Chapter 2 Bivariate Numerical Variables

NUMERICAL VARIABLES

Variables that represent quantities or measurements which are given as a number are said to be numerical variables. Numerical variables represent a measurable quantity.
For example:

- the weight of a person
- the height of a tomato plant
- the number of mobile phones in a household
- distance between towns

Numerical data is further classified as being **discrete** or **continuous**.

Discrete Numerical Variables

Discrete numerical variables are such that their values may only be represented by particular numbers and are usually compiled by counting.

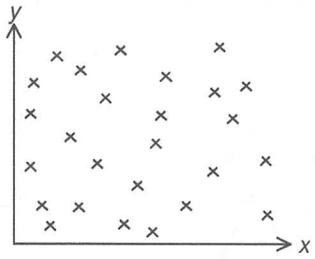
Continuous Numerical Variables are such that they can take on any real number value within a particular interval and are collected by a measuring process.

SCATTERGRAPHS or SCATTERPLOTS

The graph of a bivariate distribution having two numerical variables is called a scatter diagram, scattergram or scatterplot and gives some idea how the two variables are related. A scatterplot is the numerical equivalent of a two-way frequency table displaying categorical variables.
Consider the following scatter diagrams.

1. No identifiable relationship

The plotted points of the scatter diagram are scattered randomly and there is no obvious relationship between the two variables x and y.



2. Linear relationship

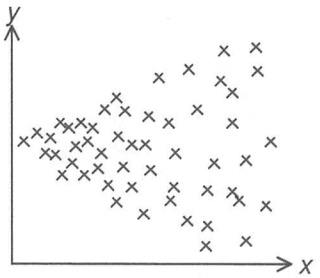
The plotted points tend to form a line. This informs us that there appears to be a linear relationship between x and y.

3. Curvilinear relationship.

The points tend to lie along a curve. In the example shown the plotted points appear to form a parabola.

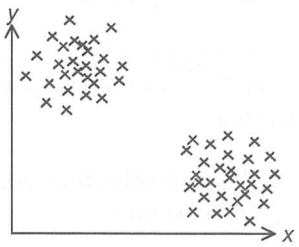
4. Delta relationship

The points tend to fan out.



5. Cluster relationship.

The points form two (or more) distinct groups. This may be because the data comes from two different populations.



Examination of the given scattergrams reveals that two variables may be related in a number of different ways.

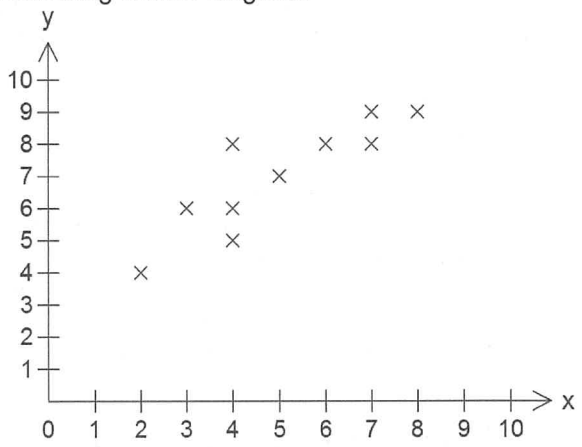
In our study we shall restrict our attention to numerical bivariate data which exhibits a **linear** trend.

LINEAR RELATIONSHIPS

Consider the following bivariate distribution.

x	2	5	4	6	4	3	7	4	7	8
y	4	7	8	8	5	6	8	6	9	9

Graphing the data contained in the table above gives us the following scatter diagram.



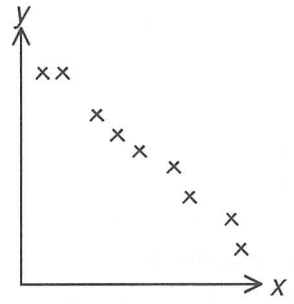
Examination of the graph reveals that a low value of x is associated with a low value of y, a middle value of x is associated with a middle value of y and a high value of x is associated with a high value of y. This examination reveals that although no precise prediction of y can be made for a given value of x, the associations between x and y are strongly related or correlated.
Linear relationships in which the points align closely to form a line are said to have a **high correlation** or a **strong linear relationship**.

Linear relationships in which the points align loosely to form a line are said to have a **low correlation** or a **weak linear relationship**.

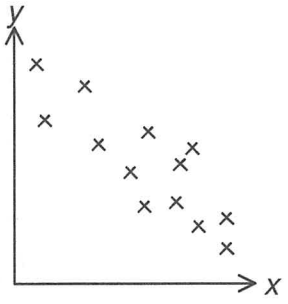
Linear relationships in which the two variables vary together, that is, as one variable increases the other increases and as one variable decreases so does the other, are said to have a **positive correlation** (as in the example above).
Linear relationships in which the two variables vary in opposite direction, that is, as one variable increases the other decreases are said to have a **negative correlation**.

Summing up we can say that a **linear relationship** is characterised by **two measures**, these being the **strength** of the relationship and the **direction** of the relationship.

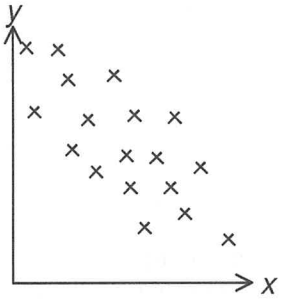
The following scatter diagrams exhibit different levels of correlation (or linear relationship):



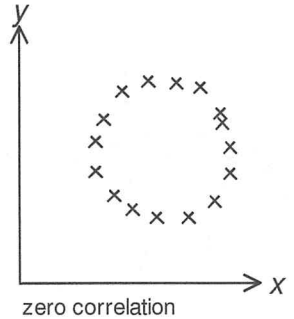
very strong negative correlation



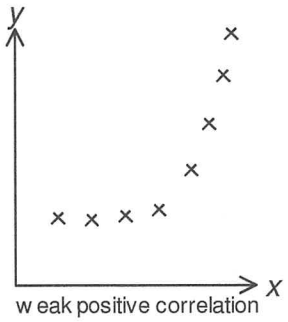
strong negative correlation



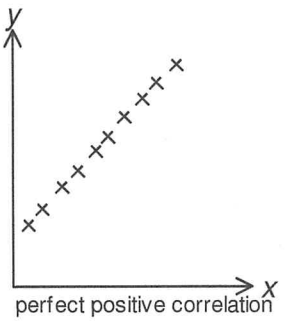
moderate negative correlation



zero correlation



w eak positive correlation



perfect positive correlation

When commenting on a scatterplot we should consider each of the following:

- **SHAPE:** Examine the shape of the distribution of the data points to see if the relationship between the variables looks **linear** or **not linear**.
- **OUTLIERS:** Check for any points which are unusual in their location compared to all the other points.
- **DIRECTION:** Decide whether the linear relationship is **positive** or **negative**.
- **STRENGTH:** Examine the distribution of the points and comment on the strength of the linear association of the variables under consideration.

NOTE: **Scatterplots may show that there is a relationship between two variables but this does not imply that a change in one of the variables causes a change in the other variable.**

EXERCISE 2A

1. Match each scatter diagram with one of the following descriptions of correlation.

(a) weak negative linear relationship

(b) very strong positive linear relationship

(c) moderate positive linear relationship

(d) perfect negative linear relationship

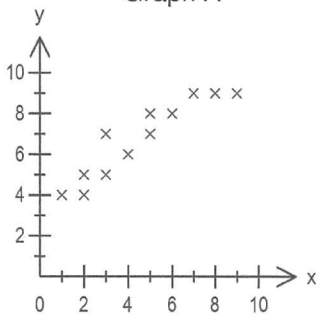
(e) perfect positive linear relationship

(f) moderate negative linear relationship

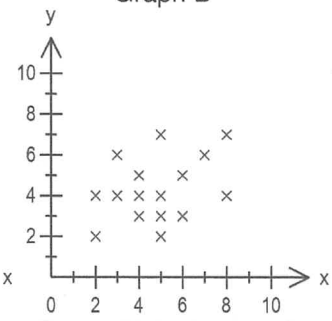
(g) very strong negative linear relationship

(h) weak positive linear relationship

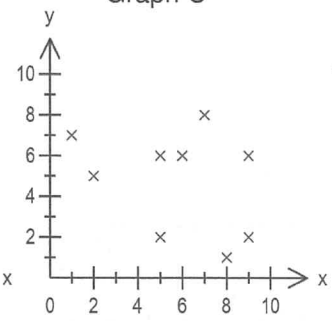
Graph A



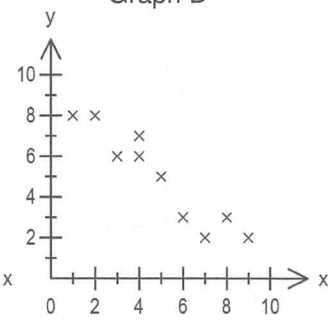
Graph B



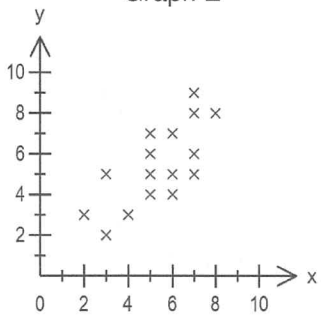
Graph C



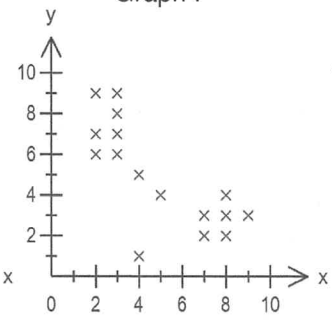
Graph D



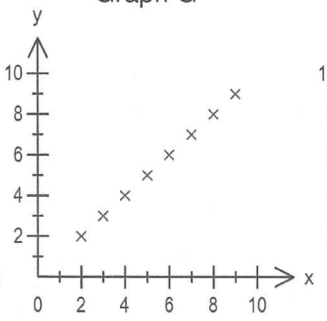
Graph E



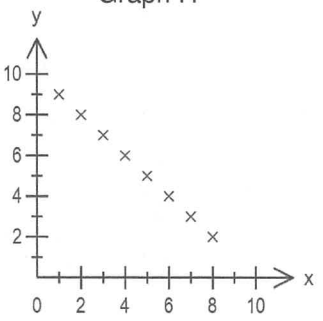
Graph F



Graph G



Graph H



2. For each of the paired variables:
(i) state which is the explanatory variable and which is the response variable.
(ii) give a brief description in terms of **strength** and **direction** of the linear relationship for each pair of the given variables.

(a) The weight of person.
The height of a person.

(b) The number of kilometres a car tyre has travelled.
The depth of the tyre tread.

(c) The number of litres of petrol purchased.
The price paid.

(d) The year of a persons birth.
The age of the person.

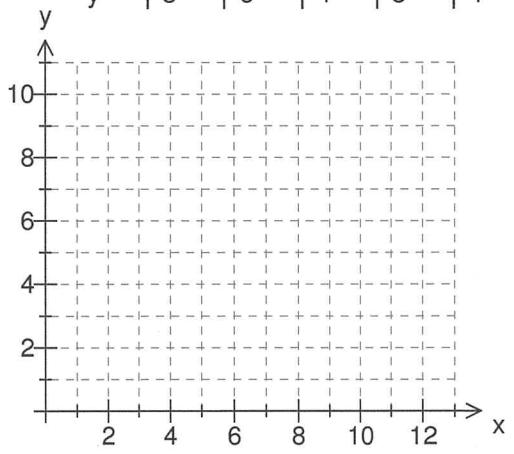
(e) Income tax on salaries.
Salaries earned.

(f) The price paid for second-hand cars.
The age of a second-hand car.

3. On the axes provided draw a scatter diagram of each bivariate distribution and then comment on the shape, strength and direction of the linear relationship between the two variables.

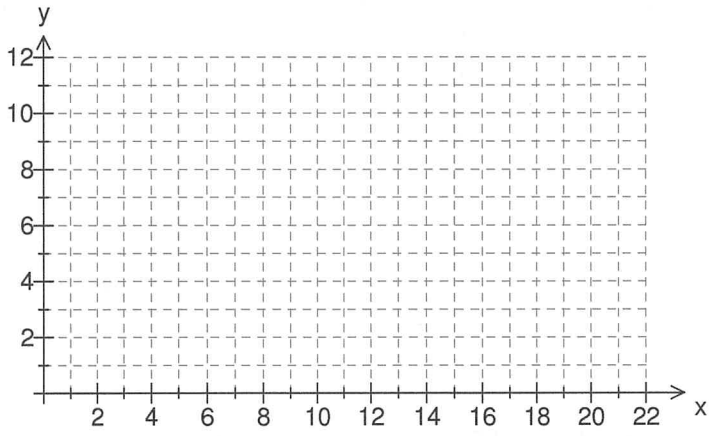
(a)

x	4	7	2	10	3	5	6	9	8	4
y	5	6	4	8	4	4	5	7	6	3



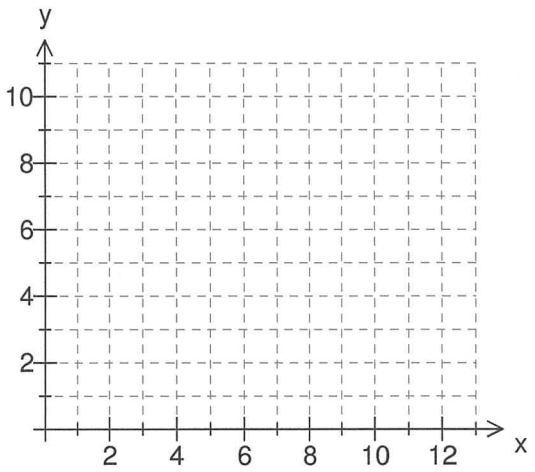
(b)

x	11	15	12	10	13	15	16	11	18	17
y	8	10	6	8	7	2	5	4	9	7



(c)

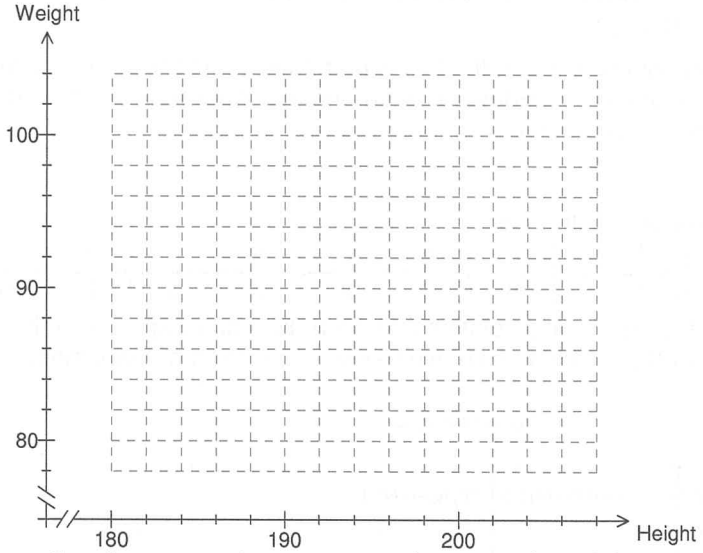
x	2	5	10	4	3	7	8	9	6	5
y	6	5	4	6	5	4	5	4	6	7



4. The table below gives the height and weight of 10 football players belonging to an AFL team.

Height (cm)	198	203	184	205	194	190	207	188	201	199
Weight (kg)	89	98	80	101	83	91	103	85	97	88

(a) On the axes provided, draw a scatter diagram of the given information.



(b) Examine your scatter diagram and comment on the strength and direction of the relationship between the height and weight of these football players.

(c) A football player belonging to this team is 200 cm tall. Explain how you can use the given information to estimate this footballers' weight. Give this estimate.

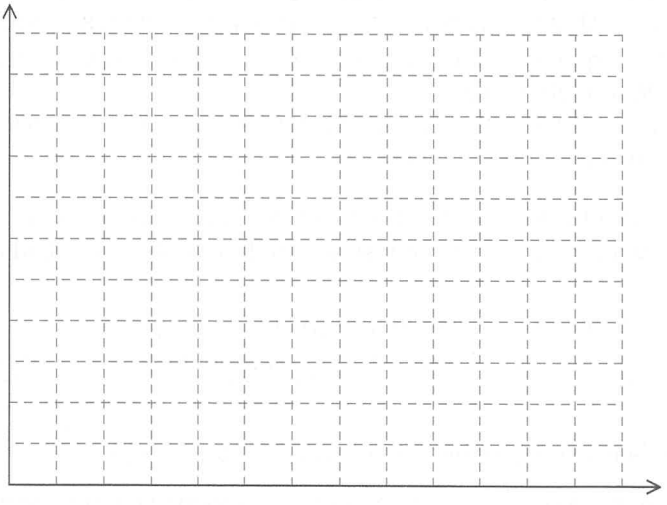
(d) A fan of this team has a height of 200 cm. Comment on the validity of using the given information to estimate this fans weight.

5. The table below gives the number of hours 10 students spent watching videos during the weekend prior to their science test and their results for that test out of a possible mark of 10.

Science test mark	9	5	7	2	8	5	4	9	4	6
Hours spent watching videos	2	8	4	10	5	6	7	3	9	4

(a) For this data set state the explanatory variable and the response variable.

(b) Draw a scatter graph for Hours spent watching videos and Science test marks.



(c) Describe in words the relationship shown in your graph.

MEASURING LINEAR RELATIONSHIP

From your experience in working through Exercise 2A you should have found that it is not very satisfactory describing linear relationships in terms of strong, moderate or weak as the terms used are subjective, a more precise measure is required.
This precise measure is **Pearson's correlation coefficient** or commonly called the **correlation coefficient** and it is denoted by r_{xy} or simply r.

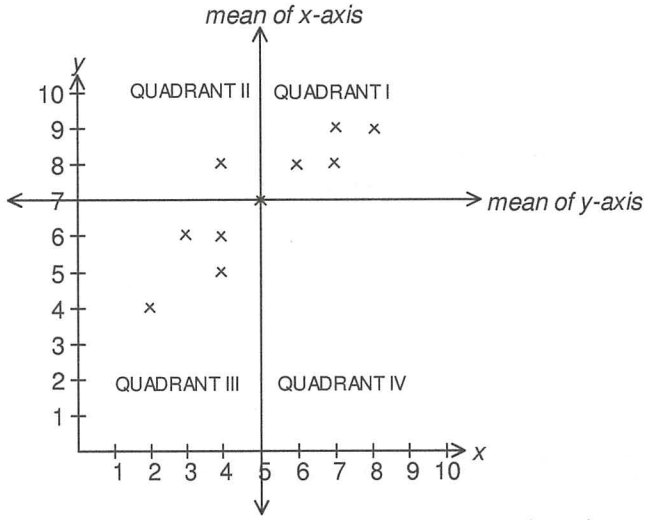
Although not a requirement of this course, the following pages outline how the correlation coefficient is calculated. This course requires the student to determine the correlation coefficient using a calculator given bivariate data in a table or in a scatterplot.

Covariance

Consider the following bivariate distribution.

x	2	5	4	6	4	3	7	4	7	8
y	4	7	8	8	5	6	8	6	9	9

To derive a numerical measure of linear relationship we draw a quadrant plot of the scatter diagram using the means of the variables x and y as axes as shown below. The mean of the variable x is 5 and the mean of the variable y is 7.



Now the mean of x-axis divides the points in the scatter diagram into those that lie above the mean of x, on the mean of x or below the mean of x. In other words into points whose mean deviations $(x\,\text{score} - \bar{x})$ are positive, zero or negative.
The mean of y-axis also divides the points in the scatter diagram into those that lie above the mean of y, on the mean of y or below the mean of y. In other words into points whose mean deviations $(y\,\text{score} - \bar{y})$ are positive, zero or negative.
If we consider the product of these mean deviations, that is $(x\,\text{score} - \bar{x})(y\,\text{score} - \bar{y})$, we can see that this product is positive in quadrants I and III, zero on the axes and negative in quadrants II and IV.

For our example, we can see that most of the points have mean deviations with the same sign, hence their products will be positive. That is, $(x\,\text{score} - \bar{x})(y\,\text{score} - \bar{y}) > 0$ this result corresponds with a high value of x being associated with a high value of y and a low value of x being associated with a low value of y, which indicates a positive linear relationship.
If the majority of points are such that $(x\,\text{score} - \bar{x})(y\,\text{score} - \bar{y}) < 0$ then this indicates that a negative linear relationship exists between the two variables.

Now on finding the *mean* of the product of the mean deviations we obtain a numerical measure of the linear relationship between two variables. This measure denoted by s_{xy}, is called the **covariance** of x and y.

Now if we have n data points (x_1,y_1), (x_2,y_2), (x_3,y_3), ... (x_n,y_n).

Then covariance, that is $s_{xy} = \dfrac{(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + (x_3-\bar{x})(y_3-\bar{y}) + \ldots + (x_n-\bar{x})(y_n-\bar{y})}{n}$

The covariance formula is usually written as follows: $s_{xy} = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{n}$

Thus we now have a numerical measure for the linear relationship between two variables. This numerical measure covariance, avoids subjective descriptions of the relationship between two variables and hence should be more acceptable.
Let us now determine the covariance for our bivariate distribution using the formula derived above.

To find the covariance examination of the formula informs us that we will need to find each of the following:

(i) $\bar{x}$ and $\bar{y}$ (ii) $(x\text{ score} - \bar{x})$ and $(y\text{ score} - \bar{y})$

(iii) $(x\text{ score} - \bar{x})(y\text{ score} - \bar{y})$ (iv) Sum of $(x\text{ score} - \bar{x})(y\text{ score} - \bar{y})$

A table will enable us to do this in an organised way.

	x score	x score $-\bar{x}$	y score	y score $-\bar{y}$	$(x\text{ score} - \bar{x})(y\text{ score} - \bar{y})$
	2	−3	4	−3	9
	5	0	7	0	0
	4	−1	8	1	−1
	6	1	8	1	1
	4	−1	5	−2	2
	3	−2	6	−1	2
	7	2	8	1	2
	4	−1	6	−1	1
	7	2	9	2	4
	8	3	9	2	6
TOTAL	**50**	**0**	**70**	**0**	**26**

Now $\bar{x} = \frac{50}{10} = 5$, $\bar{y} = \frac{70}{10} = 7$ and $s_{xy} = \frac{26}{10} = 2.6$

As our answer for the covariance for the data is positive it indicates that generally a low value of x is associated with a low value of y, a middle value of x is associated with a middle value of y and a high value of x is associated with a high value of y.

The magnitude of our answer does not really tell us very much about the strength of the linear relationship as we have nothing in place as yet to compare it to. In order to comment on the strength of the linear relationship between the variables x and y we will need to examine the scatter diagram of this distribution.

Correlation Coefficient

Consider the following data sets and their calculated covariance.

SET A

x	4	9	1	7	3
y	6	7	3	8	5

For set A $s_{xy} = 4.36$

SET B

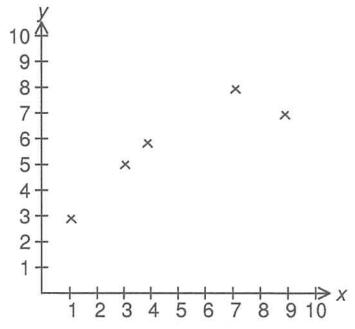
x	40	90	10	70	30
y	60	70	30	80	50

For set B $s_{xy} = 436$

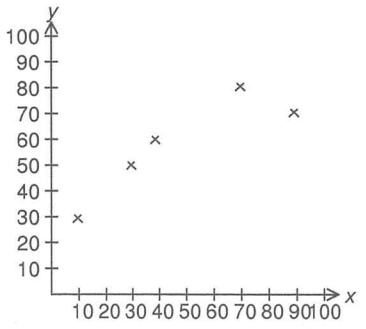
Both data sets have a positive linear relationship and it appears from the magnitude of the covariance that the linear relationship for B is far stronger than that for A.

The scatterplots of both sets are shown below:

Set A



Set B



The scatterplots of both data sets are the same, informing us that the linear relationship between x and y for both data sets is the same and it follows that the covariance for both data sets should be the same.

This example informs us that the covariance is not a reliable measure of correlation as identical scatterplots give rise to a different covariance. The reason why this occurs is because the covariance is scale dependent. Examination of the data sets reveals that data set A has been scaled by a factor of 10 to obtain data set B and consequently the covariance of B is 100 times that of set A.

This scale dependency of the covariance makes it useless as a measure of the linear relationship between two variables as distributions with the same correlation are given different numerical measures.

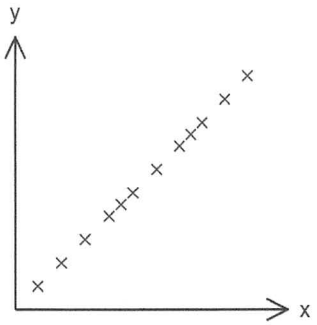
To overcome this scale dependence we divide the covariance by the standard deviation of x and the standard deviation of y. The statistic obtained is known as **Pearson's correlation coefficient** or more simply as the **correlation coefficient** is denoted by r_{xy} and calculated using the following equation: $r_{xy} = \frac{s_{xy}}{s_x s_y}$.

The correlation coefficient, r_{xy} measures the strength and the direction of the linear association between the two variables x and y and is obtained by standardising a statistic known as the covariance and hence its value will always lie in the interval $-1 \leq r \leq 1$.
If $r = 1$, then all of the points in the scatter diagram lie in a line with a positive slope, that is we have a perfect positive linear correlation.
If $r = -1$, then all the points in the scatter diagram lie in a line with a negative slope, that is we have a perfect negative linear correlation.
If r is close to zero then this indicates that there is little or no association between x and y. To confirm that there is no association between x and y the scatter diagram must be examined as there may be a non linear association between the two variables, for example a parabolic association.
Now we have a measure for linear relationship that not only gives the direction but also a meaningful measure of the strength.

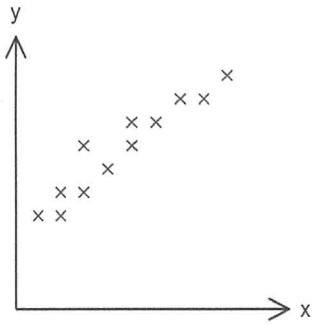
Interpreting Pearson's Correlation Coefficient
The following correlation scale gives meaning to the numerical values of the correlation coefficient.

Value of r	Degree of correlation or linear relationship
$r = 1$	perfect positive correlation
$0.7 < r < 1$	strong positive correlation
$0.5 < r \leq 0.7$	moderate positive correlation
$0.3 < r \leq 0.5$	weak positive correlation
$0 < r \leq 0.3$	no significant correlation
$r = 0$	no *linear* correlation
$-0.3 \leq r < 0$	no significant correlation
$-0.5 \leq r < -0.3$	weak negative correlation
$-0.7 \leq r < -0.5$	moderate negative correlation
$-1 < r < -0.7$	strong negative correlation
$r = -1$	perfect negative correlation

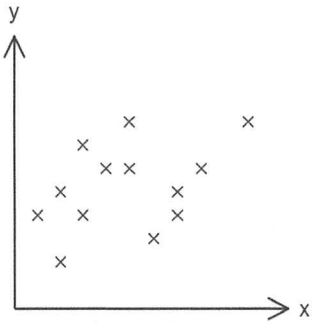
The diagrams below describe different types and strength of linear relationships that exist between two variables of a bivariate data set together with the correlation coefficient.



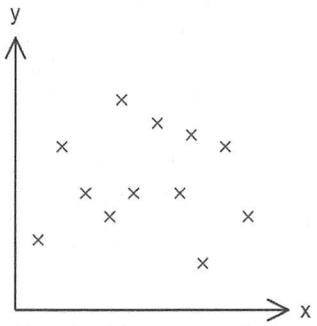
Perfect positive linear
r = 1



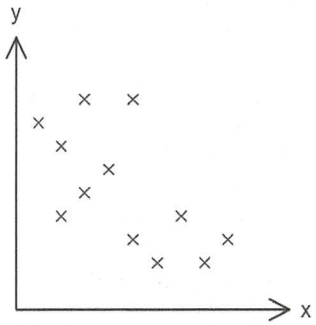
Strong positive linear
r = 0.95



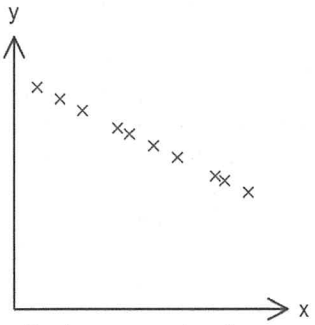
Weak positive linear
r = 0.41



No significant correlation
r = 0.01



Moderate negative linear
r = -0.62



Perfect negative linear
r = -1

Example 1

For the following set of scores:

(a) determine the correlation coefficient correct to 3 decimal places.
(b) interpret the determined correlation coefficient.
(c) write a sentence describing the relationship between the variables x and y.

x	3	9	4	7	6	2	5	1
y	5	10	5	8	9	3	6	2

Enter the given bivariate data into a calculator and then read off the required statistic.

(a) $r = 0.965$.
(b) The value of the correlation coefficient indicates a very strong positive linear relationship between x and y.
(c) The value of r indicates that the value of y should increase as the value of x increases.

Note: 1. Perfect correlation coefficients of $r = 1$ and $r = -1$ are unusual.
2. The value of r lies in the interval $-1 \leq r \leq 1$.
3. $r = 1$ does not mean that the slope of the line is 1, it indicates that the relationship between x and y is a positive one and that we have a perfect linear relationship.

EXERCISE 2B

1. For each of the given tabled values of x and y, give the following :
(i) mean and standard deviation of the variable x. (ii) mean and standard deviation of the variable y.
(iii) the correlation coefficient. (iv) a word description of the relationship between x and y.

(a)

x	10	8	8	7	7	7	6	5
y	13	15	16	12	8	6	4	3

(b)

x	1	3	5	7	8	9	11	14
y	9	8	11	7	14	13	15	16

(c)

x	46	38	34	57	46	47	38	42
y	48	48	33	54	46	47	44	43

(d)

x	2.8	2.5	1.1	1.5	2.7	2.4	2.2	1.4	1.9	2.6
y	1.1	1.3	1.1	1.2	1.3	1.7	1.5	2.3	1.1	1.4

(e)

x	-0.23	-0.34	-0.39	-0.42	-0.58	-0.59	-0.63	-0.72	-0.77	-0.89
y	11	12	14	14	16	18	19	22	20	21

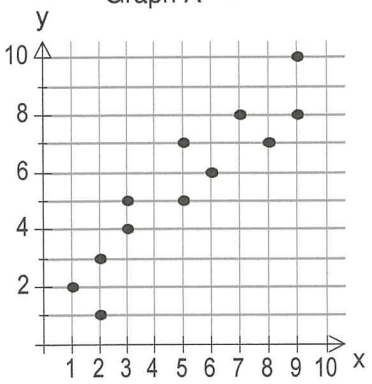
2. (a) Using the data in 1(e) above, change the sign of the variable x and then calculate r_{xy}.
How does your answer differ to that found in 1(e)?

(b) Using the data in 1(e) above, change the sign of the variable y and then calculate r_{xy}.
How does your answer differ to that found in 1(e)?

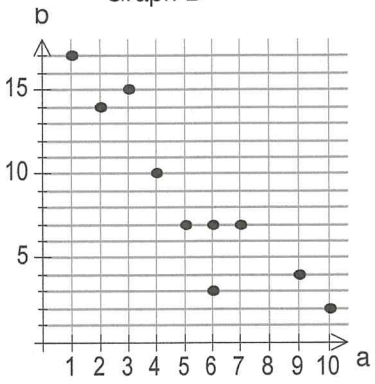
(c) Using the data in 1(e) above, change the sign of both x and y, then calculate r_{xy}.
How does your answer differ to that found in 1(e)?

3. For the given scatterplots determine the following statistics:
   (i) mean and standard deviation of the explanatory variable.
   (ii) mean and standard deviation of the response variable.
   (iii) correlation coefficient.

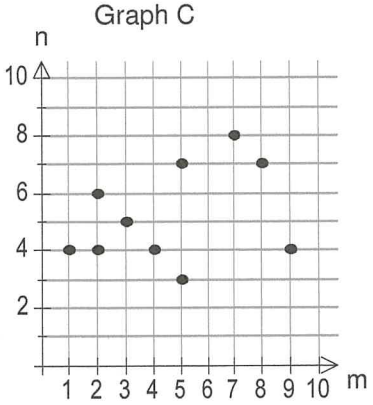
Graph A



Graph B

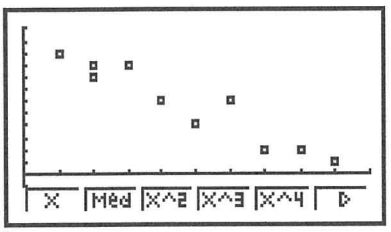


Graph C

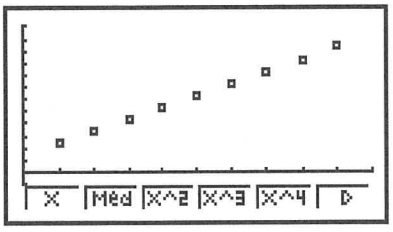


4. Match each of the following correlation coefficients with one of the graphs shown below.
   r = 0.84 r = -0.95 r = 1 r = 0.14 r =-0.32 r = -1

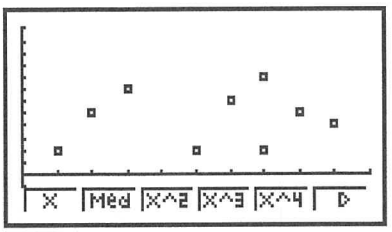
(a)



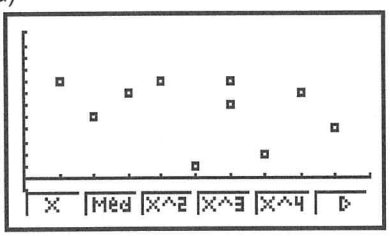
(b)



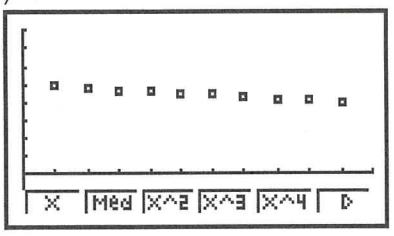
(c)



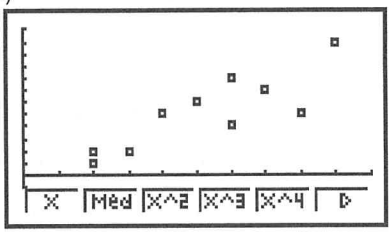
(d)



(e)



(f)



5. Draw a scatter diagram to show:
   (a) high negative correlation, and
   (b) little or no correlation.

6. The table below gives the results of two tests for a class of 18 Unit 3 Applications mathematics students.

STUDENT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
TEST 1	90	67	50	90	75	70	93	75	65	85	66	91	57	33	66	92	93	79
TEST 2	78	82	35	86	86	81	87	87	76	88	64	85	33	50	68	78	86	82

(a) Identify the response and explanatory variables.

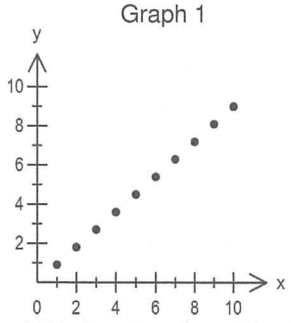
(b) Find the correlation coefficient correct to two decimal places for the test scores.

(c) What does the value of the correlation coefficient imply?

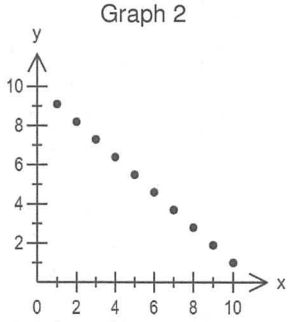
(d) Describe the relationship between the two variables.

7. Shown below are four graphs in which all of the points lie in a straight line.

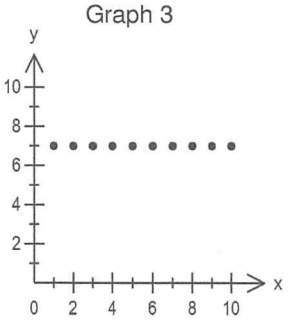
Graph 1



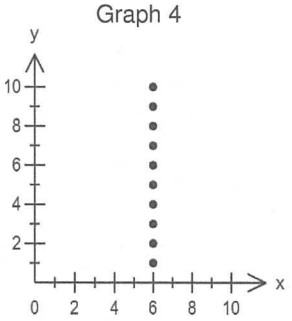
Graph 2



Graph 3



Graph 4



Comment on the statement:
"In each case above the correlation coefficient has a magnitude of one because all the given points in each graph line up to form a perfect straight line."

8. A new airline "Rex" in Western Australia advertised the following return airfares for the "off-peak" period.

From Perth to	Return Airfare ($)	Distance (km)
Albany	419	389
Alice Springs	795	1992
Broome	548	1683
Darwin	768	2648
Esperance	448	602
Kalgoorlie	450	544
Karratha	548	1279
Learmonth	496	1128
Port Headland	548	1348

(a) Identify the explanatory variable and response variable.

(b) On your computer or calculator construct a scatterplot of the given information and comment on your resulting graph.

(c) Calculate a statistic to determine the strength of the relationship between your variables. Comment on your calculated statistic.

(d) During the "peak" period Rex airlines increases all fares by 40%. Determine the correlation coefficient between "peak" fares and "off-peak" fares. Comment on your answer.

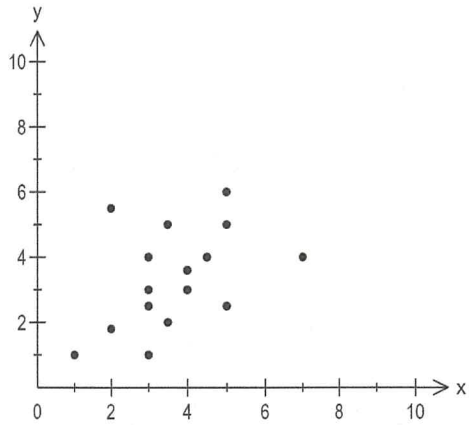
OUTLIERS AND THE CORRELATION COEFFICIENT

An outlier is an ordered pair of variables which is not in keeping with the general trend of the data. In this context the values of x and or y need not be unacceptably large or unacceptably small, but their combination places the point formed by their values well outside the given data set.
On identifying an outlier it may be removed so that any further discussion and calculations are more meaningful.
As discussed in Unit 2, the mean and standard deviation are affected by an outlier and since covariance and the correlation coefficient both rely on these statistics they are also affected by outliers.

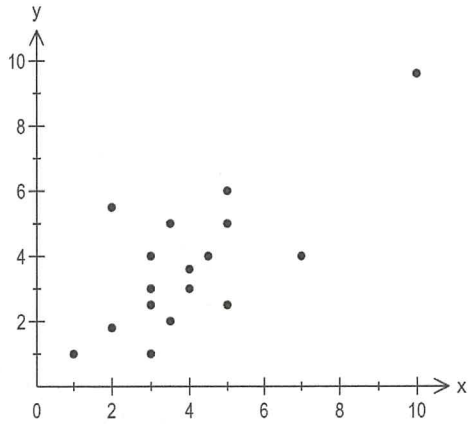
An outlier, depending upon where it falls in a data set may **increase** the correlation coefficient or **decrease** the correlation coefficient.

An outlier increasing the correlation coefficient

Consider the following scatter plots showing the same set of data.



The correlation coefficient of the scatter plot above is 0.4298 rounded to 4 decimal places. This indicates that a weak positive linear relationship exists between the variables x and y.

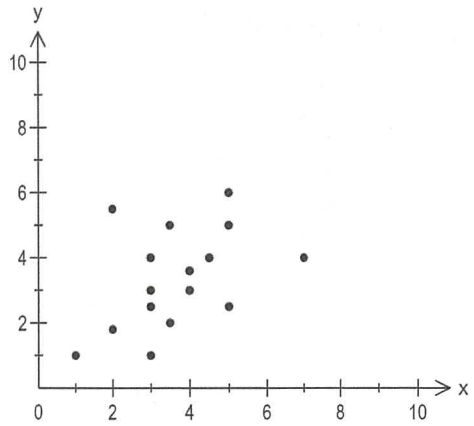


The scatter plot above shows the same data as the one on the left with an outlier included. The correlation coefficient for this data is 0.7291. This indicates that there is a strong positive linear relationship between x and y.

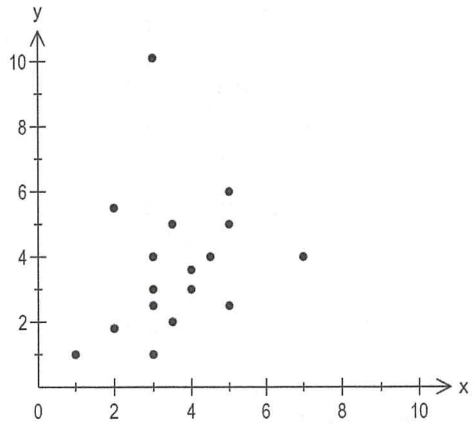
In this situation we can see that the outlier falls in the path of the general trend of the data set and as a result its inclusion has increase the magnitude of the correlation coefficient indicating a stronger linear relationship between the variables x and y.

An outlier decreasing the correlation coefficient

Consider the following scatter plots showing the same set of data.



The correlation coefficient of the scatter plot above is 0.4298 rounded to 4 decimal places. This indicates that a weak positive linear relationship exists between the variables x and y.



The scatter plot above shows the same data as the one on the left with an outlier included. The correlation coefficient for this data is 0.2047. This indicates that there is no significant relationship between x and y.

In this case the outlier falls well outside the path of the general trend of the data set and as a result its inclusion had decreased the magnitude of the correlation coefficient indicating that there is no significant relationship between the variables x and y.

Example 2 Consider the following set of scores:

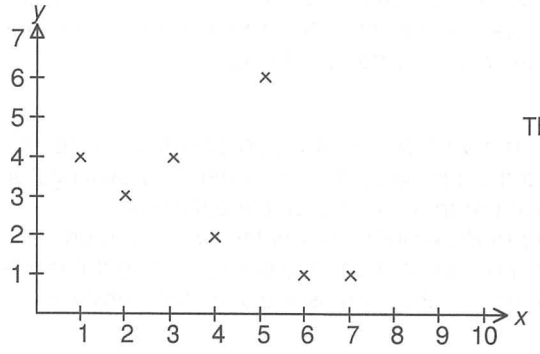
x	3	5	2	6	4	1	7
y	4	6	3	1	2	4	1

(a) Find the correlation coefficient for the given data.
(b) Draw a scattergram and identify the outlier.
(c) Find correlation coefficient after removing the outlier.
(d) Comment on effect the removal of the outlier has on the linear relationship between the variables.

Solution:

(a) Using a calculator $r_{xy} \approx -0.4648$.

(b)



The outlier is the point (5,6).

(c) Removing the outlier (5, 6) we obtain the following: $r_{xy} \approx -0.9082$

(d) When the outlier is included the linear relationship between x and y is negative and weak ($r_{xy} \approx -0.4648$), removal of the outlier results in the variables exhibiting a strong negative linear relationship ($r_{xy} \approx -0.9082$. Therefore it can be seen that outliers can significantly change the strength of a linear relationship between two variables as indicated by the correlation coefficient.

In example 2 above the presence of an outlier had the effect of reducing the magnitude of the correlation coefficient indicating a decrease in the strength of the association between the two variables.

NOTE: Describing the relationship between two variables using the correlation coefficient is only reliable if (i) the data is linear in form, (ii) the data set is not small and (iii) the data does not contain outliers.

COMMENT ON THE CORRELATION COEFFICIENT

1. r_{xy} or r is called Pearson's Correlation Coefficient or the product moment correlation coefficient or just the correlation coefficient.
2. r is a number and lies in the interval $-1 \leq r \leq 1$ and it has **no units**.
3. For any bivariate distribution r is a measure of the strength and direction of a ***linear*** relationship. For bivariate distributions with a non-linear relationship applying a value of r is not appropriate and other measures must be used to measure the non-linear relationship.
4. If the units of measurement of data points being considered change this will not affect the value of r. For example, suppose the value of r was determined using the heights and weights of a number of people in inches and pounds. If the heights were converted to centimetres or the weights converted to kilograms, or both, the value of r would remain the same value.
5. Perfect correlation does not exist for horizontal and vertical sets of data points as in both of these cases there is only one variable as the other is constant, hence the idea of r in these situations is meaningless. Revisit question 7 Exercise 2B on page 33.
6. r generally gives a good indication of the linear relationship between two variables when a large sample of continuous data pairs are considered. Situations involving small samples and/or only discrete data pairs may give a degree of correlation which may not be very reliable. In the case of small samples consulting a table of r values enables one to tell whether the relationship is unlikely to have occurred by chance.
7. r_{xy} was derived to give a mathematical measure of the strength and direction of a linear relationship between two variables. This measure, that is r, does NOT imply cause.

ASSOCIATION AND CAUSATION

When considering the observed association with two variables one must not assume a causal relationship between them. The association between the two variables could be due to either coincidence or the presence of a third (or more) variable known as a **lurking variable** or **confounding factor** or variable. Consider the following situations.

Situation 1

If we found that there was a high correlation between eating in restaurants and high cholesterol we could not conclude that eating in restaurants causes one to have high cholesterol. This high association may be due to either coincidence or the presence of another variable or variables There are many other variables in this equation that we did not take into account that may be the cause or help contribute to the cause, such as alcohol intake, type of food eaten, age of people in the survey, stress levels of people involved and so on.

The value of r cannot be the basis on which we can state that one variable is the cause of another. One must take into account the fact that most cause-effect relationships involve more than just the two variables being considered also the two variables in question may not be genuinely related, they are spuriously related and may be caused by some third variable which may or may not be obvious.

Situation 2

If a high correlation was found between the number of ice creams sold and the number of drowning deaths could we conclude that an increase in ice cream sales causes an increase in drownings or the greater the number of drownings the greater the sale of ice creams? Not a logical conclusion.
The variables ice cream sales and the number of drownings are related to a common variable, the daily temperature. The higher the temperature the more ice creams are sold and also the more people go swimming thus the more drownings. The lurking variable here is the daily temperature.

Situation 3

Over recent years sales of mobile phones and sales of athletic footwear have increased significantly and hence there is a high correlation between sales of mobile phones and sales of athletic footwear.
Just because there is a high correlation between these two variables we cannot assume that buying mobile phones causes people to buy athletic footwear or buying athletic footwear causes people to buy mobile phones.
In this case the high association between sales of mobile phones and sales of athletic footwear may be due to **coincidence.** Alternatively it may be due to higher degree of affluence or some other variable (s).

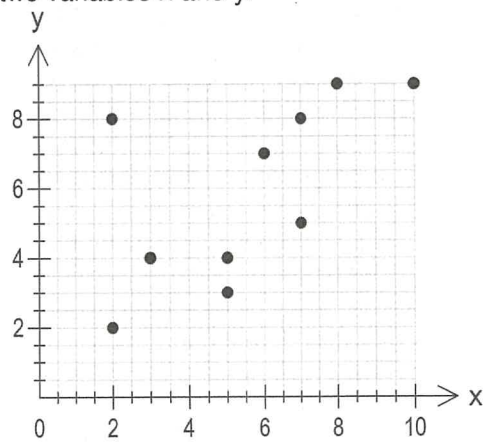
Situation 4

A newspaper reported that students who eat breakfast perform better at school. Follow up studies found that there was a correlation between eating breakfast and higher academic achievement, also the researchers found that eating breakfast did not affect the intelligence of students. The research found that the type of student who did not eat breakfast due to various circumstances in their daily lives were also the type of student that had trouble at school.
It may appear to some people that the newspaper reported on a causation when it really reported on a correlation or association.

Summing up: Just because there is a high value of r this does not mean that one variable causes the other. There could be a lurking variable(s) that actually is the cause.

EXERCISE 2C

1. The scattergram below shows the relationship between the two variables x and y.



(a) For the data shown in the scattergram find,
(i) mean of x (ii) mean of y

(iii) standard deviation of x

(iv) standard deviation of y

(v) correlation coefficient

(b) The point (2, 8) was deemed to be too far from the trend of the data and it was removed.
Recalculate the statistics in (a) after the removal of the point (2, 8).

(c) Comment on the effect the removal of the outlier had on the correlation coefficient.

2. The table below gives the number of hours 10 students spent watching videos during the weekend prior to their science test and their results for that test out of a possible mark of 10.

Hours spent watching videos	2	3	4	4	5	6	7	8	9	10
Science test mark	9	9	7	6	8	5	4	5	4	9

(a) Identify the response and explanatory variables. Justify your choice.

(b) Find the mean test mark and the mean number of hours spent watching videos.

(c) Find the median test mark.

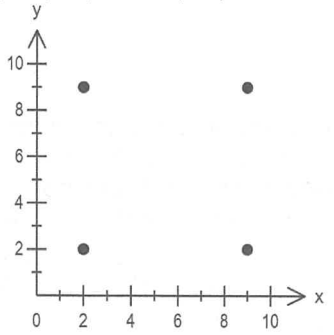
(d) Is there a relationship between the number of hours spent watching videos and the marks achieved in the Science test? Justify your answer mathematically.

On checking the test papers it was found that the mark of 9 for the student who had watched 10 hours of videos should have been a mark of 2.

(e) Find the correct mean test mark and median test mark.

(f) Does this correction alter the relationship between test marks and hours spent watching videos? Mathematically justify your answer.

3. The four points A(2,9), B(2,2), C(9,2) and D(9,9) shown on the scatter diagram form a square.



Answer the following questions **without the aid of a calculator** and check you answers using a calculator.

(a) By carefully considering the relationship between the points write down correlation coefficient for the four points.

(b) Write down the correlation coefficient if the points A and C are removed.

(c) Write down the correlation coefficient if the points B and D are removed.

(d) Find r_{xy} if A and B are removed.

(e) Find r_{xy} if A and D are removed.

If the point A only is removed the correlation coefficient of the remaining points is 0.5.

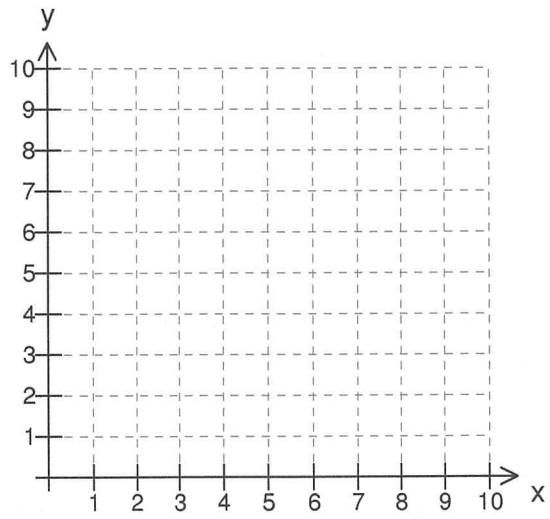
(f) If the point B only is removed, what is the value of r_{xy} ?

(g) If the point C only is removed, what is the value of r_{xy} ?

(h) If the point representing the centre of the square was introduced, what would the value of r_{xy} be?

4. (a) Sketch a scatter diagram of the following set of scores.

x	1	2	3	4	4	5	6	7	8	9
y	3	4	3	4	8	4	5	6	6	7



(b) Comment on the relationship between the variables x and y. Justify mathematically.

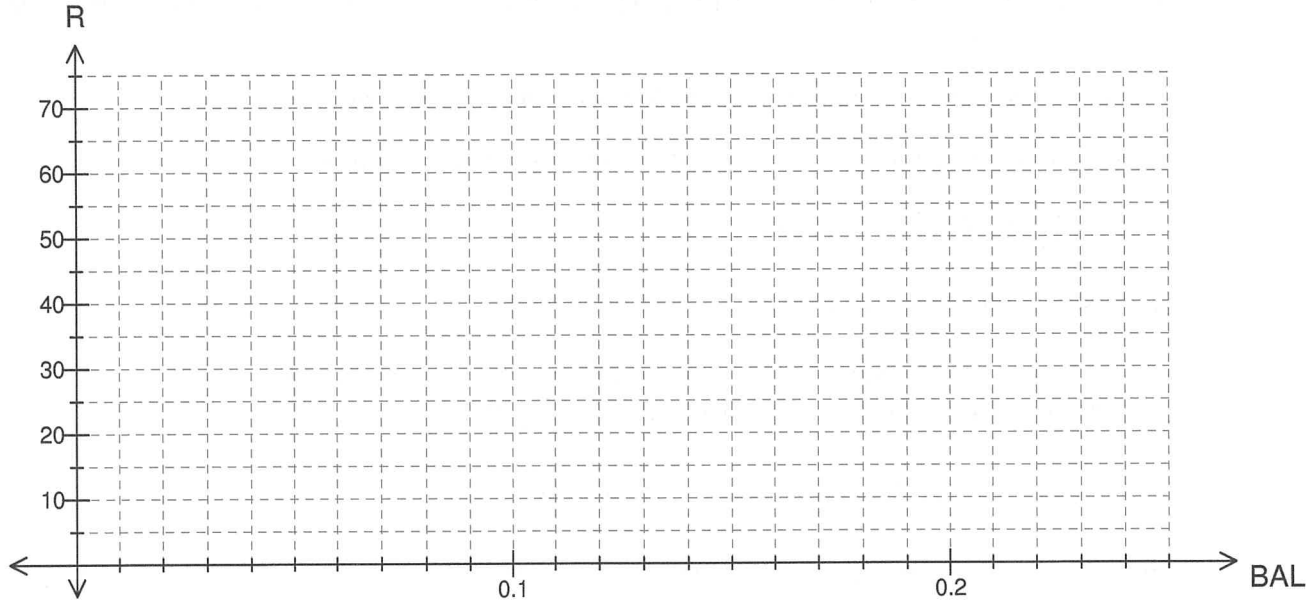
(c) It was found that the point (4, 8) should have been registered as (8, 4). Comment on the corrected relationship between the variables x and y. Justify mathematically.

(d) Compare and comment on your results with (a) and (b) above.

5. The risk of a motor vehicle accident is associated with the driver's blood alcohol level (BAL). The risk factor R is the number of times more likely it is that a driver who has alcohol in their bloodstream will have an crash than a driver with a blood alcohol level of zero.
The table below shows the relationship between blood alcohol level and R, the risk factor.

Blood Alcohol Level	0	0.04	0.08	0.10	0.15	0.17	0.18	0.20
Risk Factor R	1	2	4	15	28	40	55	70

(a) On the grid below draw the scatter plot for the given data.



(b) Calculate the value of r, the correlation coefficient, for the given data set.

(c) Describe the data using your r value.

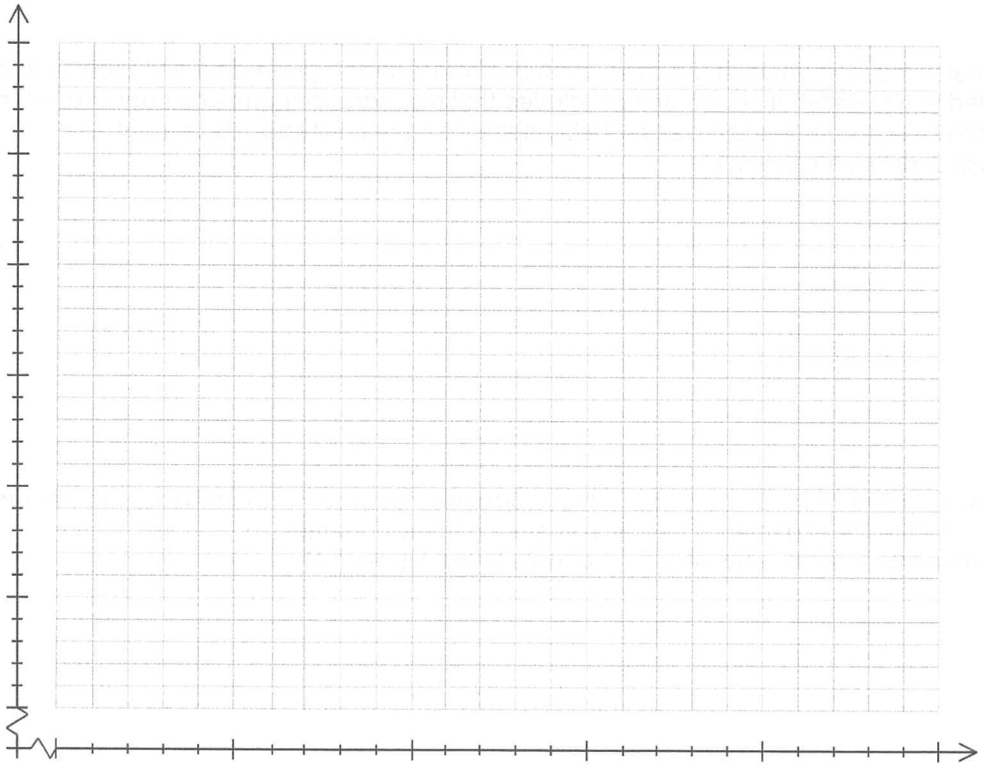
(d) Explain why the correlation coefficient is not useful in describing the given data set.

6. A group of 12 year old school children were given two fitness tests. The first test was a measure of maximum lung pressure before exercise and the second was a measure of heart rate taken one minute after strenuous exercise. The test results have been tabled below.

Child number	1	2	3	4	5	6	7	8	9	10	11	12	13
Lung pressure	48	26	45	36	40	43	30	29	38	44	35	39	33
Heart rate	80	95	84	90	85	84	82	94	90	82	88	86	92

(a) Is there mathematical evidence of an association between maximum lung pressure and recovery heart rate? Explain using a suitable statistic.

(b) On the give grid construct a scatter plot for the given data.



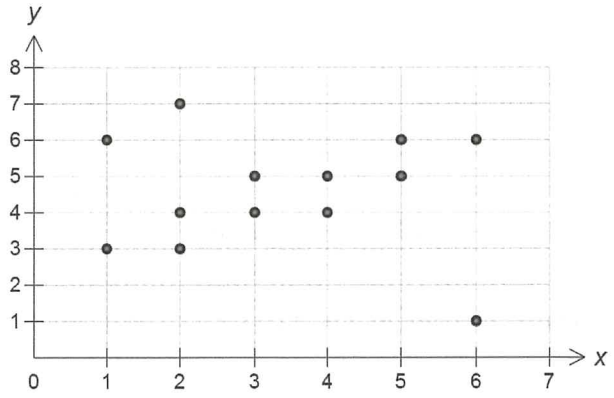
(c) From your scatter plot identify the child with the greatest variation from the main group.

(d) Would it make much difference to the correlation coefficient if the data from this child were treated as an outlier and removed? Discuss.

11. A survey of a large number of cities revealed a high correlation between the number of police officers and the number of crimes committed.
Does this mean that more police officers are causing more crimes to be committed? Discuss.

12. A study revealed a high correlation between the number of television sets per household and the life expectancy per person among many countries. Can we conclude therefore that television sets cause people to live longer? Comment.

13. The scattergram shows the relationship between the variables x and y.



(a) Calculate the value of the correlation coefficient and comment on the relationship between x and y.

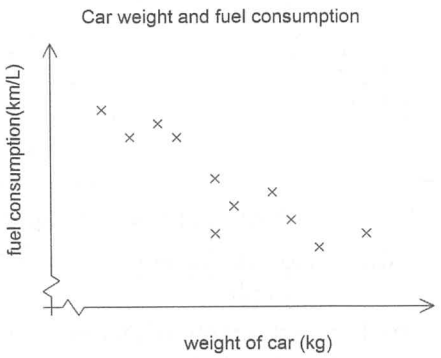
Closer examination of the scattergram reveals that three data points do not appear to fit the data set.
(b) List the outliers for this data set.

(c) Crop the data set to exclude these outliers and recalculate the correlation coefficient. Comment on the cropped relationship.

(d) If the outliers were real values of the data set of the distribution, justify their removal.

CHAPTER TWO REVIEW EXERCISE

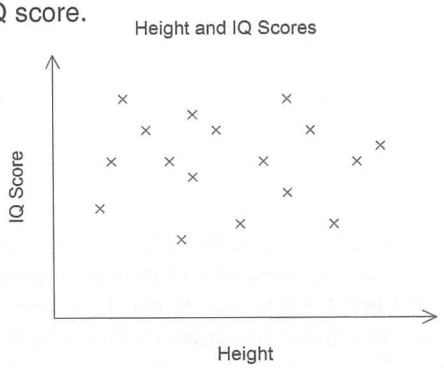
1. The scatter plot shows the weights (in kilograms) and fuel consumption (kilometres per litre) for a number of family cars.



(a) Identify the explanatory and response variables.

(b) Describe what the scatterplot shows about the relationship between these two variables.

2. The scatterplot shows the heights of students and their IQ score.



(a) Identify the explanatory and response variables.

(b) Describe what the scatterplot shows about the relationship between these two variables.

3. Consider the following real number values:

0.79, 1.05, 0.99, 1.0, -0.02, -1.99, -0.48, 0.7, -0.31, -1.0

From the given set of numbers found above select a value for the correlation coefficient r which you think is the most suitable to describe each of the following relationships between the variables under consideration. Enter your selections in the table at the bottom of the page.

(a) There is a very strong positive linear relationship between the two variables.

(b) There is a very weak negative linear relationship between the two variables.

(c) There is no significant relationship between the two variables.

(d) There is a linear relationship such that as one variable increases uniformly so does the other.

(e) There may be a weak linear relationship between the two variables where as one increases the other tends to decrease.

(f) There is a moderate relationship where as one variable decreases the other tends also to decrease.

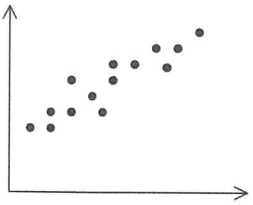
(g) There must be some error made in calculating the correlation coefficient.

(h) There is a linear relationship such that as one variable increases uniformly the other variable decreases.

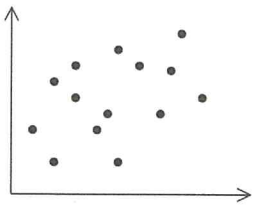
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Value of r								

4. Consider the scatter graphs shown below:

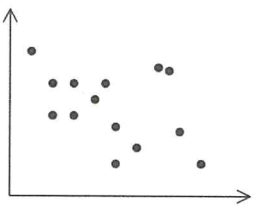
Graph A



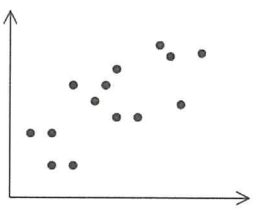
Graph B



Graph C



Graph D



(a) Match each graph with an appropriate correlation coefficient given in the table below.

Correlation coefficient	-0.4	0.4	0.7	0.9
Graph				

(b) For each graph (i) interpret the correlation coefficient and (ii) write a sentence describing the relationship between the explanatory variable and the response variable.

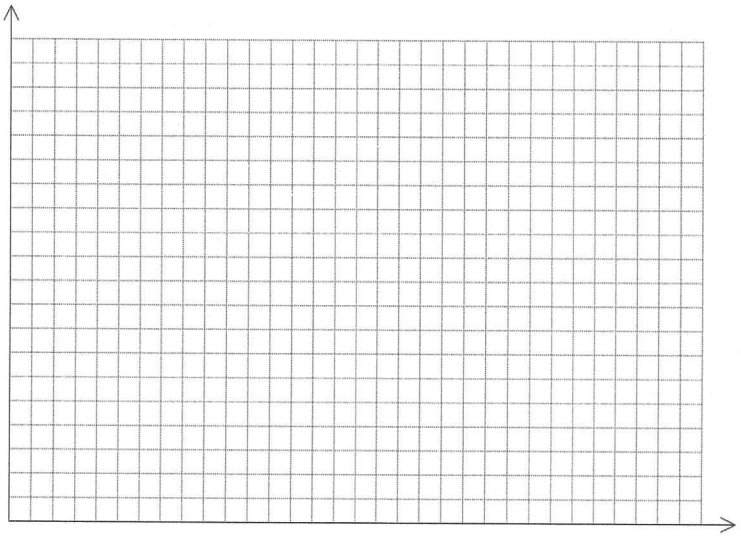
5. The head of mathematics at a senior college decided to examine whether or not there is a relationship between the first semester examination scores and the amount of time that a student studied for the examination. Tabled below are the examination marks (out of 100) and the time (in hours) spent by students studying for the examination.

Hours of study	9	13	19	4	16	11	17	7	23	11	24
Examination mark	56	77	77	54	79	74	88	59	92	60	94

Hours of study	6	20	13	27	9	3	30	8	13	21	17
Examination mark	65	85	68	98	64	62	91	66	81	88	78

(a) Identify the explanatory and response variables for this data set.

(b) On the axes below construct a scatter plot to show the tabled results of this investigation.



(c) Describe what the scatter plot shows.

6. It was found that the correlation coefficient of attendance at West Coast Eagle's games and the temperature on those days was 0.82.
(a) Identify the explanatory variable and the response variable.

(b) Interpret the correlation coefficient.

(c) Describe the relationship between the temperature and attendance.

Another study revealed that the correlation of price of admittance to the Eagle's games and the attendance was -0.91.
(d) For this study identify the explanatory and response variables.

(e) Interpret the correlation coefficient.

(f) Describe the relationship between the price of admittance and attendance.

(g) Which factor, price of attendance or temperature is the better predictor of attendance? Justify.

7. The table below show the ATAR score and the current salary of ten people aged 34.

Person	ATAR (A)	Salary (S)
1	94.05	108 000
2	77.60	73 000
3	80.00	78 000
4	96.50	110 000
5	76.50	74 000
6	87.05	92 000
7	95.55	99 000
8	77.85	70 000
9	84.25	113 000
10	87.05	85 000

(a) Jackson was an average student at school but developed a very successful online business. Which person is most likely to be Jackson. Justify your choice.

(b) Determine the correlation coefficient between the ATAR grades and salary.

(c) Comment on the relationship between ATAR grades and salary.

(d) Which person would be the best to remove from this group if the relationship between ATAR grades and salary was to be strengthened. Justify your choice.

8. Over the past 40 years divorce rates have increased significantly. During this period the price of new cars has also increased significantly. When graphed, these two variables show a very high positive linear correlation. Does this mean that divorce rates cause the price of new cars to increase? Discuss.

9. Statistical data indicated that the correlation coefficient between age and the incidence of a particular respiratory ailment, over a two year period in a country town is 0.95. Would it be correct to say that the ailment is caused as a result of age? Give reasons for your answer.

10. The following table of Climatological Data was compiled by the Bureau of Meteorology.

Month	Sunshine (S)	Cloud Cover (C)	Evaporation Rate (E)
January	12.1	28	11.2
February	11.6	29	10.7
March	10.7	33	8.5
April	9.6	40	5.3
May	8.0	51	3.6
June	7.2	56	2.8
July	7.7	53	2.9
August	8.3	54	3.4
September	9.3	47	4.8
October	10.1	46	6.8
November	11.4	37	8.5
December	12.1	30	10.4

Where: S is the mean daily hours of sunshine.
C is the percentage of the sky covered in cloud.
E is the mean daily evaporation rate in millimetres.

(a) In a study involving sunshine and cloud cover identify the explanatory and response variables. Justify your choice.

(b) In a study involving cloud cover and evaporation rate identify the explanatory and response variables. Justify your choice.

(c) In a study involving sunshine and evaporation rate identify the explanatory and response variables. Justify your choice.

(d) Calculate the following correlation coefficients rounding answers to two decimal places.
(i) r_{SC} (ii) r_{SE} (iii) r_{CE}

(e) Examine each of the following statements. Decide if the statement is true or false. Justify your response.
(i) There exist a high correlation between all three pairs of variables.

(ii) The correlation between sunshine and evaporation is much greater than the correlation between cloud cover and evaporation.

(iii) The value of r_{CE} implies that the higher the value of C, the higher the rate of evaporation will be.