VCE General Mathematics Unit 1

1.1. Investigating and comparing data distributions

1.1.1. Displaying the distribution of a numerical variable

Plotting the distribution of a numerical variable

Eighteen members of a youth group are lining up for tickets to an ‘Australian Idol’ concert. Their current ages are: 10, 13, 10, 15, 16, 11, 11, 16, 15, 10, 10, 11, 19, 16, 15, 11, 10, 15.

(a) Represent the current ages of the members in a dot plot.

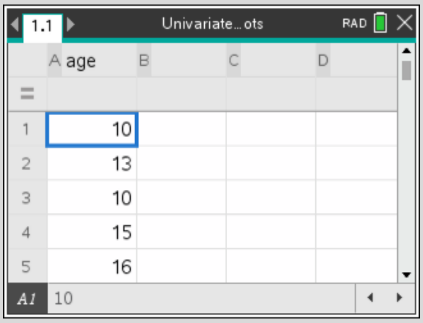
(b) Comment on the variation within the data, and what other information the dot plot reveals.

(c) Represent the current ages as a box plot.

(d) Represent the current ages as a histogram with an interval width of two years.

(a) On a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the variable name ***age***.
- Enter the data for ***age*** into column A.



1.1 Univariate...ots RAD

	A age	B	C	D
=				
1	10			
2	13			
3	10			
4	15			
5	16			

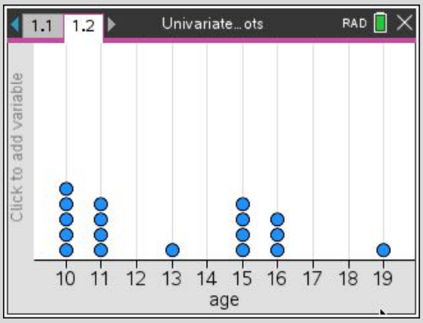
A1 10

***Note:** There are shortcuts for moving the cursor more quickly around a **Lists & Spreadsheet** page.*
Press ctrl 1 *to go to the last entry in a column.*
Press ctrl 7 *to go to the first entry in a column.*
Press ctrl 3 *to go down a page and* ctrl 9 *to go up a page.*

Add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***age***.

The default plot is a dot plot.



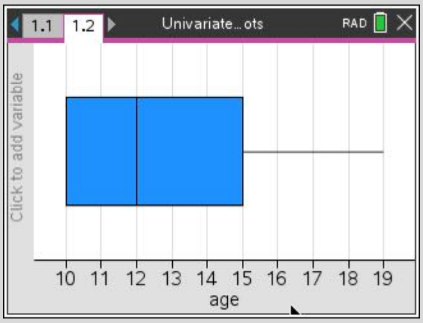
(b) Answer: The dot plot shows that:

- Student ages range from 10 to 19.
- The most common age of these students is 10, but a few other ages (11, 15) occur nearly as frequently.
- Half of the students are aged 10 or 11 years old.
- One student is at least three years older than any other student.

(c) To change the plot type to a box plot:

- Press menu > **Plot Type** > **Box Plot**.

Moving the cursor over the box plot will confirm the quartile values for the student ages.



***Note:** The plot type can also be changed by hovering over the plot window, and then pressing* ctrl menu.
***Note:** The default box plot will display outliers if any exist. To hide the display of outliers, hover over the box plot and then press* ctrl menu, *then select **Extend Box Plot Whiskers**.*

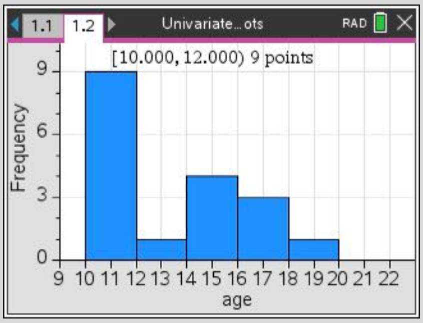
... continued

Plotting the distribution of a numerical variable (continued)

(d) To change the plot type to a histogram:

- Press menu > **Plot Type** > **Histogram**.
- Press menu > **Plot Properties** > **Histogram Properties** > **Bin Settings** > **Equal Bin Width**.
- Change **Width** to 2.
  (i.e. the width of each histogram column)
- Change **Alignment** to 10.
  (i.e. the starting age value for the histogram columns)
- Press menu > **Window/Zoom** > **Zoom – Data**.

Moving the cursor over the histogram will confirm the frequencies for each age.



Plotting a histogram when using variables with frequencies

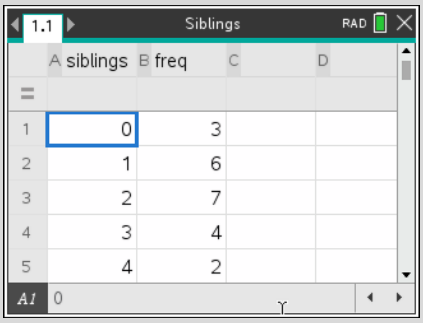
The number of siblings for each of the 26 students in a Year 11 class were recorded in a table.

No. of siblings	0	1	2	3	4	5	6
Frequency	3	6	7	4	2	3	1

Display this table as a histogram.

On a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the variable name ***siblings***.
- In the column B heading cell, enter the name ***freq***.
- Enter the data for ***siblings*** into column A.
- Enter the frequency information into column B.



	A siblings	B freq	C	D
=				
1	0	3		
2	1	6		
3	2	7		
4	3	4		
5	4	2		

A1 0

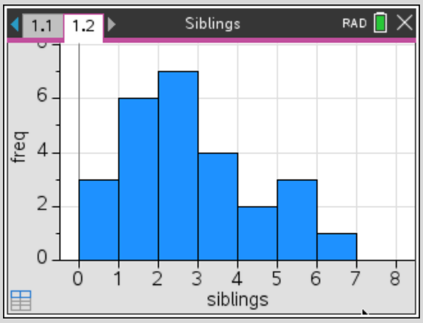
Add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***siblings***.

The default plot is a dot plot.

- Press menu > **Plot Properties** > **Add Y Summary List**.
- Select ***freq*** from the pop-up menu.

Note: For scenarios involving continuous numerical variables that have intervals, use the midpoint of the intervals as the 'typical' value.



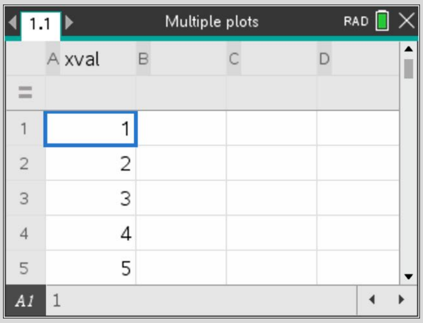
Comparing plot types of the distribution of a numerical variable

Compare the use of a dot plot and a box plot for displaying the distribution of the variable *xval*, if *xval* is the following set of twenty numbers: 1, 2, 3, 4, 5, 6, 6, 6, 6, 7, 7, 8, 8, 8, 8, 9, 9, 10, 10, 10.

On a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the variable name ***xval***.
- Enter the data for ***xval*** into column A.

Note: *The variable name 'xval' is used here to avoid any confusion with the spreadsheet column labelled 'X'.*



	A xval	B	C	D
=				
1	1			
2	2			
3	3			
4	4			
5	5			

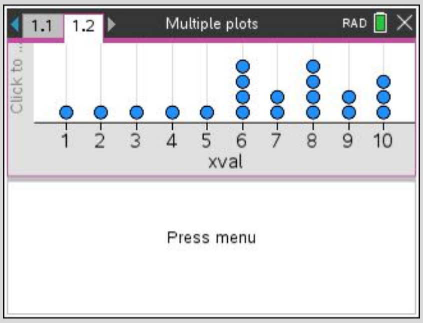
A1 1

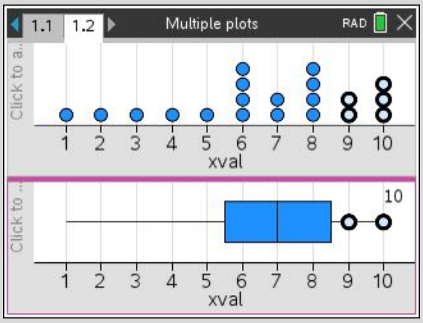
Add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***xval***.

The default plot is a dot plot. To split the current page and add a box plot of *xval* below the dot plot:

- Press doc▾ > **Page Layout** > **Select Layout** > **Layout 3**.
- Click in the lower half of the screen and press menu, then select **Add Data & Statistics**.
- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***xval***.
- Press menu > **Plot Type** > **Box Plot**.





Hover over the data. If you click on any points in one plot, the corresponding points are highlight in both plots. This makes clearer how each plot type represents the distribution of *xval*. For example, note how the top 25% of values is displayed differently in the dot plot and box plot (see right).

Note: *It is possible to 'move' a point (or points), and examine how each plot is affected. To do this, select the required point(s), then drag the point(s) to a different location. Remember that* ctrl esc *will undo such changes. Click away from the selected point(s) to deselect them.*

Constructing parallel box plots

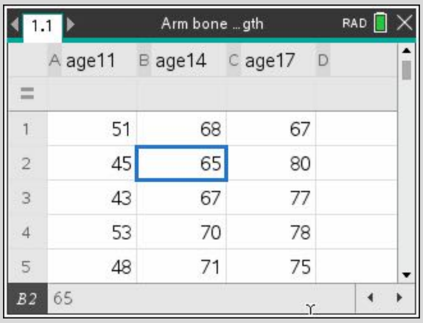
Three adolescents failed to return from a fishing trip. They were Maria aged 11, Thanh aged 14, and Sharma aged 17. Later, a shark was caught offshore, which had a human arm in its stomach. The arm had no identifying features, but the length could be measured. Police found it to be 66 cm.
As part of the investigation, data was collected about the arm lengths of students of the three ages.

Age 11	51	45	43	53	48	48	59	50	50	47	51	49	48	51	47	50
Age 14	68	65	67	70	71	67	66	71	59	62	65	62	67	64	62	60
Age 17	67	80	77	78	75	78	74	76	77	71	65	66	73	66	80	74

Use parallel box plots to help decide whether the arm belongs to one of the missing people.

On a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the name ***age11***.
- In the column B heading cell, enter the name ***age14***.
- In the column C heading cell, enter the name ***age17***.
- Enter the data for the three variables into columns A, B and C respectively.

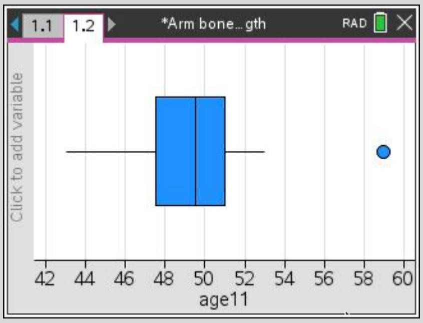


	A age11	B age14	C age17	D
=				
1	51	68	67	
2	45	65	80	
3	43	67	77	
4	53	70	78	
5	48	71	75	

B2 65

Add a **Data & Statistics** page, and then:

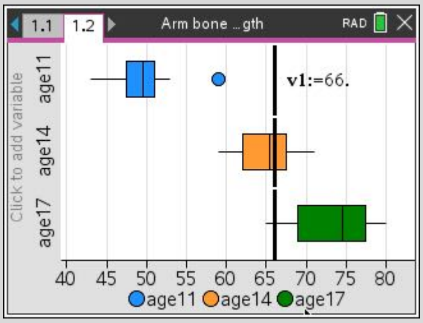
- Press `tab` to activate **Click to add variable** underneath the horizontal axis and select the variable ***age11***.
- Press `menu` > **Plot Type** > **Box Plot**.
- Press `menu` > **Plot Properties** > **Add X Variable** and select the variable ***age14***.
- Repeat the previous step to add a box plot for the variable ***age17***.



Note: The plot type can also be changed by hovering over the plot window, and then pressing `ctrl` `menu`.

To plot the arm bone length of 66 cm:

- Press `menu` > **Analyse** > **Plot Value**.
- Enter the value **66**.



Answer: Observing the parallel box plots and the plotted value of 66 cm, it is possible, but unlikely, that the arm belongs to the 11-year-old, as the maximum arm length for 11-year-olds is 59 cm. A 66 cm arm is in the lowest quartile of the 17-year-old data and so quite possible that the arm belongs to the 17-year-old. It is most likely that it is a 14-year-old's arm, as the median arm length for this group is 65.5 cm. However, more data (including gender specific data) is needed to make a more confident prediction.

Note: The plotted value can be removed by pressing `menu` > ***Analyse > Remove Plotted Value****. A box plot can be removed by pressing* `menu` ***> Plot Properties > Remove X Variable****.*

1.1.2. Summarising the distribution of a numerical variable

Calculating individual statistics for a numerical variable

The handspans (to the nearest cm) of a sample of ten students were recorded as follows:

15, 14, 17, 19, 14, 14, 13, 16, 21, 16.

Calculate the mean, median and standard deviation of the handspans of students in the sample.

On a **Calculator** page:

- Type the variable name ***handspan***.
- Press ctrl ⊞ to enter the 'Assign' symbol.
- Press ctrl) to enter the braces (set brackets).
- Enter the ***handspan*** data as shown.

To calculate the mean of the sample:

- Press menu **> Statistics > List Maths > Mean**.
- Press var and select the variable ***handspan***.

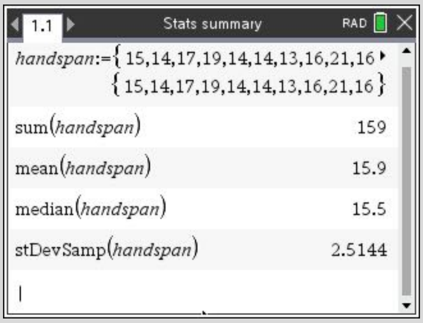
To calculate the median of the sample:

- Press menu **> Statistics > List Maths > Median**.
- Press var and select the variable ***handspan***.

To calculate the standard deviation of the sample:

- Press menu **> Statistics > List Maths > Sample Standard Deviation**.
- Press var and select the variable ***handspan***.

Answer: See statistical results in screen shown right.



```
handspan:={15,14,17,19,14,14,13,16,21,16
              {15,14,17,19,14,14,13,16,21,16}
sum(handspan)                        159
mean(handspan)                      15.9
median(handspan)                    15.5
stDevSamp(handspan)               2.5144
```

***Note**: If the calculator is set to **Auto Calculation Mode**, press ctrl enter to express answers in decimal format.*
*If the calculator is set to **Approximate Calculation Mode**, the answers will default to decimal format.*
*The **Calculation Mode** can be set via ⌂on **> Settings > Document Settings**.*

Summarising a numerical variable from a Calculator page

The handspans (to the nearest cm) of a sample of ten students were recorded as follows:

15, 14, 17, 19, 14, 14, 13, 16, 21, 16.

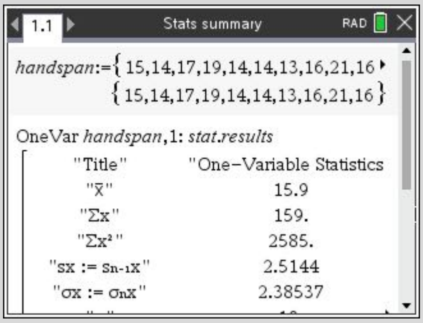
Calculate the summary statistics for the handspans of students in the sample.

On a **Calculator** page:

- Type the variable name ***handspan***.
- Press ctrl ⊞ to enter the 'Assign' symbol.
- Press ctrl) to enter the braces (set brackets).
- Enter the ***handspan*** data as shown.

To calculate the summary statistics of the sample:

- Press menu **> Statistics > Stat Calculations > One Variable Statistics**.
- Set the following values:
  - ○ **Num of Lists = 1**.
  - ○ **X1 List,** click ***handspan***.
- Press enter to calculate and display the summary statistics for ***handspan***. Use the arrow keys to scroll through the summary statistics.



```
handspan:={15,14,17,19,14,14,13,16,21,16
              {15,14,17,19,14,14,13,16,21,16}
OneVar handspan,1: stat.results
    "Title"      "One-Variable Statistics"
     "x̄"                 15.9
    "Σx"                 159.
    "Σx²"                2585.
  "sx := sn-1x"          2.5144
  "σx := σnx"            2.38537
```

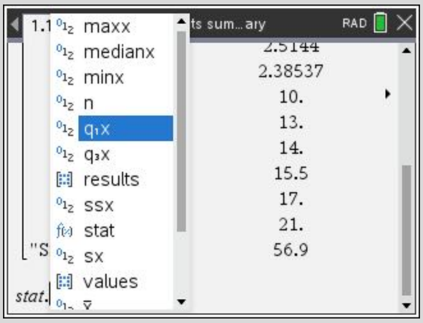
continued ...

Summarising a numerical variable from a Calculator page (continued)

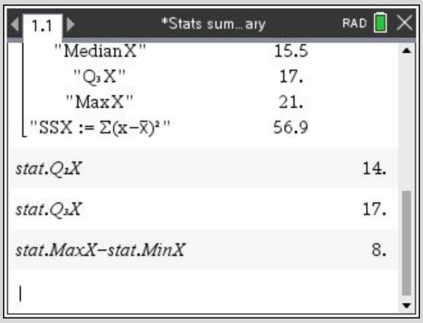
Once the summary statistics have been calculated, they can be accessed individually using the following procedure:

- Type **stat.** (including the decimal point symbol)
- Select the required summary statistic from the pop-up list (e.g. the value of Q1).
- Press enter twice to display the value of the summary statistic.

Note: *All summary statistics are stored and accessible but are only relevant to the most recent calculation of summary statistics. They are updated after any recalculation of the summary statistics*



```
1.1   ts sum…ary                 RAD
      maxx                 2.5144
      medianx              2.38537
      minx                 10.
      n                    13.
      q₁X                  14.
      q₃X                  15.5
      results              17.
      ssx                  21.
      stat                 56.9
 "S   sx
      values
stat. x̄
```



```
1.1   *Stats sum…ary                 RAD
       "MedianX"            15.5
         "Q₃X"              17.
        "MaxX"              21.
 "SSX := Σ(x−x̄)²"           56.9

stat.Q₁X                          14.
stat.Q₃X                          17.
stat.MaxX−stat.MinX               8.
```

Summarising a numerical variable from a Lists & Spreadsheet page

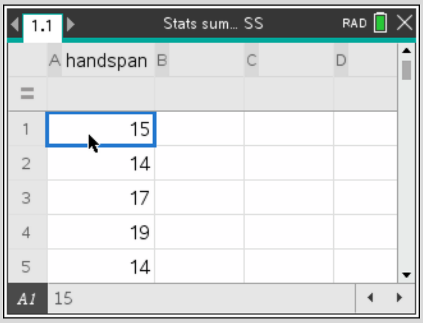
The handspans (to the nearest cm) of a sample of ten students were recorded as follows:

15, 14, 17, 19, 14, 14, 13, 16, 21, 16.

Calculate the summary statistics for the handspans of students in the sample.

On a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the variable name ***handspan***.
- Enter the data for ***handspan*** into column A.

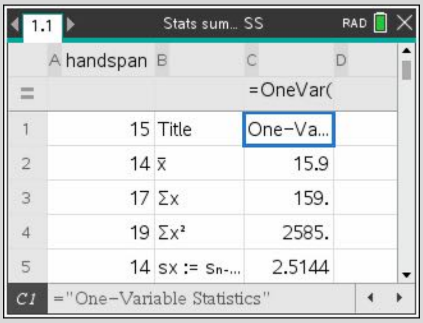


1.1 Stats sum… SS RAD

	A handspan	B	C	D
=				
1	15			
2	14			
3	17			
4	19			
5	14			

A1 15

- Press menu > **Statistics** > **Stat Calculations** > **One Variable Statistics**.
- Set the following values:
  - **Num of Lists = 1**.
  - **X1 List** = ***handspan***.
  - **1st Result Column** = **b[]** (to place statistical results starting from column B).
- Press enter to calculate and display the summary statistics for ***handspan***.



1.1 Stats sum… SS RAD

	A handspan	B	C	D
=			=OneVar(	
1	15	Title	One−Va…	
2	14	x̄	15.9	
3	17	Σx	159.	
4	19	Σx²	2585.	
5	14	sx := sn-…	2.5144	

C1 ="One−Variable Statistics"

Use the arrow keys to scroll through the summary statistics.

Note: *The summary statistics that have been calculated most recently can be accessed from a* ***Calculator*** *page (as above).*

Summarising a numerical variable with frequency information

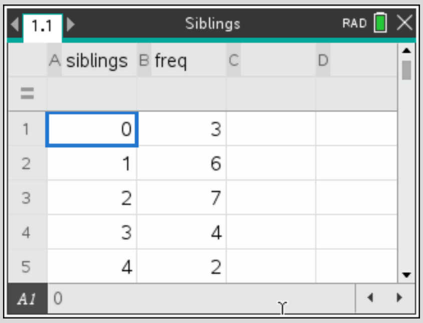
The number of siblings for each of the 26 students in a Year 11 class were recorded in a table.

No. of siblings	0	1	2	3	4	5	6
Frequency	3	6	7	4	2	3	1

Find the mean number of siblings for students in the class, correct to two decimal places.

On a **Lists & Spreadsheet** page:

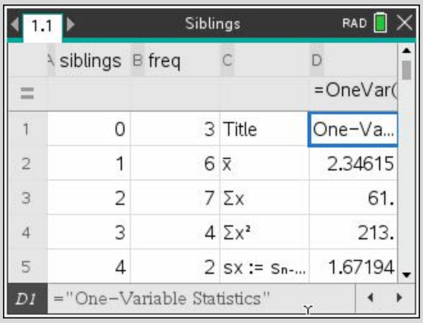
- In the column A heading cell, enter the variable name ***siblings***.
- In the column B heading cell, enter the name ***freq***.
- Enter the data for ***siblings*** into column A.
- Enter the frequency information into column B.



	A siblings	B freq	C	D
=				
1	0	3		
2	1	6		
3	2	7		
4	3	4		
5	4	2		

A1 0

- Press menu > **Statistics > Stat Calculations > One Variable Statistics.**
- Set the following values:
  - **Num of Lists = 1**.
  - **X1 List** = ***siblings***.
  - **Frequency List** = ***freq***.
  - **1st Result Column** = **c[]** (to place statistical results starting from column C).
- Press enter to calculate and display the summary statistics for ***siblings***.



	A siblings	B freq	C	D
=				=OneVar(
1	0	3	Title	One-Va...
2	1	6	$\bar{x}$	2.34615
3	2	7	Σx	61.
4	3	4	Σx^2	213.
5	4	2	sx := Sn-...	1.67194

D1 ="One-Variable Statistics"

Use the arrow keys to scroll through the summary statistics.

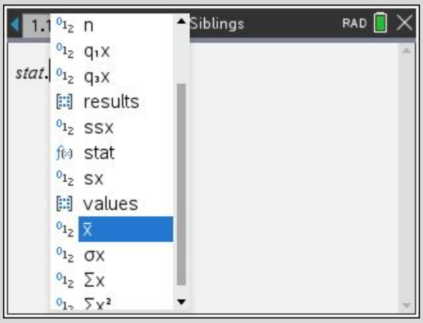
Answer: The mean number of siblings for students in the class is 2.35, correct to two decimal places.

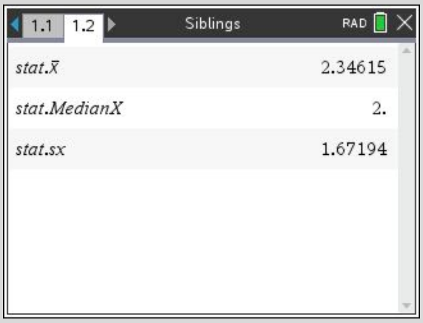
The summary statistics that have most recently been calculated can be accessed individually.

To do this, add a **Calculator** page and then:

- Type **stat.** (including the decimal point symbol) to display a pop-up menu of the calculated statistics.
- Select the desired statistic and press enter twice.

To display all the calculated statistics, enter **stat.results**.





```
stat.x̄                2.34615
stat.MedianX               2.
stat.sx               1.67194
```

Note: *For scenarios involving continuous numerical variables that have intervals, use the midpoint of the intervals as the 'typical' value. In such scenarios the summary statistics calculated will be estimates of the actual statistics.*

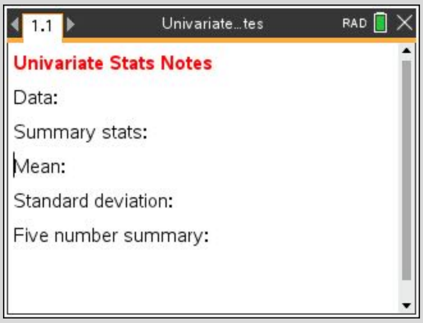
Creating a Notes page to summarise a numerical variable

A **Notes** page can be constructed to automatically calculate summary statistics for a numerical variable. We will use sample data as follows: 10, 13, 10, 15, 16, 11, 11, 16, 15.

On a **Notes** page:

- Enter the text shown in the screenshot (the colon symbol can be found via the ?!▸ key).
- Move the cursor to the right of the word 'Data' and press menu **> Insert > Maths Box** (or press ctrl M).

Repeat to insert **Maths Boxes** next to each of the other template headings.



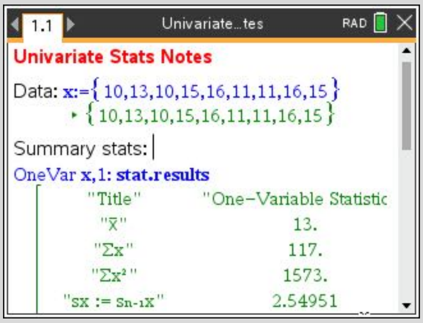
```
Univariate Stats Notes
Data:
Summary stats:
Mean:
Standard deviation:
Five number summary:
```

***Note:** To edit the text colour, select the text by holding ⇧shift and cursor keys. Then press menu **> Format > Text colour**.*

- Click on the **Maths Box** next to the word 'Data'.
- Inside the **Maths Box**, input using assign (ctrl ≔) and braces (ctrl ⦃⦄) as follows:

$$\boldsymbol{x := \{10,13,10,15,16,11,11,16,15\}}$$

- For 'Summary stats:', press menu **> Calculations > Statistics > Stat Calculations > One–Variable Statistics …** and then set the following values:
  - ○ **Num of Lists = 1.**
  - ○ **X1 List** = ***x***
- Press enter twice to calculate and display the summary statistics for ***x***.

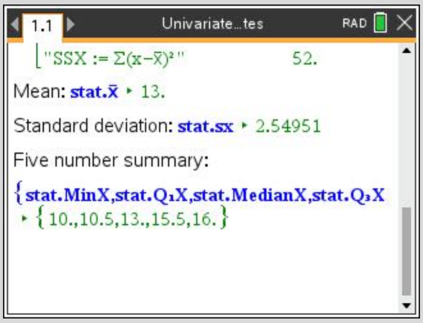


```
Univariate Stats Notes
Data: x:={10,13,10,15,16,11,11,16,15}
   ▸ {10,13,10,15,16,11,11,16,15}
Summary stats:
OneVar x,1: stat.results
   "Title"          "One–Variable Statistic
     "x̄"                     13.
     "Σx"                    117.
     "Σx²"                   1573.
   "sx := sn-1x"             2.54951
```

In the other **Maths Boxes**:

- For 'Mean', type **stat.** and then select $\bar{x}$.
- For 'Standard deviation', type **stat.** and then select sx.
- For 'Five number summary', press ctrl ⦃⦄ to enter the braces (set brackets), and then use the above method to complete the following expression:

$$\{stat.MixX, stat.Q1X, stat.MedianX, stat.Q3X, stat.MaxX\}$$



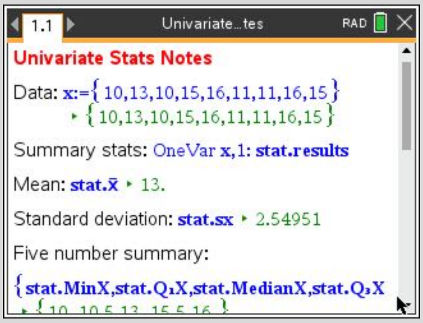
```
 "SSX := Σ(x−x̄)²"            52.
Mean: stat.x̄ ▸ 13.
Standard deviation: stat.sx ▸ 2.54951
Five number summary:
{stat.MinX,stat.Q₁X,stat.MedianX,stat.Q₃X
 ▸ {10.,10.5,13.,15.5,16.}
```

***Note**: The statistical variables from the most recent statistical calculations can also be accessed via the var key.*

To hide the output results from the 'Summary stats' maths box, click inside that maths box and then:

- Press menu **> Maths Box Options > Maths Box Attributes**.
- In the **Input & Output** attribute, select **Hide Output**.

This allows the most relevant stats to be displayed more conveniently. If the data values for *x* are changed, the statistics will be updated automatically.



```
Univariate Stats Notes
Data: x:={10,13,10,15,16,11,11,16,15}
   ▸ {10,13,10,15,16,11,11,16,15}
Summary stats: OneVar x,1: stat.results
Mean: stat.x̄ ▸ 13.
Standard deviation: stat.sx ▸ 2.54951
Five number summary:
{stat.MinX,stat.Q₁X,stat.MedianX,stat.Q₃X
 ▸ {10.,10.5,13.,15.5,16.}
```

***Note**: Changes to the attributes of a **Maths Box** are not always visible until you have clicked outside of that **Maths Box**. For example, the **Hide Output** attribute selected above is not enacted until the user clicks outside the 'Summary stats' **Maths Box**.*

Investigating the meaning of standard deviation

Jemima and Hania play for a local cricket club. Their first four lots of scores for the season are:

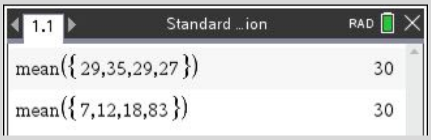
- Jemima: 29, 35, 29 and 27
- Hania: 7, 12, 18, and 83

For each batter, how does each score deviate from their mean score? Is there a way to compare the scores for each cricketer and look at the ways their scores are similar and the ways in which they are different? It also explores the usefulness of the 'standard deviation'.

To calculate the mean scores for Jemima and Hania, on a **Calculator** page:

- For Jemima's mean score, enter **mean({29,35,29,27})**.
- For Hania's mean score, enter **mean({7,12,18,83})**.

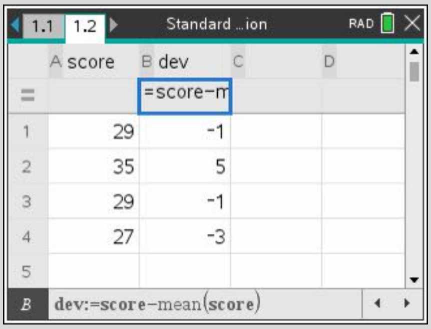
The mean scores are the same (both mean scores are 30 runs).



```
mean({29,35,29,27})      30
mean({7,12,18,83})       30
```

To calculate by how much each score deviates from the mean score for Jemima, on a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the label ***score***.
- In the column B heading cell, enter the variable name ***dev***.
- Enter the data for ***score*** into column A.
- In the column B formula cell, enter the formula ***=score – mean(score)***.



	A score	B dev	C	D
=		=score−m		
1	29	-1		
2	35	5		
3	29	-1		
4	27	-3		
5				

B dev:=score−mean(score)

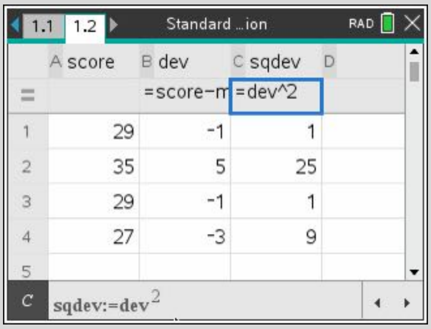
The sum of these 'deviations' is zero ($-1+5+-1+-3=0$), which is unhelpful as the 'average deviation' of the scores (This will always be the case).

The 'negative deviations' can be made positive by squaring the values of the deviations. To calculate the 'squared deviations' for Jemima, on a **Lists & Spreadsheet** page:

- In the column C heading cell, enter the label ***sqdev***.
- In the column C formula cell, enter the formula ***=dev***$^{\mathbf{2}}$.

The sum of the 'squared deviations' is 36.

$$\left(\text{i.e. } (-1)^2+(5)^2+(-1)^2+(-3)^2=36\right)$$



	A score	B dev	C sqdev	D
=		=score−m	=dev^2	
1	29	-1	1	
2	35	5	25	
3	29	-1	1	
4	27	-3	9	
5				

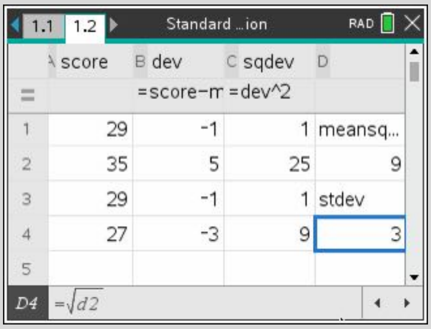
C sqdev:=dev^2

The average or mean 'squared deviation' can be calculated as follows:

- Click in the cell D1, and enter the label **meansqdev**.
- Click in the cell D2, and enter the formula **=mean(sqdev)**.

Finally, to find the average deviation (called the 'standard deviation') by this method, find the square root of the value of the average squared deviation as follows:

- Click in the cell D3, and enter the label **stdev**.
- Click in the cell D4, and enter the formula **=sqrt(D2)**.



	A score	B dev	C sqdev	D
=		=score−m	=dev^2	
1	29	-1	1	meansq...
2	35	5	25	9
3	29	-1	1	stdev
4	27	-3	9	3
5				

D4 $=\sqrt{d2}$

Now we have a *standardised* measure of the mean deviation (referred to as the standard deviation), which ignores whether the individual deviations are positive or negative.

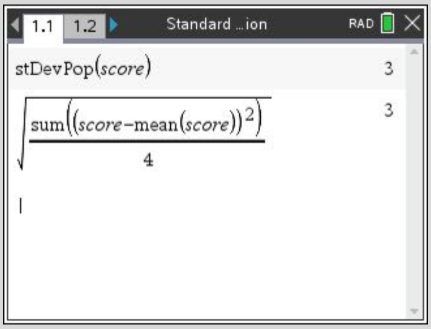
... continued

Investigating the meaning of standard deviation (continued)

To verify this measure, on a **Calculator** page:

- Press menu **> Statistics > List Maths > Population Standard Deviation**, then enter **stDevPop(*score*)** as shown.
- Enter the formula summary

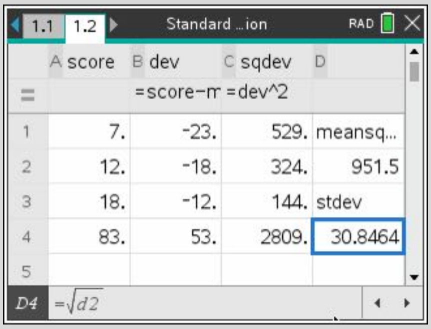
$$\sqrt{\frac{\textbf{sum}\left(\left(\textbf{\textit{score}}-\textbf{mean}(\textbf{\textit{score}})\right)^2\right)}{\textbf{4}}}$$



stDevPop(*score*) 3

$\sqrt{\frac{\text{sum}\left(\left(score-\text{mean}(score)\right)^2\right)}{4}}$ 3

For comparison, if Hania's scores are entered (see screen at right), the standard deviation is much greater (approximately 31 runs), reflecting greater variability in her scores.



	A score	B dev	C sqdev	D
=		=score-m	=dev^2	
1	7.	-23.	529.	meansq...
2	12.	-18.	324.	951.5
3	18.	-12.	144.	stdev
4	83.	53.	2809.	30.8464
5				

D4 $=\sqrt{d2}$

Further exploration using the spreadsheet might include finding four scores that have a particular standard deviation, or the 'smallest' and 'largest' possible standard deviation.

> ***Note:*** *The general formula for the standard deviation of a numerical variable x which has n values is:*
>
> $$\textbf{\textit{Standard deviation}} = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n}}$$
>
> *In senior mathematics courses where the standard deviation of a sample is used as an estimate of the population standard deviation, it is usual to use **n – 1** rather than **n** in the definition.*

VCE General Mathematics Unit 2

2.1. Investigating relationships between two numerical variables

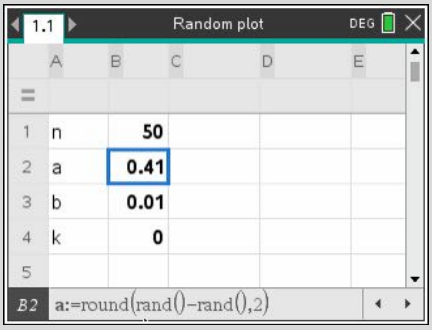
2.1.1. Analysing the association between two numeric variables

Generating random scatter plots

It can be useful to generate scatter plots when analysing association between two numeric variables. In the following example, a scatter plot is generated with the help of the random number generator, and a slider for controlling the 'extent' of this randomness to be applied.

To generate a random set of points on a **Lists & Spreadsheet** page:

- In column A, enter the letters *n*, *a*, *b*, and *k* as shown.
- In column B:
  - For *n*, enter ***n*:=50**.
  - For *a*, enter ***a*:=round(rand()–rand(),2)**.
  - For *b*, enter ***b*:=round(rand()–rand(),2)**.
  - For *k*, enter ***k*:=0**.



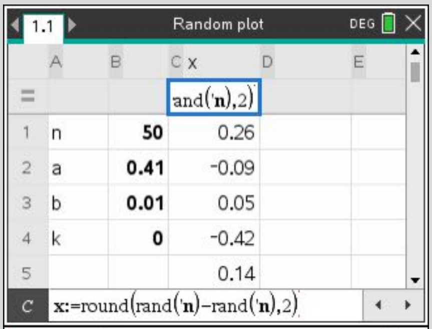
	A	B	C	D	E
=					
1	n	50			
2	a	0.41			
3	b	0.01			
4	k	0			
5					

B2 a:=round(rand()−rand(),2)

__Note__: The __rand()__ function generates a random number between 0 and 1, so the command __rand()–rand()__ will generate a number between –1 and 1.

To generate *n* random values for *x*, in column C:

- In the column C heading cell, enter the name ***x***.
- In the column C formula cell, type the formula **=round(rand('n)-rand('n),2)**



	A	B	C x	D	E
=			and('n),2)		
1	n	50	0.26		
2	a	0.41	-0.09		
3	b	0.01	0.05		
4	k	0	-0.42		
5			0.14		

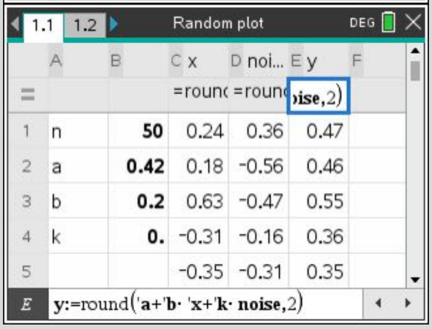
C x:=round(rand('n)−rand('n),2)

To generate values for *noise*, in column D:

- In the column D heading cell, enter the name ***noise***.
- In the column D formula cell, type the formula **=round(rand('n)-rand('n),2)**

To generate values for *y*, in column E:

- In the column E heading cell, enter the name ***y***.
- In the column E formula cell, type the formula **=round('a+'b×'x+'k×noise,2)**



	A	B	C x	D noi...	E y	F
=			=rounc	=rounc	ise,2)	
1	n	50	0.24	0.36	0.47	
2	a	0.42	0.18	-0.56	0.46	
3	b	0.2	0.63	-0.47	0.55	
4	k	0.	-0.31	-0.16	0.36	
5			-0.35	-0.31	0.35	

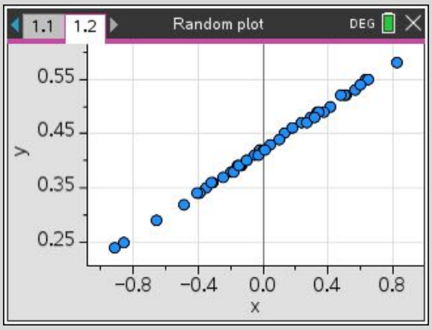
E y:=round('a+'b· 'x+'k· noise,2)

__Note__: The single 'dash' character in the above formulae can be found via the ?!▸ key. It is used here to indicate a variable reference rather than a column reference.

Add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***x***.
- Press tab to activate **Click to add variable** to the left of the vertical axis and select the variable ***y***.

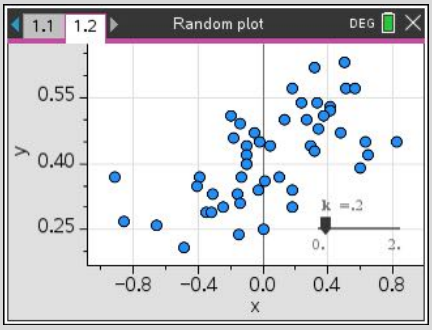
This will display a scatter plot with a very linear appearance, since *k* = 0 and any 'imperfections' will be related to the rounding of values.



... continued

To add a slider on the plot controlling the value of k:

- Press menu > **Actions** > **Add Slider**.
  In the dialog box that follows, enter the values:
  Variable: ***b*** Value: **0** Minimum: **0**
  Maximum: **2** Step Size: **0.1**
- Press enter to display the slider and position it as required.



Click on the slider and use the arrow keys to change the value of k to vary the strength of association, and press menu > **Window/Zoom** > **Zoom-Data** to redraw the window.

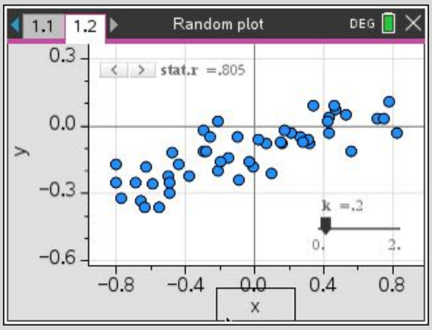
To generate a new plot:

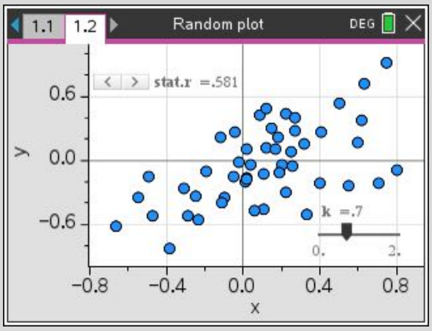
- Press ctrl ◀ to display the **Lists & Spreadsheet** page.
- Press ctrl R to recalculate all the formulae and generate another random plot.
- Press ctrl ▶ to display the new plot on the **Data & Statistics** page.
- Press menu > **Window/Zoom** > **Zoom-Data** to redraw the window if needed.

Extension

It is also possible to display the value of the correlation coefficient for each scatter plot. To do this:

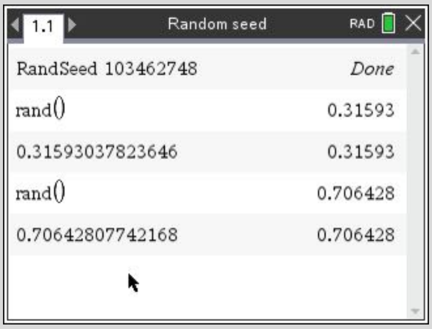
- Press ctrl ◀ to display the **Lists & Spreadsheet** page.
- Press menu > **Statistics** > **Stat Calculations** > **Two-Variable Staistics**
  - For X List, click to select **'*x***.
  - For Y List, click **'*y***.
  - Press enter to calculate and display the summary statistics for these two variables in the spreadsheet.
- Press ctrl ▶ to display the **Data & Statistics** page.
- Press menu > **Actions** > **Add Slider**.
  In the dialog box that follows, select the variable: ***stat.r*** and click the minimised option.
- Press enter to display the slider and position it as required.





If a ‘recalculation’ is done, or the value of k is changed, the new value of the correlation coefficient r is displayed. It is **not** intended that you interact with the slider directly.

Note: *The ‘random’ number generator used on the TI-Nspire CX II CAS (and most mathematical software) is a ‘pseudo-random’ generator, because it simulates a sequence of random numbers from a known sequence. One way to highlight this is to show how a calculator that has been reset will always give the same starting value as an output for its* ***rand*** *function. For this reason, it is best to ‘seed’ the random number function so that it starts from a different spot in the sequence. An example of seeding is shown in the screen shown right. The* ***RandSeed*** *command can be found by pressing* menu ***> Probability > Random > Seed****.*



```
RandSeed 103462748          Done
rand()                   0.31593
0.31593037823646         0.31593
rand()                  0.706428
0.70642807742168        0.706428
```

2.1.2. Lines of good fit

Calculating the slope and y-intercept of a line of best fit given two points

A given scatter plot shows that the line of best fit will approximately pass through the points (2,3) and (6,11). Find the slope and *y*-intercept of such a line.

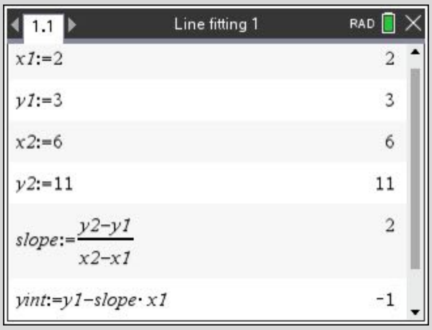
On a **Calculator** page:

- Enter the values for **($x1$, $y1$)** and **($x2$, $y2$)** as shown.

To calculate the slope and *y*-intercept of the line passing through both points:

- Enter $\boldsymbol{slope := \dfrac{y2 - y1}{x2 - x1}}$.
- Enter $\boldsymbol{yint := y1 - slope \times x1}$.

Answer: The slope is 2 and the *y*-intercept is –1.



```
x1:=2                    2
y1:=3                    3
x2:=6                    6
y2:=11                  11
slope:=(y2−y1)/(x2−x1)   2
yint:=y1−slope·x1       −1
```

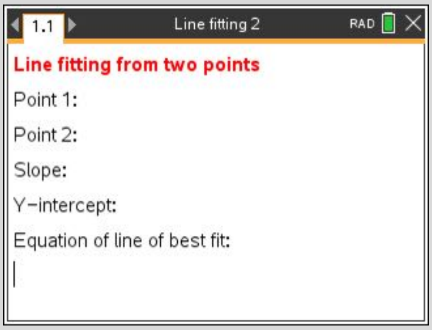
Constructing a notes template to find the equation of the line of best fit given two points

A **Notes** page can be constructed to calculate the equation of the line of best fit for a scatter plot (using the form $y = a + bx$) given two suitable points on a plot. This will be tested with the sample points (2,3) and (6,11).

On a **Notes** page:

- Enter the text shown in the screenshot.
- Move the cursor to the right of the word ‘Point 1:’ and press menu > **Insert** > **Maths Box** (or press ctrl M).

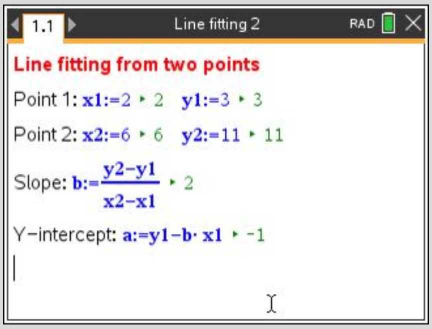
Repeat to insert **Maths Boxes** next to each of the other template headings.



Line fitting from two points
Point 1:
Point 2:
Slope:
Y−intercept:
Equation of line of best fit:

Note: *To edit the text colour, select the text by holding ⇧shift and ‘arrow’ across the text. Then press menu* ***> Format > Text colour****.*

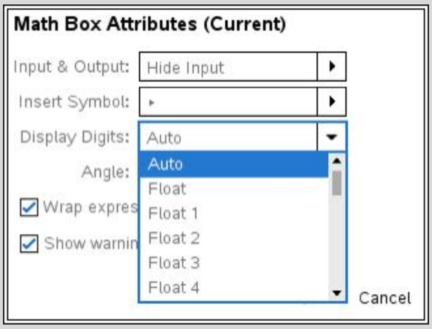
- Click on the **Maths Box** next to the word ‘Point 1’.
- Inside the **Maths Box**, input $\boldsymbol{x1 := 2}$
- Repeat this method to enter the following:
  - For the *y*-coordinate of Point 1, enter $\boldsymbol{y1 := 3}$
  - For the *x*-coordinate of Point 2, enter $\boldsymbol{x2 := 6}$
  - For the *y*-coordinate of Point 2, enter $\boldsymbol{y2 := 11}$
  - For ‘Slope:’, enter $\boldsymbol{b := \dfrac{y2 - y1}{x2 - x1}}$.
  - For ‘*Y*-intercept:’, enter $\boldsymbol{a := y1 - b \times x1}$.



Line fitting from two points
Point 1: x1:=2 ▸ 2 y1:=3 ▸ 3
Point 2: x2:=6 ▸ 6 y2:=11 ▸ 11
Slope: b:=(y2−y1)/(x2−x1) ▸ 2
Y−intercept: a:=y1−b· x1 ▸ −1

Answer: The equation of the line passing through the points with coordinates (2,3) and (6,11) is $y = -1 + 2x$.

Note: *Unless otherwise specified, the output values will be displayed using the system settings for ‘Display Digits’. However, the display precision of each Maths Box can be set individually using menu* ***> Maths Box Options > Math Box Attributes.***


Math Box Attributes (Current)

Input & Output: Hide Input
Insert Symbol: ▸
Display Digits: Auto
Angle:
☑ Wrap expres
☑ Show warnin

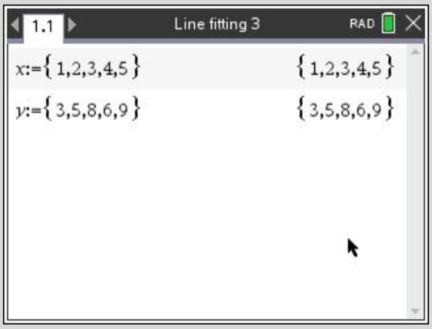
Finding the equation of the line of best fit by minimising the sum of the areas of squares

The least squares regression method can be introduced to students by visualising the process of minimising the areas of the residual squares, using the approach outlined below.

To enter sample points for a scatter plot on a **Calculator** page:

- Enter ***x*:= {1,2,3,4,5}**.
- Enter ***y*:= {3,5,6,8,9}**.

Note: *To enter the 'Assign' symbol (i.e. ':='), press* ctrl ⊞.



x:={1,2,3,4,5} {1,2,3,4,5}

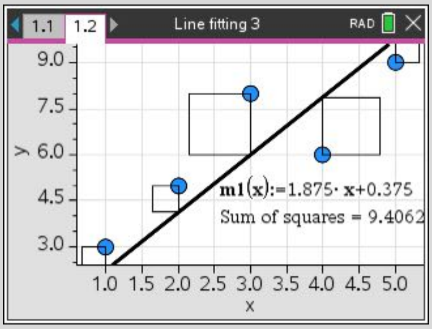
y:={3,5,8,6,9} {3,5,8,6,9}

Add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***x***.
- Press tab to activate **Click to add variable** to the left of the vertical axis and select the variable ***y***.

This will display a scatter plot. To add a 'movable' line, and display residual squares:

- Press menu > **Analyse** > **Add Movable Line**.
- Press menu > **Analyse** > **Residuals** > **Show Residual Squares**.



m1(x):=1.875·x+0.375

Sum of squares = 9.4062

This will display a line and its equation nearby the plotted points. It also displays the squares formed by the vertical distance between each point and the line, as well as the total area of these squares (the 'Sum of squares').

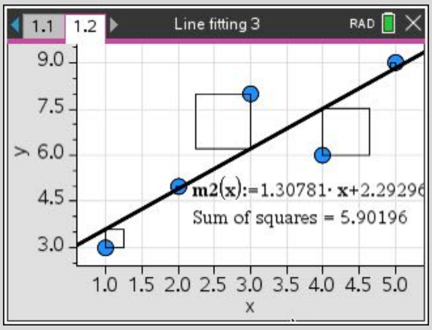
The line can be 'moved' in two possible ways.

To change the *slope* of the line:

- Hover the cursor near to the 'ends' of the line, then click and drag to change the *slope* of the line and note how the equation and total area of the squares is changed.

To change the *y-intercept* of the line:

- Hover the cursor near to the 'middle' of the line, then click and drag to change the *y-intercept* of the line and note how the equation and total area of the squares is changed.



m2(x):=1.30781·x+2.29296

Sum of squares = 5.90196

By making appropriate 'moves', it is possible to reduce the 'Sum of squares' to below 6 square units. For reference, the least squares algorithm gives the equation of the line as $y = 1.3x + 2.3$, with the minimum 'sum of squares' as 5.9.

Note: *To stop using the 'movable line', press* menu > ***Analyse > Remove Movable Line.***

VCE General Mathematics Unit 3

3.1. Investigating data distributions

3.1.1. Displaying the distribution of a numerical variable

Using a logarithmic scale to display the distribution of a numerical variable

The estimated 2025 *population size* of the 11 countries in South-East Asia is shown in the table on the right.

(a) Construct a histogram of *population size* using an appropriate linear scale.

(b) Construct a histogram using the $\log_{10}$(*population size*).

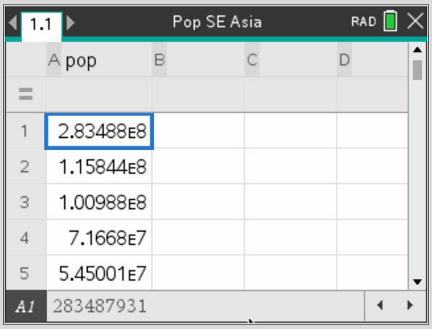
(c) Compare the two histograms constructed – which is more useful in displaying detail of the distribution of population size?

Country	*Population size*
Indonesia	283,487,931
Philippines	115,843,670
Vietnam	100,987,686
Thailand	71,668,011
Myanmar	54,500,091
Malaysia	35,557,673
Cambodia	17,638,801
Laos	7,769,819
Singapore	5,832,387
Timor-Leste	1,400,638
Brunei	462,721

On a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the variable name ***pop***.
- Enter the data for ***pop*** into column A.

***Note:** The data for population size may be displayed using scientific notation (see right). For instance, the screen shown right is from a calculator with display digits set to FLOAT 6. In this instance, values of population size greater than 6 digits will be displayed using scientific notation, with up to 6 digits in the mantissa. The display digits setting can be altered by pressing* ⌂on *> **Settings** > **Document Settings**.*



	A pop	B	C	D
=				
1	2.83488E8			
2	1.15844E8			
3	1.00988E8			
4	7.1668E7			
5	5.45001E7			

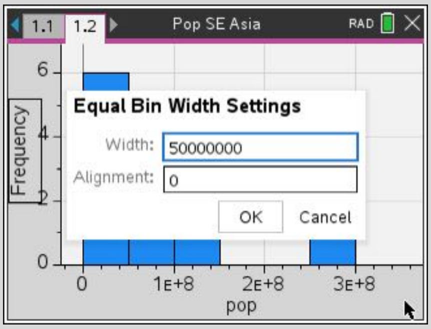
A1 283487931

(a) Add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***pop***.

The default plot is a dot plot. Observing the data, a histogram column width of 50 million seems reasonable. To change the plot type to a histogram:

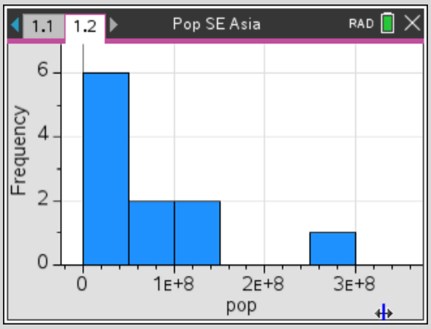
- Press menu > **Plot Type** > **Histogram**.
- Press menu > **Plot Properties** > **Histogram Properties** > **Bin Settings** > **Equal Bin Width**.
- Change **Width** to 50,000,000 (i.e. the width of each histogram column).
- Change **Alignment** to 0 (i.e. the starting value of ***pop*** for the histogram columns).
- Press menu > **Window/Zoom** > **Zoom – Data**.



Equal Bin Width Settings

Width: 50000000

Alignment: 0



***Note:** The **Bin Width** of a histogram is the width of each column. The **Bin Alignment** of a histogram is the value from which the histogram columns should start.*

Moving the cursor over the histogram will confirm the frequencies for each ***pop*** interval of 50,000,000.

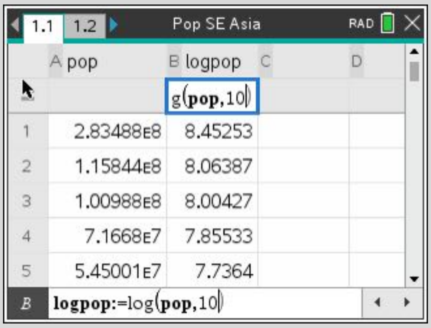
… continued

Using a logarithmic scale to display the distribution of a numerical variable (continued)

(b) To construct a histogram for the logarithm (using base 10) of *population size*, move back to the **Lists & Spreadsheet** page and then:

- In the column B heading cell, enter the name ***logpop***.
- In the column B formula cell, type **=log(*pop*,10)**.

This will calculate and display the base 10 logarithmic numbers associated with each country's population size. For instance, Indonesia's population of 283,487,931 is approximately $10^{8.45253}$, and so the associated logarithm value is approximately 8.45253.



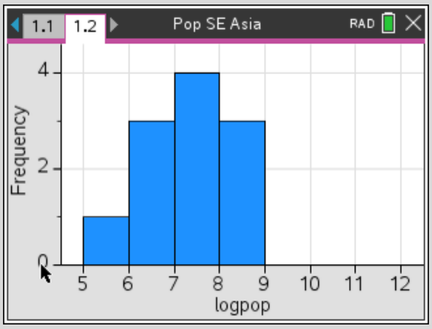
	A pop	B logpop	C	D
		g(pop,10)		
1	2.83488E8	8.45253		
2	1.15844E8	8.06387		
3	1.00988E8	8.00427		
4	7.1668E7	7.85533		
5	5.45001E7	7.7364		

B **logpop**:=log(**pop**,10)

Note: *If the* ***Calculation Mode*** *is set to* ***Auto*** *mode (via* ⌂ on *>* ***Settings*** *>* ***Document Settings****), the calculator will display the values in the* ***logpop*** *column in exact form (e.g.* log_{10}*(283,487,931)) rather than approximate decimal form (e.g. 8.45253). To display in approximate form, set the* ***Calculation Mode*** *to* ***Approx****, or include a decimal point in the formula (i.e.* ***=log(pop,10.0)****)*

To display the histogram for the logarithm of *population size*, move back to the **Data & Statistics** page and then:

- Click on the variable ***pop*** underneath the horizontal axis and select the variable ***logpop***.
- Press menu > **Plot Properties > Histogram Properties > Bin Settings > Equal Bin Width**.
- Change **Width** to 1
  (i.e. the width of each histogram column).
- Change **Alignment** to 5
  (i.e. the starting **logpop** value for the histogram columns).
- Press menu > **Window/Zoom > Zoom – Data**.



Note: *The horizontal scale can be adjusted directly by clicking on the horizontal axis and then using the left or right arrows to adjust the scale.*

(c) Answer: The histogram using a linear display shows that the values of *population size* are spread out from just under 500,000 to just under 300,000,000, with 6 countries with *population size* between 0 and 50,000,000. It is difficult to view details of the entire distribution in this histogram.
In contrast to this, the histogram displaying the logarithmic scale brings the columns more closely together highlights that only one country has a population between 10^5 and 10^6 (i.e. between 100,000 and 1,000,000), whereas 3 countries have a population between 10^6 and 10^7 (i.e. between one and ten million). The histogram displaying the logarithmic scale makes it easier to display values on the horizontal axis, as they span between logarithmic values of 5 and 9.

3.1.2. Modelling with the normal distribution

Visualising the normal distribution

On *Able Farm*, the eggs produced are known to have a mean weight of 70 g, with a standard deviation of 10 g. Let A be the weight of eggs from *Able Farm* and assume that A can be modelled by a normal distribution.

(a) Create a plot of the distribution of A in a suitable viewing window.

(b) Shade the area under the graph associated with egg weights between 50 and 90 g.

(c) Use the graph to find the approximate percentage of eggs for which

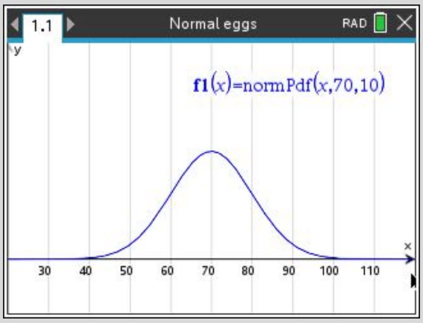
(i) $50 \le A \le 90$ **(ii)** $40 \le A \le 100$ **(iii)** $60 \le A \le 90$.

(a) To plot the distribution of A, on a **Graphs** page:

- Press 🕮 N and select **normPdf()**.
- Enter $f1(x)=$ **normPdf(*x*,70,10)**.
- Press menu > **Window/Zoom** > **Window Settings**

In the dialog box that follows, enter the following values:

XMin = 20	Xmax = 120	XScale = 10
YMin = −0.02	YMax = 0.08	YScale = 1

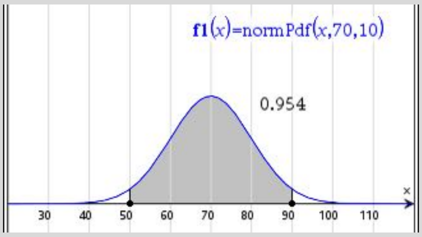


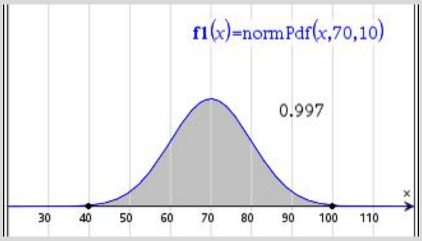
Note: The screen shown right has been enhanced to display a lined grid (via menu *>* ***View*** *>* ***Grid*** *>* ***Lined Grid****). The axis values are labelled, which can be displayed by hovering over one of the axes, then pressing* ctrl menu *>* ***Attributes****. In the popup menu, press* ▼ *to select the* ***Tick Labels*** *attribute, then press* ◀ *to select* ***Multiple Labels****. Press* enter *to save this attribute.*

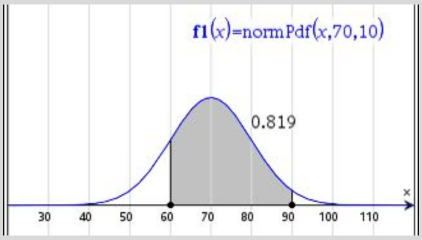
(b) To shade and calculate the area under the normal curve:

- Press menu > **Analyse Graph** > **Integral**.
- For 'lower bound', type **50** and then press enter.
- For 'upper bound', type **90** and then press enter.

The relevant area under the normal curve for part (c)(i) will be shaded, and an approximate value for the area is calculated and displayed. Press ctrl Z to remove the shading before each new area calculation.







(c) Answers:

(i) It is expected that approximately 95.4% of eggs will have weight between 50 and 90 g.

(ii) It is expected that the 99.7% of eggs will have a weight between 40 and 100 g.

(iii) It is expected that the 81.9% of eggs will have a weight between 60 and 90 g.

Note: *The percentages given in the answers here are* ***only*** *approximate values, as are the values referenced by the '68/95/99.7% rule', which are even more approximate! The example here is used mainly to help students visualise the approximate percentages based on the relevant area under the normal curve.*

... continued

Visualising the normal distribution (continued)

On *Bonza Farm*, the eggs produced are known to have a mean weight of 70 g, with a standard deviation of 5 g. Let *B* be the weight of eggs from *Bonza Farm*, and assume that *B* can also be modelled by a normal distribution.

(d) Add a plot of the distribution of *B* underneath the plot of the distribution of *A*.
(e) Use the two plots to help explain why an egg weight of greater than 80 g might be more common on *Able Farm* than *Bonza Farm*.

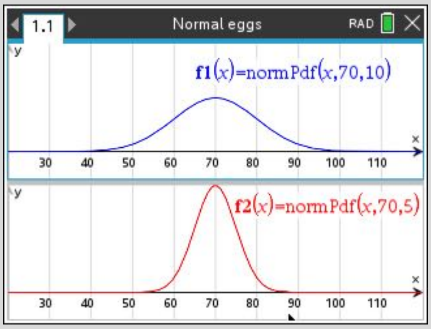
(d) To add a plot of the distribution of *B*, on the **Graphs** page:

- Press doc▾ > **Page Layout** > **Select Layout** > **Layout 3**
- Click in the bottom half of the screen and add another **Graphs** page.
- Enter $f2(x)=$ **normPdf(*x*,70,5)**.
- Press menu > **Window/Zoom** > **Window Settings**

  In the dialog box that follows, enter the following values:
  XMin = 20 Xmax = 120 XScale = 10
  YMin = −0.02 YMax = 0.08 YScale = 1

- Press menu > **View** > **Grid** > **Lined Grid**.
- Hover over the horizontal axis, then press ctrl menu > **Attributes**. In the popup menu, press ▼ to select the **Tick Labels** attribute, then press ◀ to select **Multiple Labels**. Press enter to save this attribute.

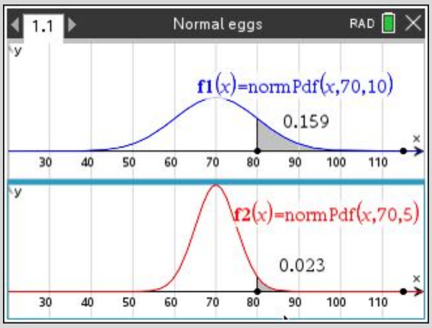
This displays the distribution of *A* and *B* on the same scale.



***Note**: It is not a trvial skill for students to find stuitable boundaries for the viewing window. Discussing the value of the mean and standard deviation will help, as is the need to adjust the vertical scale. **Zoom-Fit** is helpful here.*

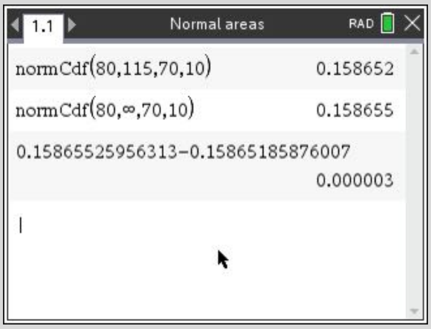
(e) To shade and calculate the area under each of the normal curves:

- Click in the top half of the screen (*Able Farm* curve)
- Press menu > **Analyse Graph** > **Integral**.
- For 'lower bound', type **80** and then press enter.
- For 'upper bound', type **115** and then press enter.*
- Repeat the above steps for the lower half of the screen (for *Bonza Farm* curve)



Answer: Comparing the areas, approximately 16% of the eggs on *Able Farm* will have weights greater than 80 g, whereas approximately 2% of the eggs on *Bonza Farm* will have weights greater than 80 g. Therefore, an egg weight of greater than 80 g will be more common on *Able Farm* than *Bonza Farm*. An 80 g egg from *Able Farm* represents a weight only one standard deviation above the mean weight, whereas an 80 g egg from *Bonza Farm* represents a weight two standard deviations above the mean weight.

****Note:** To estimate an area associated with the statements 'greater than 80 g', it is reasonable to select an upper bound which is sufficiently large (e.g. 115 g). The screen right highlights the negligible difference in the area for a chosen upper bound of 115 g rather than the theoretical upper bound of 'infinity' grams!*



```
normCdf(80,115,70,10)                  0.158652
normCdf(80,∞,70,10)                    0.158655
0.15865525956313−0.15865185876007
                                       0.000003
```

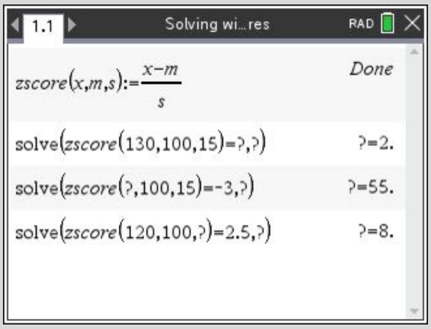
Solving with z-scores and a 'wildcard' symbol

Test scores on the 'Weschler Adult Intelligence Scale' (often used in so-called 'IQ tests') are normally distributed with mean 100 marks and standard deviation 15 marks.

(a) What z–score corresponds to a test score of 130?

(b) What test score corresponds to a z–score of –3?

(c) In a modified version of the above test, test scores are normally distributed with mean 100, but the standard deviation is different. If a test score of 120 now corresponds to a z–score of 2.5, what is the standard deviation for the modified test?



$zscore(x,m,s):=\frac{x-m}{s}$ Done

solve($zscore$(130,100,15)=?,?) ?=2.

solve($zscore$(?,100,15)=-3,?) ?=55.

solve($zscore$(120,100,?)=2.5,?) ?=8.

The formula for standard normal score (z–score) is $z=\frac{x-\bar{x}}{s_x}$.

To construct a calculator formula for the z–score, on a **Calculator** page:

- Enter $\boldsymbol{zscore(x,m,s):=\frac{x-m}{s}}$.

(a) To find the z–score for a test score of 130:

- Enter $\boldsymbol{solve(zscore(130,100,15)=?,?)}$.

(b) To find the *test score* for a z–score of –3:

- Enter $\boldsymbol{solve(zscore(?,100,15)=-3,?)}$.

(c) To find the new standard deviation for a test score of 120:

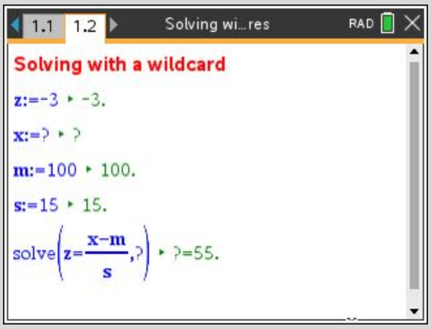
- Enter $\boldsymbol{solve(zscore(120,100,?)=2.5,?)}$.

Answer: (a) $z = 2$ **(b)** Test score = 55 **(c)** $s_x = 8$

Note: *The '?' symbol can be found by pressing* ?!▸*.*
It is referred to here as a 'wildcard' symbol, as it can be used to represent whichever one of the variables has an unknown value. In this way the standard formula can be used, substituting the wildcard symbol for the unknown variable in the solve command.

To create a **Notes** page utilising the same approach, add a **Notes** page and then (using the example from part **(b)** above):

- Enter the title text as shown.
- Press ctrl M to insert a Maths Box and then:
  - Enter $\boldsymbol{z:=-3}$
  - Enter $\boldsymbol{x:=?}$
  - Enter $\boldsymbol{m:=100}$
  - Enter $\boldsymbol{s:=15}$
  - Enter $\boldsymbol{solve(z=\frac{x-m}{s},?)}$.



Solving with a wildcard

z:=-3 ▸ -3.

x:=? ▸ ?

m:=100 ▸ 100.

s:=15 ▸ 15.

solve$\left(z=\frac{x-m}{s},?\right)$ ▸ ?=55.

To solve for a particular z–score, test score, mean or standard deviation, enter the known values, and then enter '?' for the unknown value. The solution will be visible in the last line.

Note: *This method of using the wildcard symbol (?) can be used for solving most formulas used in the General Mathematics course.*

3.2. Investigating association between two variables

3.2.1. Correlation

Finding the correlation coefficient

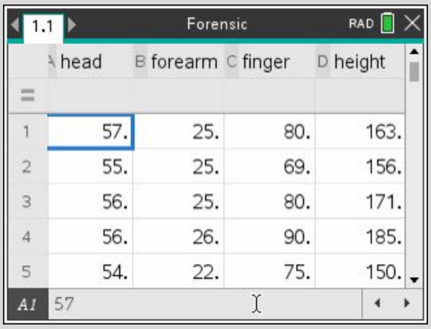
Four body measurements were taken from 15 students. The following table displays this data.

Head circumference (cm)	57	55	56	56	54	58	57	57	59	59	57	61	58	59	57
Forearm length (cm)	25	25	25	26	22	27	25	27	26	23	24	28	27	23	24
Middle finger length (mm)	80	69	80	90	75	89	80	82	79	80	78	85	95	74	85
Height (cm)	163	156	171	185	150	169	150	172	175	169	166	188	179	162	170

(a) Find the correlation between each pair of variables (using Pearson's correlation coefficient *r*).
(b) Which pair of variables have the highest correlation?

(a) To enter the above data, on a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the name ***head***.
- In the column B heading cell, enter the name ***forearm***.
- In the column C heading cell, enter the name ***finger***.
- In the column D heading cell, enter the name ***height***.
- Enter the data for the four variables into columns A to D.



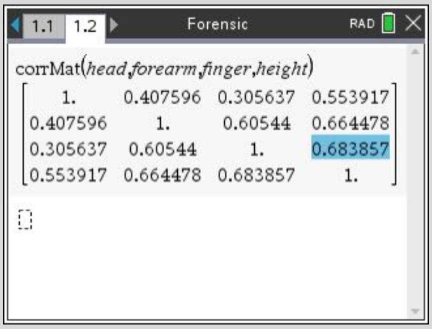
	A head	B forearm	C finger	D height
=				
1	57.	25.	80.	163.
2	55.	25.	69.	156.
3	56.	25.	80.	171.
4	56.	26.	90.	185.
5	54.	22.	75.	150.

A1 57

To calculate the values of *r*, add a **Calculator** page and then:

- Press menu > **Statistics** > **Stat Calculations** > **Correlation Matrix.**
- Press var and select each variable in turn, separated by the comma symbol (as shown right).

This will display a matrix of the values of *r* for each pair of variables (in the order that the variables were entered).



```
corrMat(head,forearm,finger,height)
[    1.     0.407596  0.305637  0.553917]
[0.407596     1.      0.60544   0.664478]
[0.305637  0.60544       1.     0.683857]
[0.553917  0.664478  0.683857      1.   ]
```

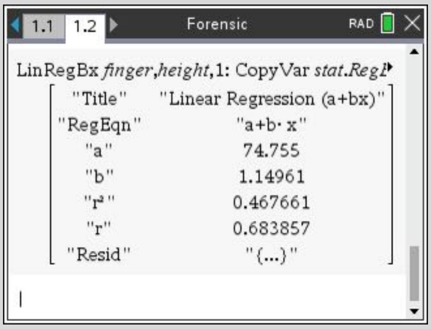
Note: For 4 variables, there will be 6 variable pairings. The main diagonal values of r reflect perfect correlations, as they are pairings of the same variable. There are repeated values due to the symmetry properties of a square matrix (e.g. if X is the correlation matrix, then $X_{1,2} = X_{2,1}$*).*

(b) From the correlation matrix screen shown right, the highlighted value of $r = 0.683857$ is in a cell located in column 4 and row 3, and so this value represents the correlation between the variables with the highest correlation, *height* and *finger*.

The value of the correlation coefficient for a pair of numerical variables can also be found by calculating the least squares regression line. For instance, for the two variables *height* (response variable) and *finger* (explanatory variable):

- Press menu > **Statistics** > **Stat Calculations** > **Linear Regression (a+bx)**.
- For 'X List', select ***finger***.
- For 'Y List', select ***height***.

This will display both the equation of the least squares regression line and the values of *r* and r^2.



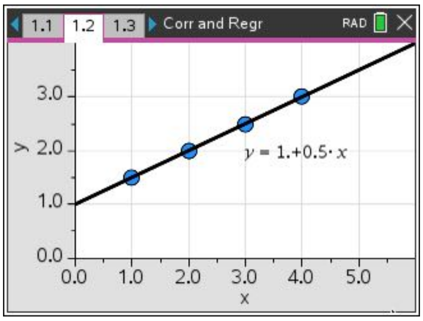
```
LinRegBx finger,height,1: CopyVar stat.RegE▸
[  "Title"    "Linear Regression (a+bx)" ]
[ "RegEqn"           "a+b· x"            ]
[   "a"              74.755              ]
[   "b"              1.14961             ]
[   "r²"             0.467661            ]
[   "r"              0.683857            ]
[  "Resid"           "{...}"             ]
```

3.2.2. Least squares regression

Linking correlation and least-squares regression

Consider the following dataset, where the regression line (using form $y = a + bx$) with equation $y = 1 + 0.5x$ is a perfect linear model (that is, $r = 1$).

x	1	2	3	4
y	1.5	2	2.5	3



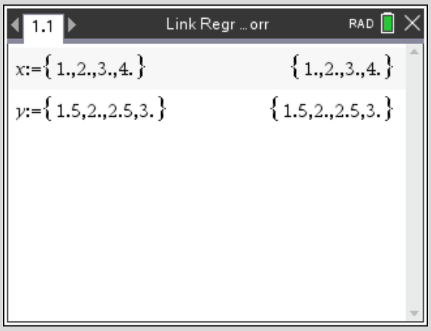
The scatter plot and regression line are shown right, highlighting that the line passes perfectly through the four points.

(a) Calculate the values of s_x and s_y, then calculate the ratio $\frac{s_y}{s_x}$. What do you notice?

(b) Calculate the value of $\bar{x}$ and $\bar{y}$. then calculate the value of $\bar{y} - 0.5\,\bar{x}$. What do you notice?

To enter the data, on a **Calculator** page:

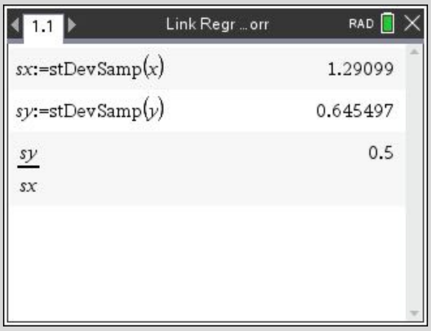
- Type the variable name ***x***.
- Press ctrl ⊞ to enter the 'Assign' symbol.
- Press ctrl) to enter the braces (set brackets).
- Enter the ***x*** data as shown.
- Repeat the above steps for the variable ***y***.



```
x:={1.,2.,3.,4.}                {1.,2.,3.,4.}
y:={1.5,2.,2.5,3.}          {1.5,2.,2.5,3.}
```

(a) To calculate the standard deviation of ***x*** and ***y***:

- Type the variable name ***sx***.
- Press ctrl ⊞ to enter the 'Assign' symbol.
- Press menu > **Statistics > List Maths > Sample Standard Deviation.**
- Press var and select the variable ***x***.
- Repeat these steps for the variable ***sy***.
- Enter ***sy***/***sx*** to calculate the value of the ratio $\frac{s_y}{s_x}$

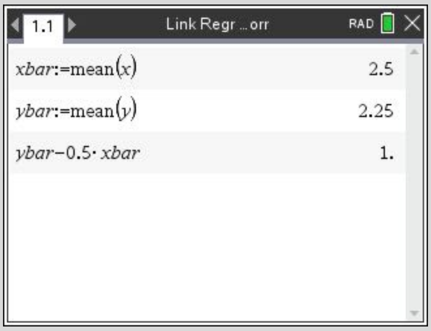


```
sx:=stDevSamp(x)         1.29099
sy:=stDevSamp(y)        0.645497
sy/sx                        0.5
```

Answer: The value of the ratio $\frac{s_y}{s_x} = 0.5$. This is the same as the value of the gradient b of the regression line.

(b) To calculate the mean of ***x*** and ***y***:

- Type the variable name ***xbar***.
- Press ctrl ⊞ to enter the 'Assign' symbol.
- Press menu > **Statistics > List Maths > Mean.**
- Press var and select the variable ***x***.
- Repeat these steps for the variable ***ybar***.



```
xbar:=mean(x)          2.5
ybar:=mean(y)         2.25
ybar−0.5·xbar           1.
```

To calculate the value if $\bar{y} - 0.5\,\bar{x}$:

- Enter ***ybar – 0.5xbar***.

Answer: The value of $\bar{y} - 0.5\,\bar{x} = 1$. This is the same as the value of the y-intercept a of the regression line.

... continued

Linking correlation and least-squares regression (continued)

***Note**: The following comments and screens are included to illustrate how selected summary statistics can be used to show the link between correlation and the associated least squares regression equation. It may be of interest to show students this link.*

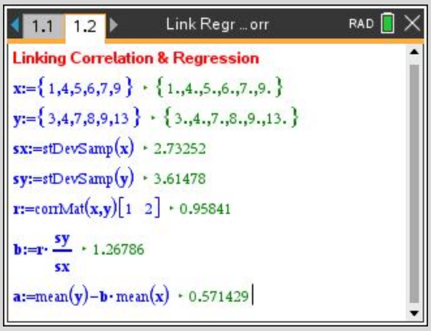
For the preceding 'perfect' correlation example, the values of least-squares regression parameters for the slope (b) and y-intercept (a) were linked, with $b = \frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$.

In the more common scenario where the correlation between the two numeric variables is not perfect, the following linking formulae can be applied.

$$b = r\frac{s_y}{s_x} \text{ and } a = \bar{y} - b\bar{x}$$

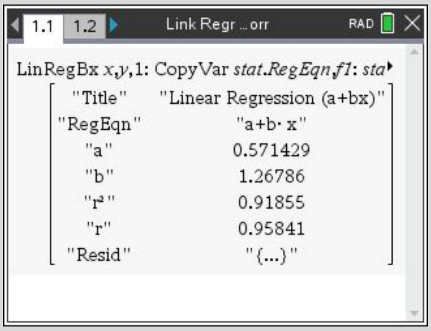
In the screen top right, a **Notes** page has been constructed to calculate the values of the slope (b) and y-intercept (a) based on the mean and standard deviation values of x and y, along with the value of the correlation coefficient r.

For comparison, the screen right, shows the same results on a **Calculator** page, obtaining by pressing menu **> Statistics > Stat Calculations > Linear Regression (a + bx).** This shows that the same values for the slope (b) and y-intercept (a) are obtained via the least-squares regression algorithm directly.



Linking Correlation & Regression

```
x:={1,4,5,6,7,9} ▸ {1.,4.,5.,6.,7.,9.}
y:={3,4,7,8,9,13} ▸ {3.,4.,7.,8.,9.,13.}
sx:=stDevSamp(x) ▸ 2.73252
sy:=stDevSamp(y) ▸ 3.61478
r:=corrMat(x,y)[1 2] ▸ 0.95841
b:=r·sy/sx ▸ 1.26786
a:=mean(y)−b·mean(x) ▸ 0.571429
```



```
LinRegBx x,y,1: CopyVar stat.RegEqn,f1: sta▸
 "Title"     "Linear Regression (a+bx)"
 "RegEqn"    "a+b·x"
 "a"         0.571429
 "b"         1.26786
 "r²"        0.91855
 "r"         0.95841
 "Resid"     "{...}"
```

Finding the least-square regression line using the List & Spreadsheet App

Length measurements of the femur bone (in the leg) and humerus (upper arm) were made on fossils of a particular species. A researcher is interested in whether the length of the femur is a good predictor of the length of the humerus (so the length of the femur is the explanatory variable).

Length of femur (cm)	59	56	64	38	74	50
Length of humerus (cm)	70	63	72	41	84	53

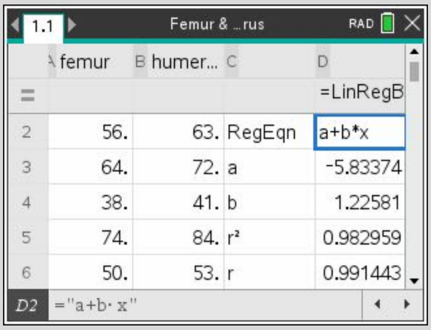
Find the least squares regression equation which can be used to predict humerus length from femur length. Then use this to predict the length of the humerus for a fossil whose femur has a known length of 80 cm.

To enter the above data, on a **Lists & Spreadsheet** page:

- In the column A heading cell, enter the name ***femur***.
- In the column B heading cell, enter the name ***humerus***.
- Enter the data for the four variables into columns A to B.

To find the equation of the least squares regression line:

- Press menu > **Statistics > Stat Calculations > Linear Regression (a + bx).**
- For 'X List', click and select the variable **femur**.
- For 'Y List', click and select the variable **humerus**.
- Click **Ok** to display the results in columns C and D



	A femur	B humer...	C	D
=				=LinRegB
2	56.	63.	RegEqn	a+b*x
3	64.	72.	a	-5.83374
4	38.	41.	b	1.22581
5	74.	84.	r²	0.982959
6	50.	53.	r	0.991443

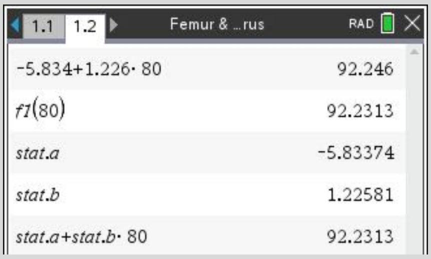
D2 ="a+b· x"

Answer: The equation of the least squares regression line is approximately (to 4 significant figures):

$$\textit{length of humerus} = -5.834 + 1.226 \times \textit{length of femur}\,.$$

Using a **Calculator** page, if the length of the femur is 80 cm, the predicted length of the humerus is:

$$\textit{length of humerus} = -5.834 + 1.226 \times (80) \approx 92 \text{ cm}.$$



```
-5.834+1.226· 80          92.246
f1(80)                    92.2313
stat.a                   -5.83374
stat.b                    1.22581
stat.a+stat.b· 80         92.2313
```

***Note**: The least squares regression equation by default will be stored in the function variable **f1**. On a **Calculator** page, entering **f1(80)** will calculate the predicted humerus length for a femur length of 80 cm. The answer obtained will be more accurate than the answer obtained using rounded values for **a** and **b**. Note also that you can access the values of **a** and **b** via the* var *key (see screen right). The variables **stat.a** and **stat.b** contain the values of the y-intercept and slope respectively obtained from the most recent statistical calculation.*

The importance of visualising for understanding

Here are four famous datasets (*Anscombe's Quartet*) that highlight the need to visualise the association between two numerical variables, rather than solely relying on summary statistics.

Dataset 1

*x*1	10	8	13	9	11	14	6	4	12	7	5
*y*1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Dataset 2

*x*2	10	8	13	9	11	14	6	4	12	7	5
*y*2	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Dataset 3

*x*3	10	8	13	9	11	14	6	4	12	7	5
*y*3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Dataset 4

*x*4	8	8	8	8	8	8	8	19	8	8	8
*y*4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

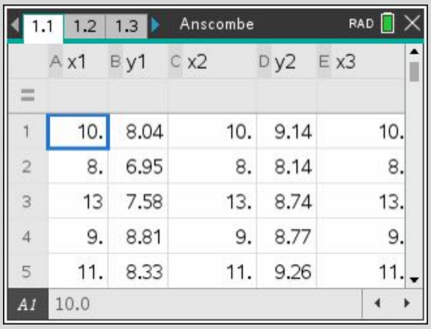
(a) For each of the datasets shown, find:

- the mean and standard deviation of each of the variables in the dataset
- the correlation coefficient and the equation of the least squares regression line

(b) Construct a scatter plot for each data set and compare these plots.

(a) To enter the above data, on a **Lists & Spreadsheet** page:

- In the column heading cells A to H, enter the variable names ***x*1**, ***y*1**, ***x*2**, ***y*2**, ***x*3**, ***y*3**, ***x*4**, ***y*4**.
- Enter the data for the eight variables into columns A to H.



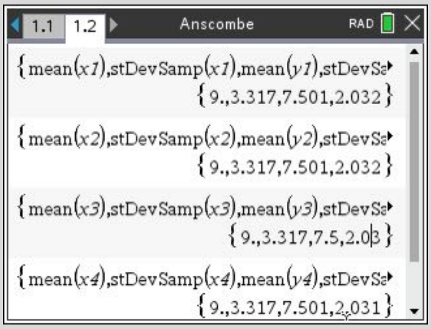
	A x1	B y1	C x2	D y2	E x3
=					
1	10.	8.04	10.	9.14	10.
2	8.	6.95	8.	8.14	8.
3	13	7.58	13.	8.74	13.
4	9.	8.81	9.	8.77	9.
5	11.	8.33	11.	9.26	11.

A1 10.0

To calculate the mean and standard deviation of each variable, add a **Calculator** page and then:

- Press ctrl) to enter a set of braces (set brackets).
- Enter the command **{mean(*x*1),stDevSamp(*x*1),mean(*y*1),stDevSamp(*y*1)}**.

Note: *The mean and sample standard deviation commands can be found via* menu ***> Statistics > List Maths***.

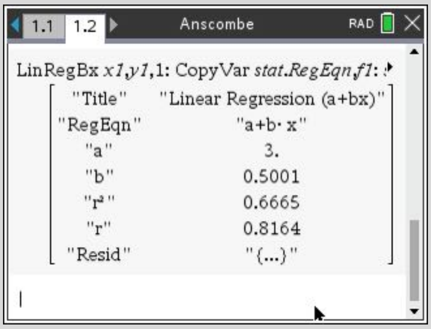


```
{mean(x1),stDevSamp(x1),mean(y1),stDevSa▸
                        {9.,3.317,7.501,2.032}
{mean(x2),stDevSamp(x2),mean(y2),stDevSa▸
                        {9.,3.317,7.501,2.032}
{mean(x3),stDevSamp(x3),mean(y3),stDevSa▸
                          {9.,3.317,7.5,2.03}
{mean(x4),stDevSamp(x4),mean(y4),stDevSa▸
                        {9.,3.317,7.501,2.031}
```

Answer: The means of the *x*-variables have the same mean and standard deviation $\left(\bar{x} = 9, s_x \approx 3.32\right)$, and the *y*-variables have approximately the same mean and standard deviation $\left(\bar{y} = 7.5, s_y \approx 2.03\right)$.

To calculate the correlation coefficients and the equation of the least squares regression lines:

- Press menu > **Statistics** > **Stat Calculations** > **Linear Regression (a+bx)**.
- For X list, select ***x1***.
- For Y List, select ***y1***.
- Repeat this for each dataset.



```
LinRegBx x1,y1,1: CopyVar stat.RegEqn,f1: ▸
   [   "Title"    "Linear Regression (a+bx)" ]
   [ "RegEqn"            "a+b· x"             ]
   [   "a"                 3.                 ]
   [   "b"               0.5001               ]
   [   "r²"              0.6665               ]
   [   "r"               0.8164               ]
   [ "Resid"             "{...}"              ]
```

Answer: The values of the correlation coefficient $(r \approx 0.816)$ and least squares regression equation ($y = 3 + 0.5x$) for each dataset are approximately equal.

... continued

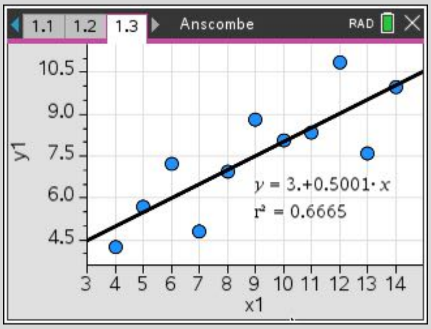
The importance of visualising for understanding (continued)

(b) To display the scatter plot for Dataset 1, add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***x1***.
- Press tab to activate **Click to add variable** to the left of the vertical axis and select the variable ***y1***.
- To show the least squares regression line, press menu > **Analyse** > **Regression** > **Show Linear (a + bx)**.

This will display a scatter plot for Dataset 1, with the least squares line drawn over the plot.

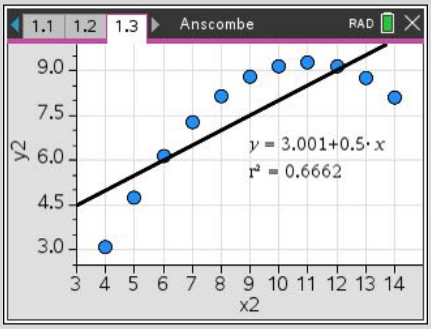
Note: *Click on the line to display the equation of the least squares regression line and the coefficient of determination. Press* menu *>* ***Settings*** *and adjust the* ***Display Digits*** *as required. The* ***Diagnostics*** *tick box is to show/hide the coefficient of determination.*



To display the scatter plot for Dataset 2:

- Click the variable underneath the horizontal axis and select the variable ***x2***.
- Click the variable to the left of the vertical axis and select the variable ***y2***.

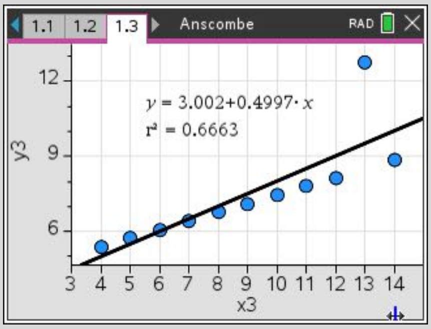
This scatter plot reveals a clear non-linear pattern not obvious from the summary statistics calculated.



To display the scatter plot for Dataset 3:

- Click the variable underneath the horizontal axis and select the variable ***x3***.
- Click the variable to the left of the vertical axis and select the variable ***y3***.

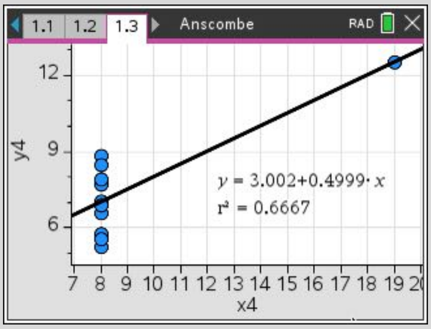
This scatter plot reveals the impact of an outlier point on the summary statistics for an otherwise perfect linear pattern.



To display the scatter plot for Dataset 4:

- Click the variable underneath the horizontal axis and select the variable ***x4***.
- Click the variable to the left of the vertical axis and select the variable ***y4***.

This scatter plot reveals clustering of all but one point around a set of points with the same value of the explanatory variable.



3.3. Investigating and modelling linear associations

3.3.1. Analysing linear models

Constructing a template file for analysing linear association

Consider the following bivariate dataset. Assume that *x* is the explanatory variable.

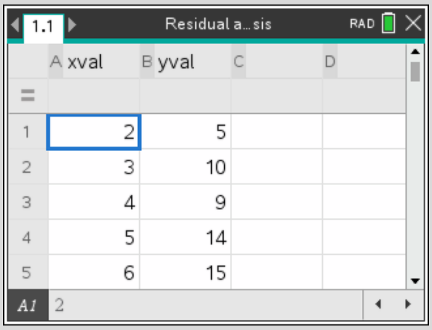
x	2	3	4	5	6
y	5	10	9	14	15

(a) Construct a scatter plot for each data set and compare these plots.
(b) Find the equation of the least squares regression line and the coefficient of determination
(c) Display a plot of the residual values, and comment on whether a linear model is reasonable.

(a) To enter the above data, on a **Lists & Spreadsheet** page:

- In the column heading cells A and B, enter the variable names ***xval*** and ***yval***.
- Enter the data for the two variables into columns A and B.

***Note**: If column(s) A and B are not empty (that is, there is data in column under **xval** and/or **yval**), press the ▲ key until the entire column is selected, and then press* menu *> **Data** > **Clear Data**. This ensures that the previous data is removed.*

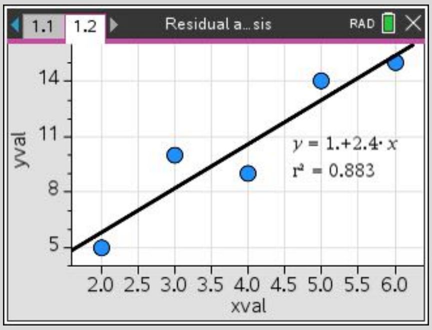


	A xval	B yval	C	D
=				
1	2	5		
2	3	10		
3	4	9		
4	5	14		
5	6	15		

A1 2

(b) To display the scatter plot and add a least squares regression line, add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***xval***.
- Press tab to activate **Click to add variable** to the left of the vertical axis and select the variable ***yval***.
- To show the least squares regression line, press menu > **Analyse** > **Regression** > **Show Linear (a + bx)**

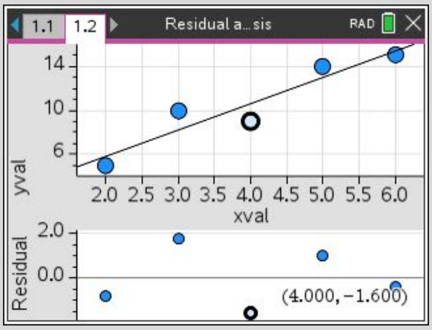


This will display a scatter plot, with the least squares line drawn over the plot. Click on the line to display the equation of the least squares regression line and the coefficient of determination.

***Note**: If the coefficient of determination is not displayed, press* menu *> **Settings** and click the **Diagnostics** option.*

(c) To display a residual plot, press menu > **Analyse** > **Residuals** > **Show Residual Plot.**

***Note**: Clicking a point on the residual plot highlights the associated point on the scatter plot, and displays the residual plot. For example, in the screen shown right, the point at (4,9) on the scatter plot has an associated residual value of –1.6, meaning that the actual y-value is 1.6 units below the predicted value for x = 4.*



Answer: The residuals appear to be small and randomly placed about zero, and so a linear model is appropriate.

3.3.2. Analysing non-linear models

Transforming variables to identify a better model

A stone is dropped from a bridge that is 50 metres above the water, and a photo is taken (from the side) every 0.5 seconds, yielding the data shown about the height of the stone above the water over the first three seconds.

Time elapsed (*s*)	0.5	1.0	1.5	2.0	2.5	3.0
Height of stone (*m*)	49	45	39	30	20	6

(a) Construct a scatterplot and fit a regression line to this data. Plot the residuals and use this to comment on the suitability of the linear regression model.

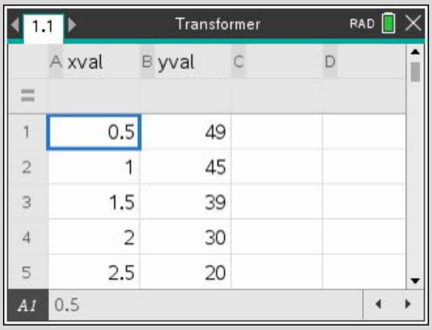
(b) Construct formulas to transform the explanatory and response variables so that non-linear regression models can be checked. Include squared, logarithmic and reciprocal transformations of both the explanatory and response variables.

(c) It has been suggested that a non-linear model *height* = $a + b \times$ *time*^{2} may fit the model well. Check this model and comment on its suitability.

(a) To enter the above data, on a **Lists & Spreadsheet** page:

- In the column heading cells A and B, enter the variable names ***xval*** and ***yval***.
- Enter the data for the two variables into columns A and B.

Note: *If column(s) A and B are not empty (that is, there is data in column under* ***xval*** *and/or* ***yval****), press the ▲ key until the entire column is selected, and then press* menu *>* ***Data*** *>* ***Clear Data****. This ensures that the previous data is removed.*



	A xval	B yval	C	D
=				
1	0.5	49		
2	1	45		
3	1.5	39		
4	2	30		
5	2.5	20		

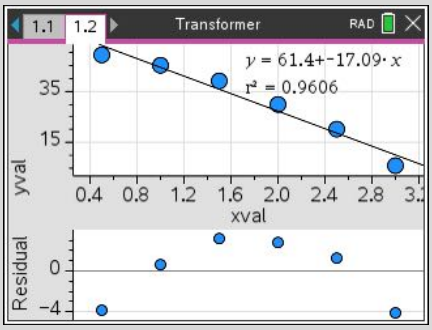
A1 0.5

To display the scatter plot and add a least squares regression line, add a **Data & Statistics** page, and then:

- Press tab to activate **Click to add variable** underneath the horizontal axis and select the variable ***xval***.
- Press tab to activate **Click to add variable** to the left of the vertical axis and select the variable ***yval***.

To show the least squares regression line and display the residual plot:

- Press menu > **Analyse** > **Regression** > **Show Linear (a+bx)**.
- Press menu > **Analyse** > **Residuals** > **Show Residual Plot**.



This will display a scatter plot, with the least squares line drawn over the plot. Click on the line to display the equation of the least squares regression line and the coefficient of determination.

Note: *If the coefficient of determination is not displayed, press* menu *>* ***Settings*** *and click the* ***Diagnostics*** *option.*

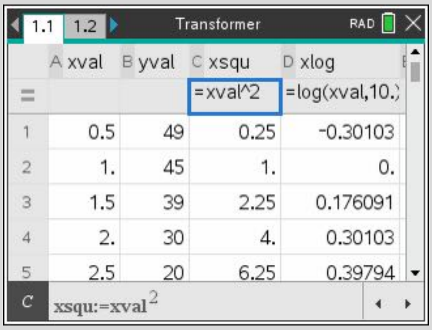
Answer: Despite the high value of $r^2 = 0.96$, the residuals form a curved pattern, suggesting that a non-linear model may better fit the data.

... continued

Transforming variables to identify a better model (continued)

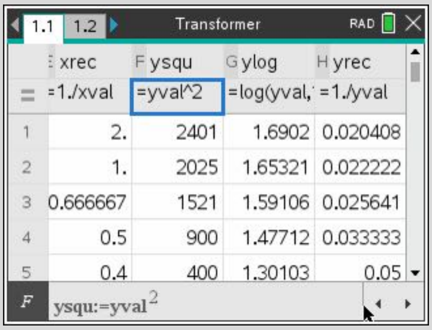
(b) Press ctrl ◀ to return to the **Lists & Spreadsheet** page, and then:

- Add the variable names for the transformed variables:
  - In the heading cell for column C, enter the name ***xsqu***
  - In the heading cell for column D, enter the name ***xlog***
  - In the heading cell for column E, enter the name ***xrec***
  - In the heading cell for column F, enter the name ***ysqu***
  - In the heading cell for column G, enter the name ***ylog***
  - In the heading cell for column H, enter the name ***yrec***.
- Add the column formulas for each variable as shown:
  - In the formula cell for column C, enter the formula ***xsqu:=xval²***
  - In the formula cell for column D, enter the formula ***xlog:=log(xval,10.0)***
  - In the formula cell for column E, enter the formula ***xrec:=1.0/xval***
  - In the formula cell for column F, enter the formula ***ysqu:=yval²***
  - In the formula cell for column G, enter the formula ***ylog:=log(yval,10.0)***
  - In the formula cell for column H, enter the formula ***yrec:=1.0/yval***.



	A xval	B yval	C xsqu	D xlog
=			=xval^2	=log(xval,10.)
1	0.5	49	0.25	-0.30103
2	1.	45	1.	0.
3	1.5	39	2.25	0.176091
4	2.	30	4.	0.30103
5	2.5	20	6.25	0.39794

C xsqu:=xval²



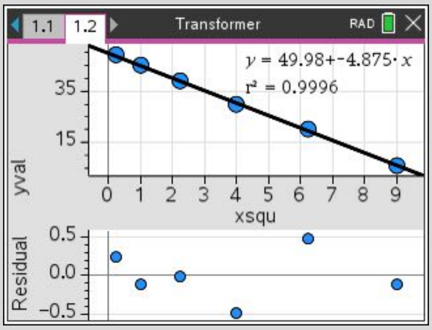
	E xrec	F ysqu	G ylog	H yrec
=	=1./xval	=yval^2	=log(yval,	=1./yval
1	2.	2401	1.6902	0.020408
2	1.	2025	1.65321	0.022222
3	0.666667	1521	1.59106	0.025641
4	0.5	900	1.47712	0.033333
5	0.4	400	1.30103	0.05

F ysqu:=yval²

Note: The use of decimal point in some of the formulas is just to force the data to be displayed as decimal approximations. If the calculator setting is in APPROX calculation mode, then this is not necessary.

(c) Press ctrl ▶ to return to the **Data & Statistics** page, and then:

- Click on the variable label ***xval*** (the variable on the horizontal axis).
- In the pop-up box that follows, select the variable ***xsqu***.
- Press menu > **Window/Zoom** > **Zoom – Data**.



Answer: The plot of height values against the (time)2 values linearises the plot very well, and the coefficient of determination is very high ($r^2 = 0.9996$). The residuals seem small and randomly distributed around zero, so the transformed model *height = 49.98 – 4.875 × time*^{2} seems to fit the data well.

Note: The number of display digits given for the regression equation and the coefficient of determination can be changed by pressing menu *> **Settings** and modifying the number of display digits.*