Statistics 11

11.1 Introduction EMCCU

In this chapter, you will use the mean, median, mode and standard deviation of a set of data to identify whether the data is normally distributed or whether it is skewed. You will learn more about populations and selecting different kinds of samples in order to avoid bias. You will work with lines of best fit, and learn how to find a regression equation and a correlation coefficient. You will analyse these measures in order to draw conclusions and make predictions.

See introductory video: VMhwi at www.everythingmaths.co.za

11.2 Normal Distribution EMCCV

Activity:

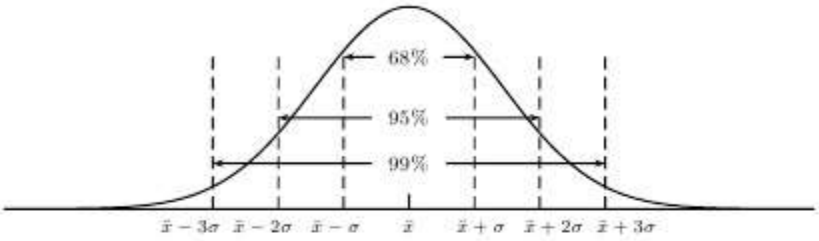
You are given a table of data below.

75	67	70	71	71	73	74	75
80	75	77	78	78	78	78	79
91	81	82	82	83	86	86	87

1. Calculate the mean, median, mode and standard deviation of the data.
2. What percentage of the data is within one standard deviation of the mean?
3. Draw a histogram of the data using intervals $60 \leq x < 64$; $64 \leq x < 68$; etc.
4. Join the midpoints of the bars to form a frequency polygon.

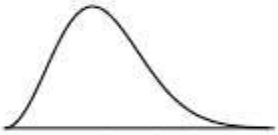
If large numbers of data are collected from a population, the graph will often have a bell shape. If the data was, say, examination results, a few learners usually get very high marks, a few very low marks and most get a mark in the middle range. We say a distribution is *normal* if

- the mean, median and mode are equal.
- it is symmetric around the mean.
- $\pm 68\%$ 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.

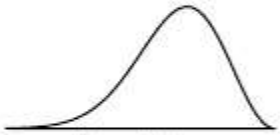


What happens if the test was very easy or very difficult? Then the distribution may not be symmetrical. If extremely high or extremely low scores are added to a distribution, then the mean tends to shift towards these scores and the curve becomes skewed.

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


Skewed right

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


Skewed left

Exercise 11 - 1

1. Given the pairs of normal curves below, sketch the graphs on the same set of axes and show any relation between them. An important point to remember is that the area beneath the curve corresponds to 100%.

   (a) Mean = 8, standard deviation = 4 and Mean = 4, standard deviation = 8

   (b) Mean = 8, standard deviation = 4 and Mean = 16, standard deviation = 4

   (c) Mean = 8, standard deviation = 4 and Mean = 8, standard deviation = 8

2. After a class test, the following scores were recorded:

Test Score	Frequency
3	1
4	7
5	14
6	21
7	14
8	6
9	1
Total	64
Mean	6
Standard Deviation	1,2

   (a) Draw the histogram of the results.

   (b) Join the midpoints of each bar and draw a frequency polygon.

(c) What mark must one obtain in order to be in the top 2% of the class?

(d) Approximately 84% of the pupils passed the test. What was the pass mark?

(e) Is the distribution normal or skewed?

3. In a road safety study, the speed of 175 cars was monitored along a specific stretch of highway in order to find out whether there existed any link between high speed and the large number of accidents along the route. A frequency table of the results is drawn up below.

Speed (km.h^{-1})	Number of cars (Frequency)
50	19
60	28
70	23
80	56
90	20
100	16
110	8
120	5

The mean speed was determined to be around 82 km·hr^{-1}while the median speed was worked out to be around 84,5 km·hr^{-1}.

(a) Draw a frequency polygon to visualise the data in the table above.

(b) Is this distribution symmetrical or skewed left or right? Give a reason fro your answer.

 More practice  video solutions  or help at www.everythingmaths.co.za

(1.) 01aw (2.) 01ax (3.) 01ay

11.3 Extracting a Sample Population

 EMCCW

Suppose you are trying to find out what percentage of South Africa's population owns a car. One way of doing this might be to send questionnaires to peoples homes, asking them whether they own a car. However, you quickly run into a problem: you cannot hope to send every person in the country a questionnaire, it would be far to expensive. Also, not everyone would reply. The best you can do is send it to a few people, see what percentage of these own a car, and then use this to estimate what percentage of the entire country own cars. This smaller group of people is called the *sample population*.

The sample population must be carefully chosen, in order to avoid biased results. How do we do this?
First, it must be *representative*. If all of our sample population comes from a very rich area, then almost all will have cars. But we obviously cannot conclude from this that almost everyone in the country has a car! We need to send the questionnaire to rich as well as poor people.
Secondly, the *size* of the sample population must be large enough. It is no good having a sample population consisting of only two people, for example. Both may very well not have cars. But we obviously cannot conclude that no one in the country has a car! The larger the sample population size, the more likely it is that the statistics of our sample population corresponds to the statistics of the entire population.

So how does one ensure that ones sample is representative? There are a variety of methods available, which we will look at now.

Random Sampling. Every person in the country has an equal chance of being selected. It is unbiased and also independent, which means that the selection of one person has no effect on the selection on another. One way of doing this would be to give each person in the country a number, and then ask a computer to give us a list of random numbers. We could then send the questionnaire to the people corresponding to the random numbers.

Systematic Sampling. Again give every person in the country a number, and then, for example, select every hundredth person on the list. So person with number 1 would be selected, person with number 100 would be selected, person with number 200 would be selected, etc.

Stratified Sampling. We consider different subgroups of the population, and take random samples from these. For example, we can divide the population into male and female, different ages, or into different income ranges.

Cluster Sampling. Here the sample is concentrated in one area. For example, we consider all the people living in one urban area.

Exercise 11 - 2

1. Discuss the advantages, disadvantages and possible bias when using
   (a) systematic sampling
   (b) random sampling
   (c) cluster sampling
2. Suggest a suitable sampling method that could be used to obtain information on:
   (a) passengers views on availability of a local taxi service.
   (b) views of learners on school meals.
   (c) defects in an item made in a factory.
   (d) medical costs of employees in a large company.
3. 2% of a certain magazines' subscribers is randomly selected. The random number 16 out of 50, is selected. Then subscribers with numbers $16; 66; 116; 166; \ldots$ are chosen as a sample. What kind of sampling is this?

A+ More practice ▶ video solutions ? or help at www.everythingmaths.co.za

(1.) 01az (2.) 01b0 (3.) 01b1

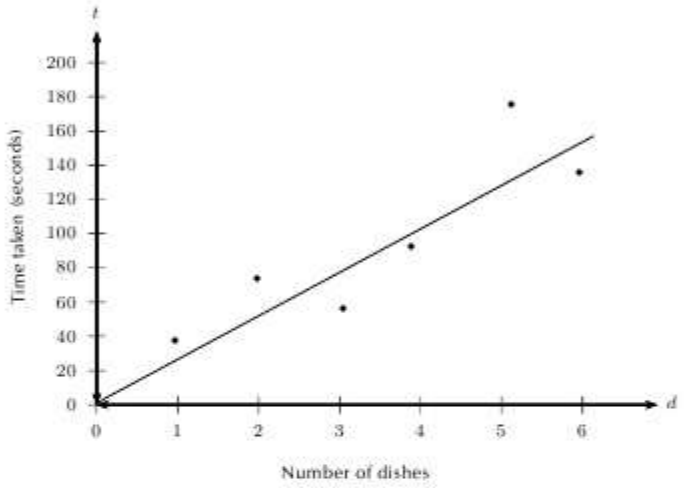
11.4 Function Fitting and Regression Analysis

EMCCX

In Grade 11 we recorded two sets of data (bivariate data) on a scatter plot and then we drew a line of best fit as close to as many of the data items as possible. Regression analysis is a method of finding

out exactly which function best fits a given set of data. We can find out the equation of the regression line by drawing and estimating, or 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 a more familiar equation $y = mx + c$.

Suppose you are doing an experiment with washing dishes. You count how many dishes you begin with, and then find out how long it takes to finish washing them. So you plot the data on a graph of time taken versus number of dishes. This is plotted below.



If t is the time taken, and d the number of dishes, then it looks as though t is proportional to d, ie. $t = m \,.\, d$, where m is the constant of proportionality. There are two questions that interest us now.

1. How do we find m? One way you have already learnt, is to draw a line of best-fit through the data points, and then measure the gradient of the line. But this is not terribly precise. Is there a better way of doing it?
2. How well does our line of best fit really fit our data? If the points on our plot don't all lie close to the line of best fit, but are scattered everywhere, then the fit is not "good", and our assumption that $t = m \,.\, d$ might be incorrect. Can we find a quantitative measure of how well our line really fits the data?

In this chapter, we answer both of these questions, using the techniques of *regression analysis*. ⊙ See simulation: VMibv at www.everythingmaths.co.za)

Example 1: *Fitting by hand*

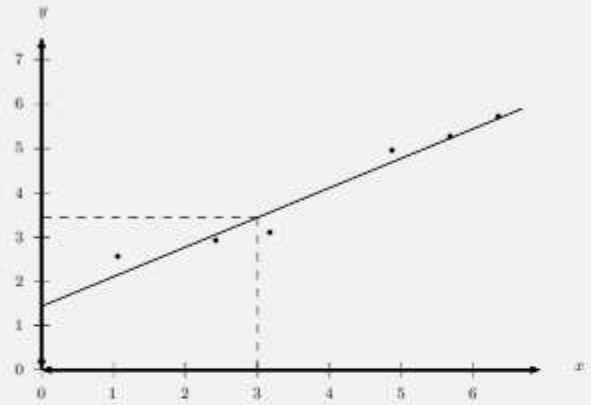
QUESTION

Use the data given to draw a scatter plot and line of best fit. Now write down the equation of the line that best seems to fit the data.

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 : ***Drawing the graph***
The first step is to draw the graph. This is shown below.



Step 2 : ***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

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

So we finally have that the equation of the line of best fit is

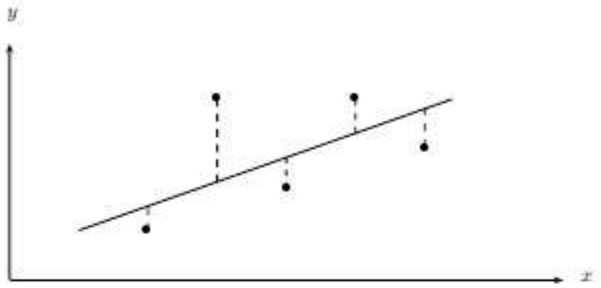
$$y = \frac{2}{3}x + 1{,}5$$

The Method of Least Squares

EMCCY

We now come to a more accurate method of finding the line of best-fit. The method is very simple. Suppose we guess a line of best-fit. Then at at every data point, we find the distance between the data point and the line. If the line fitted the data perfectly, this distance should 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.
Suppose we have a data set of n points $\{(x_1;y_1),(x_2;y_2),\ldots,(x_n;y_n)\}$. 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

$$\text{distance} = y_1 - f(x_1) = y_1 - (mx_1 + c)$$

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 m_0 and c_0. Then the line of best-fit is $f(x) = m_0x + c_0$. We can find m_0 and c_0 using calculus, but it is tricky, and we will just give you the result, which is that

$$\begin{aligned} m_0 &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ c_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{m_0}{n}\sum_{i=0}^{n} x_i = \bar{y} - m_0\bar{x} \end{aligned}$$

See video: VMhyr at www.everythingmaths.co.za

Example 2: *Method of Least Squares*

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.

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	11400	900
Total	80	1160	23000	1650

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{5 \times 23000 - 80 \times 1160}{5 \times 1650 - 80^2} = 12$$

$$a = \bar{y} - b\bar{x} = \frac{1160}{5} - \frac{12 \times 80}{5} = 40$$

$$\therefore \quad \hat{y} = 40 + 12x$$

Using a Calculator EMCCZ

Example 3: *Using the Sharp EL-531VH calculator*

QUESTION

Find a regression equation 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

SOLUTION

Step 1 : ***Getting your calculator ready***
Using your calculator, change the mode from normal to "Stat xy". This mode enables you to type in bivariate data.

Step 2 : ***Entering the data***
Key in the data as follows:

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	$n = 5$

Step 3 : ***Getting regression results from the calculator***

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

RCL	a	gives	$a = 0{,}9$
RCL	b	gives	$b = 0{,}55$

$$\therefore \hat{y} = 0.9 + 0{,}55x$$

Example 4: *Using the CASIO fx-82ES Natural Display calculator*

QUESTION

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

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?

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	$ln\ X$
5	eX	6	$A \cdot B^X$
7	$A \cdot X^B$	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	
2	55	
3		

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

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

Now press [SHIFT][1] to get the stats computations screen shown below. Choose Regression by pressing [7].

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

Step 3 : ***Getting regression results from the calculator***

a) 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][7][2][=]. The calculator gives $b = 1{,}169\ldots = 1{,}17$ (to two decimal places)

The equation of the line of regression is thus:
$\hat{y} = -5{,}07 + 1{,}17x$

b) Press [AC][65][SHIFT][1][7][5][=]
This gives a (predicted) Accounting mark of $= 70{,}94 = 71\%$

Exercise 11 - 3

1. The table below lists the exam results for five students in the subjects of Science and Biology.

Learner	1	2	3	4	5
Science %	55	66	74	92	47
Biology %	48	59	68	84	53

   (a) Use the formulae to find the regression equation coefficients a and b.

   (b) Draw a scatter plot of the data on graph paper.

   (c) Now use algebra to find a more accurate equation.

2. Footlengths and heights of seven students are given in the table below.

Height (cm)	170	163	131	181	146	134	166
Footlength (cm)	27	23	20	28	22	20	24

(a) Draw a scatter plot of the data on graph paper.

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

(c) Find the equation of the least squares line by using algebraic methods and draw the line on your graph.

(d) Use your equation to predict the height of a student with footlength 21,6 cm.

(e) Use your equation to predict the footlength of a student 176 cm tall.

3. Repeat the data in Question 2 and find the regression line using a calculator.

More practice video solutions or help at www.everythingmaths.co.za

(1.) 01b2 (2.) 01b3 (3.) 01b4

Correlation Coefficients

EMCDA

Once we have applied regression analysis to a set of data, we would like to have a number that tells us exactly how well the data fits the function. A correlation coefficient, r, is a tool that tells us to what degree there is a relationship between two sets of data. The correlation coefficient $r \in [-1; 1]$ when $r = -1$, there is a perfect negative correlation, when $r = 0$, there is no correlation and $r = 1$ is a perfect positive correlation.

Positive, strong $r \approx 0,9$ | Positive, fairly strong $r \approx 0,7$ | Positive, weak $r \approx 0,4$ | No association $r = 0$ | Negative, fairly strong $r \approx -0,7$

We often use the correlation coefficient r^2 in order to examine the strength of the correlation only.

In this case:

$r^2 = 0$	no correlation
$0 < r^2 < 0,25$	very weak
$0,25 < r^2 < 0,5$	weak
$0,5 < r^2 < 0,75$	moderate
$0,75 < r^2 < 0,9$	strong
$0,9 < r^2 < 1$	very strong
$r^2 = 1$	perfect correlation

The correlation coefficient r can be calculated using the formula

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

- where n is the number of data points,
- s_x is the standard deviation of the x-values and
- s_y is the standard deviation of the y-values.

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

Chapter 11 End of Chapter Exercises

1. Below is a list of data concerning 12 countries and their respective carbon dioxide (CO_2) emmission levels per person and the gross domestic product (GDP is a measure of products produced and services delivered within a country in a year) per person.

	CO_2 emmissions per capita (x)	GDP per capita (y)
South Africa	8,1	3 938
Thailand	2,5	2 712
Italy	7,3	20 943
Australia	17,0	23 893
China	2,5	816
India	0,9	463
Canada	16,0	22 537
United Kingdom	9,0	21 785
United States	19,9	31 806
Saudi Arabia	11,0	6 853
Iran	3,8	1 493
Indonesia	1,2	986

   (a) Draw a scatter plot of the data set and your estimate of a line of best fit.
   (b) Calculate equation of the line of regression using the method of least squares.
   (c) Draw the regression line equation onto the graph.
   (d) Calculate the correlation coefficient r.
   (e) What conclusion can you reach, regarding the relationship between CO_2 emission and GDP per capita for the countries in the data set?

2. A collection of data on the peculiar investigation into a foot size and height of students was recorded in the table below. Answer the questions to follow.

Length of right foot (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 and your estimate of a line of best fit.
   (b) Calculate equation of the line of regression using the method of least squares or your calculator.
   (c) Draw the regression line equation onto the graph.

(d) Calculate the correlation coefficient r.

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

3. 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.

(a) Is there a strong association 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.

Learner	Test 1 (Full marks: 50)	Test 2 (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

4. A fast food company produces hamburgers. The number of hamburgers made, and the costs are recorded over a week.

Hamburgers made	Costs
495	R2 382
550	R2 442
515	R2 484
500	R2 400
480	R2 370
530	R2 448
585	R2 805

(a) Find the linear regression function that best fits the data.

(b) If the total cost in a day is R2 500, estimate the number of hamburgers produced.

(c) What is the cost of 490 hamburgers?

5. The profits of a new shop are recorded over the first 6 months. The owner wants to predict his future sales. The profits so far have been R90 000; R93 000; R99 500; R102 000; R101 300; R109 000.

(a) For the profit data, calculate the linear regression function.

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

(c) The owner wants a profit of R130 000. Estimate how many months this will take.

6. 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.

 More practice  video solutions  or help at www.everythingmaths.co.za

(1.) 01b5 (2.) 01b6 (3.) 01b7 (4.) 01b8 (5.) 01b9 (6.) 01ba