CHAPTER 9

Statistics

9.1 Revision EMCJK

Terminology EMCJM

Measures of central tendency:

Provide information on the data values at the centre of the data set.

- The **mean** is the 'average' value of a data set. It is calculated as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  where the x_i are the data and n is the number of data entries. We read $\bar{x}$ as "x bar".

- The **median** is the middle value of an ordered data set. To find the median, we first sort the data in ascending or descending order and then pick out the value in the middle of the sorted list. If the middle is in between two values, the median is the average of those two values.

Measures of dispersion:

Tell us how spread out a data set is. If a measure of dispersion is small, the data are clustered in a small region. If a measure of dispersion is large, the data are spread out over a large region.

- The **range** is the difference between the maximum and minimum values in the data set.

- The **inter-quartile range** is the difference between the first and third quartiles of the data set. The quartiles are computed in a similar way to the median. The median is halfway into the ordered data set and is sometimes also called the second quartile. The first quartile is one quarter of the way into the ordered data set, whereas the third quartile is three quarters of the way into the ordered data set.

  If you begin numbering your ordered data set with the number 1, the formulae for the location of each quartile are as follows:

$$\text{Location of } Q_1 = \frac{1}{4}(n-1)+1$$
$$\text{Location of } Q_2 = \frac{1}{2}(n-1)+1$$
$$\text{Location of } Q_3 = \frac{3}{4}(n-1)+1$$

- The **variance** of the data is the average squared distance between the mean and each data value.

  The variance of the data is

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

  in a population of n elements, $\{x_1; x_2; \ldots; x_n\}$, with a mean of $\bar{x}$.

- The **standard deviation** measures how spread out the values in a data set are around the mean. More precisely, it is a measure of the average distance between the values of the data in the set and the mean.

  The standard deviation of the data is

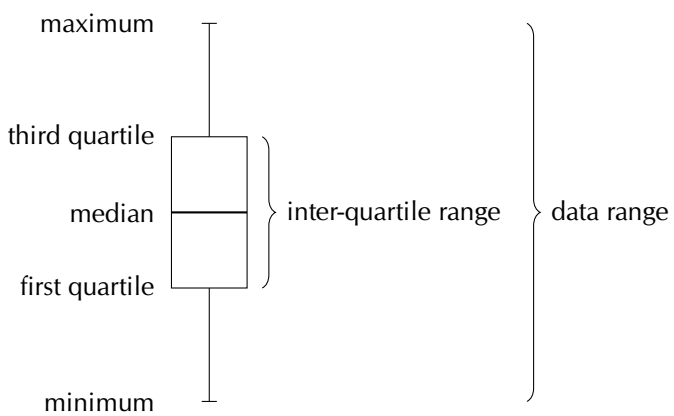
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

  in a population of n elements, $\{x_1; x_2; \ldots; x_n\}$, with a mean of $\bar{x}$.

The **five number summary** combines a measure of central tendency, the median, with measures of dispersion, namely the range and the inter-quartile range. More precisely, the five number summary is written in the following order:

- minimum;
- first quartile;
- median;
- third quartile;
- maximum.

The five number summary is often presented visually using a **box and whisker diagram**, illustrated below.



See video: 29BS at www.everythingmaths.co.za

Worked example 1: Five number summary

QUESTION

Draw a box and whisker diagram for the following data set:

$$1{,}25\ ;\ 1{,}5\ ;\ 2{,}5\ ;\ 2{,}5\ ;\ 3{,}1\ ;\ 3{,}2\ ;\ 4{,}1\ ;\ 4{,}25\ ;\ 4{,}75\ ;\ 4{,}8\ ;\ 4{,}95\ ;\ 5{,}1$$

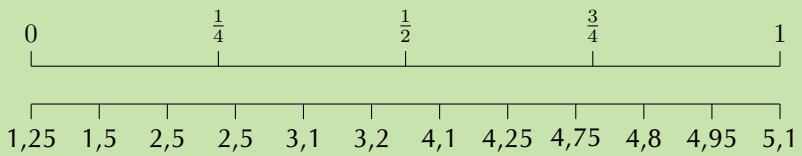
SOLUTION

Step 1: Determine the minimum and maximum

Since the data set is already ordered, we can read off the minimum as the first value ($1{,}25$) and the maximum as the last value ($5{,}1$).

Step 2: Determine the quartiles

There are 12 values in the data set. We can use the figure below or the formulae to determine where the quartiles are located.



Using the figure above we can see that the median is between the sixth and seventh values. We can confirm this using the formula:

$$\begin{aligned}\text{Location of } Q_2 &= \frac{1}{2}(n-1)+1\\ &= \frac{1}{2}(11)+1\\ &= 6{,}5\end{aligned}$$

Therefore, the value of the median is

$$\frac{3{,}2+4{,}1}{2} = 3{,}65$$

The first quartile lies between the third and fourth values. We can confirm this using the formula:

$$\begin{aligned}\text{Location of } Q_1 &= \frac{1}{4}(n-1)+1\\ &= \frac{1}{4}(11)+1\\ &= 3{,}75\end{aligned}$$

Therefore, the value of the first quartile is

$$Q_1 = \frac{2{,}5+2{,}5}{2} = 2{,}5$$

The third quartile lies between the ninth and tenth values. We can confirm this using the formula:

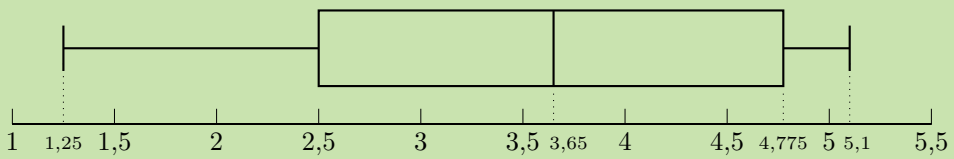
$$\begin{aligned}\text{Location of } Q_3 &= \frac{3}{4}(n-1)+1\\ &= \frac{3}{4}(11)+1\\ &= 9{,}25\end{aligned}$$

Therefore, the value of the third quartile is

$$Q_3 = \frac{4{,}75+4{,}8}{2} = 4{,}775$$

Step 3: Draw the box and whisker diagram

We now have the five number summary as $(1{,}25;\ 2{,}5;\ 3{,}65;\ 4{,}775;\ 5{,}1)$. The box and whisker diagram representing the five number summary is given below.



See video: 29BT at www.everythingmaths.co.za

Worked example 2: Variance and standard deviation

QUESTION

You flip a coin 100 times and it lands on heads 44 times. You then use the same coin and do another 100 flips. This time in lands on heads 49 times. You repeat this experiment a total of 10 times and get the following results for the number of heads.

$$\{44; 49; 52; 62; 53; 48; 54; 49; 46; 51\}$$

For the data set above:

- Calculate the mean.
- Calculate the variance and standard deviation using a table.
- Confirm your answer for the variance and standard deviation using a calculator.

SOLUTION

Step 1: Calculate the mean

The formula for the mean is

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In this case, we sum the data and divide by 10 to get $\overline{x} = 50{,}8$.

Step 2: Calculate the variance using a table

The formula for the variance is

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

We first subtract the mean from each data point and then square the result.

x_i	44	49	52	62	53	48	54	49	46	51
$x_i - \overline{x}$	$-6{,}8$	$-1{,}8$	$1{,}2$	$11{,}2$	$2{,}2$	$-2{,}8$	$3{,}2$	$-1{,}8$	$-4{,}8$	$0{,}2$
$(x_i - \overline{x})^2$	$46{,}24$	$3{,}24$	$1{,}44$	$125{,}44$	$4{,}84$	$7{,}84$	$10{,}24$	$3{,}24$	$23{,}04$	$0{,}04$

The variance is the sum of the last row in this table divided by 10, so $\sigma^2 = 22{,}56$.

Step 3: Calculate the variance using a calculator

Using the SHARP EL-531VH calculator:

Using your calculator, change the mode from normal to "Stat x ". Do this by pressing [2ndF] and then 1. This mode enables you to type in univariate data.

Key in the data, row by row:

Enter:	Press:	See:
44	**DATA**	n = 1
49	**DATA**	n = 2
52	**DATA**	n = 3
62	**DATA**	n = 4
53	**DATA**	n = 5
48	**DATA**	n = 6
54	**DATA**	n = 7
49	**DATA**	n = 8
46	**DATA**	n = 9
51	**DATA**	n = 10

Note: The **[DATA]** button is the same as the [M+] button.

Get the value for σ_x:

Press:	Press:	See:
RCL	σx	$\sigma x = \pm 4{,}75$

$\therefore \sigma_x = \pm 4{,}75$ and ${\sigma_x}^2 = (4{,}75)^2 = 22{,}56$

Using the CASIO fx-82ZA PLUS calculator:

Switch on the calculator. Press [MODE] and then select STAT by pressing [2]. The following screen will appear:

1:	$1 - VAR$	2:	$A + BX$
3:	$_{+}CX^2$	4:	lnX
5:	eX	6:	$A.BX$
7:	$A.XB$	8:	$1/X$

Now press [1] for variance and standard deviation. Your screen should look something like this:

	X
1	
2	
3	

Press [44] and then [=] to enter the first x-value under x. Then enter the other values in the same way.

	X
1	44
2	49
3	52

Then press [AC]. The screen clears but the data remains stored.

Now press [SHIFT][1] to get the stats computations screen shown below.

1:	Type	2:	Data
3:	Sum	4:	Var
5:	MinMax		

Choose variance by pressing [4].

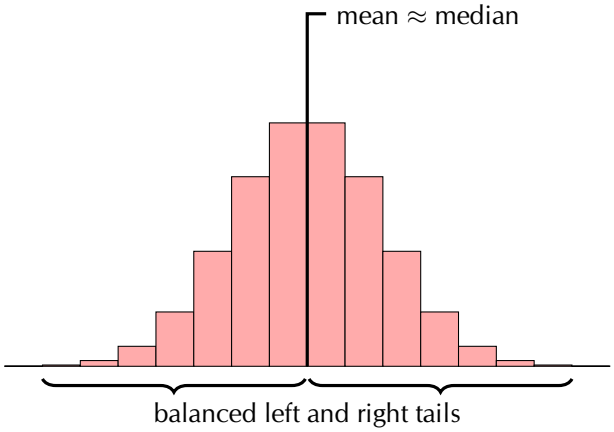
1:	n	2:	$\bar{x}$
3:	σx	4:	sx

Get the value for σ_x:

Press [3] and [=] to get the value of σx $\therefore \sigma_x = \pm 4{,}75$ and ${\sigma_x}^2 = (4{,}75)^2 = 22{,}56$

Last year you learnt about three shapes of data distribution: symmetric, left skewed and right skewed.

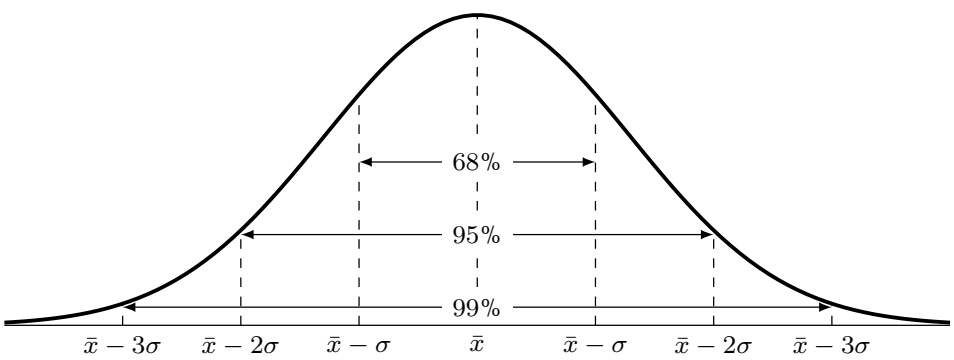
A *symmetric* distribution is one where the left and right hand sides of the distribution are roughly equally balanced around the mean. The histogram below shows a typical symmetric distribution.



For symmetric distributions, the mean is approximately equal to the median and the left and right tails are equally balanced, meaning that they have about the same length.

If large numbers of data are collected from a population, the graph will often have a bell shape. If the data was, say, examination results, a few learners usually get very high marks, a few very low marks and most get a mark in the middle range. This is a common type of symmetric data known as a *normal* distribution. We say a distribution is normal if

- the mean, median and mode are equal.
- it is symmetric around the mean.
- 68% of the sample lies within one standard deviation of the mean, 95% within two standard deviations and 99% within three standard deviations of the mean.



See video: 29BV at www.everythingmaths.co.za

What happens if the test was very easy or very difficult? Then the distribution may not be symmetrical. If extremely high or extremely low scores are added to a distribution, then the mean and median tend to shift towards these scores and the curve becomes skewed.

If the test was very difficult, the mean and median scores are shifted to the left. In this case, we say the distribution is *positively skewed*, or *skewed right*.

A distribution that is skewed right has the following characteristics:

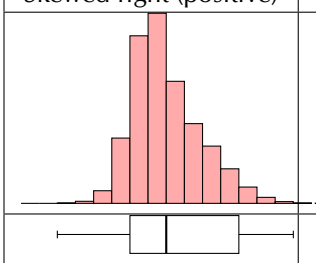
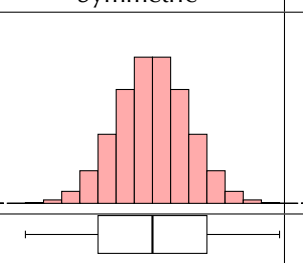
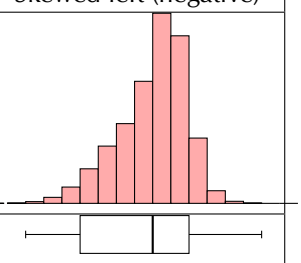
- the mean is typically more than the median;
- the tail of the distribution is longer on the right hand side than on the left hand side; and
- the median is closer to the first quartile than to the third quartile.

If the test was very easy, then many learners would get high scores, and the mean and median of the distribution would be shifted to the right. We say the distribution is *negatively skewed*, or *skewed left*.

A distribution that is skewed right has the following characteristics:

- the mean is typically less than the median;
- the tail of the distribution is longer on the left hand side than on the right hand side; and
- the median is closer to the third quartile than to the first quartile.

The table below summarises the different categories visually.

Skewed right (positive)	Symmetric	Skewed left (negative)
		
mean $>$ median	mean $\approx$ median	mean $<$ median

Worked example 3: Skewed and symmetric data

QUESTION

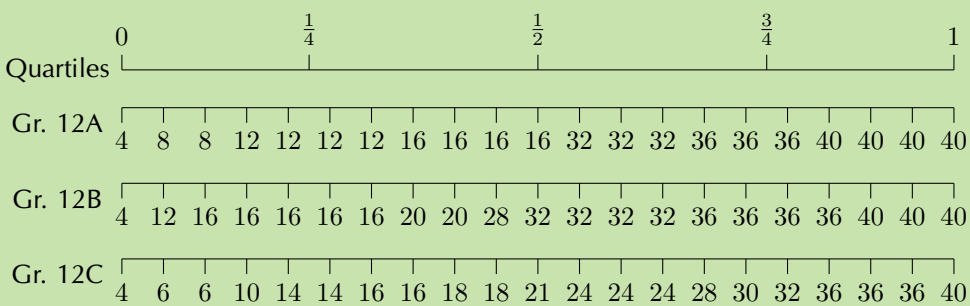
Three Matric classes wrote a Mathematics test. The test is out of 40 marks and each class has 21 learners. The results of the test are shown in the table on the next page.

Gr. 12A	Gr. 12B	Gr. 12C
4	4	4
8	12	6
8	16	6
12	16	10
12	16	14
12	16	14
12	16	16
16	20	16
16	20	18
16	28	18
16	32	21
32	32	24
32	32	24
32	32	24
36	36	28
36	36	30
36	36	32
40	36	36
40	40	36
40	40	36
40	40	40

1. For each class, determine the five number summary and draw a box and whisker diagram on the same set of axes using an appropriate scale.
2. Determine the mean and standard deviation for each class.
3. Comparing the mean and median values for each class, comment on the distribution of the test marks for each class.

SOLUTION

1. First, we order the data from smallest to largest. This has already been done for us. Then, we divide our data into quartiles:



The minimum of each data set is 4. The maximum of each data set is 40.

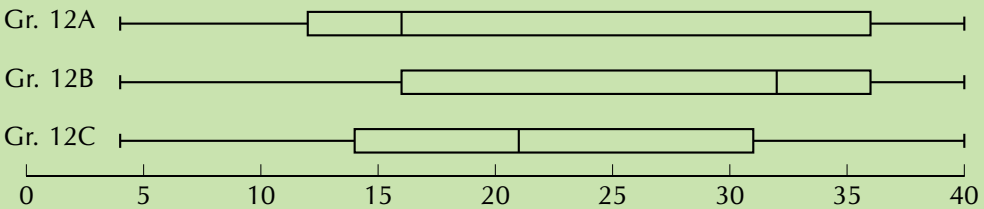
Since there are 21 values in the data set, the median lies on the eleventh mark, making it equal to 16 for Gr. 12A, 32 for Gr. 12B and 21 for Gr. 12C.

The first quartile lies between the fifth and sixth values, making it equal to 12 for Gr. 12A, 16 for Gr. 12B and 14 for Gr. 12C.

The third quartile lies between the 16^{th} and 17^{th} values, making it equal to 36 for Gr. 12A and Gr. 12B, and $\frac{30+32}{2} = 31$ for Gr. 12C.

Therefore, we are able to formulate the following five number summaries and subsequent box and whisker plots:

- Gr. 12A $= [4; 12; 16; 36; 40]$
- Gr. 12B $= [4; 16; 32; 36; 40]$
- Gr. 12C $= [4; 14; 21; 31; 40]$



2. Gr. 12A:

$$\text{mean } (\bar{x}) = \frac{496}{21} = 23{,}6$$

$$\text{standard deviation } (\sigma) = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \pm 12{,}70$$

Gr. 12B:

$$\text{mean } (\bar{x}) = \frac{556}{21} = 26{,}5$$

$$\text{standard deviation } (\sigma) = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \pm 10{,}65$$

Gr. 12C:

$$\text{mean } (\bar{x}) = \frac{453}{21} = 21{,}6$$

$$\text{standard deviation } (\sigma) = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \pm 10{,}54$$

3. If the mean is greater than the median, the data is typically positively skewed and if the mean is less than the median, the data is typically negatively skewed.

   Gr. 12A: mean $-$ median $= 23{,}6 - 16 = 7{,}6$. The marks for 12A are therefore positively skewed, meaning that there were many low marks in the class with the high marks being more spread out.

   Gr. 12B: mean $-$ median $= 26{,}5 - 32 = -5{,}5$. The marks for 12B are therefore negatively skewed, meaning that there were many high marks in the class with the low marks being more spread out.

   Gr. 12C: mean $-$ median $= 21{,}6 - 21 = 0{,}6$. The marks for 12C are therefore normally distributed, meaning that there are as many low marks in the class as there are high marks.

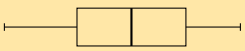
Exercise 9 – 1: Revision

1. State whether each of the following data sets are symmetric, skewed right or skewed left.

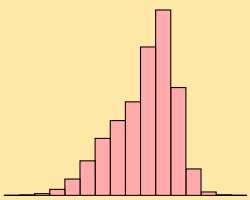
   a) A data set with this distribution:

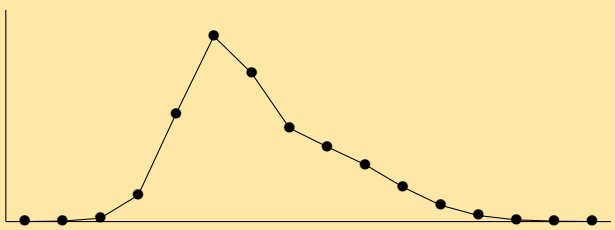
   b) A data set with this box and whisker plot:

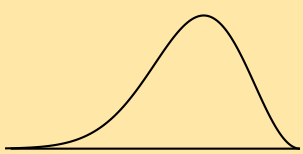
   c) A data set with this histogram:

   d) A data set with this frequency polygon:

   e) A data set with this distribution:

   f) The following data set:
   $105\ ;\ 44\ ;\ 94\ ;\ 149\ ;\ 83\ ;\ 178\ ;\ -4\ ;\ 112\ ;\ 50\ ;\ 188$

2. For the following data sets:
   - Determine the mean and five number summary.
   - Draw the box and whisker plot.
   - Determine the skewness of the data.

   a) $40\ ;\ 45\ ;\ 12\ ;\ 6\ ;\ 9\ ;\ 16\ ;\ 11\ ;\ 7\ ;\ 35\ ;\ 7\ ;\ 31\ ;\ 3$
   b) $65\ ;\ 100\ ;\ 99\ ;\ 21\ ;\ 8\ ;\ 27\ ;\ 21\ ;\ 31\ ;\ 33\ ;\ 31\ ;\ 38\ ;\ 16$
   c) $65\ ;\ 57\ ;\ 77\ ;\ 92\ ;\ 77\ ;\ 58\ ;\ 90\ ;\ 46\ ;\ 11\ ;\ 81$
   d) $1\ ;\ 99\ ;\ 76\ ;\ 76\ ;\ 50\ ;\ 74\ ;\ 83\ ;\ 91\ ;\ 41\ ;\ 17\ ;\ 33$
   e) $0{,}5;\ -0{,}9\ ;\ -1{,}8\ ;\ 3\ ;\ -0{,}2\ ;\ -5{,}2\ ;\ -1{,}8\ ;\ 0{,}1\ ;\ -1{,}7\ ;\ -2\ ;\ 2{,}2\ ;\ 0{,}5;\ -0{,}5$
   f) $86\ ;\ 64\ ;\ 25\ ;\ 71\ ;\ 54\ ;\ 44\ ;\ 97\ ;\ 31\ ;\ 78\ ;\ 46\ ;\ 60\ ;\ 86$

3. For the following data sets:
   - Determine the mean.
   - Use a table to determine the variance and the standard deviation.
   - Determine the percentage of data points within one standard deviation of the mean. Round your answer to the nearest percentage point.

   a) $\{9{,}1; 0{,}2; 2{,}8; 2{,}0; 10{,}0; 5{,}8; 9{,}3; 8{,}0\}$
   b) $\{9; 5; 1; 3; 3; 5; 7; 4; 10; 8\}$
   c) $\{81; 22; 63; 12; 100; 28; 54; 26; 50; 44; 4; 32\}$

4. Use a calculator to determine the
   - mean,
   - variance,
   - and standard deviation

   of the following data sets:

   a) $8 ; 3 ; 10 ; 7 ; 7 ; 1 ; 3 ; 1 ; 3 ; 7$
   b) $4 ; 4 ; 13 ; 9 ; 7 ; 7 ; 2 ; 5 ; 15 ; 4 ; 22 ; 11$
   c) $4{,}38 ; 3{,}83 ; 4{,}99 ; 4{,}05 ; 2{,}88 ; 4{,}83 ; 0{,}88 ; 5{,}33 ; 3{,}49 ; 4{,}10$
   d) $4{,}76 ; -4{,}96 ; -6{,}35 ; -3{,}57 ; 0{,}59 ; -2{,}18 ; -4{,}96 ; -3{,}57 ; -2{,}18 ; 1{,}98$
   e) $7 ; 53 ; 29 ; 42 ; 12 ; 111 ; 122 ; 79 ; 83 ; 5 ; 69 ; 45 ; 23 ; 77$

5. Xolani surveyed the price of a loaf of white bread at two different supermarkets. The data, in rands, are given below.

Supermarket A	3,96	3,76	4,00	3,91	3,69	3,72
Supermarket B	3,97	3,81	3,52	4,08	3,88	3,68

   a) Find the mean price at each supermarket and then state which supermarket has the lower mean.
   b) Find the standard deviation of each supermarket's prices.
   c) Which supermarket has the more consistently priced white bread? Give reasons for your answer.

6. The times for the 8 athletes who swam the 100 m freestyle final at the 2012 London Olympic Games are shown below. All times are in seconds.
   $47{,}52 ; 47{,}53 ; 47{,}80 ; 47{,}84 ; 47{,}88 ; 47{,}92 ; 48{,}04 ; 48{,}44$

   a) Calculate the mean time.
   b) Calculate the standard deviation for the data.
   c) How many of the athletes' times are more than one standard deviation away from the mean?

7. The following data set has a mean of $14{,}7$ and a variance of $10{,}01$.

   $$18 ; 11 ; 12 ; a ; 16 ; 11 ; 19 ; 14 ; b ; 13$$

   Calculate the values of a and b.

8. The height of each learner in a class was measured and it was found that the mean height of the class was $1{,}6$ m. At the time, three learners were absent. However, when the heights of the learners who were absent were included in the data for the class, the mean height did not change.

   If the heights of two of the learners who were absent are $1{,}45$ m and $1{,}63$ m, calculate the height of the third learner who was absent. [NSC Paper 3 Feb-March 2013]

9. There are 184 students taking Mathematics in a first-year university class. The marks, out of 100, in the half-yearly examination are normally distributed with a mean of 72 and a standard deviation of 9. [NSC Paper 3 Feb-March 2013]

   a) What percentage of students scored between 72 and 90 marks?

   b) Approximately how many students scored between 45 and 63 marks?

10. More questions. Sign in at Everything Maths online and click 'Practise Maths'.

Check answers online with the exercise code below or click on 'show me the answer'.

1a. 29BW	1b. 29BX	1c. 29BY	1d. 29BZ	1e. 29C2	1f. 29C3
2a. 29C4	2b. 29C5	2c. 29C6	2d. 29C7	2e. 29C8	2f. 29C9
3a. 29CB	3b. 29CC	3c. 29CD	4a. 29CF	4b. 29CG	4c. 29CH
4d. 29CJ	4e. 29CK	5. 29CM	6. 29CN	7. 29CP	8. 29CQ
9. 29CR					

 **www.everythingmaths.co.za**  **m.everythingmaths.co.za**

9.2 Curve fitting EMCJP

Intuitive curve fitting EMCJQ

In Grade 11, we used various means, such as histograms, frequency polygons and ogives, to visualise our data. These are very useful tools to depict **univariate** data, i.e. data with only one variable such as the height of learners in a class.

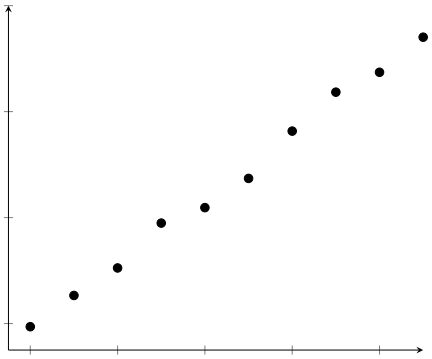
Last year we also learnt about a visual tool called scatter plots. **Scatter plots** are a common way to visualise **bivariate** data, i.e. data with two variables. This allows us to identify the *direction* and *strength* of a relationship between two variables.

We identify the nature of a relationship between two variables by examining if the points on the scatter plot conform to a linear, exponential, quadratic or some other function. The process of fitting functions to data is known as **curve fitting**.

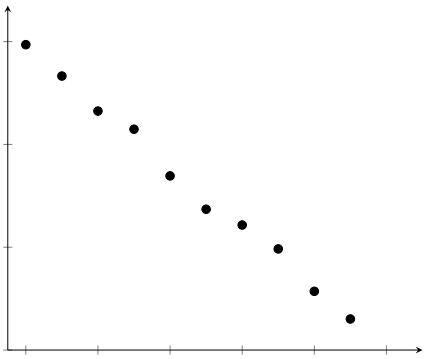
The strength of a relationship can be described as *strong* if the data points conform closely to a function or *weak* if they are further away.

In the case of linear functions, the direction of a relationship is *positive* if high values of one variable occur with high values of the other or *negative* if high values of one variable occur with low values of the other.

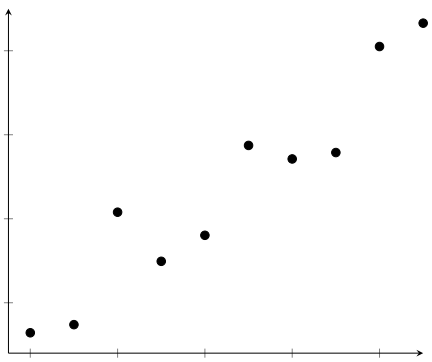
The different relationships are summarised below:



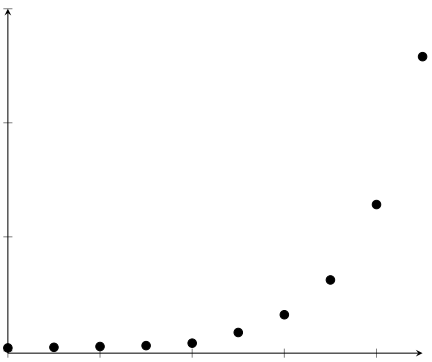
Strong, positive linear relationship



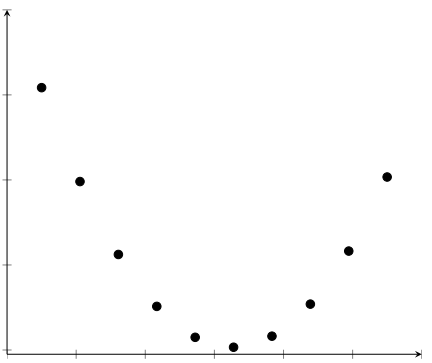
Strong, negative linear relationship



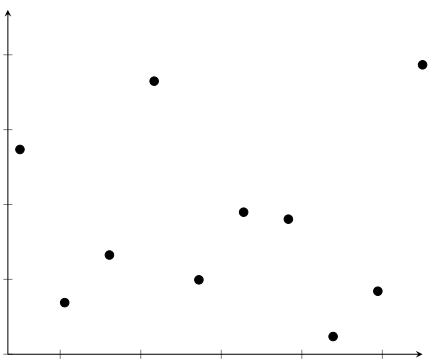
Weak, positive linear relationship



Exponential relationship



Quadratic relationship

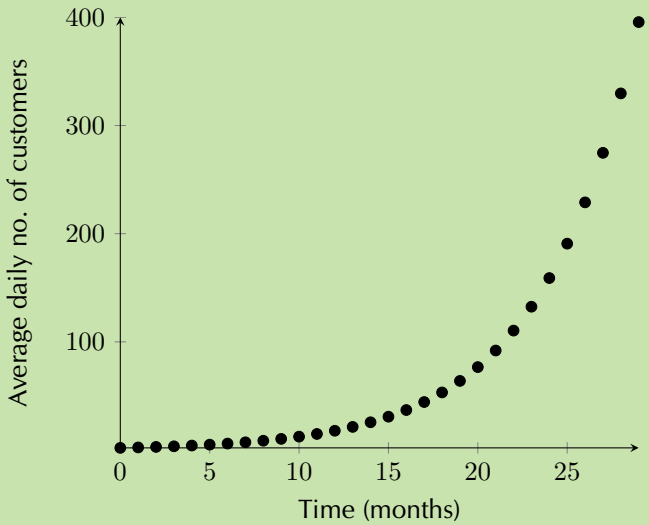


No relationship

Worked example 4: Intuitive curve fitting

QUESTION

Examine the scatter plot below of data collected from a new shop:



- What are the two variables being compared?
- What type of function best fits the data?
- Is the relationship between the two variables strong or weak?
- Is the relationship between the two variables positive or negative?
- Using your answers above, describe the relationship between the two variables in one sentence.

SOLUTION

- The variables being compared are average daily number of customers and time in months.
- The data fit an exponential function.
- The data points appear to fit the curve close to perfectly, so the relationship can be described as very strong.
- As time increases, the number of customers increases, so the relationship can be described as positive.
- There is a very strong, positive, exponential relationship between average daily customers and time in the new shop.

In the worked example above, by plotting the average daily customers and time data of a new shop on a scatter plot, we were able to identify the relationship between the two variables. Once we know the relationship between two variables, we are able to do another very useful thing we can predict values where no data exist.

DEFINITION: *Interpolation and extrapolation*

When we predict values that fall within the range of our data, this is known as **interpolation**. When we predict the values of a variable beyond the range of our data, this is known as **extrapolation**.

Extrapolation must be done with caution unless it is known that the observed relationship continues beyond the range of our data. For example, an exponential function may look linear if we only have the first few data points available but if we extrapolate far enough beyond the initial data points, our predictions will be inaccurate.

In order to interpolate or extrapolate values, we need to find the equation of the function which best fits the data. For linear data, we draw a straight line through the data which best approximates the available data points. This line is known as the **line of best fit** or trend line. Let us try our hand at this in the following example.

Worked example 5: Fitting by hand

QUESTION

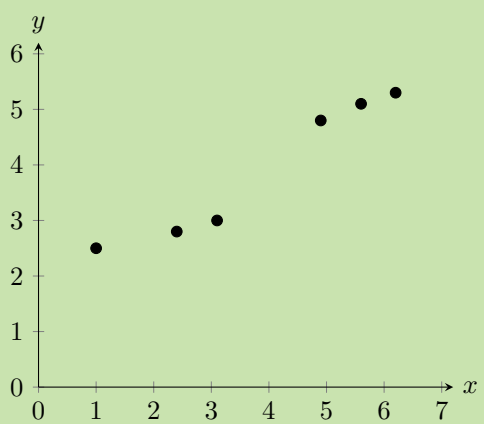
- Use the data below to draw a scatter plot and line of best fit.
- Write down the equation of the line that best seems to fit the data.
- Use your equation to calculate the estimated value for y if $x = 4$.
- Use your equation to calculate the estimated value for x if $y = 6$.

x	1,0	2,4	3,1	4,9	5,6	6,2
y	2,5	2,8	3,0	4,8	5,1	5,3

SOLUTION

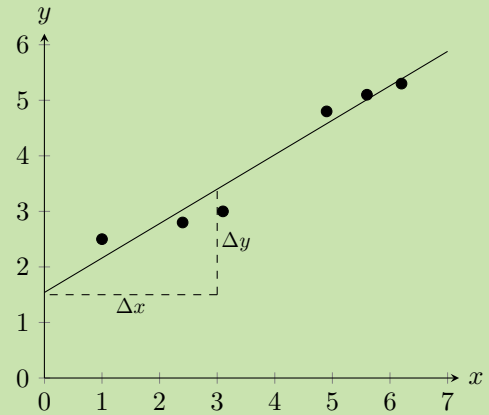
Step 1: Draw the graph

1. Choose a suitable scale for the axes.
2. Draw the axes.
3. Plot the points.



Step 2: Drawing the line of best fit

The next step is to draw a straight line which goes as close to as many points as possible. It is generally best to have as many points above the line as below the line.



Step 3: Calculating the equation of the line

The equation of the line is $y = mx + c$

From the graph we have drawn, we estimate the y-intercept to be $1{,}5$. We estimate that $y = 3{,}5$ when $x = 3$. So we have that points $(3; 3{,}5)$ and $(0; 1{,}5)$ lie on the line. The gradient of the line, m, is given by

$$\begin{aligned} m &= \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} \\ &= \frac{3{,}5 - 1{,}5}{3 - 0} \\ &= \frac{2}{3} \end{aligned}$$

So we finally have that the equation of the line of best fit is $y = \frac{2}{3}x + 1{,}5$

Step 4: Calculate the unknown values

The equation of the line is $y = \frac{2}{3}x + 1{,}5$ so in order to find the unknown values, we insert the known values into our equation.

For $x = 4$:

$$\begin{aligned} y &= \frac{2}{3} \cdot 4 + 1{,}5 \\ &= 4{,}17 \end{aligned}$$

Since this x-value is within the data range, this is **interpolation**.

For $y = 6$:

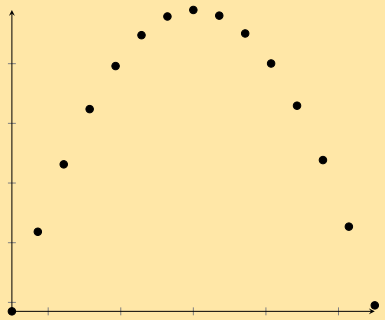
$$\begin{aligned} 6 &= \frac{2}{3} \cdot x + 1{,}5 \\ \therefore x &= (6 - 1{,}5) \times \frac{3}{2} \\ &= 6{,}75 \end{aligned}$$

Since this y-value is outside the data range, this is **extrapolation**.

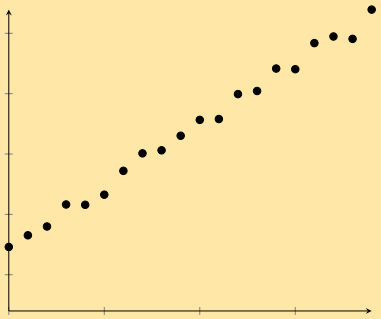
Exercise 9 – 2: Intuitive curve fitting

1. Identify the function (linear, exponential or quadratic) which would best fit the data in each of the scatter plots below:

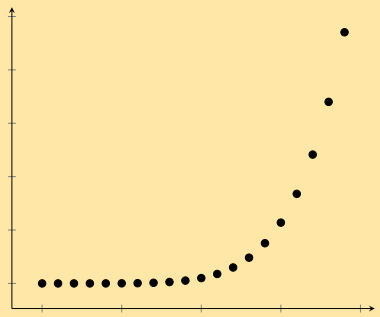
a)



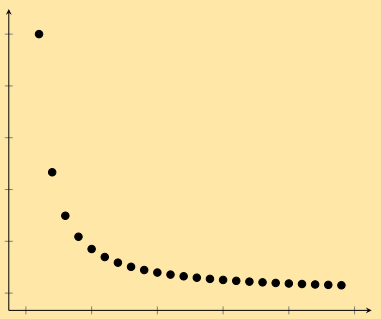
d)



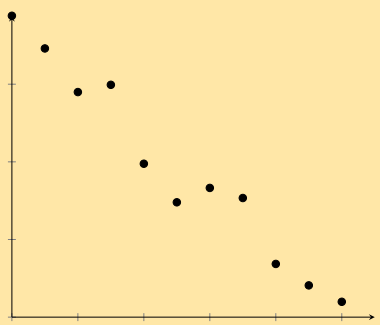
b)



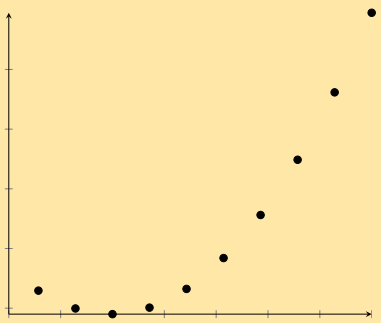
e)



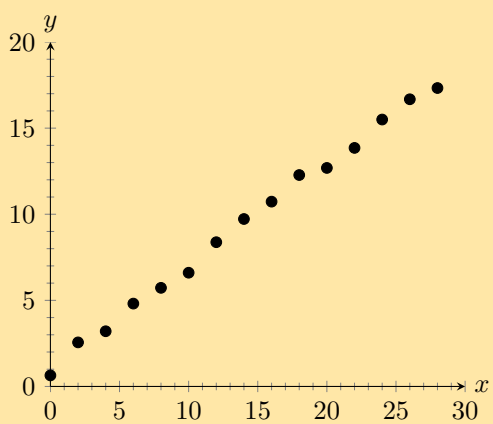
c)



f)



2. Given the scatter plot below, answer the questions that follow on the next page.



a) What type of function fits the data best? Comment on the fit of the function in terms of strength and direction.

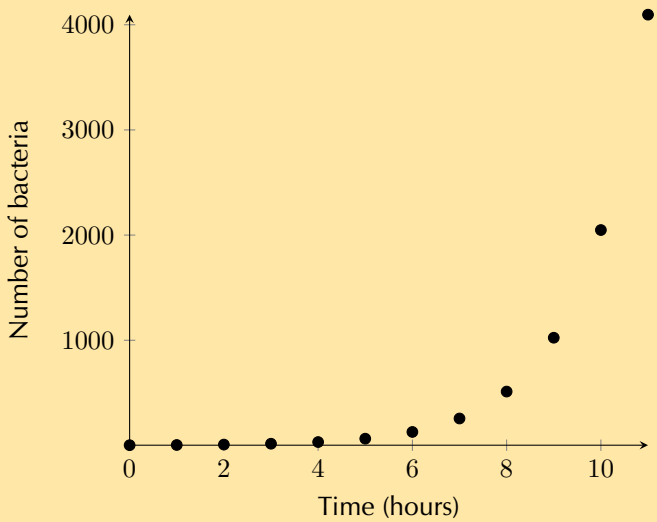
b) Draw a line of best fit through the data and determine the equation for your line.

c) Using your equation, determine the estimated y-value where $x = 25$.

d) Using your equation, determine the estimated x-value where $y = 25$.

3. Tuberculosis (TB) is a disease of the lungs caused by bacteria which are spread through the air when an infected person coughs or sneezes. Drug-resistant TB arises when patients do not take their medication properly. Andile is a scientist studying a new treatment for drug-resistant TB.

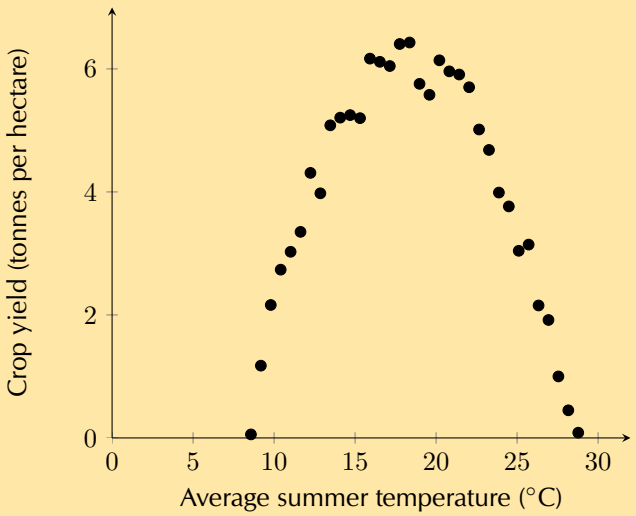
   For his research, he needs to grow the TB bacterium. He takes two bacteria and puts them on a plate with nutrients for their growth. He monitors how the number of bacteria increases over time. Look at his data in the scatter plot below and answer the questions that follow.



a) What type of function do you think fits the data best?

b) The equation for bacterial growth is $x_n = x_0(1 + r)^t$ where x_0 is the initial number of bacteria, r is the growth rate per unit time as a proportion of 1, t is time in hours, and x_n is the number of bacteria at time, t. Determine the number of bacteria grown by Andile after 24 hours if the number of bacteria doubles every hour (i.e. the growth rate is 100% per hour).

4. Marelize is a researcher at the Department of Agriculture. She has noticed that farmers across the country have very different crop yields depending on the region. She thinks that this has to do with the different climate in each region. In order to test her idea, she collected data on crop yield and average summer temperatures from a number of farmers. Examine her data on the opposite page and answer the questions that follow.



a) Identify what type of function would fit the data best.

b) Marelize determines that the equation for the function which fits the data best is $y = -0{,}06x^2 + 2{,}2x - 14$. Determine the optimal temperature to grow wheat and the respective crop yield. Round your answer to two decimal places.

5. Dr Dandara is a scientist trying to find a cure for a disease which has an 80% mortality rate, i.e. 80% of people who get the disease will die. He knows of a plant which is used in traditional medicine to treat the disease. He extracts the active ingredient from the plant and tests different dosages (measured in milligrams) on different groups of patients. Examine his data below and complete the questions that follow.

Dosage (mg)	0	25	50	75	100	125	150	175	200
Mortality rate (%)	80	73	63	49	42	32	25	11	5

a) Draw a scatter plot of the data

b) Which function would best fit the data? Describe the fit in terms of strength and direction.

c) Draw a line of best fit through the data and determine the equation of your line.

d) Use your equation to estimate the dosage required for a 0% mortality rate.

e) Dr Dandara decided to administer the estimated dosage required for a 0% mortality rate to a group of infected patients. However, he still found a mortality rate of 5%. Name the statistical technique Dr Dandara used to estimate a mortality rate of 0% and explain why his equation did not accurately predict his experimental results.

6. More questions. Sign in at Everything Maths online and click 'Practise Maths'.

Check answers online with the exercise code below or click on 'show me the answer'.

1a. 29CS 1b. 29CT 1c. 29CV 1d. 29CW 1e. 29CX 1f. 29CY
2. 29CZ 3. 29D2 4. 29D3 5. 29D4

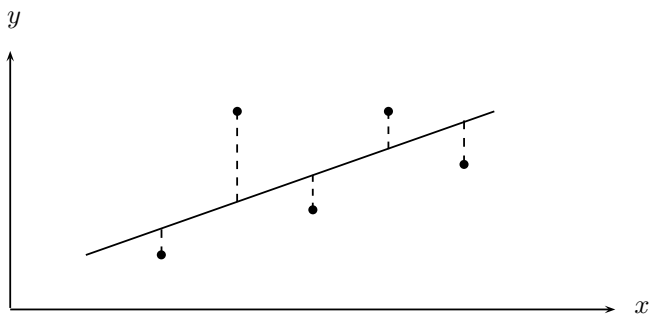
 **www.everythingmaths.co.za**  **m.everythingmaths.co.za**

In the previous worked example and exercises, you drew the line of best fit by hand. This can give us a reasonable approximation of which function best fits the data when the data points are close together. However, you and your classmates may have found that you obtained slightly different answers from one another. In the next section, we will learn about a more precise way of fitting a linear function to data.

Linear regression EMCJR

Linear regression analysis is a statistical technique for finding out exactly which linear function best fits a given set of data. We can find out the equation of the regression line by using an algebraic method called the **least squares method**, available on most scientific calculators. The linear regression equation is written $\hat{y} = a + bx$ (we say y-hat) or $y = A + Bx$. Of course these are both variations of the more familiar equation $y = mx + c$.

The least squares method is very simple. Suppose we guess a line of best fit, then at every data point, we find the distance between the data point and the line. If the line fitted the data perfectly, this distance would be zero for all the data points. The worse the fit, the larger the differences. We then square each of these distances, and add them all together.



The best-fit line is then the line that minimises the sum of the squared distances.

See video: 29D5 at www.everythingmaths.co.za

Suppose we have a data set of n points $\{(x_1; y_1), (x_2; y_2), ..., (x_n; y_n)\}$. We also have a line $f(x) = mx + c$ that we are trying to fit to the data. The distance between the first data point and the line, for example, is distance $= y_1 - f(x_1) = y_1 - (mx_1 + c)$

We now square each of these distances and add them together. Lets call this sum $S(m, c)$. Then we have that

$$\begin{aligned} S(m, c) &= (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \ldots + (y_n - f(x_n))^2 \\ &= \sum_{i=1}^{n} (y_i - f(x_i))^2 \end{aligned}$$

Thus our problem is to find the value of m and c such that $S(m, c)$ is minimised. Let us call these minimising values b and a respectively. Then the line of best-fit is $f(x) = a + bx$. We can find a and b using calculus, but it is tricky, and we will just give you the result, which is that

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} x_i = \bar{y} - b\bar{x}$$

Worked example 6: Method of least squares by hand

QUESTION

In the table below, we have the records of the maintenance costs in rands compared with the age of the appliance in months. We have data for five appliances. Determine the equation for the least squares regression line by hand.

Appliance	1	2	3	4	5
Age (x)	5	10	15	20	30
Cost (y)	90	140	250	300	380

SOLUTION

Appliance	x	y	xy	x^2
1	5	90	450	25
2	10	140	1400	100
3	15	250	3750	225
4	20	300	6000	400
5	30	380	11 400	900
Total	80	1160	23 000	1650

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{5 \times 23000 - 80 \times 1160}{5 \times 1650 - 80^2}$$
$$= 12$$
$$a = \bar{y} - b\bar{x} = \frac{1160}{5} - \frac{12 \times 80}{5}$$
$$= 40$$
$$\therefore \hat{y} = 40 + 12x$$

Worked example 7: Using the SHARP EL-531VH calculator

QUESTION

Using a calculator, find the equation of the least squares regression line for the following data:

Days (x)	1	2	3	4	5
Growth in m (y)	1,00	2,50	2,75	3,00	3,50

NB. If you have a CASIO calculator, do the next worked example first. Come back to this worked example once you are done and see if you get the same answer on your calculator.

SOLUTION

Step 1: Getting your calculator ready

Using your calculator, change the mode from normal to "Stat xy ". Do this by pressing [2ndF] and then 2. This mode enables you to type in bivariate data.

Step 2: Entering the data

Key in the data row by row:

Enter:	Press:	Enter:	Press:	See:
1	(x, y)	1	**DATA**	n = 1
2	(x, y)	2,5	**DATA**	n = 2
3	(x, y)	2,75	**DATA**	n = 3
4	(x, y)	3,0	**DATA**	n = 4
5	(x, y)	3,5	**DATA**	n = 5

Note: The [(x, y)] button is the same as the [**STO**] button and the [**DATA**] button is the same as the [M+] button.

Step 3: Getting regression results from the calculator

Ask for the values of the regression coefficients a and b.

Press:	Press:	See:
RCL	a	$a = 0{,}9$
RCL	b	$b = 0{,}55$

$\therefore \hat{y} = 0{,}9 + 0{,}55x$

Worked example 8: Using the CASIO fx-82ZA PLUS calculator

QUESTION

Using a calculator determine the least squares line of best fit for the following data set.

Learner	1	2	3	4	5
Chemistry (%)	52	55	86	71	45
Accounting (%)	48	64	95	79	50

For a Chemistry mark of 65%, what mark does the least squares line predict for Accounting?

NB. If you have a SHARP calculator, ensure that you have done the previous worked example first. Once you have completed the previous worked example, attempt this example using your calculator and see if you get the same answer.

SOLUTION

Step 1: Getting your calculator ready

Switch on the calculator. Press [MODE] and then select STAT by pressing [2]. The following screen will appear:

1:	$1 - VAR$	2:	$A + BX$
3:	$_{+}CX^2$	4:	lnX
5:	eX	6:	$A.BX$
7:	$A.XB$	8:	$1/X$

Now press [2] for linear regression. Your screen should look something like this:

	X	Y
1		
2		
3		

Step 2: Entering the data

Press [52] and then [=] to enter the first mark under x. Then enter the other values, in the same way, for the x-variable (the Chemistry marks) in the order in which they are given in the data set. Then move the cursor across and up and enter 48 under y opposite 52 in the x-column. Continue to enter the other y-values (the Accounting marks) in order so that they pair off correctly with the corresponding x-values.

	X	Y
1	52	48
2	55	64
3	86	95

Then press [AC]. The screen clears but the data remains stored.

Now press [SHIFT][1] to get the stats computations screen shown below.

1:	Type	2:	Data
3:	Sum	4:	Var
5:	Reg	6:	MinMax

Choose Regression by pressing [5].

1:	A	2:	B
3:	r	4:	$\hat{x}$
5:	$\hat{y}$		

Step 3: Getting regression results from the calculator

1. Press [1] and [=] to get the value of the y-intercept, $a = -5{,}065\ldots = -5{,}07$ (to two decimal places)
   Finally, to get the slope, use the following key sequence: [SHIFT][1][5][2][=]. The calculator gives $b = 1{,}169\ldots = 1{,}17$ (to two decimal places)
   The equation of the line of regression is thus:
   $\hat{y} = -5{,}07 + 1{,}17x$
2. Press [AC][65][SHIFT][1][5][5][=]
   This gives a (predicted) Accounting mark of $= 70{,}94 = 71\%$

Exercise 9 – 3: Least squares regression analysis

1. Determine the equation of the least-squares regression line using a table for the data sets below. Round a and b to two decimal places.

a)

x	10	4	9	11	11	6	8	18	9	13
y	1	0	6	3	9	5	9	8	7	15

b)

x	8	12	12	7	6	14	8	14	14	17
y	−5	4	3	−3	−5	−6	−2	0	−4	3

c)

x	−9	3	4	7	13	6	0	8	1	14
y	0	−12	−10	−14	−31	−32	−41	−52	−51	−63

2. Use your calculator to determine the equation of the least squares regression line for the following sets of data:

a)

x	0,16	0,32	3	2,6	6,12	7,68	6,16	8,56	11,24	11,96
y	5,48	10,56	13,4	15,96	15,44	16,6	17,2	22,28	22,04	24,32

b)

x	−3,5	5,5	4	1	5,5	5	3,5	5,5	7,5	8,5
y	−10	−20,5	−30,5	−46	−46,5	−64,5	−67	−76,5	−83,5	−94

c)

x	2,5	4,5	−2	9	8,5	10	7,5	3	8	15
y	−2	6	11	11,5	17	21	21	30,5	32,5	33,5

d)

x	7,24	8,24	5,34	1,66	0,32	11,46	9,34	14,24	12,9	12,34
y	−3,2	−18,78	−21,1	−32	−31,2	−53,02	−53	−65,46	−74,8	−80,24

e)

x	−0,28	2,32	0,12	4,64	3,08	7,92	5,08	8,96	10,28	7,12
y	−6,88	−0,32	3,68	4,8	11,68	19,2	20,96	24,96	29,28	33,28

f)

x	1	1,1	4,8	3,55	2,75	1,95	6,1	8,9	10,35	9,55
y	−8,45	−5,95	−4,35	0,85	−2,95	−1,8	0,25	0,05	4,8	−3,05

g)

x	1,9	1,1	$-1,5$	1,3	0,95	8,25	10,6	6,2	8,1	8,65
y	7	8,45	0,9	0,1	2,45	4,35	2,2	1,4	0,15	2,05

h)

x	$-81,8$	73,1	84	92,2	$-69,7$	$-56,1$	8,8	80,9	68,4	$-40,4$
y	10,6	16,1	3,6	4,6	11,9	18,3	16,6	17,6	17,7	24,1

i)

x	2,8	7,4	$-2,4$	4	11,3	6,9	2,5	1,7	5,4	8,2
y	12,4	13,4	15,3	15,4	16,4	19,2	21,1	19,4	21,3	25

j)

x	5	1,2	8	6	7,4	7,4	6,7	8,7	12,2	14,3
y	$-4,2$	$-13,7$	$-23,7$	$-33,5$	$-43,8$	$-54,2$	$-63,9$	$-73,9$	$-84,5$	$-93,5$

3. Determine the equation of the least squares regression line given each set of data values below. Round a and b to two decimal places in your final answer.
   a) $n = 10;\ \sum x = 74;\ \sum y = 424;\ \sum xy = 4114{,}51;\ \sum(x^2) = 718{,}86$
   b) $n = 13;\ \bar{x} = 8{,}45;\ \bar{y} = 17{,}83;\ \sum xy = 1879{,}25;\ \sum(x^2) = 855{,}45$
   c) $n = 10;\ \bar{x} = 5{,}77;\ \bar{y} = 17{,}03;\ \overline{xy} = 133{,}817;\ \sigma_x = \pm 3{,}91$
   (Hint: multiply the numerator and denominator of the formula for b by $\frac{1}{n^2}$)

4. The table below shows the average maintenance cost in rands of a certain model of car compared to the age of the car in years.

Age (x)	1	3	5	6	8	9	10
Cost (y)	1000	1500	1600	1800	2000	2400	2600

   a) Draw a scatter plot of the data.
   b) Complete the table below, filling in the totals of each column in the final row:

Age (x)	**Cost** (y)	xy	x^2
1	1000		
3	1500		
5	1600		
6	1800		
8	2000		
9	2400		
10	2600		
$\sum = \ldots$	$\sum = \ldots$	$\sum = \ldots$	$\sum = \ldots$

   c) Use your table to determine the equation of the least squares regression line. Round a and b to two decimal places.
   d) Use your equation to estimate what it would cost to maintain this model of car in its 15^{th} year.
   e) Use your equation to estimate the age of the car in the year where the maintenance cost totals over R 3000 for the first time.

5. Miss Colly has always maintained that there is a relationship between a learner's ability to understand the language of instruction and their marks in Mathematics. Since she teaches Mathematics through the medium of English, she decides to compare the Mathematics and English marks of her learners in order to investigate the relationship between the two marks. Examine her data below and answer the quesitons on the following page.

English % (x)	28	33	30	45	45	55	55	65	70	76	65	85	90
Mathematics % (y)	35	36	34	45	50	40	60	50	65	85	70	80	90

a) Complete the table below, filling in the totals of each column in the final row:

English % (x)	Mathematics % (y)	xy	x^2
28	35		
33	36		
30	34		
45	45		
45	50		
55	40		
65	50		
70	65		
76	85		
65	70		
85	80		
90	90		
$\sum = \ldots$	$\sum = \ldots$	$\sum = \ldots$	$\sum = \ldots$

b) Use your table to determine the equation of the least squares regression line. Round a and b to two decimal places.

c) Use your equation to estimate the Mathematics mark of a learner who obtained 50% for English, correct to two decimal places.

d) Use your equation to estimate the English mark of a learner who obtained 75% for Mathematics, correct to two decimal places.

6. Foot lengths and heights of ten students are given in the table below.

Height (cm)	170	163	131	181	146	134	166	172	185	153
Foot length (cm)	27	23	20	28	22	20	24	26	29	22

a) Using foot length as your x-variable, draw a scatter plot of the data.

b) Identify and describe any trends shown in the scatter plot.

c) Find the equation of the least squares line using the formulae and draw the line on your graph. Round a and b to two decimal places in your final answer.

d) Confirm your calculations above by finding the least squares regression line using a calculator.

e) Use your equation to predict the height of a student with a foot length of $21{,}6$ cm.

f) Use your equation to predict the foot length of a student 190 cm tall, correct to two decimal places.

7. More questions. Sign in at Everything Maths online and click 'Practise Maths'.

Check answers online with the exercise code below or click on 'show me the answer'.

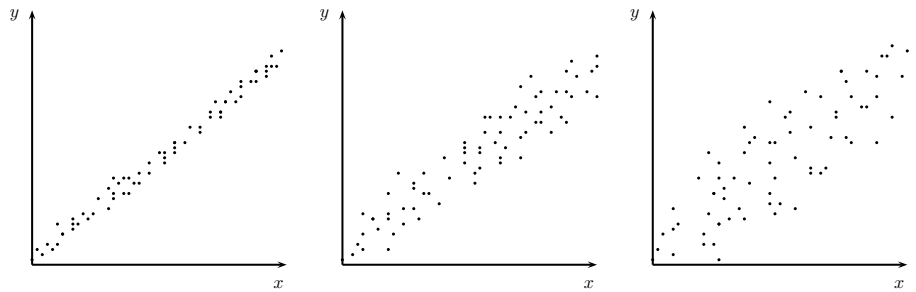
1a. 29D6 1b. 29D7 1c. 29D8 2a. 29D9 2b. 29DB 2c. 29DC
2d. 29DD 2e. 29DF 2f. 29DG 2g. 29DH 2h. 29DJ 2i. 29DK
2j. 29DM 3a. 29DN 3b. 29DP 3c. 29DQ 4. 29DR 5. 29DS
6. 29DT

 www.everythingmaths.co.za  m.everythingmaths.co.za

Now that we have a precise technique for finding the line of best fit, we still do not know how well our line of best fit really fits our data. We can fit a least squares regression line to any bivariate data, even if the two variables do not show a linear relationship. If the fit is not "good", our assumption of the a and b values in $\hat{y} = a + bx$ might be incorrect. Next, we will learn of a quantitative measure to determine how well our line really fits our data.

9.3 Correlation EMCJS

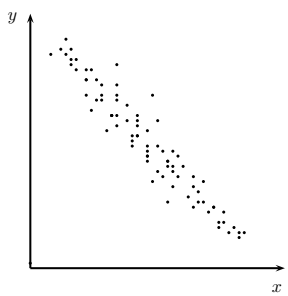
The linear correlation coefficient, r, is a measure which tells us the strength and direction of a relationship between two variables. The correlation coefficient $r \in [-1; 1]$. When $r = -1$, there is perfect negative correlation, when $r = 0$, there is no correlation and when $r = 1$ there is perfect positive correlation.



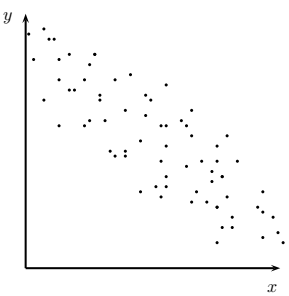
Positive, strong
$r \approx 0{,}9$

Positive, fairly strong
$r \approx 0{,}7$

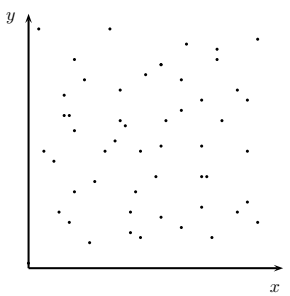
Positive, weak
$r \approx 0{,}4$



Negative, fairly strong
$r \approx -0{,}7$



Negative, weak
$r \approx -0{,}4$



No correlation
$r = 0$

The linear correlation coefficient r can be calculated using the formula $r = b\dfrac{\sigma_x}{\sigma_y}$

- where b is the gradient of the least squares regression line,
- σ_x is the standard deviation of the x-values and
- σ_y is the standard deviation of the y-values.

This is known as the Pearson's product moment correlation coefficient. It is much easier to do on a calculator where you simply follow the procedure to calculate the regression equation, and go on to find r.

In general:

Positive	Strength	Negative
$r = 0$	no correlation	$r = 0$
$0 < r < 0{,}25$	very weak	$-0{,}25 < r < 0$
$0{,}25 < r < 0{,}5$	weak	$-0{,}5 < r < -0{,}25$
$0{,}5 < r < 0{,}75$	moderate	$-0{,}75 < r < -0{,}5$
$0{,}75 < r < 0{,}9$	strong	$-0{,}9 < r < -0{,}75$
$0{,}9 < r < 1$	very strong	$-1 < r < -0.9$
$r = 1$	perfect correlation	$r = -1$

NOTE:
Correlation does not imply causation! Just because two variables are correlated does not mean that they are causally linked, i.e. if A and B are correlated, that does not mean A causes B, or vice versa. This is a common mistake made by many people, especially journalists looking for their next juicy story.

For example, ice cream sales and shark attacks are correlated. This does not mean that the sale of ice cream is somehow causing more shark attacks. Instead, a simpler explanation is that the warmer it is, the more likely people are to buy ice cream and the more likely people are to go to the beach as well, thus increasing the likelihood of a shark attack.

See video: 29DV at www.everythingmaths.co.za

Worked example 9: The correlation coefficient

QUESTION

A cardiologist wanted to test the relationship between resting heart rate and the peak heart rate during exercise. Heart rate is measured in beats per minute (bpm). The following set of data was generated from 12 study participants after they had run on a treadmill at 10 km/h for 10 minutes.

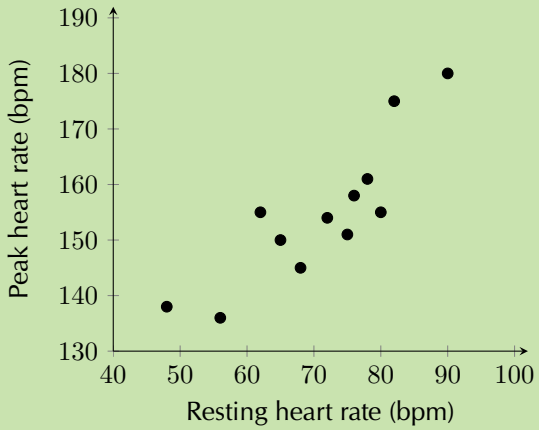
Resting heart rate	48	56	90	65	75	78	80	72	82	76	68	62
Peak heart rate	138	136	180	150	151	161	155	154	175	158	145	155

- Draw a scatter plot of the data. Use resting heart rate as your x-variable.
- Use your calculator to determine the equation of the line of best fit.
- Estimate what the heart rate of a person with a resting heart rate of 70 bpm will be after exercise.
- Without using your calculator, find the correlation coefficient, r. Confirm your answer using your calculator.
- What can you conclude regarding the relationship between resting heart rate and the heart rate after exercise?

SOLUTION

Step 1: Draw the scatter plot

1. Choose a suitable scale for the axes.
2. Draw the axes.
3. Plot the points.



Step 2: Calculate the equation of the line of best fit

As you learnt previously, use your calculator to determine the values for a and b.

$a = 86{,}75$

$b = 0{,}96$

Therefore, the equation for the line of best fit is $y = 86{,}75 + 0{,}96x$

Step 3: Calculate the estimated value for y

If $x = 70$, using our equation, the estimated value for y is:

$$y = 86{,}75 + 0{,}96 \times 70 = 153{,}95$$

Step 4: Calculate the correlation co-efficient

The formula for r is:

$$r = b\frac{\sigma_x}{\sigma_y}$$

We already know the value of b and you know how to calculate b by hand from worked example 5, so we are just left to determine the value for σ_x and σ_y. The formula for standard deviation is:

$$\sigma_x = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{n}$$

First, you need to determine $\bar{x}$ and $\bar{y}$ and then complete a table like the one below.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 71$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 154{,}83 \text{ (rounded to two decimal places)}$$

Resting heart rate (x)	**Peak heart rate (y)**	$(x - \bar{x})^2$	$(y - \bar{y})^2$
48	138	529	283,25
56	136	225	354,57
90	180	361	633,53
65	150	36	23,33
75	151	16	14,67
78	161	49	38,07
80	155	81	0,03
72	154	1	0,69
82	175	121	406,83
76	158	25	10,05
68	145	9s	96,63
62	155	81	0,03
$\sum = 852$	$\sum = 1858$	$\sum = 1534$	$\sum = 1861{,}68$

$$\sigma_x = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{n} = \frac{\sqrt{1534}}{12} = \pm 3{,}26$$

$$\sigma_y = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{n} = \frac{\sqrt{1861{,}68}}{12} = \pm 3{,}60$$

$$b = 0{,}96$$

$$\therefore r = 0{,}96 \times \frac{3{,}26}{3{,}60}$$

$$= 0{,}87$$

Step 5: Confirm your answer using your calculator

Once you know the method for finding the equation of the best line of fit on your calculator, finding the value for r is trivial. After you have entered all your x and y values into your calculator, in STAT mode:

- on a SHARP calculator: press [RCL] then [r] (the same key as [÷])
- on a CASIO calculator: press [SHIFT] then [STAT], [5], [3] then [=]

Step 6: Comment on the correlation coefficient

$$r = 0{,}87$$

Therefore, there is a strong, positive, linear relationship between resting heart rate and peak heart rate during exercise. This means that the higher your resting heart rate, the higher your peak heart rate during exercise is likely to be.

Exercise 9 – 4: Correlation coefficient

1. Determine the correlation coefficient by hand for the following data sets and comment on the strength and direction of the correlation. Round your answers to two decimal places.

a)

x	5	8	13	10	14	15	17	12	18	13
y	5	8	3	8	7	5	3	-1	4	-1

b)

x	7	3	11	7	7	6	9	12	10	15
y	13	23	32	45	50	55	67	69	85	90

c)

x	3	10	7	6	11	16	17	15	17	20
y	6	24	30	38	53	56	65	75	91	103

2. Using your calculator, determine the value of the correlation coefficient to two decimal places for the following data sets and describe the strength and direction of the correlation.

a)

x	0,1	0,8	1,2	3,4	6,5	3,9	6,4	7,4	9,9	8,5
y	-5,1	-10	-17,3	-24,9	-31,9	-38,6	-42	-55	-62	-64,8

b)

x	-26	-34	-51	-14	50	-57	-11	-10	36	-35
y	-66	-10	-26	-51	-58	-56	45	-142	-149	-30

c)

x	101	-398	103	204	105	606	807	-992	609	-790
y	-300	98	-704	-906	-8	690	-12	686	984	-18

d)

x	101	82	-7	-6	45	-94	-23	78	-11	0
y	111	-74	21	106	51	26	21	86	-29	66

e)

x	-3	5	-4	0	-2	9	10	11	17	9
y	24	18	21	30	31	39	48	59	56	54

3. Calculate and describe the direction and strength of r for each of the sets of data values below. Round all r-values to two decimal places.

a) $b = -1{,}88$; $\sigma_x^2 = 48{,}62$; $\sigma_y^2 = 736{,}54$.

b) $a = 32{,}19$; $x = 4{,}3$; $\bar{y} = 36{,}6$; $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 620{,}1$; $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 2636{,}4$.

4. The geography teacher, Mr Chadwick, gave the data set below to his class to illustrate the concept that average temperature depends on how far a place is from the equator (known as the latitude). There are 90 degrees between the equator and the North Pole. The equator is defined as 0 degrees. Examine the data set below and answer the questions that follow.

City	**Degrees N (x)**	**Ave. temp. (y)**	xy	x^2	$(x-\bar{x})^2$	$(y-\bar{y})^2$
Cairo	43	22				
Berlin	53	19				
London	40	18				
Lagos	6	32				
Jerusalem	31	23				
Madrid	40	28				
Brussels	51	18				
Istanbul	39	23				
Boston	43	23				
Montreal	45	22				
Total:						

a) Copy and complete the table.

b) Using your table, determine the equation of the least squares regression line. Round a and b to two decimal places in your final answer.

c) Use your calculator to confirm your equation for the least squares regression line.

d) Using your table, determine the value of the correlation coefficient to two decimal places.

e) What can you deduce about the relationship between how far north a city is and its average temperature?

f) Estimate the latitude of Paris if it has an average temperature of $25°$C

5. A taxi driver recorded the number of kilometres his taxi travelled per trip and his fuel cost per kilometre in Rands. Examine the table of his data below and answer the questions that follow.

Distance (x)	3	5	7	9	11	13	15	17	20	25	30
Cost (y)	2,8	2,5	2,46	2,42	2,4	2,36	2,32	2,3	2,25	2,22	2

a) Draw a scatter plot of the data.

b) Use your calculator to determine the equation of the least squares regression line and draw this line on your scatter plot. Round a and b to two decimal places in your final answer.

c) Using your calculator, determine the correlation coefficient to two decimal places.

d) Describe the relationship between the distance travelled per trip and the fuel cost per kilometre.

e) Predict the distance travelled if the cost per kilometre is R $1{,}75$.

6. The time taken, in seconds, to complete a task and the number of errors made on the task were recorded for a sample of 10 primary school learners. The data is shown in the table below. [Adapted from NSC Paper 3 Feb-March 2013]

Time taken to complete task (in seconds)	23	21	19	9	15	22	17	14	21	18
Number of errors made	2	4	5	9	7	3	7	8	3	5

   a) Draw a scatter plot of the data.
   b) What is the influence of more time taken to complete the task on the number of errors made?
   c) Determine the equation of the least squares regression line and draw this line on your scatter plot. Round a and b to two decimal places in your final answer.
   d) Determine the correlation coefficient to two decimal places.
   e) Predict the number of errors that will be made by a learner who takes 13 seconds to complete this task.
   f) Comment on the strength of the relationship between the variables.

7. A recording company investigates the relationship between the number of times a CD is played by a national radio station and the national sales of the same CD in the following week. The data below was collected for a random sample of 10 CDs. The sales figures are rounded to the nearest 50. [NSC Paper 3 November 2012]

Number of times CD is played	47	34	40	34	33	50	28	53	25	46
Weekly sales of the CD	3950	2500	3700	2800	2900	3750	2300	4400	2200	3400

   a) Draw a scatter plot of the data.
   b) Determine the equation of the least squares regression line.
   c) Calculate the correlation coefficient.
   d) Predict, correct to the nearest 50, the weekly sales for a CD that was played 45 times by the radio station in the previous week.
   e) Comment on the strength of the relationship between the variables.

8. More questions. Sign in at Everything Maths online and click 'Practise Maths'.

Check answers online with the exercise code below or click on 'show me the answer'.

1a. 29DW 1b. 29DX 1c. 29DY 2a. 29DZ 2b. 29F2 2c. 29F3
2d. 29F4 2e. 29F5 3a. 29F6 3b. 29F7 4. 29F8 5. 29F9
6. 29FB 7. 29FC

 **www.everythingmaths.co.za**  **m.everythingmaths.co.za**

- **Curve fitting** is the process of fitting functions to data.
- **Intuitive** curve fitting is performed by visually interpreting if the points on the scatter plot conform to a linear, exponential, quadratic or some other function.
- The **line of best fit** or trend line is a straight line through the data which best approximates the available data points. This allows for the estimation of missing data values.
- **Interpolation** is the technique used to predict values that fall within the range of the available data.
- **Extrapolation** is the technique used to predict the value of variables beyond the range of the available data.
- **Linear regression analysis** is a statistical technique of finding out exactly which linear function best fits a given set of data.
- The **least squares method** is an algebraic method of finding the linear regression equation. The linear regression equation is written $\hat{y} = a + bx$, where

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} x_i = \bar{y} - b\bar{x}$$

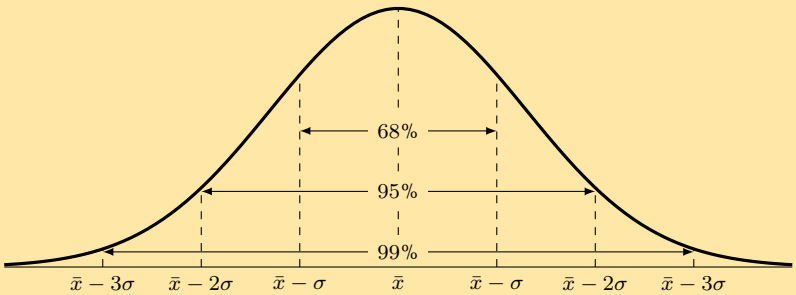
- The **linear correlation coefficient**, r, is a measure which tells us the strength and direction of a relationship between two variables, determined using the equation:

$$r = b(\frac{\sigma_x}{\sigma_y})$$

- The correlation coefficient $r \in [-1; 1]$. When $r = -1$, there is perfect negative correlation, when $r = 0$, there is no correlation and when $r = 1$, there is perfect positive correlation.

Exercise 9 – 5: End of chapter exercises

1. The number of SMS messages sent by a group of teenagers was recorded over a period of a week. The data was found to be normally distributed with a mean of 140 messages and a standard deviation of 12 messages. [NSC Paper 3 Feb-March 2012]



   Answer the following questions with reference to the information provided in the graph:

   a) What percentage of teenagers sent less than 128 messages?

   b) What percentage of teenagers sent between 116 and 152 messages?

2. A company produces sweets using a machine which runs for a few hours per day. The number of hours running the machine and the number of sweets produced are recorded.

Machine hours	Sweets produced
3,80	275
4,23	287
4,37	291
4,10	281
4,17	286

   Find the linear regression equation for the data, and estimate the machine hours needed to make 300 sweets.

3. The profits of a new shop are recorded over the first 6 months. The owner wants to predict his future sales. The profits by month so far have been R 90 000; R 93 000; R 99 500; R 102 000; R 101 300; R 109 000.

   a) Calculate the linear regression function for the data, using profit as your y-variable. Round a and b to two decimal places.

   b) Give an estimate of the profits for the next two months.

   c) The owner wants a profit of R 130 000. Estimate how many months this will take.

4. A fast food company produces hamburgers. The number of hamburgers made and the costs are recorded over a week. Examine the data below and answer the questions on the following page.

Hamburgers made	Costs
495	R 2382
550	R 2442
515	R 2484
500	R 2400
480	R 2370
530	R 2448
585	R 2805

a) Find the linear regression function that best fits the data. Use hamburgers made as your x-variable and round a and b to two decimal places.

b) Calculate the value of the correlation coefficient, correct to two decimal places, and comment on the strength and direction of the correlation.

c) If the total cost in a day is R 2500, estimate the number of hamburgers produced. Round your answer down to the nearest whole number.

d) What is the cost of 490 hamburgers?

5. A collection of data related to an investigation into biceps length and height of students was recorded in the table below. Answer the questions to follow.

Length of right biceps (cm)	Height (cm)
25,5	163,3
26,1	164,9
23,7	165,5
26,4	173,7
27,5	174,4
24	156
22,6	155,3
27,1	169,3

a) Draw a scatter plot of the data set.

b) Calculate equation of the line of regression.

c) Draw the regression line onto the graph.

d) Calculate the correlation coefficient r

e) What conclusion can you reach, regarding the relationship between the length of the right biceps and height of the students in the data set?

6. A class wrote two tests, and the marks for each were recorded in the table below. Full marks in the first test was 50, and the second test was out of 30.

Learner	Test 1	Test 2
	(Full marks: 50)	(Full marks: 30)
1	42	25
2	32	19
3	31	20
4	42	26
5	35	23
6	23	14
7	43	24
8	23	12
9	24	14
10	15	10
11	19	11
12	13	10
13	36	22
14	29	17
15	29	17
16	25	16
17	29	18
18	17	
19	30	19
20	28	17

a) Is there a strong correlation between the marks for the first and second test? Show why or why not.

b) One of the learners (in Row 18) did not write the second test. Given her mark for the first test, calculate an expected mark for the second test. Round the mark up to the nearest whole number.

7. Lindiwe works for Eskom, the South African power distributor. She knows that on hot days more electricity than average is used to cool houses. In order to accurately predict how much more electricity needs to be produced, she wants to determine the precise nature of the relationship between temperature and electricity usage.

   The data below shows the peak temperature in degrees Celsius on ten consecutive days during summer and the average number of units of electricity used by a number of households. Examine her data and answer the questions that follow.

Peak temp. (y)	32	40	30	28	25	38	36	20	24	26
Ave. no. of units (x)	37	45	35	30	20	40	38	15	20	22

a) Draw a scatter plot of the data.

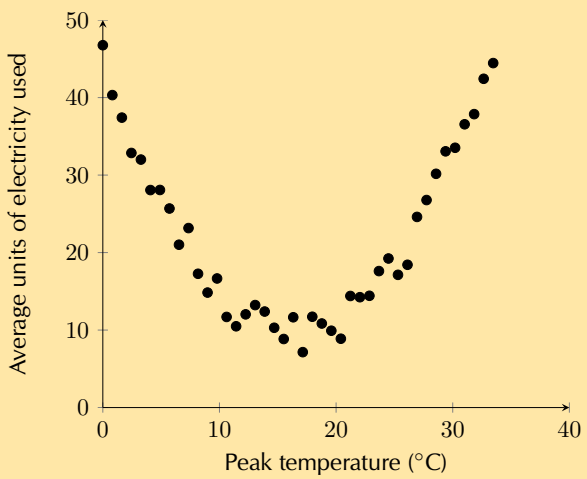
b) Using the formulae for a and b, determine the equation of the least squares line.

c) Determine the value of the correlation coefficient, r, by hand.

d) What can Lindiwe conclude about the relationship between peak temperature and the number of electricity units used?

e) Predict the average number of units of electricity used by a household on a day with a peak temperature of 45°C. Give your answer correct to the nearest unit and identify what this type of prediction is called.

f) Lindiwe suspected that the relationship between temperature and electricity consumption was not linear for all temperatures. She then decided to collect data for peak temperatures down to 0°C. Examine the graph of her data below and identify which type of function would best fit the data and describe the nature of the relationship between temperature and electricity for the newly available data.



g) Lindiwe is asked by her superiors to determine which day is best to perform maintenance on one of their power plants. She determined that the equation $y = 0{,}13x^2 - 4{,}3x + 45$ best fit her data. Use her equation to estimate the peak temperature and average no. of units used on the day when the least amount of electricity generation is required.

8. Below is a list of data concerning 12 countries and their respective carbon dioxide (CO_2) emission levels per person per annum (measured in tonnes) and the gross domestic product (GDP is a measure of products produced and services delivered within a country in a year) per person (in US dollars). Data sourced from the World Bank and the US Department of Energy's Carbon Dioxide Information Analysis Center.

	CO_2 emmissions per capita (x)	GDP per capita (y)
South Africa	8,8	11 440
Thailand	4,1	9815
Italy	7,5	32 512
Australia	18,3	44 462
China	5,3	9233
India	1,4	3876
Canada	15,3	42 693
United Kingdom	8,5	35 819
United States	17,2	49 965
Saudi Arabia	16,1	24 571
Iran	7,3	11 395
Indonesia	1,8	4956

a) Draw a scatter plot of the data set.

b) Draw your estimate of the line of best fit on your scatter plot and determine the equation of your line of best fit.

c) Use your calculator to determine the equation for the least squares regression line. Round a and b to two decimal places in your final answer.

d) Use your calculator to determine the correlation coefficient, r. Round your answer to two decimal places.

e) What conclusion can you reach regarding the relationship between CO_2 emissions per annum and GDP per capita for the countries in the data set?

f) Kenya has a GDP per capita of \$1712. Use your equation of the least squares regression line to estimate the annual CO_2 emissions of Kenya correct to two decimal places.

9. A group of students attended a course in Statistics on Saturdays over a period of 10 months. The number of Saturdays on which a student was absent was recorded against the final mark the student obtained. The information is shown in the table below. [Adapted from NSC Paper 3 Feb-March 2012]

Number of Saturdays absent	0	1	2	2	3	3	5	6	7
Final mark (as %)	96	91	78	83	75	62	70	68	56

a) Draw a scatter plot of the data.

b) Determine the equation of the least squares line and draw it on your scatter plot.

c) Calculate the correlation coefficient.

d) Comment on the trend of the data.

e) Predict the final mark of a student who was absent for four Saturdays.

10. Grant and Christie are training for a half-marathon together in 8 weeks time. Christie is much fitter than Grant but she has challenged him to beat her time at the race. Grant has begun a rigid training programme to try and improve his time.

    Time taken to complete a half marathon was recorded each Sunday. The first recorded Sunday is denoted as week 1. The half-marathon takes place on the eighth Sunday, i.e. week 8. Examine the data set in the table below and answer the questions the follow.

Week	1	2	3	4	5	6
Grant's time (HH:MM)	02:01	01:59	01:55	01:53	01:47	01:42
Christie's time (HH:MM)	01:40	01:42	01:38	01:39	01:37	01:35

    a) Draw a scatter plot of the data sets. Include Grant and Christie's data on the same set of axes. Use a • to denote Grant's data points and × to denote Christie's data points. Convert all times to minutes.
    b) Comment on and compare any trends that you observe in the data.
    c) Determine the equations of the least squares regression lines for Grant's data and Christie's data. Draw these lines on your scatter plot. Use a different colour for each.
    d) Calculate the correlation coefficient and comment on the fit for each data set.
    e) Assuming the observed trends continue, will Grant beat Christie at the race?
    f) Assuming the observed trends continue, extrapolate the week in which Grant will be able to run a half-marathon in less time than Christie.

11. More questions. Sign in at Everything Maths online and click 'Practise Maths'.

Check answers online with the exercise code below or click on 'show me the answer'.

1. 29FD 2. 29FF 3. 29FG 4. 29FH 5. 29FJ 6. 29FK
7. 29FM 8. 29FN 9. 29FP 10. 29FQ

 **www.everythingmaths.co.za**  **m.everythingmaths.co.za**