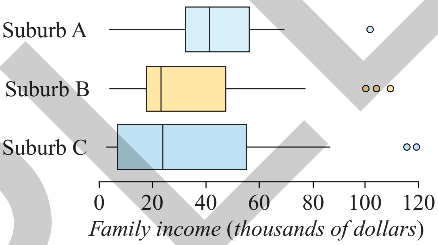
a On average, did the teacher tend to overestimate or underestimate her students' marks? Explain.

b Were the teacher's marks more or less variable then the actual marks? Explain.

c Compare the two distributions in terms of shape (including outliers, if any), centre and spread. Give appropriate values at a level of accuracy that can be read from the plot.

d Comment on how the predicted marks of the teacher compared to the students' actual marks.

6 A random sample of 250 families from three different suburbs was used in a study to try to identify factors that influenced a family's decision about taking out private health insurance. One variable investigated was family income. The information gathered on family incomes is presented opposite in the form of parallel box plots.



a In which suburb was the median household income the greatest?

b In which suburb were family incomes most variable?

c What do the outliers represent?

d Which of the following statements are true?

i 'At least 75% of the families in Suburb A have an income that exceeds the median family income in Suburb B.'

ii 'More than 50% of the families in Suburb A have incomes less than $45 000.'

iii 'The distribution of family incomes in Suburb C is approximately symmetric.'

iv 'The *mean* family income in Suburb B is greater than the *median* family income in Suburb B.'