5 Plan the Investigation

A Describing Variables and Managing Sources of Variation

As part of your plan for your investigation you will have sourced data and identified the variables of interest. You should also state the source of your data in your investigative report and consider the sources of variation that exist as part of the collection of the data. You may not have collected the data yourself and so you will need to consider the possible sources of variation that may have occurred in the original collection process and how they may have been managed.

Examples of possible sources of variation could be : the weather, the amount of assistance given, the instructions provided, the equipment that is used, and the ability of the person taking the measurements.

Sources of Variation Categories

Sources of variation can be considered in five different categories :

1 Natural variation : this is the differences that occur due to nature, such as the variation in weights or head circumference between different people.

2 Occasion to occasion variation : this type of variation occurs when repeated measurements are taken over time. For example, a person's blood sugar levels will not be the same at different times of day.

3 Measurement variation : when taking a measurement, variation can occur in the method. For example, when measuring a height, it would be important to make sure the person being measured removes their shoes as this could add a different amount to each height due to the different sized soles and heels on shoes worn by different people.

4 Induced variation : different factors may affect the same measurement. For example, the price of fuel in two different petrol stations in different areas of the country may be affected by their location relative to the nearest port. They could also be part of different companies.

5 Sampling variation : this always occurs when a sample is taken. Unless the entire population of interest is measured, sampling variation will be present. This means that each time a different sample is taken, variation will occur in the individuals present in the sample and therefore the sample statistics will vary between samples.

Example :

For the investigative question : *'What is the relationship between the hand spans and heights of students' at Goldbury High.*

a) Describe the variables and how they will be measured.

b) Describe possible sources of variation in the collection of the data and what actions (if any) could be taken to ensure the data is consistent.

c) Where possible describe what would happen if these actions were not taken.

Possible Answers :

a) *The two variables are the student's hand span and the height. Hand spans (in cm) will be measured by getting people to fully stretch their hand along a ruler and measuring from the outside of the little finger to the outside of the thumb. Heights (in cm) will be measured by standing students against a wall where a tape measure has been stuck.*

b) *When measuring heights all students will be asked to take their shoes off and stand upright against the wall. The height will be measured by placing a book horizontally on the person's head. When measuring hand spans people need to stretch their hand as much as possible. Measurements need to be taken from the same part from the finger to the same part of the thumb each time. These steps will all help minimise* measurement variation.

c) *If students do not stretch their hand out fully then that could result in the measurements taken being less than they should be. Inconsistency in the postion of the finger or thumb measured from could lead to results being under or over what they should be. Keeping shoes off could make the students' height measurement too great, if they do not stand up straight, the measurement could be less than it should be. Inconsistency in the placing of the book could lead to measurements being greater or less than what they should be.*

1 For each of these investigations state how the variables would have to be measured. Discuss possible sources of variation and actions that would have to be taken to ensure the data is consistent. What would happen if these sources of variation were not controlled?

a) *'What is the relationship between the recovery time and the distance a student runs?'*

i) Describe the variables and how they will be measured.

.....

.....

.....

ii) Discuss possible sources of variation in the measurements and actions that could be taken to ensure the data is consistent. What would happen if these sources of variation were not controlled?

.....

.....

.....

.....

9 Data Display 2

A Think About the Scales

1 This set of bivariate data shows the shoe sizes and heights of 25 senior students in a school.

Data :

Shoe size	10.5	7	6	7.5	7	12.5	7	14	6.5	8	9.5	13	6.5
Height (cm)	175	157	157	157	152	180	160	183	152	168	175	188	157

Shoe size	8	10	10	7.5	12	12	6	9.5	11	8.5	11	7.5
Height (cm)	163	178	173	168	193	178	147	175	183	178	185	165

a) Give a reason why the variable *'shoe size'* should go on the **x**-axis of this scatter diagram.

.....

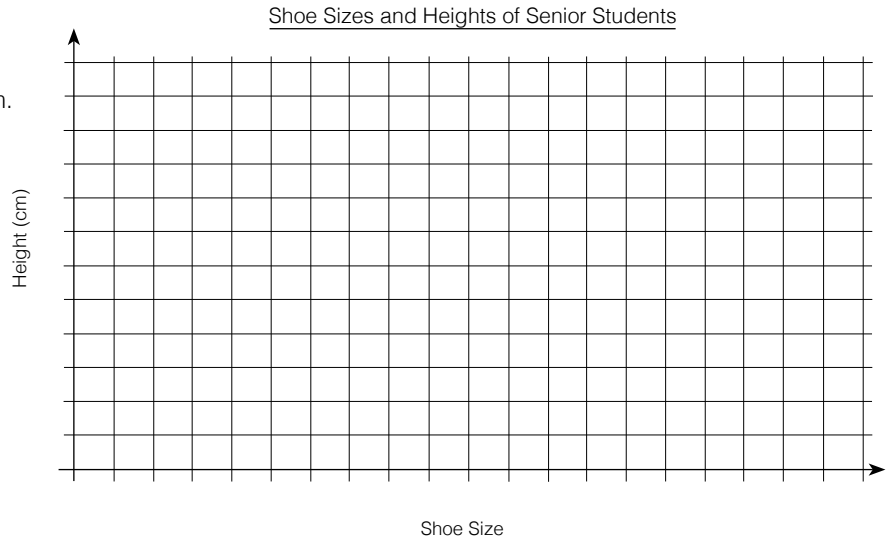
.....

.....

.....

.....

b) Select an appropriate scale for the axes and plot the scatter diagram.



2. The heat setting on the element of a stove is a continuous number between 0 and 6, with 6 being the hottest. A litre of tap water is put in a pot and the time it takes to boil is measured. This is done 20 times with a range of heat settings. The resulting bivariate data is shown in this table :

Data :

Setting on Element	4.5	5.0	4.0	5.5	6.0	2.0	3.0	3.0	3.5	1.5
Time to Boil (minutes)	10	4	15	5	3	18	22	15	11	23

Setting on Element	5.5	2.5	6.0	4.5	4.0	2.5	2.0	5.0	3.5	1.5
Time to Boil (minutes)	8	17	4	12	10	13	20	8	16	21

a) Give a reason why the variable *'heat setting on element'* should go on the **x**-axis of this scatter diagram.

.....

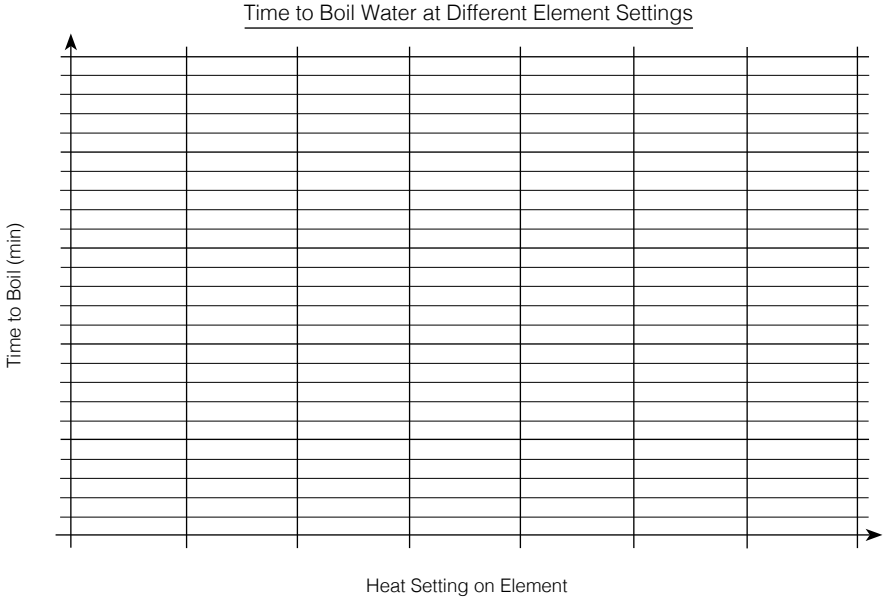
.....

.....

.....

.....

b) Select an appropriate scale for the axes and plot the scatter diagram.

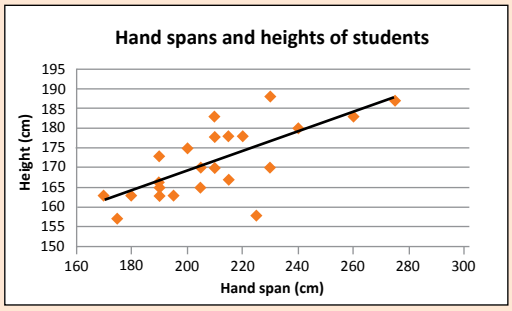


A Fitting Trend Lines Using Excel

On page 11 we discussed how a spreadsheet could be used to draw your scatter plots. If you want to put a trend line on a graph you have drawn using a spreadsheet then click on a data point and select the *add trend line* option.

You are then given a number of trend line options. If you think your relationship is a linear one select *linear*; if you think the relationship is non-linear then you could investigate the other options but you must check that any line you select is a good visual fit.

A trend line can also easily be fitted if you are using **Google Sheets**, **NZ Grapher** or **CODAP**.



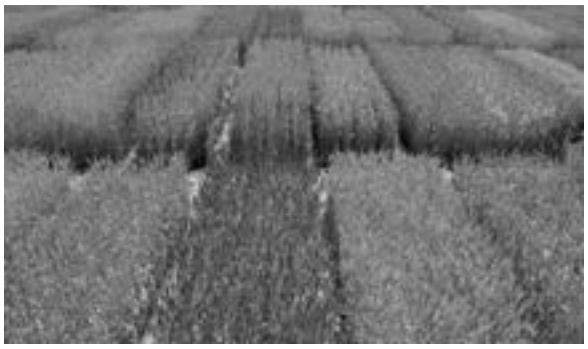
1 Use computer software to plot scatter diagrams with trend lines for the following sets of bivariate data.

a) Data : Crop yields for different applications of fertilizer.

Fertiliser (kg/ha)	10	10	10	20	20	20	30	30	30	40	40	40
Yield (kg/ha)	1250	1450	890	1760	1540	1230	1550	1686	2088	2100	2310	1900

Fertiliser (kg/ha)	50	50	50	60	60	60	70	70	70	80	80	80
Yield (kg/ha)	2245	2008	2587	2870	3120	2046	3065	3415	2887	3458	2889	2316

Paste a print-out of your computer generated scatter diagram here.



b) Data : The time it takes different concentrations of a drug to take effect..

Concentration (g/L)	20	5	54	41	229	9	46	13	25	19
time (minutes)	12	25	5	11	16	23	12	17	12	17

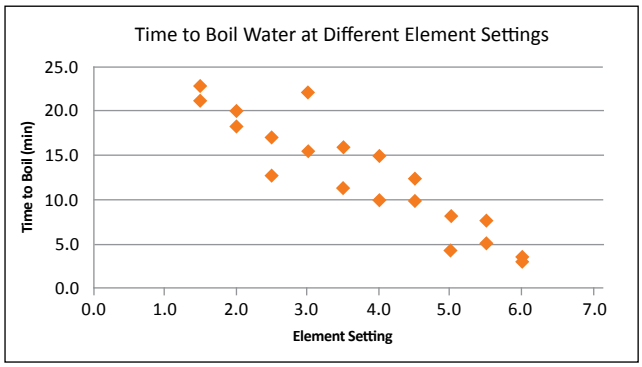
Concentration (g/L)	45	30	13	49	10	23	30	9	37	32
time (minutes)	7	10	22	8	10	15	115	18	9	10

Paste a print-out of your computer generated scatter diagram here.

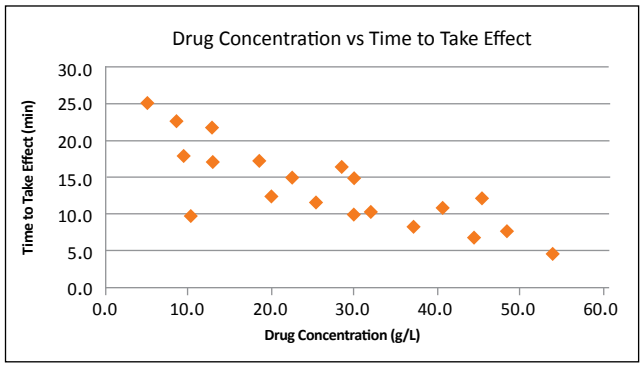
19 Analysing Data 2 - continued

A Features of your Scatter Plot in Context - continued

c)



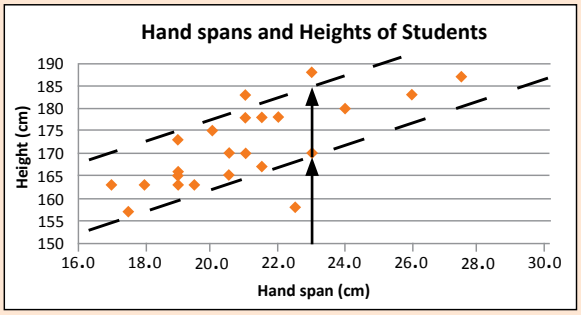
d)

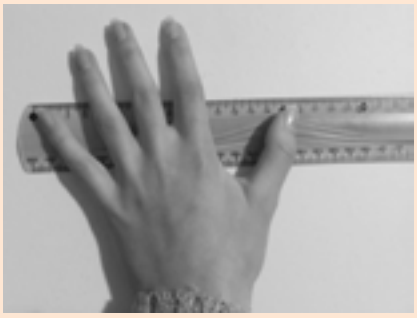


A Making Predictions

As part of the investigation for this standard you must make a prediction using the relationship that you have found. In order to do this, you can add rough "train tracks" as in the diagram below. The train tracks should contain the majority of the data between them and head in the same direction as the trend line you added previously. You should discuss how sensible your prediction is.

Example 1 : *'What is the relationship between the hand spans and heights of students?'*





Prediction

An estimate for the height of a student with a hand span of 23 cm would be somewhere roughly between 170cm and 185cm. Because there is a reasonable amount of scatter about the trend line I would treat this as a rough estimate only.

1 Use the plots in the previous exercise (answering the relationship question) to answer the following questions. Discuss how sensible your prediction is.

a) Estimate the recovery time for the student if he runs 90 m. (See graph on page 21, Exercise 1a)

……

……

……

……

……

……

……

……

……

……

b) Estimate the height of a senior student with a shoe size of nine. (See graph on page 21, Exercise 1b)

……

……

……

……

……

……

……

……

……

……

c) Estimate the time it would take to boil the pot of water if the element setting was 3.3. (see graph on page 22, Exercise 1c)

……

……

……

……

……

……

……

……

Practice Investigation 2 - continued

A Test Investigation - 'Leaves' - continued

Data and Graphs :

.....

.....

.....

.....

.....

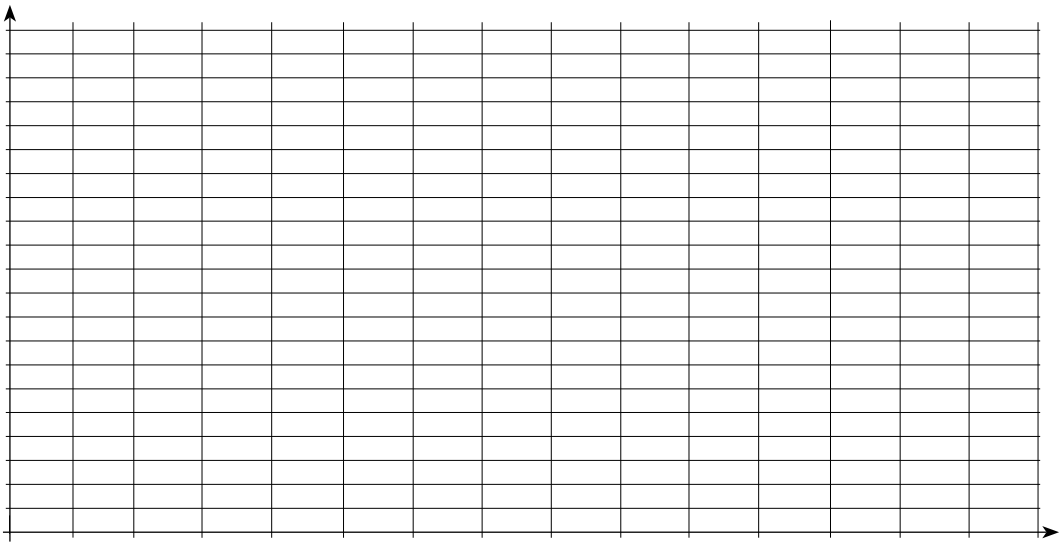
.....

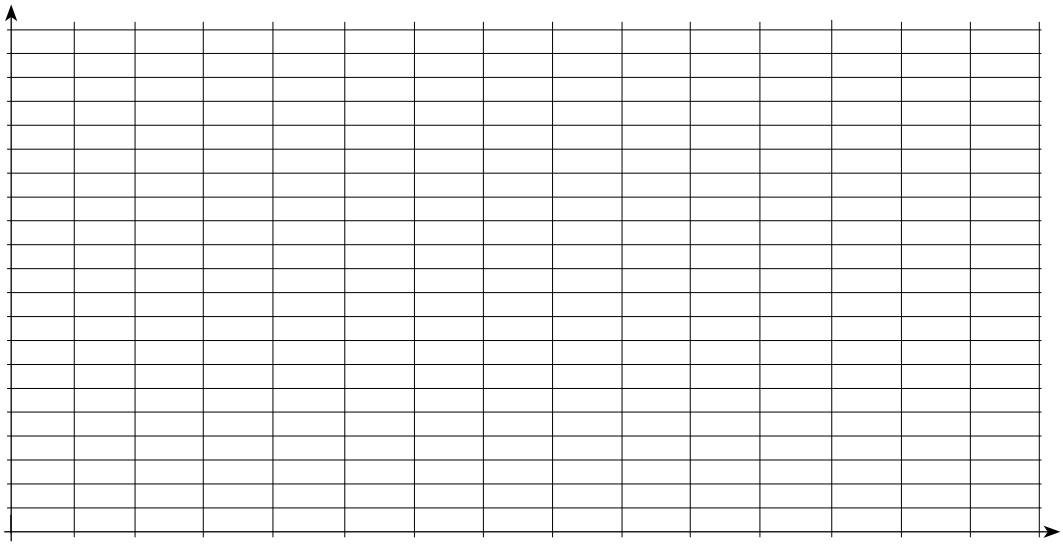
.....

.....

.....

.....





Pages 21-23 - Form a Conclusion 1 - continued

A1 c) Example answer :
Pots on higher element settings **tend** to take shorter times to boil. This can be seen by the negative gradient of the trend time. You would expect this result as higher element settings produce more heat and this reduces the time for the pot to boil.

d) Example answer :
Higher drug concentrations **tend** to take shorter times to take effect. This can be seen by the negative gradient of the trend line. While you need to know a lot about the drug to be sure, you would expect using a higher concentration would result in a faster reaction.

e) Example answer :
As points on the plot appear to be random there is no relationship between the time to run 200 m and time spent on HW. There is no reason why people who run faster would spend more time on HW and so this result is reasonable.

f) Example answer :
Longer trout **tend** to weigh more. The upward curve in the trend line shows this. You would expect this because as lengths increase, other body proportions (and hence weight) would also increase (the weight/ length relationship is likely to be cubic).

g) Example answer :
Higher yields **tend** to result with higher fertilizer applications. This can be seen by the positive gradient of the trend line. You would expect this but there would be a limit, too much fertilizer could damage the crop. This might be why the points appear to be levelling out.

Pages 24-25 - Form a Conclusion 2

Answers for pages 24-25 are examples only :

A1 a) Roughly between 175 and 265 secs. There is more variation in the data for distances around 90 m than for shorter distances so I would be less confident in this prediction than, for example, one at 30 m.

b) Roughly between 162 and 175 cm. The scatter about the trend line is moderate but fairly constant so I think the estimate is reasonable, but not highly accurate.

c) Roughly between 12 and 17 minutes. The scatter about the trend line is moderate but fairly constant so I think the estimate is reasonable, but not highly accurate.

d) The estimate will depend on the selected value. While there is some scatter the variation is relatively constant and so it is reasonable to make estimates.

e) It is not sensible to make an estimate. As there is no relationship you should not use the analysis to make estimates.

f) The estimate will depend on the selected value. Any estimate should be reasonably accurate because the relationship is a strong one.

B1 a) i) roughly between 1700 and 2900 kg/ha
ii) roughly between 2700 and 4000 kg/ha
iii) roughly between 3700 and 4700 kg/ha

b) The first prediction (for 45 kg/ha) is likely to be the most accurate. It lies within the given data values and is at a point where there is not much scatter about the trend line.
The second prediction is likely to be less accurate than for a) because it just lies outside the data range and there is more variation in the yields for higher application rates.
I would not be very confident in d). 110 kg/ha is well outside the data range and it appears that crop yields could be levelling out for high application rates.

Pages 28-29 - Practice Investigations 1

Answers for pages 28-29 are examples only :

A2 Plan :

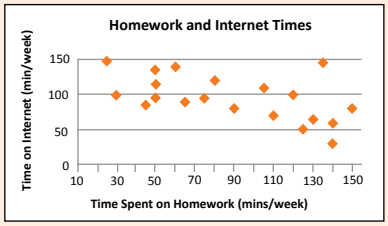
a) The two variables are the homework time (minutes / wk) and the free time spent on the internet (minutes / wk).

b) Homework times would be measured by asking students to record the time (in minutes) they spent on homework. For each night of a school week they would do this by noting the time on a clock that they start and finish their homework. Internet times would be recorded the same way.

c) Students may not be able to remember times so a daily recording sheet for how much time they spent on homework (start time, finish time, no breaks) would be used. The recording sheet would also be used to record start and finish times whenever the student went on the internet.

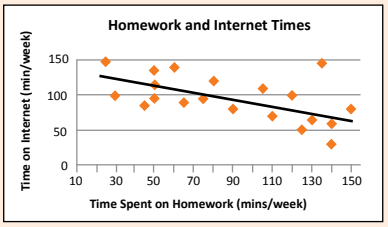
d) There would be a separate recording sheet for each night of the school week. Data from these sheets would be collated by calculating the daily times and adding them to get a total.

A4 Analysis :



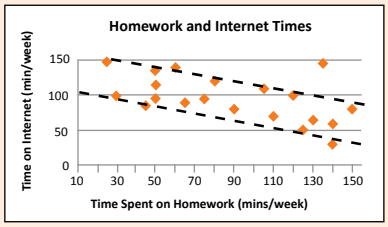
There is a **negative linear** relationship between homework and internet times. There is one unusual point, a student who spent 135 minutes on homework and more than the data would suggest (145 minutes) on the internet. To have a student who spends a lot of time on both the internet and homework is not surprising, some would do this and perhaps spend less time on other things such as sport. The data is relatively well spread from the trend line, suggesting a **moderate to weak** relationship.

A5 Conclusion :



Students who spend more time on homework **tend** to spend less free time on the internet. This can be seen by the **negative gradient** of the trend line.
You would expect this because students who spend a lot of free time on the internet would have less time available for other things, for example homework.
For student spending 90 minutes on homework per week, we would predict that they would spend roughly between 70 and 120 minutes per week on the internet. As the data is very scattered around the trend line, this prediction is a very rough estimate.

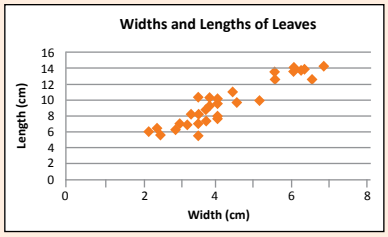
The fact that the data is quite spread from the trend line is not surprising because, for example, you would expect some students to make relatively more time for both homework and the internet.



Pages 30-32 - Practice Investigation 2

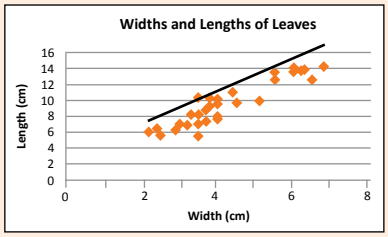
Answers for pages 30-32 are examples only :
The variables are the length (cm) and width (cm) of the leaves. It does not matter which way the variables are plotted on the graph because both are continuous and the data is observational.
The leaves will be placed flat on a desk and a ruler used to measure both the widths and lengths of them. When measuring the leaf width the widest part of the leaf needs to be used each time. For the length any curvature in the stem will be ignored and the length from where the leaf starts to form on the stem to its end point will be measured. A variety of leaf sizes will be selected by taking leaves from different parts of the tree; this means my data should reflect all leaves on the tree. Enough leaves need to be picked to ensure a good reflection of all leaves on the tree is obtained, but you do not want to take so many that damage to the tree occurs. About 30 will be picked. Data will be recorded in a table that shows both the width and length of each leaf.

Your data will depend on the tree you use.
[These answers use data obtained from a bush at the back of a home garden.]



The relationship is a **positive linear** one. There appears to be two groups, one with widths below 4.5 cm and one above 5.5 cm (and only one width of 5.1 cm in the gap). This could relate to the position of the leaves on the bush, the larger ones coming from the top of the bush where they were exposed to the sun.

Points have some scatter about the trend line but are relatively close to it, suggesting that the relationship is a **moderate** one.



Wider leaves **tend** to be longer. This can be seen by the positive gradient of the trend line. Lengths of leaves are a bit more than twice their width. For example (3, 7) and (6, 13) are points on the trend line. This conclusion applies to this bush only but you would think it could be applied to all bushes of the same species. For a leaf with width 4cm, we would predict a length of roughly between 6.5cm and 10.5cm.
As the data is quite close to the trend line, we can be fairly confident about this prediction.

In general you would think that wider leaves would **tend** to be longer for any species because they grow in proportion.