4A Scatterplots

The manager of a small ski resort has a problem. He wishes to be able to predict the number of skiers using his resort each weekend in advance so that he can organise additional resort staffing and catering if needed. He knows that good deep snow will attract skiers in big numbers but scant covering is unlikely to attract a crowd. To investigate the situation further he collects the following data over twelve consecutive weekends at his resort.

<table>
<thead>
<tr>
<th>Depth of snow (m)</th>
<th>Number of skiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>120</td>
</tr>
<tr>
<td>0.8</td>
<td>250</td>
</tr>
<tr>
<td>2.1</td>
<td>500</td>
</tr>
<tr>
<td>3.6</td>
<td>780</td>
</tr>
<tr>
<td>1.4</td>
<td>300</td>
</tr>
<tr>
<td>1.5</td>
<td>280</td>
</tr>
<tr>
<td>1.8</td>
<td>410</td>
</tr>
<tr>
<td>2.7</td>
<td>320</td>
</tr>
<tr>
<td>3.2</td>
<td>640</td>
</tr>
<tr>
<td>2.4</td>
<td>540</td>
</tr>
<tr>
<td>2.6</td>
<td>530</td>
</tr>
<tr>
<td>1.7</td>
<td>200</td>
</tr>
</tbody>
</table>

Areas of study

- Scatterplots
- Informal interpretation of patterns and features of scatterplots
- Correlation and regression:
  - use and interpretation of the quadrant, \( q \), correlation coefficient
  - fitting a line to data with an appropriate linear association for a dependent variable with respect to a given independent variable, by eye and using the two-mean method, determining the equation of this line, and using this equation for prediction. Informal consideration of closeness of fit (how close the data points are to the fitted line).
The data in this example are known as \textit{bivariate} data. For each item (weekend), two variables are considered (\textit{depth of snow} and \textit{number of skiers}). When analysing bivariate data we are interested in examining the relationship between the two variables. In the case of the ski resort data we might be interested in finding out:

- Are visitor numbers related to depth of snow?
- If there is a relationship, then is it always true or is it just a guide? In other words, how strong is the relationship?
- Is it possible to predict the likely number of skiers if the depth of snow is known?
- How much confidence could be placed in the prediction?

To help answer these questions the data can be graphed on a \textit{scatterplot} (or \textit{scattergraph}) as shown at right.

Each of the data points is represented by a single visible point on the graph.

When drawing a scatterplot, we need to choose the correct variable to assign to each of the axes. The convention is to place the \textit{independent} variable on the \textit{x}-axis and the \textit{dependent} variable on the \textit{y}-axis.

The independent variable in an experiment or investigation is the variable that is deliberately controlled or adjusted by the investigator. The dependent variable is the variable that responds to changes in the independent variable.

Neither of the variables involved in the ski resort data were controlled directly by the investigator but ‘\textit{Number of skiers}’ would be considered the dependent variable because it could be considered to respond to \textit{depth of snow} (rather than the snow depth somehow obliging changes in the numbers of skiers). As ‘\textit{Number of skiers}’ is the dependent variable we graph it on the \textit{y}-axis and the ‘\textit{Depth of snow}’ on the \textit{x}-axis.

Notice how the scatterplot for the ski resort data shows a general upward trend. It is not a perfectly straight line but it is still clear that a general trend or relationship has formed: as the depth of snow increases, so too the number of skiers also increases.

\textbf{When one variable increases with another it is said that there is a positive correlation between the variables.}

Relationships can show either positive or negative correlation. Consider the following example in which 10 Year 11 students were surveyed to find out the amount of time that they spent doing exercise each week. This was compared with their blood cholesterol level.

<table>
<thead>
<tr>
<th>Period of exercise (h)</th>
<th>6</th>
<th>8</th>
<th>12</th>
<th>16</th>
<th>2</th>
<th>0</th>
<th>5</th>
<th>8</th>
<th>7</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood cholesterol level (mM)</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

In this example there seems to be a general downward trend. As the amount of exercise increases, the blood cholesterol decreases.

\textbf{When one variable decreases while the other is increasing it is said that a negative correlation exists between the variables.}
Notice that in this case the points are not as closely aligned as in the previous example. We could say that the relationship (or correlation) between the variables is only a weak relationship. In general terms, the closer the points are to forming a straight line, the stronger the relationship is between the variables.

Sometimes we find that there is no relationship between the variables. Consider the example shown at right in which a researcher was looking for a link between people’s heights and their IQs (intelligence quotients). The points appear to be randomly dispersed across the scatterplot. In cases like this it can be concluded that there is no clear relationship between the variables.

**Drawing conclusions/causation**

When data are graphed, we can often estimate by eye (rather than measure) the type of correlation involved. Our ability to make these qualitative judgements can be seen from the following examples, which summarise the different types of correlation that might appear in a scatterplot.

- **Strong positive correlation**
  Conclusion: The greater the level of \( x \) the greater the level of \( y \).

- **Moderate positive correlation**
  Conclusion: There is evidence to show that the greater the level of \( x \) the greater the level of \( y \).

- **Weak positive correlation**
  Conclusion: There is a little evidence to show that the greater the level of \( x \) the greater the level of \( y \).

- **No correlation**
  Conclusion: There appears to be no relationship between the variables \( x \) and \( y \).

- **Weak negative correlation**
  Conclusion: There is a little evidence to show that the greater the level of \( x \) the smaller the level of \( y \).

- **Moderate negative correlation**
  Conclusion: There is evidence to show that the greater the level of \( x \) the smaller the level of \( y \).

- **Strong negative correlation**
  Conclusion: The greater the level of \( x \), the smaller the level of \( y \).
Notice how the conclusion drawn for each of the scatterplots is slightly different. If the correlation is strong then the resulting conclusion can be more definite than if it were weak.

When drawing conclusions from a scatterplot or in summarising its trend it is important to avoid using statements like ‘x causes y’. Just because there is a strong relationship between two variables it does not mean that one variable causes the other. In fact, a strong correlation might have resulted because variable y is causing changes in x, or it could be that there is some third factor that is causing changes in both variables x and y.

To illustrate this point, a Dutch researcher compared the human birth rate (births per 1000 population) in different areas with the stork population in those areas. He found that there was a strong positive correlation between the stork population in the different areas and the human birth rate in those areas. What could he conclude? That storks cause babies? Absolutely not! His conclusion could only be along the lines of ‘the greater the stork population, the greater the human birth rate’. In this case there was a third factor that was causing the apparent relationship. Storks have a preference for nesting in rural areas and for social-demographic reasons rural dwellers tend to have larger families than their cosmopolitan counterparts. In other words, the land usage of the areas was causing changes in both the stork population and the human birth rate.

A statement as bold as the warning on this cigarette packet — ‘Smoking causes lung cancer’ — cannot be made on the basis of statistics alone. In this case scientific evidence would have been considered as well. The statement is drawn from medical understanding of how nicotine affects the lung system rather than upon statistical analysis.

To avoid trouble when framing conclusions from a scatterplot, it might be best to stick closely to the wording used in the scatterplots on the previous page. In the case of the ski resort data we might conclude: ‘There is a moderate positive relationship. There is evidence to show that the greater the depth of snow, the greater the number of skiers at the resort’. In the case of the exercise–blood-cholesterol data we might conclude: ‘There is a weak negative correlation. There is a little evidence to show that the greater the amount of exercise, the lower the blood cholesterol’. You might be asked to provide further explanations or reasons for a trend in a graph. Such reasons would be based entirely upon conjecture.

---

**WORKED EXAMPLE 1**

The operators of a casino keep records of the number of people playing a ‘Jackpot’ type game. The table below shows the number of players for different prize amounts.

<table>
<thead>
<tr>
<th>Number of players</th>
<th>260</th>
<th>280</th>
<th>285</th>
<th>340</th>
<th>390</th>
<th>428</th>
<th>490</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prize ($)</td>
<td>1000</td>
<td>1500</td>
<td>2000</td>
<td>2500</td>
<td>3000</td>
<td>3500</td>
<td>4000</td>
</tr>
</tbody>
</table>

a Draw a scatterplot of the data.
b State the type of correlation that the scatterplot shows and draw a suitable conclusion from the graph.
c Suggest why the plot is not perfectly linear.
‘Number of players’ is the dependent variable, so plot it on the y-axis against ‘prize money’ on the x-axis. On the Statistics screen, label list1 as ‘prize’ and enter the prize values. Label list2 as ‘players’ and enter the number of players.

To draw the scatterplot of the points, tap:
- SetGraph
- Setting
  Set:
  Type: Scatter
  XList: main\prize
  YList: main\players
  Freq: 1
  Mark: Square
- Set
- y

Copy the scatterplot into your workbook.

Study the graph and make a judgement on the type of correlation.

With this type of more open question you can bring in new ideas and information from your own general knowledge.

There is a strong positive correlation. Therefore, the greater the prize money, the greater the number of players.

Although the correlation is strong, the graph is not perfectly linear because factors other than the prize money may affect the number of people playing. For example, the time of evening might have an effect: there might be fewer players late in the evening even if the prize is high.
1. Bivariate data result from measurements being made on each of the two variables for a given set of items.
2. When representing bivariate data on a scatterplot, put the independent variable on the \(x\)-axis and the dependent variable on the \(y\)-axis.
3. The independent variable is the one which is deliberately controlled or adjusted by the investigator.
4. The dependent variable is the one which responds to changes in the independent variable.
5. The pattern of the scatterplot gives an indication of the level of association (correlation) between the variables.
6. The correlation can be negative or positive.
7. When one variable increases with another, there is a positive correlation between them.
8. When one variable decreases while the other increases, there is negative correlation.
9. Sometimes the scatterplot indicates that there is no correlation between variables.
10. The level of correlation can be strong, moderate or weak.
11. Strong correlation between two variables does not necessarily mean that one variable ‘causes’ the other.

**EXERCISE 4A Scatterplots**

1. For each of the following examples, state which of the two variables should be considered the independent variable and which should be considered the dependent variable. Also state which variable should be plotted on the \(x\)-axis of a scattergraph and which variable should be plotted on the \(y\)-axis.
   - a An experimenter measures the acidity level (pH) of some soil samples taken from different cornfields and also records the height of the corn grown in each field.
   - b A market researcher surveys the population, asking individuals the amount of product that they would buy at different prices.
   - c A medical scientist administers a drug to a patient and then at different times measures the amount of drug that is present in the bloodstream.
   - d An economist demonstrates the value of replacing equipment by plotting a graph which shows how the running costs of the machine change as it gets older.
   - e A scientist investigates the relationship between the circumference of an athlete’s biceps and the athlete’s ability to perform a lift test.

2. Match each of the following scattergraphs with the correlation that it shows.
3 A pie seller at a football match notices that there seems to be a relationship between the number of pies that he sells and the temperature of the day. He records the following data.

<table>
<thead>
<tr>
<th>Daily temperature (°C)</th>
<th>12</th>
<th>22</th>
<th>26</th>
<th>11</th>
<th>8</th>
<th>18</th>
<th>14</th>
<th>16</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of pies sold</td>
<td>620</td>
<td>315</td>
<td>295</td>
<td>632</td>
<td>660</td>
<td>487</td>
<td>512</td>
<td>530</td>
<td>546</td>
<td>492</td>
</tr>
</tbody>
</table>

a Draw a scatterplot of the data.
b State the type of correlation that the scatterplot shows and draw a suitable conclusion from the graph.
c Suggest why the plot is not perfectly linear.

4 A researcher is investigating the effect of living in airconditioned buildings upon general health. She records the following data.

<table>
<thead>
<tr>
<th>Hours spent each week in airconditioned buildings</th>
<th>2</th>
<th>13</th>
<th>6</th>
<th>48</th>
<th>40</th>
<th>0</th>
<th>10</th>
<th>0</th>
<th>2</th>
<th>5</th>
<th>18</th>
<th>10</th>
<th>12</th>
<th>26</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of days sick due to flu and colds</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>15</td>
<td>13</td>
<td>8</td>
<td>14</td>
<td>1</td>
<td>16</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>14</td>
<td>15</td>
</tr>
</tbody>
</table>

a Plot the data on a scattergraph.
b State the type of correlation the graph shows and draw a suitable conclusion from it.
c The researcher finishes her experimental report by concluding that air-conditioning is the cause of poor health. Is she correct to say: ‘... is the cause’ of poor health? What other factors could have influenced the relationship shown by the scatterplot?

5 The following table gives the forecasted maximum temperature (°C) for each day in February and the actual maximum that was recorded.

<table>
<thead>
<tr>
<th>Date</th>
<th>Forecast maximum</th>
<th>Actual maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Feb</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>2 Feb</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>3 Feb</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>4 Feb</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>5 Feb</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>6 Feb</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>7 Feb</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>8 Feb</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>9 Feb</td>
<td>33</td>
<td>35</td>
</tr>
<tr>
<td>10 Feb</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>11 Feb</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>12 Feb</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>13 Feb</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>14 Feb</td>
<td>22</td>
<td>21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th>Forecast maximum</th>
<th>Actual maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 Feb</td>
<td>25</td>
<td>22</td>
</tr>
<tr>
<td>16 Feb</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>17 Feb</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>18 Feb</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>19 Feb</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>20 Feb</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>21 Feb</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>22 Feb</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td>23 Feb</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>24 Feb</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>25 Feb</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>26 Feb</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>27 Feb</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>28 Feb</td>
<td>25</td>
<td>21</td>
</tr>
</tbody>
</table>
a Plot the data on a scattergraph. (You may like to attempt this using a CAS calculator.)
b State the type of correlation the graph shows and draw a suitable conclusion from it.
c What can you say about the ability of weather forecasters to ‘get it right’?

6 The data below show the population and area of the Australian States and territories.

<table>
<thead>
<tr>
<th>State</th>
<th>Area ($\times 1000$ km$^2$)</th>
<th>Population ($\times 1000$)</th>
<th>State</th>
<th>Area ($\times 1000$ km$^2$)</th>
<th>Population ($\times 1000$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vic.</td>
<td>228</td>
<td>4645</td>
<td>NT</td>
<td>1346</td>
<td>211</td>
</tr>
<tr>
<td>NSW</td>
<td>802</td>
<td>6372</td>
<td>WA</td>
<td>2526</td>
<td>1851</td>
</tr>
<tr>
<td>ACT</td>
<td>2</td>
<td>312</td>
<td>SA</td>
<td>984</td>
<td>1467</td>
</tr>
<tr>
<td>Qld</td>
<td>1727</td>
<td>3655</td>
<td>Tas.</td>
<td>68</td>
<td>457</td>
</tr>
</tbody>
</table>

a Plot the data on a scattergraph.
b State the type of correlation the graph shows and draw a suitable conclusion from it.

7 In an experiment, 12 people were administered different doses of a drug. When the drug had taken effect, the time taken for each person to react to a set stimulus was measured. The results are detailed below.

<table>
<thead>
<tr>
<th>Amount of drug (mg)</th>
<th>Reaction time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.030</td>
</tr>
<tr>
<td>0.2</td>
<td>0.025</td>
</tr>
<tr>
<td>0.3</td>
<td>0.028</td>
</tr>
<tr>
<td>0.4</td>
<td>0.036</td>
</tr>
<tr>
<td>0.5</td>
<td>0.040</td>
</tr>
<tr>
<td>0.6</td>
<td>0.052</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amount of drug (mg)</th>
<th>Reaction time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.046</td>
</tr>
<tr>
<td>0.8</td>
<td>0.068</td>
</tr>
<tr>
<td>0.9</td>
<td>0.085</td>
</tr>
<tr>
<td>1.0</td>
<td>0.092</td>
</tr>
<tr>
<td>1.1</td>
<td>0.084</td>
</tr>
<tr>
<td>1.2</td>
<td>0.096</td>
</tr>
</tbody>
</table>

a Plot the data on a scattergraph.
b State the type of correlation the graph shows and draw a suitable conclusion from it.
c What factors might be preventing the graph from being perfectly linear?

8 A researcher administers different amounts of growth hormone to newborn chickens then measures their size after six weeks. He presents his findings using a scattergraph. The researcher should plot:
A size on the $x$-axis because it is the independent variable and amount of hormone on the $y$-axis because it is the dependent variable
B size on the $y$-axis because it is the independent variable and amount of hormone on the $x$-axis because it is the dependent variable
C size on the $x$-axis because it is the dependent variable and amount of hormone on the $y$-axis because it is the independent variable
D size on the $y$-axis because it is the dependent variable and amount of hormone on the $x$-axis because it is the independent variable
E the variables on either axis because it makes no difference.

9 What type of correlation is shown by the graph at right?
A Strong positive correlation
B Moderate positive correlation
C No correlation
D Moderate negative correlation
E Strong negative correlation

10 What type of correlation is shown by the graph at right?
A Strong positive correlation
B Moderate positive correlation
C No correlation
D Moderate negative correlation
E Strong negative correlation
A researcher finds that there is a moderate negative correlation between the amount of pollution found in water samples collected from different sites and the fish population of the sites. The researcher can conclude that:

A. Pollution is causing strain on the fish population.
B. There is evidence to show that the greater the level of pollution, the lower the fish population.
C. There is evidence to show that the greater the level of pollution, the greater the fish population.
D. Dead fish are causing the sites to become polluted.
E. Pollution has a negative effect on the environment.

The correlation coefficient

So far, our assessment of the type of correlation shown between two variables has been based upon subjective judgement. However, for more advanced work, we need a quantitative measurement of the strength of the relationship shown between a pair of variables. This measurement is known as a correlation coefficient.

There are several methods used for obtaining the correlation coefficient. The method which is used in this course is called the *q*-correlation coefficient. It can be found by using the steps detailed below.

Step 1: Draw a scatterplot of the data.
Step 2: Find the position of the median. Remember the median is the \( \frac{n+1}{2} \) th score if the scores are arranged in order of size.
Step 3: Draw a vertical line on the scatterplot of the data to indicate the position of the median of the \( x \)-values. There should be an equal number of points on each side of this line. Then draw a horizontal line to indicate the position of the median of the \( y \)-values. Note that these lines may or may not pass through one or more of the data points.
Step 4: The scatterplot is now divided into four sections or quadrants (hence the name ‘*q*-correlation coefficient’) that are labelled A, B, C and D as in the diagram at right. (The order of the labelling is important.)
Step 5: Count the number of data points that lie in each of the quadrants. Points which lie on either of the median lines are omitted.
Step 6: The correlation coefficient is calculated using the formula:

\[
q = \frac{(a + c) - (b + d)}{a + b + c + d}
\]

where 
- \( a \) is the number of data points in A
- \( b \) is the number of data points in B
- \( c \) is the number of data points in C
- \( d \) is the number of data points in D.
Step 7 The correlation coefficient may be interpreted by using the scale at right.

The method has been illustrated by applying it to the ski resort data that were introduced at the start of the chapter.

In the ski resort data there are 12 data points so the median is the \( \frac{12+1}{2} \)th score. That is, the 6.5th score, or halfway between the 6th and 7th score.

For these data: \( a = 5, b = 1, c = 5, d = 1 \).

\[
q = \frac{(a+c)-(b+d)}{a+b+c+d} \\
= \frac{(5+5)-(1+1)}{5+1+5+1} \\
= 0.67 \text{ (to 2 decimal places)}
\]

A \( q \)-correlation coefficient of 0.67 indicates a moderate positive correlation. The correlation coefficient can be interpreted in the way discussed in the previous section. In this case we would conclude: There is evidence to show that the greater the depth of snow, the greater the number of skiers.

Notice that the scale for correlation coefficients runs from -1 to 1. Every correlation coefficient must lie within this range. Any answer outside the range would indicate an error in workings.

It is also important to note that a \( q \)-correlation coefficient of 1 (or -1) does not necessarily indicate that all the points are in a perfectly straight line. A correlation coefficient of 1 could be obtained by having a cluster of points in quadrant A and a cluster in quadrant C (just as long as there are no points to be found in the other quadrants).

This is a limitation of the \( q \)-correlation coefficient. The \( q \)-correlation coefficient is useful because it is easy to apply to a scatterplot but it is not the most sophisticated or reliable way of finding a measure of linear association. You might like to research other methods for finding a correlation coefficient for a set of bivariate data.

**WORKED EXAMPLE 2**

For the graph shown at right,

a calculate the \( q \)-correlation coefficient

b state what type of correlation is involved.
First find the position of the median lines. (There are 12 data points.)

Position the vertical and horizontal median lines between the 5th and 6th points. In this case the vertical line will pass through two of the points. Label the quadrants A, B, C and D in an anticlockwise direction, beginning with the upper right quadrant.

Count the number of data points in each quadrant. Ignore those that lie on a line.

Use the formula to calculate the $q$-correlation coefficient.

Interpret the correlation coefficient using the scale on page 172.

The median is the $\frac{12+1}{2} = 6.5$th score.

$\begin{align*}
\text{Position:} & \quad \text{vertical line through 5th and 6th points, horizontal line through midpoints.} \\
\text{Quadrants:} & \quad A, B, C, D. \\
\text{Counting:} & \quad a = 4, b = 1, c = 4, d = 1. \\
\text{Formula:} & \quad q = \frac{(a+c) - (b+d)}{a + b + c + d} \\
& \quad = \frac{(4 + 4) - (1 + 1)}{4 + 1 + 4 + 1} \\
& \quad = 0.6. \\
\text{Interpretation:} & \quad \text{There is a moderate positive correlation.}
\end{align*}$

A manufacturer who is interested in minimising the cost of training gives 15 of his machine operators different amounts of training. He then measures the number of machine errors made by each of the operators. The results are shown in the table below.

<table>
<thead>
<tr>
<th>Hours spent training</th>
<th>2</th>
<th>5</th>
<th>7</th>
<th>3</th>
<th>8</th>
<th>9</th>
<th>16</th>
<th>18</th>
<th>10</th>
<th>4</th>
<th>6</th>
<th>12</th>
<th>10</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of errors during first week</td>
<td>43</td>
<td>16</td>
<td>25</td>
<td>31</td>
<td>24</td>
<td>18</td>
<td>6</td>
<td>8</td>
<td>24</td>
<td>34</td>
<td>27</td>
<td>15</td>
<td>22</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

- **a** Draw a scatterplot of the data.
- **b** Find the $q$-correlation coefficient.
- **c** Write a suitable conclusion for the manufacturer.

It is important to place the independent variable on the $x$-axis and the dependent variable on the $y$-axis. As we are interested to see if the number of errors changed with the hours of training (that is, errors depend on hours), errors will be placed on the $y$-axis and hours on the $x$-axis.
1. First find the position of the medians.

2. Draw vertical and horizontal lines to indicate the positions of the medians. Label the quadrants A, B, C and D (anticlockwise, from the upper right quadrant).

3. Count the number of points in each quadrant, ignoring those on median lines.

4. Use the formula to find $q$.

5. Interpret the correlation coefficient using the scale presented on page 172.

**REMEMBER**

1. The strength of the relationship between two variables can be measured using the $q$-correlation coefficient.
2. To find the $q$-correlation coefficient:
   (a) divide the scatterplot into quadrants by drawing horizontal and vertical lines representing the positions of the medians
   (b) count the number of points in each quadrant (do not include points that are on the lines)
   (c) calculate the coefficient by using the formula $q = \frac{(a + c) - (b + d)}{a + b + c + d}$.
3. The value of the $q$-correlation coefficient can be interpreted as follows:

   
<table>
<thead>
<tr>
<th>$q$</th>
<th>Strong</th>
<th>Moderate</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.00$</td>
<td>$0.75$</td>
<td>$0.50$</td>
<td>$0.25$</td>
</tr>
<tr>
<td>$0.25$</td>
<td>$0$</td>
<td>$-0.25$</td>
<td>$-0.50$</td>
</tr>
<tr>
<td>$-0.75$</td>
<td>$-1.00$</td>
<td>$-0.50$</td>
<td>$-0.25$</td>
</tr>
</tbody>
</table>

   Strong positive, Moderate positive, Weak positive, No correlation, Weak negative, Moderate negative, Strong negative

4. A limitation of the $q$-correlation coefficient is that if $q = 1$ (or $-1$), it does not necessarily mean that all points are in a perfect straight line.

**EXERCISE**

4B) The correlation coefficient

1. What type of correlation would be represented by scatterplots that had the following correlation coefficients?

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.0</td>
<td>b</td>
<td>0.4</td>
<td>c</td>
</tr>
<tr>
<td>d</td>
<td>-0.7</td>
<td>e</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>0.21</td>
<td>g</td>
<td>0.75</td>
<td>h</td>
</tr>
<tr>
<td>i</td>
<td>-0.25</td>
<td>j</td>
<td>-1.0</td>
<td></td>
</tr>
</tbody>
</table>
2 For each of the following graphs:
   i. calculate the \( r \)-correlation coefficient
   ii. state the type of correlation involved.

3 For each of the following graphs:
   i. calculate the \( r \)-correlation coefficient
   ii. state the type of correlation involved.
   (Hint: You need to work out the number of data points so that you can calculate the median line.)

4 The \( r \)-correlation coefficient for the scatterplot at right is:
   A. 1.00
   B. 0.67
   C. 0.85
   D. 9.08
   E. 0.72

5 The \( r \)-correlation coefficient for the scatterplot at right is:
   A. -0.36
   B. 0.50
   C. -0.50
   D. 1.25
   E. 0.36

6 A researcher who is investigating the proposition that ‘tall mothers have tall sons’ measures the height of twelve mothers and the height of their adult sons. The results are detailed in the following table.

<table>
<thead>
<tr>
<th>Height of mother (cm)</th>
<th>Height of son (cm)</th>
<th>Height of mother (cm)</th>
<th>Height of son (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>185</td>
<td>188</td>
<td>159</td>
<td>160</td>
</tr>
<tr>
<td>152</td>
<td>162</td>
<td>154</td>
<td>148</td>
</tr>
<tr>
<td>168</td>
<td>168</td>
<td>168</td>
<td>178</td>
</tr>
<tr>
<td>166</td>
<td>172</td>
<td>148</td>
<td>152</td>
</tr>
<tr>
<td>173</td>
<td>179</td>
<td>162</td>
<td>184</td>
</tr>
<tr>
<td>172</td>
<td>182</td>
<td>171</td>
<td>180</td>
</tr>
</tbody>
</table>
7 A teacher who is interested in the amount of time students spend doing homework asks 15 students to record the amount of time that they spent performing different activities on a particular evening. Part of her results are shown below.

<table>
<thead>
<tr>
<th>Hours watching television</th>
<th>3</th>
<th>0</th>
<th>2</th>
<th>5</th>
<th>6</th>
<th>4</th>
<th>2</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>5</th>
<th>2</th>
<th>6</th>
<th>4</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours on homework</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

\( a \) Draw a scattergraph of the data.
\( b \) Find the \( q \)-correlation coefficient for the data.
\( c \) Interpret this figure in terms of the experimental variables.
\( d \) Suggest reasons why the graph is not perfectly linear.

8 An agricultural scientist who is interested in the effects of rainfall upon grain production produces the following table of results.

<table>
<thead>
<tr>
<th>Annual rainfall (mm)</th>
<th>Grain production (tonnes ( \times 1000 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>589</td>
<td>389</td>
</tr>
<tr>
<td>380</td>
<td>285</td>
</tr>
<tr>
<td>620</td>
<td>460</td>
</tr>
<tr>
<td>360</td>
<td>350</td>
</tr>
<tr>
<td>425</td>
<td>410</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annual rainfall (mm)</th>
<th>Grain production (tonnes ( \times 1000 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>480</td>
<td>380</td>
</tr>
<tr>
<td>552</td>
<td>425</td>
</tr>
<tr>
<td>602</td>
<td>420</td>
</tr>
<tr>
<td>511</td>
<td>389</td>
</tr>
</tbody>
</table>

\( a \) Draw a scattergraph of the data.
\( b \) Find the \( q \)-correlation coefficient for the data.
\( c \) Interpret this figure in terms of the experimental variables.
\( d \) Suggest reasons why the graph is not perfectly linear.

9 The following table shows the number of hotels and the number of churches in 12 country towns.

<table>
<thead>
<tr>
<th>Number of hotels</th>
<th>Number of churches</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of hotels</th>
<th>Number of churches</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

\( a \) Draw a scattergraph of the data.
\( b \) Find the \( q \)-correlation coefficient for the data.
\( c \) Interpret this figure in terms of the experimental variables.
\( d \) Suggest reasons why the graph is not perfectly linear.

10 A psychologist asked 20 people to rate their ‘level of contentment’ on a scale of 0 to 10 (10 representing ‘perfectly content’). Their responses are detailed in the following table which also shows the annual income of each person.
<table>
<thead>
<tr>
<th>Annual salary</th>
<th>Level of contentment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$45,000</td>
<td>2</td>
</tr>
<tr>
<td>$30,000</td>
<td>7</td>
</tr>
<tr>
<td>$52,000</td>
<td>4</td>
</tr>
<tr>
<td>$36,000</td>
<td>6</td>
</tr>
<tr>
<td>$31,000</td>
<td>8</td>
</tr>
<tr>
<td>$43,000</td>
<td>8</td>
</tr>
<tr>
<td>$25,000</td>
<td>3</td>
</tr>
<tr>
<td>$49,000</td>
<td>7</td>
</tr>
<tr>
<td>$35,000</td>
<td>10</td>
</tr>
<tr>
<td>$27,000</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annual salary</th>
<th>Level of contentment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$29,000</td>
<td>6</td>
</tr>
<tr>
<td>$38,000</td>
<td>7</td>
</tr>
<tr>
<td>$40,000</td>
<td>3</td>
</tr>
<tr>
<td>$68,000</td>
<td>2</td>
</tr>
<tr>
<td>$23,000</td>
<td>5</td>
</tr>
<tr>
<td>$28,000</td>
<td>6</td>
</tr>
<tr>
<td>$31,000</td>
<td>7</td>
</tr>
<tr>
<td>$37,000</td>
<td>7</td>
</tr>
<tr>
<td>$32,000</td>
<td>8</td>
</tr>
<tr>
<td>$51,000</td>
<td>5</td>
</tr>
</tbody>
</table>

11 An experimenter who is investigating the relationship between exercise and obesity measures the weights of 30 boys (of equal height) and also documents the amount of physical exercise that the boys completed each week. The results are detailed in the following table.

<table>
<thead>
<tr>
<th>Duration of exercise (h)</th>
<th>12</th>
<th>15</th>
<th>10</th>
<th>8</th>
<th>4</th>
<th>2</th>
<th>6</th>
<th>8</th>
<th>20</th>
<th>0</th>
<th>13</th>
<th>16</th>
<th>18</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight (kg)</td>
<td>80</td>
<td>72</td>
<td>70</td>
<td>85</td>
<td>87</td>
<td>90</td>
<td>80</td>
<td>85</td>
<td>72</td>
<td>93</td>
<td>68</td>
<td>75</td>
<td>69</td>
<td>84</td>
<td>68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Duration of exercise (h)</th>
<th>7</th>
<th>9</th>
<th>12</th>
<th>3</th>
<th>6</th>
<th>8</th>
<th>12</th>
<th>16</th>
<th>18</th>
<th>12</th>
<th>10</th>
<th>8</th>
<th>9</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight (kg)</td>
<td>77</td>
<td>94</td>
<td>66</td>
<td>83</td>
<td>72</td>
<td>85</td>
<td>88</td>
<td>65</td>
<td>69</td>
<td>58</td>
<td>75</td>
<td>81</td>
<td>79</td>
<td>71</td>
<td>86</td>
</tr>
</tbody>
</table>

a Draw a scattergraph of the data.
b Find the $q$-correlation coefficient for the data.
c Interpret this figure in terms of the experimental variables.
d Do these data confirm the adage that ‘Money cannot buy you happiness’?

12 A researcher is interested in the association between the work rate of production workers and the level of incentive that they are offered under a certain scheme. After drawing a scatterplot she calculates the correlation between the two variables at 0.82. The researcher can conclude that:

A there is a strong positive correlation between the variables; the greater the incentive, the lower the work rate
B there is a strong positive correlation between the variables; the greater the incentive, the greater the work rate
C there is a strong negative correlation between the variables; the greater the incentive, the lower the work rate
D there is a strong positive correlation; incentives cause an increase in the work rate
E none of the conclusions above can be drawn without careful examination of the scatterplot itself

13 The $q$-correlation coefficient for a particular experiment was calculated at $-1.34$. It can be concluded that:

A there is an extremely strong positive relationship between the variables
B there is an extremely strong negative relationship between the variables
C there is no relationship between the variables
D the experiment was a failure
E there was an error in calculations

14 a Calculate the \( r \)-correlation coefficient for the scatterplot at right.

b Are the data truly linear (a totally straight line)? Explain the discrepancy that has occurred.

15 The data in the table below show the population density and unemployment rates for nine countries. Use the data to test the hypothesis that overcrowding causes unemployment.

<table>
<thead>
<tr>
<th>Country</th>
<th>Population density (people/sq km)</th>
<th>Unemployment %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>2</td>
<td>9.4</td>
</tr>
<tr>
<td>Canada</td>
<td>3</td>
<td>10.4</td>
</tr>
<tr>
<td>France</td>
<td>102</td>
<td>9.4</td>
</tr>
<tr>
<td>Germany</td>
<td>217</td>
<td>4.3</td>
</tr>
<tr>
<td>Italy</td>
<td>189</td>
<td>10</td>
</tr>
<tr>
<td>Japan</td>
<td>328</td>
<td>2.1</td>
</tr>
<tr>
<td>New Zealand</td>
<td>13</td>
<td>10.2</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>235</td>
<td>8.9</td>
</tr>
<tr>
<td>USA</td>
<td>27</td>
<td>6.8</td>
</tr>
</tbody>
</table>

4C Linear modelling

If a linear relationship exists between a pair of variables then it is useful to be able to summarise the relationship in terms of an equation. This equation can then be used to make predictions about the levels of one variable given the value of the other.

The process of finding the equation is known as linear modelling. An equation can be found to represent the line which passes through any two points by using two coordinate geometry formulas.

The gradient of the line, passing through \((x_1, y_1)\) and \((x_2, y_2)\) is given by:

\[
m = \frac{y_2 - y_1}{x_2 - x_1}.
\]

The equation of a straight line with the gradient \(m\) and passing through \((x_1, y_1)\) is given by:

\[
y - y_1 = m(x - x_1).
\]

**WORKED EXAMPLE 4**

Find the equation of the line passing through the points \((2, 6)\) and \((5, 12)\).

**THINK**

1 First find the gradient between the points. In order not to get muddled with the \(x\)- and \(y\)-coordinates, write \((x_1, y_1)\) and \((x_2, y_2)\) above each of the points.

**WRITE**

\[
(x_1, y_1) = (2, 6), (x_2, y_2) = (5, 12)
\]

\[
m = \frac{y_2 - y_1}{x_2 - x_1}
\]

\[
= \frac{12 - 6}{5 - 2}
\]

\[
= \frac{6}{3}
\]

\[
= 2
\]
Substitute the value of the gradient and the coordinates of one of the points into the gradient–point formula.

\[ y - y_1 = m(x - x_1) \]
\[ y - 6 = 2(x - 2) \]
\[ y - 6 = 2x - 4 \]
\[ y = 2x - 4 + 6 \]
\[ y = 2x + 2 \]

Multiply out the brackets and transpose it to the more familiar \( y = mx + c \) form.

\[ y - 6 = 2x - 4 \]
\[ y = 2x - 4 + 6 \]
\[ y = 2x + 2 \]

To apply this technique to a scatterplot that consists of many points we need to fit a straight line through the whole set of points.

The process of fitting a line to a set of points is often referred to as regression. The regression line or trend line may be placed on a scatterplot by eye or by using the two-mean method. Its equation can then be found by using the method in the previous example by choosing any two points that are on the line.

When we apply this technique to the ski resort data presented earlier in the chapter, it can be seen that it is not possible to rule a single straight line through all the points. We are looking for the line of best fit; that is, the straight line which most closely fits the data. The positioning of this line by eye will clearly rely upon some careful judgement and experience. The line drawn on the scatterplot at right is just one of many that could have equally well been drawn by sight. As a guide to help in positioning the line, try to manoeuvre the ruler on the graph until you think that the line appears to minimise the total distance of the points from the line.

The equation that represents the relationship between the number of skiers and the depth of snow can be found by using the method shown in the previous worked example. The two points \((x_1, y_1)\) and \((x_2, y_2)\) may be any two points that lie on the trend line. Neither of them needs to be one of the original data points. It might be that the line of best fit does not actually pass through any of the original data points. For accuracy it is best to choose two points — one towards each end of the line of best fit.

For the ski resort data: \((x_1, y_1)\) \((x_2, y_2)\)
\((0.5, 120)\) \((3.2, 635)\)
\[ m = \frac{y_2 - y_1}{x_2 - x_1} \]
\[ = \frac{635 - 120}{3.2 - 0.5} \]
\[ = 190.74 \]
\[ y - y_1 = m(x - x_1) \]
\[ y - 120 = 190.74(x - 0.5) \]
\[ y - 120 = 190.74x - 95.37 \]
\[ y = 190.74x + 24.63 \]
\[ y = 191x + 25 \]

Rounding the answer appropriately for the example.

In other words, the number of skiers can be found for any snow depth by applying the equation \(N = 191s + 25\) where \(N\) is the number of skiers and \(s\) is the depth of snow in metres.

There are two ways that the equation can be checked for correctness. Firstly, the \(y\)-intercept and gradient should be appropriate to the scatterplot. In the case of the ski resort data, the equation indicates a positive gradient and a \(y\)-intercept of 25. Both of these can be checked quickly by inspecting the graph. (In cases where it is necessary to break either of the axes in order to draw the graph, direct confirmation of the \(y\)-intercept is not possible.) As an
additional check on the equation it is sensible to substitute some \( x \)-values into the equation and see if the points generated by the corresponding \( y \)-values appear to lie on the line of best fit that has been drawn on the graph.

It is important to understand the meaning of the \( y \)-intercept and gradient of the line. The \( y \)-intercept is the value of \( y \) when the level of \( x \) is zero. In the case of the ski resort equation the \( y \)-intercept is 25, indicating that when the depth of snow is 0 metres there will be 25 skiers at the resort. The gradient of the equation represents the rate of change of variable \( y \) with changing \( x \). In the case of the ski resort data, the gradient of 191 indicated that for every extra metre of snow there will be 191 additional skiers.

Sometimes after drawing a scatterplot it is clear that the points represent a relationship that is not linear. The relationship might be one of the non-linear types shown below.

In such cases it is not appropriate to try to model the data by attempting to fit a straight line through the points and find its equation. It is similarly inappropriate to attempt to fit a linear model (straight line) through a scatterplot if it shows that there is no correlation between the variables.

**WORKED EXAMPLE 5**

The following table shows the fare charged by a bus company for journeys of differing length.

<table>
<thead>
<tr>
<th>Distance (km)</th>
<th>Fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>$2.10</td>
</tr>
<tr>
<td>0.5</td>
<td>$2.00</td>
</tr>
<tr>
<td>7.5</td>
<td>$4.50</td>
</tr>
<tr>
<td>6</td>
<td>$4.00</td>
</tr>
<tr>
<td>6</td>
<td>$4.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance (km)</th>
<th>Fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>$2.60</td>
</tr>
<tr>
<td>0.5</td>
<td>$2.00</td>
</tr>
<tr>
<td>8</td>
<td>$4.50</td>
</tr>
<tr>
<td>4</td>
<td>$3.50</td>
</tr>
<tr>
<td>3</td>
<td>$3.00</td>
</tr>
</tbody>
</table>

\( \text{a} \) Represent the data using a scatterplot and place in the trend line by eye.

\( \text{b} \) Find an equation which relates fare, \( F \), to distance travelled, \( d \).

\( \text{c} \) Explain in words the meaning of the \( y \)-intercept and gradient of the line.

**THINK**

\( \text{a} \) 1. Distance is the independent variable so it will be assigned to the \( x \)-axis and fare to the \( y \)-axis.

2. Plot the data and place the trend line on your scatterplot.

Try to position the trend line so that the total distance of the points from the line is minimised. It doesn’t matter if the trend line doesn’t actually pass through any of the points.

\( \text{b} \) 1. Pick two points that are reasonably well separated on the trend line: (0.5, 2.00) and (6, 4.00) will do.

(In this case, these are actually original data points but it wouldn’t matter if they weren’t, so long as they were points that were on the trend line.)
2. Find the gradient of the line.

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{4.00 - 2.00}{6 - 0.5} = 0.364$$

3. Substitute the value of the gradient and coordinates of one of the points into the gradient–point formula.

$$y - y_1 = m(x - x_1)$$

$$y - 2.00 = 0.364(x - 0.5)$$

$$y = 0.364x + 1.818$$

4. Expand and then transpose to make $y$ the subject.

5. Exchange $x$ and $y$ for the variables in the question and round off all numbers to 2 decimal places, as we are talking about dollars and cents.

The gradient is the rate of change of variable $y$ with changing variable $x$.

The $y$-intercept represents the level of $y$ when $x = 0$.

The gradient represents the extra amount charged for each extra km travelled (that is, the rate at which the fare increases with increasing distance).

The $y$-intercept shows the fare charged for a trip of distance 0 km! (that is, the minimum charge).

As seen in worked example 1, a CAS calculator can be used to help you draw scatterplots. Once the scatterplot is on the screen, the trend line can be fitted by eye. That is, a movable line can be added to the scatterplot and rotated until it fits the data well. The equation of the line will appear on the screen as shown in the following example.

**WORKED EXAMPLE 6**

The table below gives the times (in hours) spent by 8 students studying for a measurement test and the marks (in %) obtained on the test.

<table>
<thead>
<tr>
<th>Time spent studying (h)</th>
<th>2</th>
<th>1</th>
<th>7</th>
<th>5</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test mark (%)</td>
<td>54</td>
<td>48</td>
<td>98</td>
<td>74</td>
<td>60</td>
<td>64</td>
<td>89</td>
<td>65</td>
</tr>
</tbody>
</table>

a. Draw the scatterplot to represent the data.

b. Find the equation of the line of best fit. Write your equation in terms of the variables: *time spent studying* and *test mark*.

**THINK**

a. On the Statistics screen, label **list1** as ‘time’ and enter in the time spent studying.

b. Label **list2** as ‘mark’ and enter in the test marks.

**WRITE/DISPLAY**

a. On the Statistics screen, label **list1** as ‘time’ and enter in the time spent studying. Label **list2** as ‘mark’ and enter in the test marks.
To draw the scatterplot, of the points, tap:
- **SetGraph**
- **Setting**
  Set:
  Type: Scatter
  XList: main\time
  YList: main\mark
  Freq: 1
  Mark: Square
- **Set**

To determine the line of best fit, tap:
- **Calc**
- **Linear Reg**
  Set:
  XList: main\r
  YList: main\w
  Freq: 1
  Then tap **OK**.

b  
Copy the equation from the screen into your workbook.

b  
Equation of the line of best fit is:
\[
y = 6.39 \times x + 42
\]

Test mark = 6.39 \times time spent studying + 42

---

**The two-mean regression line**

The equation of the regression (or trend) line can also be found using the *two-mean method*. This method also involves finding the equation of the straight line using two points \((x_1, y_1)\) and \((x_2, y_2)\). However, unlike the method of fitting the line by eye, where *any* two points can be selected, the coordinates of the two points used here need to be calculated in a certain way, as follows.

1. First, we need to ensure the data points are arranged so that the *x*-coordinates are in *increasing order*.
2. Next, we divide the data into two sets, upper and lower.
3. We then calculate the mean value of the *x*- and *y*-coordinates of each half, to obtain the two points \((\bar{x}_L, \bar{y}_L)\) and \((\bar{x}_U, \bar{y}_U)\). (That is, we find the mean value of the *x*-coordinates of all data points in the lower half, \(\bar{x}_L\), and the mean value of the *y*-coordinates of those points, \(\bar{y}_L\).)
4. We then repeat the process for the data points in the upper half to obtain the mean values \(\bar{x}_U\) and \(\bar{y}_U\).)
5. Once the coordinates of the two points are obtained, we find the equation of the line, using the same method as in the previous two worked examples. That is, we first find the gradient,

\[
m = \frac{y_2 - y_1}{x_2 - x_1},
\]

where \((x_1, y_1)\) is \((\bar{x}_L, \bar{y}_L)\) and \((x_2, y_2)\) is \((\bar{x}_U, \bar{y}_U)\).

6. We then substitute the value of the gradient and the coordinates of either of the two points into gradient–point formula, \(y - y_1 = m(x - x_1)\) to find the equation of the regression line.

*Note*: If there is an even number of data points, we can easily divide the data into two sets by placing an equal number of points in each half. If, however, the number of data points is odd, *we have to include the middle point in both sets*. This is shown in the following worked example.
For each of the following data sets, find the equation of the regression line using the two-mean method.

\[ \begin{array}{cccccccc}
  x & 20 & 50 & 70 & 30 & 60 & 80 & 40 \\
  y & 72 & 77 & 66 & 68 & 70 & 62 & 61 \\
\end{array} \]

\[ \begin{array}{cccccccc}
  x & 10 & 15 & 20 & 25 & 30 & 35 & 40 & 45 & 50 \\
  y & 33 & 42 & 54 & 61 & 72 & 83 & 95 & 101 & 113 \\
\end{array} \]

THINK

1. Rearrange the data points so that the \( x \)-coordinates are in increasing order.
2. Divide the data into two sets. Since there are 8 data points, place the first four in the lower half and the last four in the upper half.
3. Calculate the mean of the \( x \)-values and the mean of the \( y \)-values for the lower half.
4. Write the mean values in coordinate form.
5. Repeat the procedure for the upper half of the data points. That is, calculate the mean of the \( x \)-values and the mean of the \( y \)-values and write them in coordinate form.
6. The two points, \((x_1, y_1)\) and \((x_2, y_2)\), can now be used to find the equation of the regression line. First write these points side by side and denote them as \((x_1, y_1)\) and \((x_2, y_2)\).
Find the gradient of the line.

\[ m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{64.25 - 69.5}{75 - 35} = \frac{-5.25}{40} = -0.13125 \]

\[ y - 69.5 = -0.13125(x - 35) \]

Substitute the value of the gradient and the coordinates of one of the points into the gradient-point formula.

Expand and then transpose to make \( y \) the subject.

\[ y = -69.5 - 0.13125x + 4.59375 \]

\[ y = -0.13125x + 74.09375 \]

The \( x \)-coordinates are in increasing order, so we can now divide the data into two sets. There is an odd number of data points (9), so the middle point (30, 72) must be included in both the lower and the upper halves.

Calculate the mean of the \( x \)-values and the mean of the \( y \)-values for the lower half.

\[ \bar{x}_L = \frac{10 + 15 + 20 + 25 + 30}{5} = 20 \]

\[ \bar{y}_L = \frac{33 + 42 + 54 + 61 + 72}{5} = 52.4 \]

Write the mean values in coordinate form.

\[ (\bar{x}_L, \bar{y}_L) = (20, 52.4) \]

Repeat the procedure for the upper half of the data points. That is, calculate the mean of the \( x \)-values and the mean of the \( y \)-values and write them in coordinate form.

\[ \bar{x}_U = \frac{30 + 35 + 40 + 45 + 50}{5} = 40 \]

\[ \bar{y}_U = \frac{72 + 83 + 95 + 101 + 113}{5} = 92.8 \]

\[ (\bar{x}_U, \bar{y}_U) = (40, 92.8) \]

The two points, \((\bar{x}_L, \bar{y}_L)\) and \((\bar{x}_U, \bar{y}_U)\), can now be used to find the equation of the regression line. First write these points side by side and denote them as \((x_1, y_1)\) and \((x_2, y_2)\).
Find the gradient of the line.

\[ m = \frac{y_2 - y_1}{x_2 - x_1} \]

\[ = \frac{92.8 - 52.4}{40 - 20} \]

\[ = \frac{40.4}{20} \]

\[ = 2.02 \]

Substitute the value of the gradient and the coordinates of one of the points into the gradient-point formula.

\[ y - y_1 = m(x - x_1) \]

\[ y - 52.4 = 2.02(x - 20) \]

Expand and then transpose to make \( y \) the subject.

\[ y - 52.4 = 2.02x - 40.4 \]

\[ y = 2.02x + 12 \]

REMEMBER

1. When fitting a line of best fit (or a trend line) on a scatterplot by eye, try to minimise the total distance of the points from the line.
2. The trend line should not necessarily pass through any of the original data points.
3. The equation of the line passing through the points \((x_1, y_1)\) and \((x_2, y_2)\) can be found using the formula

\[ y - y_1 = m(x - x_1) \]

where

\[ m = \frac{y_2 - y_1}{x_2 - x_1} \]

4. It is best to choose two points that are not too close to each other.
5. The \( y \)-intercept is the value of \( y \) when the level of \( x \) is 0.
6. The gradient represents the rate of change of \( y \) with changing \( x \).
7. To find the equation of the regression line using the two-mean method:
   (a) rearrange (if necessary) the data points so that the \( x \)-coordinates are in increasing order
   (b) divide the data into halves (upper and lower)
   (c) calculate the mean value of the \( x \)- and \( y \)-coordinates of each half, to obtain the two points \((\bar{x}_L, \bar{y}_L)\) and \((\bar{x}_U, \bar{y}_U)\)
   (d) use the coordinates of the two points to find the equation of the regression line.
8. If there is an even number of data points, place an equal number of points in each set. If the number of data points is odd, include the middle point in both sets.

EXERCISE 4C

Linear modelling

1. Find the equations of the straight line passing through the following pairs of points.

   a. \((2, 1)\) and \((6, 3)\)
   b. \((5, 4)\) and \((7, 0)\)
   c. \((5, 8)\) and \((10, 18)\)
   d. \((12, 17)\) and \((6, 27)\)
   e. \((10, 32)\) and \((25, 67)\)
   f. \((14, 52)\) and \((23, 18)\)
   g. \((5.2, 12.9)\) and \((7.4, 8.6)\)
   h. \((31.5, 82.3)\) and \((101.8, 32.6)\)
   i. \((6, 8)\) and \((-2, 4)\)
   j. \((5.2, -7.4)\) and \((23.7, -36.8)\)
2. The equation which joins the points (3, 8) and (12, 26) is:

A) \( y = 2.8x - 0.4 \)
B) \( y = 2x - 14 \)
C) \( y = 2x + 2 \)
D) \( y = 0.5x + 6.5 \)
E) None of these

3. Which of the following could possibly be the equation of the graph at right?

A) \( y = -2x + 20 \)
B) \( y = -2x - 20 \)
C) \( y = 2x + 20 \)
D) \( y = 2x - 20 \)
E) \( y = 20x + 2 \)

4. Which of the following could possibly be the equation of the graph at right?

A) \( y = -2x + 20 \)
B) \( y = -2x - 20 \)
C) \( y = 2x + 20 \)
D) \( y = 2x - 20 \)
E) \( y = 20x - 2 \)

5. Position the straight line of best fit by eye through each of the following graphs and find the equation of each.

a

b

c

6. As part of an experiment, a student measures the length of a spring when different masses are attached to it. Her results are shown below.

<table>
<thead>
<tr>
<th>Mass (g)</th>
<th>Length of spring (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>220</td>
</tr>
<tr>
<td>100</td>
<td>225</td>
</tr>
<tr>
<td>200</td>
<td>231</td>
</tr>
<tr>
<td>300</td>
<td>235</td>
</tr>
<tr>
<td>400</td>
<td>242</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mass (g)</th>
<th>Length of spring (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>246</td>
</tr>
<tr>
<td>600</td>
<td>250</td>
</tr>
<tr>
<td>700</td>
<td>254</td>
</tr>
<tr>
<td>800</td>
<td>259</td>
</tr>
<tr>
<td>900</td>
<td>264</td>
</tr>
</tbody>
</table>
a. Draw a scatterplot of the data and place in the trend line by eye.

b. Find an equation which relates the spring length, $L$, and the mass attached, $m$.

c. Explain in words the meaning of the $y$-intercept and the gradient of the line.

7 A scientist who measures the volume of a gas at different temperatures provides the following table of values.

<table>
<thead>
<tr>
<th>Temperature ($^\circ$C)</th>
<th>Volume (litres)</th>
</tr>
</thead>
<tbody>
<tr>
<td>−40</td>
<td>1.2</td>
</tr>
<tr>
<td>−30</td>
<td>1.9</td>
</tr>
<tr>
<td>−20</td>
<td>2.4</td>
</tr>
<tr>
<td>0</td>
<td>3.1</td>
</tr>
<tr>
<td>10</td>
<td>3.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature ($^\circ$C)</th>
<th>Volume (litres)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>4.1</td>
</tr>
<tr>
<td>30</td>
<td>4.8</td>
</tr>
<tr>
<td>40</td>
<td>5.3</td>
</tr>
<tr>
<td>50</td>
<td>6.1</td>
</tr>
<tr>
<td>60</td>
<td>6.7</td>
</tr>
</tbody>
</table>

a. Draw a scatterplot of the data.

b. Give the equation of the line of best fit. Write your equation in terms of the variables volume of gas, $V$, and its temperature, $T$.

c. Explain the meaning of the $y$-intercept and the gradient of the line.

8 MC The length (in metres) and wingspan (in metres) of eight commercial aeroplanes are displayed in the table below.

<table>
<thead>
<tr>
<th>Length</th>
<th>70.7</th>
<th>70.7</th>
<th>63.7</th>
<th>58.4</th>
<th>54.9</th>
<th>39.4</th>
<th>36.4</th>
<th>33.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wingspan</td>
<td>64.4</td>
<td>59.6</td>
<td>60.3</td>
<td>60.3</td>
<td>47.6</td>
<td>35.8</td>
<td>28.9</td>
<td>28.9</td>
</tr>
</tbody>
</table>

The equation of the regression line for this data is 

$$wingspan = -2.99 + 0.96 \times length$$

From this equation it can be concluded that, on average, for these aeroplanes, wingspan:

A decreases by 2.03 metres with each one metre increase in length  
B increases by 0.96 metres with each one metre increase in length  
C decreases by 0.96 metres with each one metre increase in length  
D increases by 2.99 metres with each one metre increase in length  
E decreases by 2.99 metres with each one metre increase in length

9 WE6 The following pair of tables gives the heights of 12 males at age two years together with their adult height, and the same information for 12 females.

<table>
<thead>
<tr>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height at age two years (cm)</td>
<td>Adult height (cm)</td>
</tr>
<tr>
<td>81</td>
<td>164</td>
</tr>
<tr>
<td>82</td>
<td>163</td>
</tr>
<tr>
<td>83</td>
<td>168</td>
</tr>
<tr>
<td>84</td>
<td>170</td>
</tr>
<tr>
<td>84</td>
<td>169</td>
</tr>
<tr>
<td>86</td>
<td>173</td>
</tr>
<tr>
<td>87</td>
<td>176</td>
</tr>
<tr>
<td>87</td>
<td>180</td>
</tr>
<tr>
<td>90</td>
<td>179</td>
</tr>
<tr>
<td>91</td>
<td>185</td>
</tr>
<tr>
<td>91</td>
<td>183</td>
</tr>
<tr>
<td>93</td>
<td>189</td>
</tr>
</tbody>
</table>
a Draw two separate scatterplots.

b For each graph find the equation of the line of best fit. Write your equation in terms of the variables adult height, \( A \), and baby height, \( b \).

For each of the following data sets, find the equation of the regression line, using the two-mean method.

a

<table>
<thead>
<tr>
<th>( x )</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>24</td>
<td>35</td>
<td>43</td>
<td>57</td>
<td>66</td>
<td>75</td>
<td>89</td>
<td>94</td>
</tr>
</tbody>
</table>

b

<table>
<thead>
<tr>
<th>( x )</th>
<th>15</th>
<th>40</th>
<th>25</th>
<th>5</th>
<th>35</th>
<th>10</th>
<th>30</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>79</td>
<td>45</td>
<td>65</td>
<td>90</td>
<td>40</td>
<td>84</td>
<td>61</td>
<td>70</td>
</tr>
</tbody>
</table>

c

<table>
<thead>
<tr>
<th>( x )</th>
<th>30</th>
<th>32</th>
<th>34</th>
<th>36</th>
<th>38</th>
<th>40</th>
<th>42</th>
<th>44</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>40</td>
<td>53</td>
<td>61</td>
<td>68</td>
<td>82</td>
<td>90</td>
<td>104</td>
<td>111</td>
<td>119</td>
</tr>
</tbody>
</table>

d

<table>
<thead>
<tr>
<th>( x )</th>
<th>78</th>
<th>81</th>
<th>80</th>
<th>84</th>
<th>77</th>
<th>83</th>
<th>76</th>
<th>79</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>28</td>
<td>43</td>
<td>37</td>
<td>52</td>
<td>25</td>
<td>47</td>
<td>44</td>
<td>21</td>
</tr>
</tbody>
</table>

An anthropologist is interested in finding a rule for establishing the height of a person from the length of one of that person’s bones. He measures the heights of 12 skeletons and the length of the femur in each.

<table>
<thead>
<tr>
<th>Length of femur (cm)</th>
<th>Height of skeleton (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>150</td>
</tr>
<tr>
<td>40</td>
<td>154</td>
</tr>
<tr>
<td>41</td>
<td>160</td>
</tr>
<tr>
<td>41</td>
<td>162</td>
</tr>
<tr>
<td>42</td>
<td>168</td>
</tr>
<tr>
<td>44</td>
<td>170</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Length of femur (cm)</th>
<th>Height of skeleton (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>171</td>
</tr>
<tr>
<td>46</td>
<td>178</td>
</tr>
<tr>
<td>46</td>
<td>180</td>
</tr>
<tr>
<td>47</td>
<td>181</td>
</tr>
<tr>
<td>47</td>
<td>185</td>
</tr>
<tr>
<td>49</td>
<td>190</td>
</tr>
</tbody>
</table>

a Draw a scatterplot of the data.

b Determine the \( q \)-correlation coefficient for the data and interpret it in words.

c Establish the rule for finding the height of a person, \( H \), given the length of the femur bone, \( L \), using the two-mean method.

A sports scientist is interested in the importance of muscle bulk to strength. He measures the biceps circumference of ten people and tests their strength by asking them to complete a lift test. His results are detailed in the following table.

<table>
<thead>
<tr>
<th>Circumference of biceps (cm)</th>
<th>Lift test (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>25</td>
<td>52</td>
</tr>
<tr>
<td>27</td>
<td>58</td>
</tr>
<tr>
<td>28</td>
<td>51</td>
</tr>
<tr>
<td>30</td>
<td>60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Circumference of biceps (cm)</th>
<th>Lift test (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>62</td>
</tr>
<tr>
<td>31</td>
<td>53</td>
</tr>
<tr>
<td>33</td>
<td>62</td>
</tr>
<tr>
<td>34</td>
<td>61</td>
</tr>
<tr>
<td>36</td>
<td>66</td>
</tr>
</tbody>
</table>
a. Draw a scatterplot of the data.

b. Determine the \( r \)-correlation coefficient for the data and interpret it in words.

c. Find a rule for determining the ability of a person to complete a lift test, \( S \), from the circumference of their biceps, \( B \), using the two-mean method.

13. MC Which of the following graphs would be unsuitable for modelling by a linear equation?

\[ \begin{array}{cc}
\text{A} & \text{B} \\
\text{a, b and c} & \text{b only}
\end{array} \]

\[ \begin{array}{cc}
\text{D} & \text{C}
\end{array} \]

\[ \begin{array}{c}
\text{b and c}
\end{array} \]

\[ \begin{array}{c}
\text{d only}
\end{array} \]

4D Making predictions

The equation of the trend line may be used to make predictions about the variables.

WORKED EXAMPLE 8

It is found that the relationship between the number of people playing a casino Jackpot game and the prize money offered is given by the equation \( N = 0.07p + 220 \), where \( N \) is the number of people playing and \( p \) is the prize money.

a. Find the number of people playing when the prize money is $2500.

b. Find the likely prize on offer if there were 500 people playing.

THINK

Method 1 : Technology-free

a. 1 We are told the value of \( p \) and are asked to find out the value of \( N \).

2 Substitute the value of \( p \) into the equation and evaluate.

3 Write the final answer in a sentence.

b. 1 This time we are asked to find \( p \) when we know \( N \).

2 Substitute the value of \( N \) into the equation.

3 Solve for \( p \).

4 Write the final answer in a sentence.
Method 2: Technology-enabled

1. On the Main screen, tap:
   - Action
   - Advanced
   - solve

Complete the entry lines as:
solve\((n = 0.07p + 220, n) \mid p = 2500\)
solve\((n = 0.07p + 220, p) \mid n = 500\)
Then press \(\mathbb{E}\).

2. Write the answers.

\[ a \quad \text{When the prize money is } \$2500, \text{ the number of people playing is } 395. \]
\[ b \quad \text{When there are } 500 \text{ people playing, the prize money is } \$4000. \]

Predictions of one variable given the other may be made by using an equation, as in the previous worked example, but can also be made by reading information from the trend line on the graph.

In the previous section we found that the equation representing the relationship between the variables in the ski resort data was \(N = 191s + 25\), where \(N\) is the number of skiers and \(s\) is the depth of snow.

Suppose that we wished to predict the number of skiers when the snow depth was 3 m. This could be done by substituting \(s = 3\) into the equation:

\[
N = 191s + 25
= 191(3) + 25
= 598\text{ skiers}
\]

Alternatively, the prediction could be made from the graph’s trend line.

**Worked Example 9**

The scatterplot at right shows the depth of snow and the corresponding number of skiers.

From the graph’s trend line find:

\[ a \quad \text{the number of skiers when snow depth was } 3\text{ m} \]
\[ b \quad \text{the depth of snow that would attract about } 400\text{ people.} \]

**Think**

\[ a \quad 1 \quad \text{Find } 3\text{ m on the } \text{‘Depth of snow’ axis.} \]

Rule a line vertically to the trend line, then horizontally to the ‘Number of skiers’ axis.

The line meets the axis at about 600.
2 Write the answer.

b 1 Find 400 m on the ‘Number of skiers’ axis. Rule a line horizontally to the trend line, then vertically to the ‘Depth of snow’ axis. The line meets the axis at about 2.

2 Write the answer.

When the snow is 3 m deep, the number of skiers is about 600.

It would take a snow cover of about 2 m to draw a crowd of 400 skiers

**Interpolation and extrapolation**

We use the term *interpolation* when we make predictions from a graph’s trend line from within the bounds of the original experimental data.

Data may be interpolated either algebraically or graphically as shown by the methods above. Both of the predictions made for the ski resort data were examples of interpolation. In the first case the number of skiers for a snow depth of 3 m was sought. This is interpolation because the original data involved recordings between the bounds of 0.5 m and 3.6 m. The second prediction (the snow depth required to draw 400 skiers), is also an example of interpolation because the result, 2 m, is within the bounds of the original recordings (0.5 m–3.6 m).

We use the term *extrapolation* when we make predictions from a graph’s trend line from outside the bounds of the original experimental data.

Data can also be extrapolated either algebraically or graphically. Suppose we wished to find the number of skiers if the snow depth was 4.5 m. This could be done algebraically by substituting \( s = 4.5 \) into the equation:

\[
N = 191s + 25
\]

\[
= 191(4.5) + 25
\]

\[
= 885 \text{ skiers.}
\]

The data could also be extrapolated graphically by first extending the trend line to make the prediction possible.

The graph shows that about 900 skiers would be attracted to the resort if the snow depth was 4.5 m.

**Reliability of predictions**

Results predicted (whether algebraically or graphically) from the trend line of a scatterplot can be considered reliable only if:

1. a reasonably large number of points were used to draw the scatterplot,
2. a reasonably strong correlation was shown to exist between the variables (the stronger the correlation, the greater the confidence in predictions),
3. the predictions were made using interpolation and not extrapolation.
Extrapolated results can never be considered to be reliable because when extrapolation is used we are assuming that the relationship holds true for untested values. In the case of the ski resort data, we would be assuming that the relationship \( N = 191s + 25 \) continues to hold true even when the snow depth is extreme. This may or may not be the case. It might be that there are only a certain number of skiers in the population — which places a limit on the maximum number visiting the resort. In this situation the trend line, if continued, may look like that in the graph above.

Alternatively, it might be that when the snow depth is extreme then skiers can not even get to the slopes. Then, the trend line would fall.

**REMEmBER**

1. Predictions of the level of one variable, given the level of the other variable, can be made algebraically (using the equation of the trend line), or graphically (using the graph of a trend line).
2. *Interpolation* is making predictions within the bounds of the original data.
3. *Extrapolation* is making predictions beyond the bounds of the original data.
4. Predictions can be considered reliable if they were obtained using interpolation from the scatterplot, indicating reasonably strong correlation and containing a large number of points.
Making predictions

1. A cab company adjusts its taxi meters so that the fare is charged according to the following equation: \( F = 1.2d + 3 \) where \( F \) is the fare in dollars and \( d \) is the distance travelled in km.
   a. Find the fare charged for a distance of 12 km.
   b. Find the fare charged for a distance of 4.5 km.
   c. Find the distance that could be covered on a fare of $26.
   d. Find the distance that could be covered on a fare of $13.20.

2. Detectives can use the equation \( H = 6.1f - 5 \) to determine the height of a burglar who leaves footprints behind. (\( H \) is the height of the burglar in cm and \( f \) is the length of the footprint.)
   a. Find the height of a burglar whose footprint is 27 cm in length.
   b. Find the height of a burglar whose footprint is 30 cm in length.
   c. Find the footprint length of a burglar of height 185 cm.
   d. Find the footprint length of a burglar of height 152 cm.

3. Computer equipment devalues according to the equation \( V = -500t + 2500 \), where \( V \) is the value of the equipment after \( t \) years.
   a. Find the value of the equipment after 4 years.
   b. Find the value of the equipment after 2 years and 9 months.
   c. Find the value of the equipment when it was new.
   d. How old will the equipment be when its value is $1000?
   e. How old will the equipment be when it becomes worthless?

4. A football match pie seller finds that the number of pies that she sells is related to the temperature of the day. The situation could be modelled by the equation \( N = -23t + 870 \), where \( N \) is the number of pies sold and \( t \) is the temperature of the day.
   a. Find the number of pies sold if the temperature was 5 degrees.
   b. Find the number of pies sold if the temperature was 25 degrees.
   c. Find the likely temperature if 400 pies were sold.
   d. How hot would the day have to be before the pie seller sold no pies at all?

5. An electronics repair shop charges according to the equation \( C = 40h + 35 \), where \( C \) is the cost of the repairs and \( h \) is the number of hours spent working on the repair.
   a. Find the cost of repairing an amplifier if the job took 2 hours.
   b. Find the cost of repairing a television if the job took 1 hour and 15 minutes.
   c. How long would the repairers have to spend on a job before the charge exceeded $175?
   d. How long would the repairers have to spend on a job before the charge exceeded $215?

6. The linear relationship between two variables, \( X \) and \( Y \), is given by the equation \( Y = -6.2X + 20 \). Predict the level of \( X \) when \( Y = 4.5 \).

   A. \( X = -7.9 \)
   B. \( X = -96.1 \)
   C. \( X = -20.73 \)
   D. \( X = 2.5 \)
   E. The level of \( X \) cannot be found using this equation. This equation gives the level of \( Y \) only.

7. Use the graph to predict:
   i. the value of \( y \) when \( x = 12 \)
   ii. the value of \( y \) when \( x = 6 \)
   iii. the value of \( y \) when \( x = 23 \)
   iv. the value of \( x \) when \( y = 600 \)
   v. the value of \( x \) when \( y = 1000 \)
   vi. the value of \( x \) when \( y = 320 \).
   b. Find the equation of the trend line.
   c. Now use the equation to confirm the answers given in part a.
8 The scatterplot below shows the average number of children (per family) and the birth rate (per 100 000) for 10 different countries.

The equation of the regression line shown on the graph is:

\[ \text{Average number of children} = -0.48 + 0.146 \times \text{birth rate}. \]

Using this equation, we predict that, for a country with a birth rate of 60 (per 100 000), the average number of children (per family) will be:

- A 8.28
- B 8.76
- C 9.24
- D 20.04
- E 59.66

EXAM TIP 82% of students answered this question correctly.

9 The following table shows the average annual costs of running a car. It includes all fixed costs (registration, insurance etc.) as well as running costs (petrol, repairs, etc.).

<table>
<thead>
<tr>
<th>Distance (km)</th>
<th>Annual cost ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 000</td>
<td>4 000</td>
</tr>
<tr>
<td>10 000</td>
<td>6 400</td>
</tr>
<tr>
<td>15 000</td>
<td>8 400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance (km)</th>
<th>Annual cost ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 000</td>
<td>10 400</td>
</tr>
<tr>
<td>25 000</td>
<td>12 400</td>
</tr>
<tr>
<td>30 000</td>
<td>14 400</td>
</tr>
</tbody>
</table>

a Draw a scatterplot of the data.
b Draw in the line of best fit by eye. How close are the data points to your line?
c Find an equation which represents the relationship between the cost of running a vehicle, \( C \), and the distance travelled, \( d \).
d i What is the \( y \)-intercept of the line? ii Interpret the significance of the \( y \)-intercept.
e i What is the gradient of the line? ii Interpret the significance of the gradient.
f Find the annual cost of running a car if it is driven 15 000 km. (Answer this question from your equation and from your graph.)
g Find the annual cost of running a car if it is driven 1000 km. (Answer this question from your equation and from your graph.)
h Find the likely number of kilometres driven if the annual costs were $8000.
i Find the likely number of kilometres driven if the annual costs were $16 000.
j Which of your answers to parts f to i do you consider reliable? Why?
10 A market researcher finds that the number of people who would purchase ‘Wise-up’ (the thinking man’s deodorant) is related to its price. He provides the following table of values.

<table>
<thead>
<tr>
<th>Price</th>
<th>Weekly sales (× 1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.40</td>
<td>105</td>
</tr>
<tr>
<td>$1.60</td>
<td>101</td>
</tr>
<tr>
<td>$1.80</td>
<td>97</td>
</tr>
<tr>
<td>$2.00</td>
<td>93</td>
</tr>
<tr>
<td>$2.20</td>
<td>89</td>
</tr>
<tr>
<td>$2.40</td>
<td>85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Price</th>
<th>Weekly sales (× 1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2.60</td>
<td>81</td>
</tr>
<tr>
<td>$2.80</td>
<td>77</td>
</tr>
<tr>
<td>$3.00</td>
<td>73</td>
</tr>
<tr>
<td>$3.20</td>
<td>69</td>
</tr>
<tr>
<td>$3.40</td>
<td>65</td>
</tr>
</tbody>
</table>

a Draw a scatterplot of the data.
b Draw in the line of best fit by eye.
c Find an equation which represents the relationship between the number of cans of ‘Wise-up’ sold, \( N \), and its price, \( p \).
d Use the equation to predict the number of cans sold each week if the price was $3.10.
e Use the equation to predict the number of cans sold each week if the price was $4.60.
f At what price should ‘Wise-up’ be sold if the manufacturers wished to sell 80 000 cans?
g Given that the manufacturers of ‘Wise-up’ can produce only 100 000 cans each week, at what price should it be sold to maximise production?
h Which of your answers to parts d to g do you consider reliable? Why?

11 The following table gives the adult return business class air fares between some Australian cities.

<table>
<thead>
<tr>
<th>City</th>
<th>Distance (km)</th>
<th>Price ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melbourne–Sydney</td>
<td>713</td>
<td>580</td>
</tr>
<tr>
<td>Perth–Melbourne</td>
<td>2728</td>
<td>1490</td>
</tr>
<tr>
<td>Adelaide–Sydney</td>
<td>1172</td>
<td>790</td>
</tr>
<tr>
<td>Brisbane–Melbourne</td>
<td>1370</td>
<td>890</td>
</tr>
<tr>
<td>Hobart–Melbourne</td>
<td>559</td>
<td>520</td>
</tr>
<tr>
<td>Hobart–Adelaide</td>
<td>1144</td>
<td>820</td>
</tr>
<tr>
<td>Adelaide–Melbourne</td>
<td>669</td>
<td>570</td>
</tr>
</tbody>
</table>

a Draw a scatterplot of the data.
b Find the \( r \)-correlation coefficient for the data and interpret it in terms of the variables in the question.
c Suggest reasons for the points on the scatterplot not being perfectly linear.
d Find an equation that represents the relationship between the air fare, \( A \), and the distance travelled, \( d \), using the two-mean method.
e Use the equation to predict the likely air fare from Sydney to the Gold Coast (671 km).
f Use the equation to predict the likely air fare from Perth to Adelaide (2125 km).
g Use the equation to predict the likely air fare from Hobart to Sydney (1024 km).
h Use the equation to predict the likely air fare from Perth to Sydney (3295 km).
i Which of your answers to parts e to h do you consider reliable? Why?

12 You might like to attempt this question using a CAS calculator or set it up on a spreadsheet then prepare the scatterplot using the charting feature of the spreadsheet package.
Rock lobsters (crayfish) are sized according to the length of their carapace (main body shell).

The table below gives the age and carapace length of 15 male rock lobsters.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Length of carapace (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>65</td>
</tr>
<tr>
<td>2.5</td>
<td>59</td>
</tr>
<tr>
<td>4.5</td>
<td>80</td>
</tr>
<tr>
<td>3.25</td>
<td>68</td>
</tr>
<tr>
<td>7.75</td>
<td>130</td>
</tr>
<tr>
<td>8</td>
<td>150</td>
</tr>
<tr>
<td>6.5</td>
<td>112</td>
</tr>
<tr>
<td>12</td>
<td>200</td>
</tr>
</tbody>
</table>

The scatterplot at right has been used as shown to predict the level of $y$ for a given level of $x$. What can be said about the reliability of the result obtained?

A. The result is reliable as the graph is drawn precisely.
B. The result is reliable as the prediction involved interpolation.
C. The result is unreliable as the result involved extrapolation.

The Fisheries Department wishes to set minimum size restrictions so that the rock lobsters have three full years from the time of sexual maturity in which to breed before they can be legally caught. What size should govern the taking of male crayfish?
D The result is unreliable as the correlation between the variables is poor.
E The result is unreliable as it was taken from the scatterplot without using an equation.

14 MC The scatterplot at right has been used as shown to predict the level of $y$ for a given level of $x$. What can be said about the reliability of the result obtained?
A The result is reliable as the graph is drawn precisely.
B The result is reliable as the prediction involved interpolation.
C The result is unreliable as the result involved extrapolation.
D The result is unreliable as the correlation between the variables is poor.
E The result is unreliable as it was taken from the scatterplot without using an equation.

15 MC Which of the following is the best explanation of why extrapolated results can never be considered as reliable?
A Extrapolation is unreliable because it uses too few data points to establish a general rule.
B Extrapolation is unreliable because it assumes a cause and effect relationship between the variables.
C Extrapolation is unreliable because it assumes that the relationship between the variables holds true for untested values.
D Extrapolation is unreliable because it relies upon the accuracy of plotted points.
E Extrapolation is unreliable because it can be used to predict only the level of $y$ from $x$ and not $x$ from $y$.

16 The table below shows world population during the years 1955 to 1985.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>World population (millions)</td>
<td>2757</td>
<td>3037</td>
<td>3354</td>
<td>3696</td>
<td>4066</td>
<td>4432</td>
<td>4828</td>
</tr>
</tbody>
</table>

a Use a CAS calculator to create a scatterplot of the data.
b Does the scatterplot of the Population-versus-Year data appear to be linear?
c Investigate how the CAS calculator can be used to fit a regression line to the data; hence find the equation of the line and graph it on the scatterplot.
d Use the equation to predict the world population in 1950.
e Use the equation to predict the world population in 1982.
f Use the equation to predict the world population in 1997.
g Which of your answers to parts d to f do you consider reliable? Why?
h In fact the world population in 1997 was 5840 million. Account for the discrepancy between this and your answer to part f.
**Scatterplots**

- Bivariate data results from measurements being made on each of two variables for a given set of items. Bivariate data can be represented on a scatterplot.
- The pattern of the scatterplot gives an indication of the strength of the relationship or level of association between the variables. This level of association is called *correlation*.

![Scatterplot examples](image)

- A strong correlation between variables does not imply that one variable *causes* the other to occur.

**The correlation coefficient**

- Correlation can be quantified by using a correlation coefficient.
- The $q$-correlation coefficient can be found by dividing the scatterplot into quadrants by positioning vertical and horizontal lines which represent the positions of the medians.
- The coefficient is determined using the formula:

$$q = \frac{(a + c) - (b + d)}{a + b + c + d}$$

where

- $a$ is the number of data points in $A$
- $b$ is the number of data points in $B$
- $c$ is the number of data points in $C$
- $d$ is the number of data points in $D$.

- The $q$-correlation coefficient may be interpreted as follows:

  - $0.75 \leq q \leq 1$ Strong positive correlation
  - $0.5 \leq q < 0.75$ Moderate positive correlation
  - $0.25 \leq q < 0.5$ Weak positive correlation
  - $-0.25 < q < 0.25$ No correlation
  - $-0.5 < q \leq -0.25$ Weak negative correlation
  - $-0.75 < q \leq -0.5$ Moderate negative correlation
  - $-1 \leq q \leq -0.75$ Strong negative correlation

- The correlation coefficient will always be a number between $-1$ and $1$.
- If $q = 1$ (or $-1$) it does not necessarily indicate that all points are in a perfect straight line.
Linear modelling

- A line of best fit or trend line may be fitted through the points on a scatterplot.
- The equation of the trend line can be found using (a) any two points \((x_1, y_1)\) and \((x_2, y_2)\) that lie on the line and (b) the gradient–point form of the equation of a straight line, 
  \[
  y - y_1 = m(x - x_1),
  \]
  where 
  \[
  m = \frac{y_2 - y_1}{x_2 - x_1}
  \]
- The two points used to find the equation of the trend line are not necessarily actual data points.
- It is best to choose two points that are not too close to each other.
- To find the equation of the regression line using the two-mean method:
  1. rearrange (if necessary) the data points so that the \(x\)-coordinates are in increasing order
  2. divide the data into halves (upper and lower)
  3. calculate the mean value of the \(x\)- and \(y\)-coordinates of each half, to obtain the two points \((\bar{x}_L, \bar{y}_L)\) and \((\bar{x}_U, \bar{y}_U)\)
  4. use the coordinates of the two points to find the equation of the regression line.
- If there is an even number of data points, place an equal number of points in each set. If the number of data points is odd, include the middle point in both sets.
- The \(y\)-intercept of the regression line represents the value of \(y\) when the level of \(x\) is 0.
- The gradient of the regression line represents the rate of change of \(y\) with changing \(x\).
- If the scatterplot shows no correlation between the variables, or indicates a non-linear relationship, it is not appropriate to attempt to fit a linear model through it.

Making predictions

- The equation and the graph of the trend line may be used to make predictions about the level of one variable given the level of the other variable.
- **Interpolation** is the process by which a prediction about the level of either variable is made (either from the trend line of the graph or from its equation) from within the bounds of the original data points.
- **Extrapolation** is the process by which a prediction is made about the level of either variable (either from the trend line of the graph or from its equation) from outside the bounds of the original data points.
- Results predicted from the trend line can be considered reliable only if the scatterplot indicates a reasonably strong correlation between the variables and consists of a reasonably large number of points, and if the predictions were made using interpolation.
1 A researcher administers different amounts of fertiliser to a number of trial plots of potato crop. She then measures the total weight of potatoes harvested from each plot. When graphing the data, the researcher should plot:
   A weight of harvest on the x-axis because it is the independent variable and amount of fertiliser on the y-axis because it is the dependent variable
   B weight of harvest on the y-axis because it is the independent variable and amount of fertiliser on the x-axis because it is the dependent variable
   C weight of harvest on the x-axis because it is the dependent variable and amount of fertiliser on the y-axis because it is the independent variable
   D weight of harvest on the y-axis because it is the dependent variable and amount of fertiliser on the x-axis because it is the independent variable
   E the variables on either axis, as it makes no difference

2 Which of the following graphs best depicts a strong negative correlation between variables?
   A
   B
   C
   D
   E

3 What type of correlation is shown by the graph at right?
   A Strong positive correlation
   B Moderate positive correlation
   C No correlation
   D Moderate negative correlation
   E Strong negative correlation

4 A researcher finds that there is a correlation coefficient of 0.62 between the number of pedestrian crossings in a town and the number of pedestrian accidents. The researcher can conclude that:
   A pedestrian crossings cause pedestrian accidents
   B pedestrian crossings save lives
   C there is evidence to show that the greater the number of pedestrian crossings the greater the number of pedestrian accidents
   D there is evidence to show that pedestrian crossings cause accidents
   E there is evidence to show that the greater the number of pedestrian crossings, the smaller the number of pedestrian accidents

5 A researcher who measures the time taken for production line workers to assemble a component and relates it to the number of weeks that each worker has spent on the production line finds that there is a correlation of −0.82 between the variables. He can conclude that:
   A the greater the number of weeks spent on the production line, the quicker the assembly of components
   B the greater the number of weeks spent on the production line, the slower the assembly of components
   C production line assembly causes worker fatigue
   D many weeks doing the same task causes production line workers to become efficient
   E many weeks doing the same task causes production line workers to become bored and slow as a result

6 Evaluate the r-correlation coefficient for the scatterplot below.

   A −1
   B −0.82
   C −0.69
   D 0.090
   E 0.82
7 Which of the following graphs could possibly be the graph of the equation \( y = 15x - 200? \)

A \[ \begin{array}{c}
\text{A} \\
\text{B} \\
\text{C} \\
\text{D} \\
\text{E}
\end{array} \]

\[ \begin{array}{c}
\text{200} \\
\text{0} \\
\text{0} \\
\text{0} \\
\text{15}
\end{array} \]

8 Find the equation of the line joining the points (100, 40) and (150, 8).

A \( y = -1.56x + 242.37 \)  
B \( y = -1.56x + 226.37 \)  
C \( y = 2.37x - 197 \)  
D \( y = -0.64x - 25.6 \)  
E None of these

9 The weekly costs of a small business are given by the equation \( C = 25x + 300 \) where \( C \) represents the weekly costs of the business when \( x \) items are being produced. What is the \( y \)-intercept of the graph of this equation and what does it represent?

A The \( y \)-intercept is 25. It represents the weekly costs even if no items are produced.

B The \( y \)-intercept is 300. It represents the rate at which costs increase for each extra item produced.

C The \( y \)-intercept is 25. It represents the rate at which costs increase for each extra item produced.

D The \( y \)-intercept is 300. It represents the weekly costs even if no items are produced.

E It is impossible to determine the \( y \)-intercept without seeing the data plotted on a scattergraph.

10 The lengths and diameters (in mm) of a sample of jellyfish selected from a certain location were recorded and displayed in the scatterplot below. The regression line for this data is shown. The equation of the regression line is \( \text{length} = 3.5 + 0.87 \times \text{diameter} \)

\[ \text{Length (mm)} \]

From the equation of the regression line, it can be concluded that for these jellyfish, on average:

A there is a 3.5 mm increase in diameter for each 1 mm increase in length

B there is a 3.5 mm increase in length for each 1 mm increase in diameter

C there is a 0.87 mm increase in diameter for each 1 mm increase in length

D there is a 0.87 mm increase in length for each 1 mm increase in diameter

E there is a 4.37 mm increase in diameter for each 1 mm increase in length

11 The average rainfall and temperature range at several different locations in the South Pacific region are displayed in the scatterplot below. A regression line has been fitted to the data as shown.
The equation of this line is closest to:

A. \( \text{average rainfall} = 210 - 11 \times \text{temperature range} \)
B. \( \text{average rainfall} = 210 + 11 \times \text{temperature range} \)
C. \( \text{average rainfall} = 18 - 0.08 \times \text{temperature range} \)
D. \( \text{average rainfall} = 18 + 0.08 \times \text{temperature range} \)
E. \( \text{average rainfall} = 250 - 13 \times \text{temperature range} \)

12. If a scatterplot contains 11 points, when fitting the regression line using the two-mean method, we should:

A. place 5 points in the lower set and 6 points in the upper set
B. place 6 points in the lower set and 5 points in the upper set
C. place the first 5 points in the lower set, the last 5 points in the upper set and discard the middle point
D. place the first 5 points in the lower set, the last 5 points in the upper set and place the middle point in both sets
E. not proceed, as none of the above methods are correct when dealing with an odd number of data points

13. Which of the following is true of the process of interpolation?

A. Interpolation is always reliable.
B. Interpolation is the process by which predictions can be made of one variable from another within the range of a given set of data.
C. Interpolation is rarely reliable as it assumes a relationship holds true for untested values.
D. Interpolation is an entirely graphical process and as such has limited accuracy.
E. All of the above.

14. The scatterplot at right has been used as shown to predict the level of \( y \) for a given level of \( x \). What can be said about the reliability of the result obtained?

A. The result is unreliable, as there are too few data points to assume that a general relationship holds.
B. The result is unreliable as the prediction involved interpolation.
C. The result is unreliable as the result involved extrapolation.

15. Use the scatterplot below to predict the level of \( x \) when \( y = 10 \).

A. \( x = 14 \)  
B. \( x = 3 \)  
C. \( x = 8 \)  
D. \( x = 9 \)  
E. None of these

16. The value of an asset which is depreciating year by year is given by the formula \( V = -200a + 4200 \) where \( V \) is the value of the asset after \( a \) years. Predict the amount of time taken before the item is worth half its initial value.

A. 4100 years  
B. 21 years  
C. 10.5 years  
D. 2100 years  
E. None of these

17. Eighteen students sat for a 15-question multiple-choice test. In the scatterplot below, the number of errors made by each student on the test is plotted against the time they reported studying for the test. A regression line has been determined for this data and is also displayed on the scatterplot.

The equation for the regression line is:

\( \text{number of errors} = 8.8 - 0.120 \times \text{study time} \)

Using the regression line, it can be estimated that, on average, a student reporting a study time of 35 minutes would make:

- A 4.3 errors
- B 4.6 errors
- C 4.8 errors
- D 5.0 errors
- E 13.0 errors

**SHORT ANSWER**

1. Prepare a scatterplot for the following data.
   Without attempting any calculation, state the type of correlation that you think is shown.

   \[ \begin{array}{c|c|c|c|c|c|c|c|c|c|c|c} x & 2 & 4 & 18 & 7 & 9 & 12 & 2 & 7 & 11 & 10 & 16 \\ y & 103 & 75 & 20 & 66 & 70 & 95 & 40 & 27 & 42 & 30 & 2 \end{array} \]

2. Find the correlation coefficient in each of the following cases.

   \( a \) \hspace{1cm} \( b \) \hspace{1cm} \( c \)

3. An experiment which tested the strength of wooden beams of different thickness demonstrated a correlation of 0.9 between the variables. What does this show?

4. A survey in which people were asked to state their age and the age of their car revealed that there was a correlation coefficient of \( -0.65 \) between the variables. What does this show?

5. Find the equation of the line which joins the points:
   - a (3, 9) and (7, 14)
   - b (8, 13) and (12, 2)
   - c (12.5, 258) and (14.8, 307).

6. Find the equation of the line of best fit by eye in each of the following examples.

   \( a \)

7. Cars depreciate in value over time. The table below gives the average value of a car (of the same brand and model) at different ages.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Value (dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18 100</td>
</tr>
<tr>
<td>2</td>
<td>15 050</td>
</tr>
<tr>
<td>3</td>
<td>13 900</td>
</tr>
<tr>
<td>4</td>
<td>11 900</td>
</tr>
<tr>
<td>5</td>
<td>10 400</td>
</tr>
<tr>
<td>6</td>
<td>9 600</td>
</tr>
<tr>
<td>7</td>
<td>8 900</td>
</tr>
<tr>
<td>8</td>
<td>8 500</td>
</tr>
<tr>
<td>9</td>
<td>8 400</td>
</tr>
</tbody>
</table>

   - a The data are to be used to build a mathematical relationship that will enable the average value of this brand and model of car to be predicted from its age. What is the dependent variable?
   - A scatterplot is constructed from the data and a regression line is fitted as shown.

   \( b \) Using the line shown in the scatterplot, or otherwise, determine the equation of the regression line. Write the coefficients correct to the nearest hundred.

   \( \text{EXAM TIP} \) Many students reversed the numbers. These responses scored a maximum of 1 mark so long as the negative sign was present in \(-1200\).
8 For each of the following data sets, find the equation of the regression line, using the two-mean method.

<table>
<thead>
<tr>
<th>a</th>
<th>x</th>
<th>70</th>
<th>40</th>
<th>90</th>
<th>50</th>
<th>30</th>
<th>80</th>
<th>60</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>y</td>
<td>33</td>
<td>46</td>
<td>26</td>
<td>41</td>
<td>49</td>
<td>31</td>
<td>39</td>
<td>53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>x</th>
<th>55</th>
<th>56</th>
<th>57</th>
<th>58</th>
<th>59</th>
<th>60</th>
<th>61</th>
<th>63</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>y</td>
<td>12</td>
<td>24</td>
<td>35</td>
<td>48</td>
<td>62</td>
<td>70</td>
<td>85</td>
<td>99</td>
</tr>
</tbody>
</table>

9 The equation of the line of best fit for certain data is found to be $C = 3.6d + 250$.

a Find the level of $d$ when $C = 430$.

b Find the level of $C$ when $d = 40$.

10 For the graph shown below, use extrapolation to predict:

a the level of $y$ when $x = 10$

b the level of $x$ when $y = 900$.

11 Over recent years, the salaries of Patagonian cricketers have increased rapidly. The following table gives the average salaries in dollars of a large group of these cricketers over the period 1991–2000.

<table>
<thead>
<tr>
<th>(Year)</th>
<th>Average salary ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>45 000</td>
</tr>
<tr>
<td>1992</td>
<td>47 000</td>
</tr>
<tr>
<td>1993</td>
<td>50 000</td>
</tr>
<tr>
<td>1994</td>
<td>58 000</td>
</tr>
<tr>
<td>1995</td>
<td>70 000</td>
</tr>
<tr>
<td>1996</td>
<td>78 000</td>
</tr>
<tr>
<td>1997</td>
<td>93 000</td>
</tr>
<tr>
<td>1998</td>
<td>105 000</td>
</tr>
<tr>
<td>1999</td>
<td>126 000</td>
</tr>
<tr>
<td>2000</td>
<td>142 000</td>
</tr>
</tbody>
</table>

a This data will be used to predict future average salaries of Patagonian cricketers. In this analysis, what is the independent variable?

EXAM TIP: ‘Time’ was accepted here but most students obtained the correct answer.

To begin the analysis, the years have been rescaled as $x = 1$ to $x = 10$ (1991 = 1, 1992 = 2, and so on), and average salary rescaled in thousands of dollars as the variable $y$.

<table>
<thead>
<tr>
<th>Year</th>
<th>Average salary ($1000s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>78</td>
</tr>
<tr>
<td>7</td>
<td>93</td>
</tr>
<tr>
<td>8</td>
<td>105</td>
</tr>
<tr>
<td>9</td>
<td>126</td>
</tr>
<tr>
<td>10</td>
<td>142</td>
</tr>
</tbody>
</table>

b Using the regression equation $y = 20.9 + 10.99x$ to model the increase in these cricketers’ average salaries:

i find the average salary increase for the cricketers per year.

EXAM TIP: Many students found an answer of 10.99, not realizing this indicated thousands of dollars.

ii predict their average salary in the year 2005.
The table below displays the mean surface temperature (in °C) and the mean duration of warm spell (in days) in Australia for 13 years selected at random from the period 1960 to 2005.

<table>
<thead>
<tr>
<th>Mean surface temperature (°C)</th>
<th>Mean duration of warm spell (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.2</td>
<td>21.4</td>
</tr>
<tr>
<td>13.3</td>
<td>16.3</td>
</tr>
<tr>
<td>13.3</td>
<td>27.6</td>
</tr>
<tr>
<td>13.4</td>
<td>32.6</td>
</tr>
<tr>
<td>13.4</td>
<td>28.7</td>
</tr>
<tr>
<td>13.5</td>
<td>30.9</td>
</tr>
<tr>
<td>13.5</td>
<td>45.9</td>
</tr>
<tr>
<td>13.5</td>
<td>35.5</td>
</tr>
<tr>
<td>13.6</td>
<td>40.6</td>
</tr>
<tr>
<td>13.7</td>
<td>42.8</td>
</tr>
<tr>
<td>13.7</td>
<td>49.9</td>
</tr>
<tr>
<td>13.7</td>
<td>55.8</td>
</tr>
<tr>
<td><strong>13.8</strong></td>
<td><strong>53.1</strong></td>
</tr>
</tbody>
</table>

This data set has been used to construct the scatterplot below. The scatterplot is incomplete.

- Complete the scatterplot at right by plotting the data values given in bold in the table above. Mark the point with a cross (×).

The mean surface temperature (in °C) of Australia for the period 1960 to 2005 is displayed in the time series plot below.

**EXAM TIP** This point had to be on the 13.8 coordinate line, with the value for the duration of the warm spell between 53 and 54. [Assessment report 2007]
b In what year was the lowest mean surface temperature recorded?
The regression line is fit to the time series plot.
c The equation of this regression line is found to be:
\[
\text{mean surface temperature} = -12.361 + 0.013 \times \text{year}
\]
i Use the regression line to predict the mean surface temperature (in °C) for 2010. Write your answer correct to 2 decimal places.
ii By how many degrees does the regression line predict Australia’s mean surface temperature will rise each year? Write your answer correct to 3 decimal places.

2 The following table gives the percentage of married women who use contraception and the birth rate (births per thousand population) for 15 countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>% Married women using contraception</th>
<th>Births per thousand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>76</td>
<td>14</td>
</tr>
<tr>
<td>Germany</td>
<td>75</td>
<td>10</td>
</tr>
<tr>
<td>India</td>
<td>41</td>
<td>29</td>
</tr>
<tr>
<td>Slovenia</td>
<td>92</td>
<td>10</td>
</tr>
<tr>
<td>China</td>
<td>83</td>
<td>17</td>
</tr>
<tr>
<td>Bolivia</td>
<td>45</td>
<td>36</td>
</tr>
<tr>
<td>Philippines</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>South Africa</td>
<td>53</td>
<td>27</td>
</tr>
<tr>
<td>Thailand</td>
<td>66</td>
<td>18</td>
</tr>
<tr>
<td>Iraq</td>
<td>18</td>
<td>38</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>72</td>
<td>13</td>
</tr>
<tr>
<td>United States</td>
<td>71</td>
<td>15</td>
</tr>
<tr>
<td>Mexico</td>
<td>65</td>
<td>27</td>
</tr>
<tr>
<td>Zambia</td>
<td>26</td>
<td>45</td>
</tr>
<tr>
<td>D. R. Congo</td>
<td>8</td>
<td>43</td>
</tr>
</tbody>
</table>

a Graph the data on a scatterplot.
b Find the \(q\)-correlation coefficient for the data and interpret it in terms of the variables in the question.
c Draw in the line of best fit by eye.
d Suggest reasons for the points on the scatterplot not being perfectly linear.
e Find an equation which represents the relationship between a country’s birth rate, \(B\), and the percentage, \(w\), of married women who are using contraception.
f Use the equation to find the birth rate if it is known that 25% of married women use contraception. Check your answer using the graph.
g Use the equation to find the birth rate if it is known that 38% of married women use contraception. Check your answer by using the graph.
h Can you be confident that your answers to parts f and g are reliable? Why (not)?
An entomologist conducted an experiment in which small amounts of insecticide were introduced into a container of 100 blowflies. The results are detailed below.

<table>
<thead>
<tr>
<th>Insecticide (micrograms)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of flies remaining after 2 hours</td>
<td>99</td>
<td>92</td>
<td>81</td>
<td>74</td>
<td>62</td>
<td>68</td>
<td>52</td>
<td>45</td>
<td>38</td>
<td>24</td>
</tr>
</tbody>
</table>

a Graph the data on a scatterplot.
b Find the $q$-correlation coefficient for the data and interpret it in terms of the variables in the question.
c Draw in the line of best fit by eye.
d Suggest reasons for the points on the scatterplot not being perfectly linear.
e Find an equation which represents the relationship between the number of survivors, $S$, and the amount of insecticide, $I$.
f Explain the significance of the $y$-intercept and gradient of the equation.
g Use the equation to find the number of survivors if 8.5 micrograms of insecticide is used. Check your answer using the graph.
h Use the equation to find the number of survivors if 12.5 micrograms of insecticide is used.
i How much insecticide would have to be used before no flies survived?
j Can you be confident that your answers to parts g, h and i are reliable? Why (not)?

Rachel decided to attend an intensive Russian language course before travelling to Moscow and St Petersburg early next year. The table below shows her vocabulary size, that is, the total number of Russian words that she has learned over the first two weeks of study.

<table>
<thead>
<tr>
<th>Number of days from the beginning of the intensive course</th>
<th>Total number of words learned from the beginning of the course</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
</tr>
<tr>
<td>4</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>127</td>
</tr>
<tr>
<td>6</td>
<td>149</td>
</tr>
<tr>
<td>7</td>
<td>175</td>
</tr>
<tr>
<td>8</td>
<td>196</td>
</tr>
<tr>
<td>9</td>
<td>223</td>
</tr>
<tr>
<td>10</td>
<td>248</td>
</tr>
<tr>
<td>11</td>
<td>280</td>
</tr>
<tr>
<td>12</td>
<td>300</td>
</tr>
<tr>
<td>13</td>
<td>321</td>
</tr>
<tr>
<td>14</td>
<td>350</td>
</tr>
</tbody>
</table>

a Which of the variables is independent and which is dependent?
b Represent the data on a scatterplot.
c State the type of correlation indicated by the scatterplot.
d Find an equation that represents the relationship between the number of days of studying the language, $n$, and the vocabulary size, $V$, using the two-mean method.
e Explain the meaning of the gradient of the regression line.
f Use the equation of the regression line to predict the size (to the nearest number) of Rachel’s vocabulary by the end of the third week (that is, by day 21).
g According to the linear model established in d, how long will it take Rachel to learn 500 Russian words?
h Explain whether or not your answer to g is reliable.
5 An investigation was made into the relationship between the number of working hours per week and the average weekly amount of money (in $) spent on junk and takeaway food by a group of twelve women.

<table>
<thead>
<tr>
<th>Time at work (h)</th>
<th>20</th>
<th>33</th>
<th>36</th>
<th>38</th>
<th>34</th>
<th>27</th>
<th>35</th>
<th>40</th>
<th>22</th>
<th>37</th>
<th>30</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount spent ($)</td>
<td>10</td>
<td>25</td>
<td>30</td>
<td>36</td>
<td>25</td>
<td>19</td>
<td>28</td>
<td>37</td>
<td>12</td>
<td>35</td>
<td>22</td>
<td>16</td>
</tr>
</tbody>
</table>

a Which of the variables is independent and which is dependent?
b Represent the data on a scatterplot.
c What type of correlation between the two variables does the scatterplot suggest?
d Find an equation that represents the relationship between the number of working hours per week, \( n \), and the average weekly amount of money spent on junk and take-away food, \( A \), using the two-mean method.
e Explain the meaning of the gradient of the regression line.
f If Maya worked 29 hours in one week, how much money (according to the equation obtained in d) would she spend on junk and take-away food that week?
g Did you use extrapolation, or interpolation to answer the previous question?
h Explain whether or not your answer to f is reliable.

6 Michael, a bricklayer, likes to keep records of his jobs. He always keeps track of the amount of time it takes him to complete any particular job and the amount he is paid. He also records the area (in square metres) of the walls he needs to build and the total number of bricks (including waste) used for the job. An extract from his records is shown below.

<table>
<thead>
<tr>
<th>Area (m²)</th>
<th>12</th>
<th>17</th>
<th>24</th>
<th>10</th>
<th>23</th>
<th>20</th>
<th>14</th>
<th>26</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bricks</td>
<td>686</td>
<td>965</td>
<td>1357</td>
<td>572</td>
<td>1323</td>
<td>1150</td>
<td>800</td>
<td>1502</td>
<td>863</td>
</tr>
</tbody>
</table>

a Construct a scatterplot for the data shown in the table.
b Describe the relationship as seen from the scatterplot.
c Calculate the \( r \)-correlation coefficient for the data. Does this value match your answer to the previous question? Explain.
d Use the two-mean method to establish the rule connecting the area of the brick wall, \( a \), that needs to be built and the number of bricks, \( N \), required to complete the job.
e Complete the following statement by filling in the correct number in the space provided: ‘According to the equation of the regression line, …… bricks are required to construct each extra square metre of the brick wall.’
f Michael is about to start working on a new project. According to specifications, 18 m² of bricks need to be laid and he has 1000 bricks at his disposal. Using the rule obtained in d, check whether the available number of bricks is enough to complete the job.
g Complete the following statement by selecting one of the two words offered in brackets: ‘To predict the amount of bricks needed to build a wall with the total area of 18 m², ……… can be used. (extrapolation/interpolation).’
h Explain why (or why not) your answer to f could be considered reliable.
Chapter opener

**Digital doc**
- 10 Quick Questions: Warm up with ten quick questions on bivariate data. *(page 163)*

**4A Scatterplots**

**Digital docs**
- Spreadsheet 140: Investigate two-variable statistics. *(page 169)*
- Career profile: Learn about how Roger Farrer, a data manager, uses mathematics. *(page 171)*

**4B The correlation coefficient**

**Tutorial**
- **WE3** int-0869: Watch how to draw a scatterplot, find the \( r \)-correlation and draw a suitable conclusion about the data. *(page 173)*

**Digital docs**
- SkillSHEET 4.1: Practise finding the median. *(page 175)*
- Spreadsheet 011: Investigate \( r \)-correlation. *(page 176)*
- WorkSHEET 4.1: Plot scatterplots, identify independent and dependent variables and calculate \( r \)-correlation. *(page 177)*

**4C Linear modelling**

**Interactivity**
- Transforming data int-0184: Consolidate your understanding of transforming bivariate data. *(page 178)*

**Tutorials**
- **WE5** int-0870: Watch how to draw a scatterplot, fit a trend line by eye and find its equation and explain the meaning of the \( y \)-intercept and gradient. *(page 180)*
- **WE7** int-0871: Watch how to use the two-mean method to find the regression line of two sets of bivariate data. *(page 183)*

**4D Making predictions**

**Tutorials**
- **WE8** int-0872: Watch how to make predictions about the value of variables by using the regression equation. *(page 189)*
- **WE9** int-0873: Watch how to make predictions about variables by using a linear graph. *(page 190)*

**Digital doc**
- Spreadsheet 071: Investigate making predictions about data. *(page 194)*

**Chapter review**

**Digital doc**
- Test Yourself Chapter 4: Take the end-of-chapter test to test your progress. *(page 208)*

To access eBookPLUS activities, log on to [www.jacplus.com.au](http://www.jacplus.com.au)
MUL TI P LE C HOI CE

15 minutes

Each question is worth 1 mark.

The following information relates to questions 1, 2 and 3.

Thirty-five Year 11 students were surveyed about the time they spent using their mobile phone each day for either text messaging or phone calls. The results were noted for male and female students and are recorded in the table below.

<table>
<thead>
<tr>
<th>PHONE USAGE</th>
<th>GENDER</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MALE</td>
<td>FEMALE</td>
</tr>
<tr>
<td>Less than 10 minutes</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Between 10 to 20 minutes</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>More than 20 minutes</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>TOTAL</td>
<td>18</td>
<td>17</td>
</tr>
</tbody>
</table>

1 The number of females who used their mobile phones for 20 minutes or less was:
A 4  B 6  C 7  D 8  E 10

2 The data represented in the table is best described as:
A numerical and continuous data
B numerical and discrete data
C categorical and nominal data
D categorical and ordinal data
E counted and ordered data

3 Which one of the following is incorrect?
A 25% of all students surveyed used their mobile phones less than 20 minutes each day.
B 29% of all students surveyed used their mobile phones more than 20 minutes each day.
C 34% of all students surveyed used their mobile phones less than 10 minutes each day.
D 56% of the males surveyed used their mobile phones more than 10 minutes each day.
E 76% of the females surveyed used their mobile phones more than 10 minutes each day.

4 Each month Ming pays $20 as a flat fee and then 25 cents for each phone call she makes on her mobile phone. If during one month Ming makes a total of 15 calls on her mobile phone then her monthly phone bill, in dollars, will be:
A 20.25  B 23.75  C 28.00
D 325     E 395

5 Seth changes his mobile phone plan to EasyDial. He is charged a flat fee of $45 plus 28 cents per call. If \( C \) represents the monthly mobile phone cost, in dollars, and \( n \) represents the number of monthly calls made. Which one of the following equations determines the Seth’s monthly charges?
A \( C = 45n + 28 \)
B \( C = 28n + 45 \)
C \( C = 0.45n + 28 \)
D \( C = 0.28n + 45 \)
E \( C = 45.28n \)

6 The cost of Stephen’s monthly phone plan is \( C_s = 0.27n + 50 \), where \( n \) is the number of calls per month and \( C \) is the cost per month. Which one of the following graphs represents Stephen’s monthly phone plan?

A \[50\]  
B \[50\]  
C \[50\]  
D \[50\]  
E \[0.27\]
In the Year 11 Mathematics class, 10 students who use a mobile phone were randomly selected and asked to record on a particular day the number of calls they made and the number of SMS text messages they sent. Their responses are recorded in Table 1 below.

Table 1

<table>
<thead>
<tr>
<th>STUDENT</th>
<th>SMS TEXT MESSAGES, s</th>
<th>PHONE CALLS, p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>H</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>I</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>J</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>

a) For the number of SMS text messages:
   i) Determine the mean and the standard deviation of the data. Express your answers correct to 1 decimal place.
   ii) What percentage of the data lies between 3 and 9?

b) The scatterplot in Figure A below has been constructed from the data in Table 1. On Figure A, mark with an X the point for Student J.

![Figure A](image)

The coordinates for \((x_u, y_u)\) are \((8.8, 2.6)\); determine the coordinates of \((x_L, y_L)\).

d) Using the two-mean method, determine the equation of the regression line in terms of the number of phone calls, \(p\), and the number of SMS text messages, \(s\). Express values correct to 2 decimal places.

e) The median for the number of SMS text messages is 6. Show this as a vertical line on Figure A.
   ii) Determine the median of the number of phone calls. Clearly show this as a horizontal line on Figure A.
   iii) Determine the \(q\)-correlation.
   iv) EasyDial is planning a new campaign encouraging students to use their mobile phones for both SMS text messages and phone calls. Using your results from part e, should EasyDial spent money on this new campaign?