Analysis of discrete data

Contents:
A Variables used in statistics [11.2]
B Organising and describing discrete data [11.2, 11.3]
C The centre of a discrete data set [11.4]
D Measuring the spread of discrete data [11.4]
E Data in frequency tables [11.4]
F Grouped discrete data [11.4]
G Statistics from technology [11.8]

Historical note

- Florence Nightingale (1820-1910), the famous “lady with the lamp”, developed and used graphs to represent data relating to hospitals and public health.

- Today about 92% of all nations conduct a census at regular intervals. The UN gives assistance to developing countries to help them with census procedures, so that accurate and comparable worldwide statistics can be collected.

Opening problem

A stretch of highway was notoriously dangerous, being the site of a large number of accidents each month.

In an attempt to rectify this, the stretch was given a safety upgrade. The road was resurfaced, and the speed limit was reduced.

To determine if the upgrade has made a difference, data is analysed for the number of accidents occurring each month for the three years before and the three years after the upgrade.
The results are:

**Before upgrade**

| 8 | 4 | 9 | 7 |
| 11 | 10 | 8 | 8 |
| 6 | 10 | 10 | 9 |
| 8 | 6 | 7 | 9 |
| 5 | 7 | 2 | 6 |
| 9 | 7 | 10 | 6 |
| 8 | 7 | 10 | 8 |

**After upgrade**

| 4 | 7 | 8 | 3 |
| 8 | 4 | 6 | 7 |
| 8 | 4 | 7 | 7 |
| 9 | 8 | 6 | 5 |
| 6 | 8 | 6 | 7 |
| 7 | 3 | 5 | 5 |
| 9 | 7 | 8 | 7 |
| 6 | 6 | 7 | 8 |
| 8 | 9 | 7 | 4 |

**Things to think about:**

- Can you state clearly the problem that needs to be solved?
- What is the best way of organising this data?
- What are suitable methods for displaying the data?
- How can we best indicate what happens in a typical month on the highway?
- How can we best indicate the spread of the data?
- Can a satisfactory conclusion be made?

**STATISTICS**

*Statistics* is the art of solving problems and answering questions by collecting and analysing data.

The facts or pieces of information we collect are called **data**. Data is the plural of the word *datum*, which means a single piece of information.

A list of information is called a **data set** and because it is not in an organised form it is called **raw data**.

The process of **statistical enquiry** (or *investigation*) includes the following steps:

**Step 1:** Examining a problem which may be solved using data and posing the correct question(s).

**Step 2:** Collecting data.

**Step 3:** Organising the data.

**Step 4:** Summarising and displaying the data.

**Step 5:** Analysing the data, making a conclusion in the form of a conjecture.

**Step 6:** Writing a report.

**CENSUS OR SAMPLE**

The two ways to collect data are by census or sample.

A **census** is a method which involves collecting data about every individual in a **whole population**.

The individuals in a population may be people or objects. A census is detailed and accurate but is expensive, time consuming, and often impractical.

A **sample** is a method which involves collecting data about a **part of the population** only.

A sample is cheaper and quicker than a census but is not as detailed or as accurate. Conclusions drawn from samples always involve some error.

A sample must truly reflect the characteristics of the whole population. It must therefore be **unbiased** and **sufficiently large**.

A **biased sample** is one in which the data has been unfairly influenced by the collection process and is not truly representative of the whole population.
There are two types of variables that we commonly deal with: categorical variables and quantitative variables.

A **categorical variable** is one which describes a particular quality or characteristic. It can be divided into categories. The information collected is called **categorical data**.

Examples of categorical variables are:
- **Getting to school**: the categories could be train, bus, car and walking.
- **Colour of eyes**: the categories could be blue, brown, hazel, green, and grey.

We saw examples of categorical variables in Chapter 5.

A **quantitative variable** is one which has a numerical value, and is often called a **numerical variable**. The information collected is called **numerical data**.

Quantitative variables can be either **discrete** or **continuous**.

A **quantitative discrete variable** takes exact number values and is often a result of **counting**.

Examples of discrete quantitative variables are:
- **The number of people in a household**: the variable could take the values 1, 2, 3, ...
- **The score out of 30 for a test**: the variable could take the values 0, 1, 2, 3, ..., 30.
- **The times on a digital watch**: the variable could be 12:15.

A **quantitative continuous variable** takes numerical values within a certain continuous range. It is usually a result of **measuring**.

Examples of quantitative continuous variables are:
- **The weights of newborn babies**: the variable could take any positive value on the number line but is likely to be in the range 0.5 kg to 7 kg.
- **The heights of Year 10 students**: the variable would be measured in centimetres. A student whose height is recorded as 145 cm could have exact height anywhere between 144.5 cm and 145.5 cm.

In this chapter we will focus on **discrete** variables. Continuous variables will be covered in Chapter 17.

**INTERNET STATISTICS**

There are thousands of sites worldwide which display statistics for everyone to see. Sites which show statistics that are important on a global scale include:
- [www.un.org](http://www.un.org) for the United Nations
- [www.who.int](http://www.who.int) for the World Health Organisation
EXERCISE 13A

1 Classify the following variables as either categorical or numerical:
   a the brand of shoes a person wears
   b the number of cousins a person has
   c voting intention at the next election
   d the number of cars in a household
   e the temperature of coffee in a mug
   f favourite type of apple
   g town or city where a person was born
   h the cost of houses on a street

2 Write down the possible categories for the following categorical variables:
   a gender
   b favourite football code
   c hair colour

3 State whether a census or a sample would be used for these investigations:
   a the reasons for people using taxis
   b the heights of the basketballers at a particular school
   c finding the percentage of people in a city who suffer from asthma
   d the resting pulse rates of members of your favourite sporting team
   e the number of pets in Canadian households
   f the amount of daylight each month where you live

4 Discuss any possible bias in the following situations:
   a Only Year 12 students are interviewed about changes to the school uniform.
   b Motorists stopped in peak hour are interviewed about traffic problems.
   c A phone poll where participants must vote by text message.
   d A ‘who will you vote for’ survey at an expensive city restaurant.

5 For each of the following possible investigations, classify these quantitative variables as quantitative discrete or quantitative continuous:
   a the number of clocks in each house
   b the weights of the members of a basketball team
   c the number of kittens in each litter
   d the number of bread rolls bought each week by a family
   e the number of leaves on a rose plant stem
   f the amount of soup in each can
   g the number of people who die from heart attacks each year in a given city
   h the amount of rainfall in each month of the year
   i the stopping distances of cars travelling at 80 km/h
   j the number of cars passing through an intersection each hour

B ORGANISING AND DESCRIBING DISCRETE DATA [11.2, 11.3]

In the Opening Problem on page 275, the quantitative discrete variable is the number of accidents per month.

To organise the data a tally-frequency table could be used. We count the data systematically and use a ‘|’ to indicate each data value. We use | to represent 5.
Below is the table for Before upgrade:

<table>
<thead>
<tr>
<th>No. of accidents/month</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>36</strong></td>
</tr>
</tbody>
</table>

A *vertical bar chart* could be used to display this data:

**DESCRIBING THE DISTRIBUTION OF A DATA SET**

The **mode** of a data set is the most frequently occurring value(s). Many data sets show symmetry or partial symmetry about the mode.

If we place a curve over the vertical bar chart we see that this curve shows symmetry. We say that we have a *symmetrical distribution*.

The distribution alongside is said to be **negatively skewed** because, by comparison with the symmetrical distribution, it has been ‘stretched’ on the left (or negative) side of the mode.

So, we have:

**EXERCISE 13B.1**

1. 20 students were asked “How many pets do you have in your household?” and the following data was collected: 2 1 0 3 1 2 1 3 4 0 0 2 2 0 1 1 0 1 0 1 1

   a. What is the variable in this investigation?
   b. Is the data discrete or continuous? Why?
   c. Construct a vertical bar chart to display the data. Use a heading for the graph, and add an appropriate scale and label to each axis.
   d. How would you describe the distribution of the data? Is it symmetrical, positively skewed or negatively skewed?
   e. What percentage of the households had no pets?
   f. What percentage of the households had three or more pets?
2 A randomly selected sample of shoppers was asked, ‘How many times did you shop at a supermarket in the past week?’ A column graph was constructed for the results.
   a How many shoppers gave data in the survey?
   b How many of the shoppers shopped once or twice?
   c What percentage of the shoppers shopped more than four times?
   d Describe the distribution of the data.

3 The number of toothpicks in a box is stated as 50 but the actual number of toothpicks has been found to vary. To investigate this, the number of toothpicks in a box was counted for a sample of 60 boxes:

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

   a What is the variable in this investigation?
   b Is the data continuous or discrete?
   c Construct a frequency table for this data.
   d Display the data using a bar chart.
   e Describe the distribution of the data.
   f What percentage of the boxes contained exactly 50 toothpicks?

4 Revisit the Opening Problem on page 275. Using the After upgrade data:
   a Organise the data in a tally-frequency table.
   b Is the data skewed?
   c Draw a side-by-side vertical bar chart of the data. (Use the graph on page 279.)
   d What evidence is there that the safety upgrade has made a difference?

GROUPED DISCRETE DATA

In situations where there are lots of different numerical values recorded, it may not be practical to use an ordinary tally-frequency table. In these cases, it is often best to group the data into class intervals. We can then display the grouped data in a bar chart.

For example, a local hardware store is concerned about the number of people visiting the store at lunch time.

Over 30 consecutive week days they recorded data.

The results were:

37 30 17 13 46 23 40 28 38 24 23 22 18 29 16
35 24 18 24 44 32 54 31 39 32 38 41 38 24 32
In this case we group the data into class intervals of length 10. The tally-frequency table is shown below. We use the table to construct the vertical bar chart below.

The first column represents the values from 10 to 19, the second from 20 to 29, and so on.

<table>
<thead>
<tr>
<th>Number of people</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 to 19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 to 29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 to 39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 to 49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 to 59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

**EXERCISE 13B.2**

1. Forty students were asked to count the number of red cars they saw on the way to school one morning. The results are shown alongside.
   a. Construct a tally-frequency table for this data using class intervals 0 - 9, 10 - 19, ......, 40 - 49.
   b. Display the data on a vertical bar chart.
   c. How many students saw less than 20 red cars?
   d. What is the modal class for the data?

2. The data gives the number of chairs made each day by a furniture production company over 26 days:
   a. Construct a tally and frequency table for this data.
   b. Draw a vertical bar chart to display the data.
   c. On what percentage of days were less than 40 chairs made?
   d. On how many days were at least 30 chairs made?
   e. Find the modal class for the data.

3. Over a 6 week period, a museum keeps a record of the number of visitors it receives each day. The results are:
   a. Construct a tally and frequency table for this data using class intervals 0 - 99, 100 - 199, 200 - 299, ......, 600 - 699.
   b. Draw a vertical bar chart to display the data.
   c. On how many days did the museum receive at least 500 visitors?
   d. What is the modal class for the data?
   e. Describe the distribution of the data.
We can get a better understanding of a data set if we can locate the middle or centre of the data and get an indication of its spread. Knowing one of these without the other is often of little use.

There are three statistics that are used to measure the centre of a data set. These are: the mean, the median and the mode.

**THE MEAN**

The mean of a data set is the statistical name for the arithmetic average.

\[
\text{mean} = \frac{\text{the sum of all data values}}{\text{the number of data values}}
\]

or \( \bar{x} = \frac{\sum x}{n} \) where \( \sum x \) is the sum of the data.

The mean gives us a single number which indicates a centre of the data set. It is not necessarily a member of the data set.

For example, a mean test mark of 67% tells us that there are several marks below 67% and several above it. 67% is at the centre, but it does not mean that one of the students scored 67%.

**THE MEDIAN**

The median is the middle value of an ordered data set.

An ordered data set is obtained by listing the data, usually from smallest to largest. The median splits the data in halves. Half of the data are less than or equal to the median and half are greater than or equal to it.

For example, if the median mark for a test is 67% then you know that half the class scored less than or equal to 67% and half scored greater than or equal to 67%.

For an odd number of data, the median is one of the data.

For an even number of data, the median is the average of the two middle values and may not be one of the original data.
If there are \( n \) data values, find the value of \( \frac{n + 1}{2} \). The median is the \( \left( \frac{n + 1}{2} \right) \)th data value.

For example:

If \( n = 13 \), \( \frac{13 + 1}{2} = 7 \), so the median = 7th ordered data value.

If \( n = 14 \), \( \frac{14 + 1}{2} = 7.5 \), so the median = average of 7th and 8th ordered data values.

**THE MODE**

The mode is the most frequently occurring value in the data set.

---

**Example 1**

The number of small aeroplanes flying into a remote airstrip over a 15-day period is 570346405369428. For this data set, find:

- a the mean
- b the median
- c the mode.

\[ \text{a mean} = \frac{5 + 7 + 0 + 3 + 4 + 6 + 4 + 0 + 5 + 3 + 6 + 9 + 4 + 2 + 8}{15} = \frac{66}{15} = 4.4 \text{ aeroplanes} \]

\[ \text{b median} = \text{average of } 5_{\text{th}} \text{ and } 6_{\text{th}} \text{ ordered data values} \]

\[ \text{c mode} = 4 \text{ aeroplanes} \]

Suppose that on the next day, 6 aeroplanes land on the airstrip in Example 1. We need to recalculate the measures of the centre to see the effect of this new data value.

We expect the mean to rise as the new data value is greater than the old mean.

In fact, the new mean = \( \frac{66 + 6}{16} = \frac{72}{16} = 4.5 \) aeroplanes.

The new ordered data set is: 0023344455667899 {as \( n = 15, \frac{n + 1}{2} = 8 \)}

\[ \therefore \text{ median} = 4 \text{ aeroplanes} \]

\[ \therefore \text{ mode} = 4 \text{ aeroplanes} \]

This new data set has two modes, 4 and 6 aeroplanes, and we say that the data set is bimodal.

If a data set has three or more modes, we do not use the mode as a measure of the middle.

Note that equal or approximately equal values of the mean, mode and median may indicate a symmetrical distribution of data. However, we should always check using a graph before calling a data set symmetric.
Example 2

Solve the following problems:

a  The mean of six scores is 78.5. What is the sum of the scores?

b  Find $x$ if 10, 7, 3, 6 and $x$ have a mean of 8.

a  \[
\frac{\text{sum}}{6} = 78.5 \\
\therefore \text{sum} = 78.5 \times 6 = 471 \\
\therefore \text{the sum of the scores is 471.}
\]

b  There are 5 scores.

\[
\frac{10 + 7 + 3 + 6 + x}{5} = 8 \\
\therefore \frac{26 + x}{5} = 8 \\
\therefore 26 + x = 40 \\
\therefore x = 14
\]

EXERCISE 13C

1 Find the i mean ii median iii mode for each of the following data sets:

   a  12, 17, 20, 24, 25, 30, 40
   b  8, 8, 8, 10, 11, 11, 12, 16, 20, 20, 24
   c  7.9, 8.5, 9.1, 9.2, 9.9, 10.0, 11.1, 11.2, 12.2, 12.9

2 Consider the following two data sets:
   Data set A:  5, 6, 6, 7, 7, 8, 8, 9, 10, 12
   Data set B:  5, 6, 6, 7, 7, 8, 8, 9, 10, 20

   a  Find the mean for each data set.
   b  Find the median for each data set.
   c  Explain why the mean of Data set A is less than the mean of Data set B.
   d  Explain why the median of Data set A is the same as the median of Data set B.

3 The selling price of nine houses are:
   $158 000, \quad $290 000, \quad $290 000, \quad $1.1 million, \quad $900 000, \quad $395 000,
   $925 000, \quad $420 000, \quad $760 000

   a  Find the mean, median and modal selling prices.
   b  Explain why the mode is an unsatisfactory measure of the middle in this case.
   c  Is the median a satisfactory measure of the middle of this data set?

4 The following raw data is the daily rainfall (to the nearest millimetre) for the month of February 2007 in a city in China:  0, 4, 1, 0, 0, 0, 2, 9, 3, 0, 0, 0, 8, 27, 5, 0, 0, 0, 0, 8, 1, 3, 0, 0, 15, 1, 0, 0

   a  Find the mean, median and mode for the data.
   b  Give a reason why the median is not the most suitable measure of centre for this set of data.
   c  Give a reason why the mode is not the most suitable measure of centre for this set of data.
A basketball team scored 38, 52, 43, 54, 41 and 36 points in their first six matches.

a Find the mean number of points scored for the first six matches.

b What score does the team need to shoot in their next match to maintain the same mean score?

c The team scores only 20 points in the seventh match. What is the mean number of points scored for the seven matches?

d If the team scores 42 points in their eighth and final match, will their previous mean score increase or decrease? Find the mean score for all eight matches.

The mean of 12 scores is 8.8. What is the sum of the scores?

While on a camping holiday, Lachlan drove on average, 214 km per day for a period of 8 days. How far did Lachlan drive in total while on holiday?

The mean monthly sales for a CD store are $216 000. Calculate the total sales for the store for the year.

Find x if 7, 15, 6, 10, 4 and x have a mean of 9.

Find a, given that 10, a, 15, 20, a, a, 17, 7 and 15 have a mean of 12.

Over a semester, Jamie did 8 science tests. Each was marked out of 30 and Jamie averaged 25. However, when checking his files, he could only find 7 of the 8 tests. For these he scored 29, 26, 18, 20, 27, 24 and 29. Determine how many marks out of 30 he scored for the eighth test.

A sample of 12 measurements has a mean of 8.5 and a sample of 20 measurements has a mean of 7.5. Find the mean of all 32 measurements.

In the United Kingdom, the months of autumn are September, October and November. If the mean temperature was $S^\circ C$ for September, $O^\circ C$ for October and $N^\circ C$ for November, find an expression for the mean temperature $A^\circ C$ for the whole of autumn.

The mean, median and mode of seven numbers are 8, 7 and 6 respectively. Two of the numbers are 8 and 10. If the smallest of the seven numbers is 4, find the largest of the seven numbers.

Knowing the middle of a data set can be quite useful, but for a more accurate picture of the data set we also need to know its spread.

For example, 2, 3, 4, 5, 6, 7, 8, 9, 10 has a mean value of 6 and so does 4, 5, 5, 6, 6, 7, 8. However, the first data set is more widely spread than the second one.

Three commonly used statistics that indicate the spread of a set of data are the

- **range**
- **interquartile range**
- **standard deviation**.

The standard deviation, which is the spread about the mean, will not be covered in this course.
**THE RANGE**

The range is the difference between the maximum (largest) data value and the minimum (smallest) data value.

\[
\text{range} = \text{maximum data value} - \text{minimum data value}
\]

**Example 3**

Find the range of the data set: 5, 3, 8, 4, 9, 7, 5, 6, 2, 3, 6, 8, 4.

\[
\text{range} = \text{maximum value} - \text{minimum value} = 9 - 2 = 7
\]

**THE QUARTILES AND THE INTERQUARTILE RANGE**

The median divides an ordered data set into halves, and these halves are divided in half again by the quartiles.

The middle value of the lower half is called the lower quartile. One quarter, or 25%, of the data have values less than or equal to the lower quartile. 75% of the data have values greater than or equal to the lower quartile.

The middle value of the upper half is called the upper quartile. One quarter, or 25%, of the data have values greater than or equal to the upper quartile. 75% of the data have values less than or equal to the upper quartile.

The interquartile range is the range of the middle half (50%) of the data.

\[
\text{interquartile range} = \text{upper quartile} - \text{lower quartile}
\]

The data is thus divided into quarters by the lower quartile \(Q_1\), the median \(Q_2\), and the upper quartile \(Q_3\).

So, the interquartile range, \[\text{IQR} = Q_3 - Q_1.\]

**Example 4**

For the data set: 7, 3, 4, 2, 5, 6, 7, 5, 9, 3, 8, 3, 5, 6 find the:

- **a** median
- **b** lower and upper quartiles
- **c** interquartile range.

The ordered data set is: 2 3 3 3 4 5 5 5 6 6 7 7 8 9 (15 of them)

- **a** As \(n = 15\), \(n + 1 \over 2 = 8\) \(\therefore\) the median = 8th score = 5
- **b** As the median is a data value, we now ignore it and split the remaining data into two:
  
  **lower** 2 3 3 3 4 5 5 5 6 6 7 7 8 9
  **upper**

  \(Q_1 = \text{median of lower half} = 3\)

  \(Q_2 = \text{median of upper half} = 7\)

- **c** \(\text{IQR} = Q_3 - Q_1 = 7 - 3 = 4\)
For the data set: 6, 10, 7, 8, 13, 7, 10, 8, 1, 7, 5, 4, 9, 4, 2, 5, 9, 6, 3, 2 find the:

a) median
b) lower and upper quartiles
c) interquartile range.

The ordered data set is: 1 2 2 3 4 4 5 5 6 6 7 7 8 8 9 9 10 10 10 13 (20 of them)

a) As \( n = 20 \), \( \frac{n + 1}{2} = 10.5 \)

\[ \text{median} = \frac{10\text{th value} + 11\text{th value}}{2} = \frac{6 + 7}{2} = 6.5 \]

b) As the median is not a data value we split the data into two:

lower
\[ Q_1 = 4 \]

upper
\[ Q_3 = 8.5 \]

\[ \text{IQR} = Q_3 - Q_1 = 8.5 - 4 = 4.5 \]

Note: Some computer packages (for example, MS Excel) calculate quartiles in a different way from this example.
3 For the data set given, find:
   a the minimum value
   b the maximum value
   c the median
   d the lower quartile
   e the upper quartile
   f the range

Stem | Leaf
-----|-----
2    | 0122
3    | 0014458
4    | 0234669
5    | 11458

5|1 represents 51

**E DATA IN FREQUENCY TABLES**

When the same data appears several times we often summarise it in a **frequency table**. For convenience we denote the data values by \( x \) and the frequencies of these values by \( f \).

### The mode

There are 14 of data value 6 which is more than any other data value.

The mode is therefore 6.

### The mean

A ‘Product’ column helps to add all scores.

We know the mean = \( \frac{\text{the sum of all data values}}{\text{the number of data values}} \), so for data in a frequency table,

\[
\bar{x} = \frac{\sum fx}{\sum f}
\]

In this case the mean = \( \frac{258}{40} = 6.45 \).

### The median

There are 40 data values, an even number, so there are two middle data values.

As the sample size \( n = 40 \), \( \frac{n + 1}{2} = \frac{41}{2} = 20.5 \)

\[ \therefore \] the median is the average of the 20th and 21st data values.

In the table, the blue numbers show us accumulated values.

<table>
<thead>
<tr>
<th>Data Value ((x))</th>
<th>Frequency ((f))</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>40</strong></td>
</tr>
</tbody>
</table>

Notice that we have a skewed distribution for which the mean, median and mode are nearly equal. This is why we need to be careful when we use measures of the middle to call distributions symmetric.
Example 6

Each student in a class of 20 is assigned a number between 1 and 10 to indicate his or her fitness.

Calculate the: a mean b median c mode d range of the scores.

<table>
<thead>
<tr>
<th>Score</th>
<th>Number of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
</tr>
</tbody>
</table>

The mean score \( \frac{\sum fx}{\sum f} \) = \( \frac{157}{20} \) = 7.85

b There are 20 scores, and so the median is the average of the 10th and 11th.

<table>
<thead>
<tr>
<th>Score</th>
<th>Number of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
</tr>
</tbody>
</table>

The 10th and 11th students both scored 8 \( \therefore \) median = 8.

c Looking down the ‘number of students’ column, the highest frequency is 7. This corresponds to a score of 8, so the mode = 8.

d The range = highest data value – lowest data value = 10 – 5 = 5

EXERCISE 13E

1 The members of a school band were each asked how many musical instruments they played. The results were:

<table>
<thead>
<tr>
<th>Number of instruments</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
</tr>
</tbody>
</table>

Calculate the: a mode b median c mean d range.
2 The following frequency table records the number of books read in the last year by 50 fifteen-year-olds.
   a For this data, find the:
      i mean    ii median    iii mode    iv range.
   b Construct a vertical bar chart for the data and show the position of the measures of centre (mean, median and mode) on the horizontal axis.
   c Describe the distribution of the data.
   d Why is the mean smaller than the median for this data?
   e Which measure of centre would be most suitable for this data set?

<table>
<thead>
<tr>
<th>No. of books</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

3 Hui breeds ducks. The number of ducklings surviving for each pair after one month is recorded in the table.
   a Calculate the:
      i mean    ii mode    iii median.
   b Calculate the range of the data.
   c Is the data skewed?
   d How does the skewness of the data affect the measures of the middle of the distribution?

<table>
<thead>
<tr>
<th>Number of survivors</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>76</td>
</tr>
</tbody>
</table>

4 Participants in a survey were asked how many foreign countries they had visited. The results are displayed in the vertical bar chart alongside:
   a Construct a frequency table from the graph.
   b How many people took part in the survey?
   c Calculate the:
      i mean    ii median
      iii mode    iv range of the data.

F GROUPED DISCRETE DATA [11.4]

One issue to consider when grouping data into class intervals is that the original data is lost. This means that calculating the exact mean and median becomes impossible. However, we can still estimate these values, and in general this is sufficient.

THE MEAN

When information has been grouped in classes we use the midpoint of the class to represent all scores within that interval.

We are assuming that the scores within each class are evenly distributed throughout that interval. The mean calculated will therefore be an estimate of the true value.
The table summarises the marks received by students for a Physics examination out of 50.

a Estimate the mean mark.
b What is the modal class?
c Can the range of the data be found?

<table>
<thead>
<tr>
<th>Class interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 9</td>
<td>2</td>
</tr>
<tr>
<td>10 - 19</td>
<td>31</td>
</tr>
<tr>
<td>20 - 29</td>
<td>73</td>
</tr>
<tr>
<td>30 - 39</td>
<td>85</td>
</tr>
<tr>
<td>40 - 49</td>
<td>28</td>
</tr>
</tbody>
</table>

\[
\text{Mean} = \frac{\sum fx}{\sum f} = \frac{6336.5}{217} \approx 29.2
\]

b The modal class is 30 - 39 marks.

c No, as we do not know the smallest and largest score.

THE MEDIAN

We can estimate the median of a grouped data set by using a cumulative frequency graph or ogive. This is done in Chapter 17.

The approximate median can also be calculated using a formula:

\[
\text{Median} \approx L + \frac{N}{F} \times I
\]

where \( L \) = the lower boundary for the class interval containing the median
\( N \) = the number of scores in the median class needed to arrive at the middle score
\( F \) = the frequency of the class interval containing the median
\( I \) = the class interval length.

For example, in Example 7 we notice that \( n = 217 \), so the median is the \( \frac{217 + 1}{2} = 109 \)th score.

There are \( 2 + 31 + 73 = 106 \) scores in the first 3 classes, so the median class is 30 - 39.

For discrete data, the lower and upper boundaries for the class interval 30 - 39 are 29.5 and 39.5. \( \therefore L = 29.5 \).

\( N = 109 - 106 = 3 \), \( F = 85 \) and \( I = 10 \)

\( \therefore \) the median \( \approx 29.5 + \frac{3}{85} \times 10 \approx 29.9 \)
EXERCISE 13F

1. 40 students receive marks out of 100 for an examination in Chemistry. The results were:
   70 65 50 40 58 72 39 85 90 65 53 75 83 92 66 78 82 88 56 68 43 90 80 85 78 72 59 83 75 54 68 75 89 92 81 77 59 63 80

   a. Find the exact mean of the data.
   b. Find the median of the data.
   c. Group the data into the classes 0 - 9, 10 - 19, 20 - 29, ..., 90 - 99, forming a frequency table. Include columns for the midpoint (x), and fx.
   d. Estimate from the grouped data of c the:
      i. mean
      ii. median.
   e. How close are your answers in d to exact values in a and b?

2. A sample of 100 students was asked how many times they bought lunch from the canteen in the last four weeks. The results were:

   a. Estimate the mean number of bought lunches for each student.
   b. What is the modal class?
   c. Estimate the median number of bought lunches.

3. The percentage marks for boys and girls in a science test are given in the table:

   a. Estimate the mean mark for the girls.
   b. Estimate the mean mark for the boys.
   c. What can you deduce by comparing a and b?

<table>
<thead>
<tr>
<th>Marks</th>
<th>Girls</th>
<th>Boys</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 - 30</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>31 - 40</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>41 - 50</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>51 - 60</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>61 - 70</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>71 - 80</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>81 - 90</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>91 - 100</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

G STATISTICS FROM TECHNOLOGY [11.8]

GRAPHICS CALCULATOR

A graphics calculator can be used to find descriptive statistics and to draw some types of graphs.

Consider the data set: 5 2 3 6 4 5 3 7 5 7 1 8 9 5

No matter what brand of calculator you use you should be able to:

- Enter the data as a list.
- Enter the statistics calculation part of the menu and obtain the descriptive statistics like these shown.

Instructions for these tasks can be found at the front of the book in the Graphics Calculator Instructions section.
**COMPUTER PACKAGE**

Various statistical packages are available for computer use, but many are expensive and often not easy to use. Click on the icon to use the statistics package on the CD.

Enter the data sets:

- **Set 1:** 5 2 3 3 6 4 5 3 7 5 7 1 8 9 5
- **Set 2:** 9 6 2 3 5 5 7 5 6 7 6 3 4 4 5 8 4

Examine the side-by-side bar charts.

Click on the Statistics tab to obtain the descriptive statistics.

Select Print... from the File menu to print all of these on one sheet of paper.

**EXERCISE 13G**

1. Helen and Jessica both play for the Lightning Basketball Club. The numbers of points scored by each of them over a 12-game period were:

   - **Helen:** 20, 14, 0, 28, 38, 26, 17, 24, 26, 24, 6, 3
   - **Jessica:** 18, 20, 22, 2, 18, 31, 7, 15, 17, 16, 22, 29

   **a** Calculate the mean and median number of points for both of them.
   **b** Calculate the range and interquartile range for both of them.
   **c** Which of the girls was the higher scorer during the 12-game period?
   **d** Who was more consistent?

2. Enter the Opening Problem data on page 275 for the Before upgrade data in Set 1 and the After upgrade data in Set 2 of the computer package. Print out the page of graphs and descriptive statistics. Write a brief report on the effect of the safety upgrade.

3. Use your graphics calculator to check the answers to Example 6 on page 289.

4. Use your graphics calculator to check the answers to Example 7 on page 291.

5. The heights (to the nearest centimetre) of boys and girls in a Year 10 class in Norway are as follows:

   - **Boys:** 165 171 169 169 172 171 171 180 168 168 166 168 170 165 171 173 187
     181 175 174 165 167 163 160 169 167 172 174 177 188 177 185 167 160
   - **Girls:** 162 171 156 166 168 163 170 171 177 169 168 165 156 159 165 164 154
     171 172 166 152 169 170 163 162 165 163 168 155 175 176 170 166

   **a** Use your calculator to find measures of centre (mean and median) and spread (range and IQR) for each data set.
   **b** Write a brief comparative report on the height differences between boys and girls in the class.

---

**Review set 13A**

1. Classify the following numerical variables as either discrete or continuous:
   **a** the number of oranges on each orange tree
   **b** the heights of seedlings after two weeks
   **c** the scores of team members in a darts competition.
A randomly selected sample of small businesses has been asked, “How many full-time employees are there in your business?”. A bar chart has been constructed for the results.

a How many small businesses gave data in the survey?

b How many of the businesses had only one or two full-time employees?

c What percentage of the businesses had five or more full-time employees?

d Describe the distribution of the data.

e Find the mean of the data.

The data alongside are the number of call-outs each day for a city fire department over a period of 25 days:

a Construct a tally and frequency table for the data using class intervals 0 - 9, 10 - 19, 20 - 29, and 30 - 39.

b Display the data on a vertical bar chart.

c On how many days were there at least 20 call-outs?

d On what percentage of days were there less than 10 call-outs?

e Find the modal class for the data.

For the following data set of the number of points scored by a rugby team, find:

a the mean

b the mode

c the median

d the range

e the upper and lower quartiles

f the interquartile range.

The test score out of 40 marks was recorded for a group of 30 students:

a Construct a tally and frequency table for this data.

b Draw a bar chart to display the data.

c How many students scored less than 20 for the test?

d If an ‘A ’ was awarded to students who scored 30 or more for the test, what percentage of students scored an ‘A’?

Eight scores have an average of six. Scores of 15 and x increase the average to 7. Find x.
9 The numbers of potatoes growing on each of 100 potato plants were recorded and summarised in the table below:

<table>
<thead>
<tr>
<th>No. of potatoes</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 2</td>
<td>7</td>
</tr>
<tr>
<td>3 - 5</td>
<td>11</td>
</tr>
<tr>
<td>6 - 8</td>
<td>25</td>
</tr>
<tr>
<td>9 - 11</td>
<td>40</td>
</tr>
<tr>
<td>12 - 14</td>
<td>15</td>
</tr>
<tr>
<td>15 - 17</td>
<td>2</td>
</tr>
</tbody>
</table>

Estimate the mean number of potatoes per plant.

10 Use technology to find the
   a mean
   b median
   c lower quartile
   d upper quartile

   of the following data set:
   147 152 164 159 140 155 166 180 177
   156 182 174 168 170 165 143 185 126
   148 138 141 156 142 130 182 156 182

Review set 13B

1 A class of 20 students was asked “How many children are there in your household?” and the following data was collected:

   1 2 3 3 2 4 5 4 2 3 8 1 2 1 3 2 1 2 1 2

   a What is the variable in the investigation?
   b Is the data discrete or continuous? Explain your answer.
   c Construct a frequency table for the data.
   d Construct a vertical bar chart to display the data.
   e How would you describe the distribution of the data? Is it symmetrical, or positively or negatively skewed?
   f What is the mode of the data?

2 A class of thirty students were asked how many emails they had sent in the last week. The results were:

   12 6 21 15 18 4 28 32 17 44 9 32 26 18 11
   24 31 17 52 7 42 37 19 6 20 15 27 8 36 28

   a Construct a tally and frequency table for this data, using intervals 0 - 9, 10 - 19,........50 - 59.
   b Draw a vertical bar chart to display the data.
   c Find the modal class.
   d What percentage of students sent at least 30 emails?
   e Describe the distribution of the data.

3 The local transport authority recorded the number of vehicles travelling along a stretch of road each day for 40 days. The data is displayed in the bar chart alongside:

   a On how many days were there at least 180 vehicles?
   b On what percentage of days were there less than 160 vehicles?
   c What is the modal class?
   d Estimate the mean number of vehicles on the stretch of road each day.
4 A sample of 15 measurements has a mean of 14.2 and a sample of 10 measurements has a mean of 12.6. Find the mean of the total sample of 25 measurements.

5 Determine the mean of the numbers 7, 5, 7, 2, 8 and 7. If two additional numbers, 2 and $x$, reduce the mean by 1, find $x$.

6 Jenny’s golf scores for her last 20 rounds were: 90 106 84 103 112 100 105 81 104 98 107 95 104 108 99 101 106 102 98 101
   a Find the median, lower quartile and upper quartile of the data set.
   b Find the interquartile range of the data set and explain what it represents.

7 For the data displayed in the stem-and-leaf plot find the:
   a mean
   b median
   c lower quartile
   d upper quartile
   e range
   f interquartile range

8 The given table shows the distribution of scores for a Year 10 spelling test in Austria.
   a Calculate the: i mean ii mode iii median iv range of the scores
   b The average score for all Year 10 students across Austria in this spelling test was 6.2. How does this class compare to the national average?
   c The data set is skewed. Is the skewness positive or negative?

9 Sixty people were asked: “How many times have you been to the cinema in the last twelve months?”. The results are given in the table alongside. Estimate the mean and median of the data.

10 The data below shows the number of pages contained in a random selection of books from a library:
   295 612 452 182 335 410 256 715 221 375
   508 310 197 245 411 246 606 192 487
   a Use technology to find the:
      i mean ii median
      iii lower quartile iv upper quartile
   b Find the range and interquartile range for the data.