Exercise 4.1

①

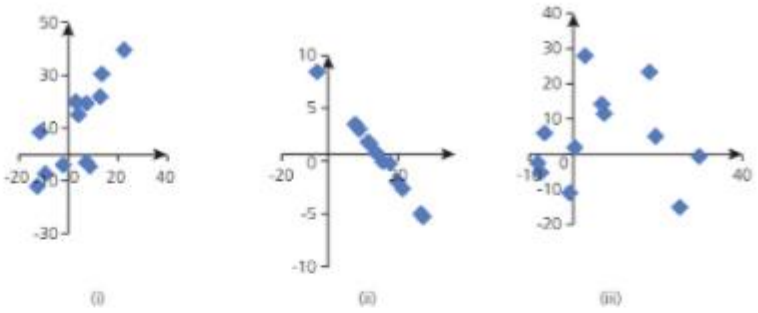

Figure 4.18

Three sets of bivariate data have been plotted on scatter diagrams, as illustrated. In each diagram the product moment correlation coefficient takes one of the values –1, –0.8, 0, 0.8, 1. Without doing any calculations, state the appropriate value of the correlation coefficient corresponding to the scatter diagrams (i), (ii) and (iii) in Figure 4.18.

② For each of the sets of data (i), (ii) and (iii) in Table 4.11 – 4.13

(a) draw a scatter diagram and comment on whether there appears to be any linear correlation.

(b) calculate the product moment correlation coefficient and compare this with your assertion based on the scatter diagram.

(i) The mathematics and physics test results of 14 students.

Table 4.11

Mathematics	45	23	78	91	46	27	41	62	34	17	77	49	55	71
Physics	62	36	92	70	67	39	61	40	55	33	65	59	35	40

(ii) The wine consumption in a country in millions of litres and the years 1993 to 2000.

Table 4.12

Year	1993	1994	1995	1996	1997	1998	1999	2000
Consumption ($\times 10^6$ litres)	35.5	37.7	41.5	46.4	44.8	45.8	53.9	62.0

(iii) The number of hours of sunshine and the monthly rainfall, in centimetres, in an eight-month period.

Table 4.13

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Sunshine (hours)	90	96	105	110	113	120	131	124
Rainfall (cm)	5.1	4.6	6.3	5.1	3.3	2.8	4.5	4.0

③ For each of these sets of data use one or more of the following to find the product moment correlation coefficient, r (note that the data for part (iv) are given in the form of a scatter diagram).

- a scientific calculator in two variable statistics mode
- a graphic calculator in two variable statistics mode
- a spreadsheet.

(i)

Table 4.14

x	10	11	12	13	14	15	16	17
y	19	16	28	20	31	19	32	35

(ii)

Table 4.15

x	12	14	14	15	16	17	17	19
y	86	90	78	71	77	69	80	73

(iii)

Table 4.16

x	56	78	14	80	34	78	23	61
y	45	34	67	70	42	18	25	50

(iv)

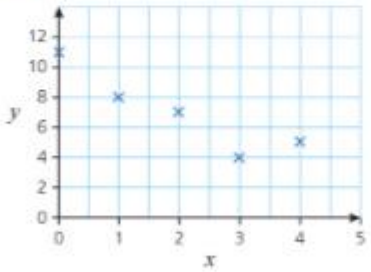

Figure 4.19

④ Find the value of r in each case below. The tables of values are not complete but, in each case, summary statistics are given for all of the data.

(i) The annual salary, in thousands of pounds, and the average number of hours worked per week by people chosen at random.

Table 4.17

Salary (× £1000)	5	7	13	
Hours worked per week	18	22	35	

$\Sigma x = 105, \Sigma y = 217, \Sigma x^2 = 3003, \Sigma y^2 = 7093, \Sigma xy = 3415, n = 7.$

(ii) The mean temperature in degrees Celsius and the amount of ice-cream sold in a supermarket in hundreds of litres.

Table 4.18

	Apr	May	Jun	Jul
Mean temperature (°C)	9	13	14	17
Ice-cream sold (100l)	11	15	17	20

$\Sigma x = 108, \Sigma y = 117, \Sigma x^2 = 1506, \Sigma y^2 = 1921, \Sigma xy = 1660, n = 8.$

(iii) The reaction times of women of various ages.

Table 4.19

Reaction time ($\times 10^{-3}$ s)	156	165	149	180	189
Age (years)	36	40	27	50	49

$\Sigma x = 1432, \Sigma y = 337, \Sigma x^2 = 259680, \Sigma y^2 = 15089, \Sigma xy = 61717, n = 8.$

⑤ A language teacher wishes to test whether students who are good at their own language are also likely to be good at a foreign language. Accordingly, she collects the marks of eight students, all native English speakers, in their end of year examinations in English and French.

Table 4.20

Candidate	A	B	C	D	E	F	G	H
English	65	34	48	72	58	63	26	80
French	74	49	45	80	63	72	12	75

(i) Calculate the product moment correlation coefficient.

(ii) State the null and alternative hypotheses.

(iii) Using the correlation coefficient as a test statistic, carry out the test at the 5% significance level.

⑥ 'You can't win without scoring goals.' So says the coach of a netball team. Jamila, who believes in solid defensive play, disagrees and sets out to prove that there is no correlation between scoring goals and winning matches. She collects the following data for the goals scored and the points gained by 12 teams in a netball league.

Table 4.21

Goals scored, x	41	50	54	47	47	49	52	61	50	29	47	35
Points gained, y	21	20	19	18	16	14	12	11	11	7	5	2

(i) Calculate the product moment correlation coefficient.

(ii) State suitable null and alternative hypotheses, indicating whose position each represents.

(iii) Carry out the hypothesis test at the 5% significance level and comment on the result.

⑦ A medical student is trying to estimate the birth weight of babies using pre-natal scan images. The actual weights, x kg, and the estimated weights, y kg, of ten randomly selected babies are given in the table below. The data are plotted in the scatter diagram.

Table 4.22

x	2.61	2.73	2 .87	2.96	3.05	3.14	3.17	3.24	3.76	4.10
y	3.2	2.6	3.5	3.1	2.8	2.7	3.4	3.3	4.4	4.1

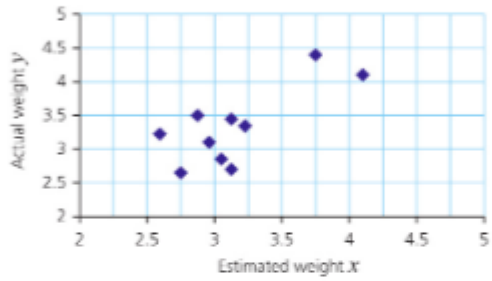

Figure 4.20

(i) The student decides to carry out a test based on the product moment correlation coefficient to investigate whether there is a positive relationship between the two variables. A friend suggests that there are two outliers so this test would not be appropriate. Explain why it may still be valid to carry out the test.

(ii) The value of the product moment correlation coefficient for these data is 0.7604. Carry out the test at the 1% significance level. [MEI]

⑧ It is widely believed that those who are good at chess are good at bridge, and vice-versa. A commentator decides to test this theory using as data the grades of a random sample of eight people who play both games.

Table 4.23

Player	A	B	C	D	E	F	G	H
Chess grade	160	187	129	162	149	151	189	158
Bridge grade	75	100	75	85	80	70	95	80

(i) Calculate the product moment correlation coefficient.

(ii) State suitable null and alternative hypotheses.

(iii) The output in Figure 4.21 over the page comes from a statistical package. Using the p-value given at the bottom of the figure, complete the hypothesis test.

Figure 4.21

⑨ The correlation matrix below shows the correlation between NVR (non-verbal reasoning),VR (verbal reasoning) and Q (quantitative reasoning) scores for a large group of students. The scores are obtained from tests given to the students on each of the areas.

Table 4.24

	VR	NVR	Q
VR	1		
NVR	0.21575396	1	
Q	0.164106922	0.2393041	1

(i) Use Cohen's interpretation to comment on the levels of correlation.

(ii) A teacher at the school suggests that as there is correlation between the scores, there is no point in giving the next intake of students all three tests but just give the verbal reasoning test and predict the results of the others from that one. Comment on this suggestion.

⑩ A biologist believes that a particular type of fish develops black spots on its scales in water that is polluted by certain agricultural fertilisers. She catches a number of fish; for each one she counts the number of black spots on its scales and measures the concentration of the pollutant in the water it was swimming in. She uses these data to test for positive linear correlation between the number of spots and the level of pollution.

Table 4.25

Fish	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Pollutant concentration (parts per million)	124	59	78	79	150	12	23	45	91	68
Number of black spots	15	8	7	8	14	0	4	5	8	8

(i) Calculate the product moment correlation coefficient.

(ii) State suitable null and alternative hypotheses.

(iii) Carry out the hypothesis test at the 2% significance level. What can the biologist conclude?

⑪ The correlation matrix below shows the correlation between four variables in different districts of a large city. The four variables are life expectancy (L), infant mortality (M), average income (A) and a measure of income inequality (I). Use Cohen's interpretation to comment on the levels of correlation.

Table 4.26

	L	M	A	I
L	1			
M	0.37	1		
A	−0.23	0.17	1	
I	0.06	−0.19	0.46	1

⑫ Andrew claims that the older you get, the slower is your reaction time. His mother disagrees, saying the two are unrelated. They decide that the only way to settle the discussion is to carry out a proper test. A few days later they are having a small party and so ask their 12 guests to take a test that measures their reaction times. The results are as follows.

Table 4.27

Age	Reaction time (s)	Age	Reaction time (s)
78	0.8	35	0.5
72	0.6	30	0.3
60	0.7	28	0.4
56	0.5	20	0.4
41	0.5	19	0.3
39	0.4	10	0.3

Carry out the test at the 5% significance level, stating the null and alternative hypotheses. Who won the argument, Andrew or his mother?

⑬ The teachers at a school have a discussion as to whether girls, in general, run faster or slower as they get older. They decide to collect data for a random sample of girls the next time the school cross country race is held (which everybody has to take part in). They collect the following data, with the times given in minutes and the ages in years (the conversion from months to decimal parts of a year has already been carried out).

Table 4.28

Age	Time	Age	Time	Age	Time
11.6	23.1	18.2	45.	13.9	29.1
15.0	24.0	15.4	23.2	18.1	21.2
18.8	45.0	14.4	26.1	13.4	23.9
16.0	25.2	16.1	29.4	16.2	26.0
12.8	26.4	14.6	28.1	17.5	23.4
17.6	22.9	18.7	45.0	17.0	25.0
17.4	27.1	15.4	27.0	12.5	26.3
13.2	25.2	11.8	25.4	12.7	24.2
14.5	26.8				

(i) State suitable null and alternative hypotheses and decide on an appropriate significance level for the test.

(ii) Calculate the product moment correlation coefficient and state the conclusion from the test.

(iii) Plot the data on a scatter diagram and identify any outliers. Explain how they could have arisen.

(iv) Comment on the validity of the test.

⑭ Apprentices joining a large company are given several tests. Two of the tests are 'Basic English' and 'Manual dexterity'. The results of these tests are labelled x and y respectively. A manager believes that there will be positive correlation between x and y.

The spreadsheet shows the first 3 and 60th rows of data, together with the sums of each of the 5 columns.

	A	B	C	D	E	F
1		x	y	x^2	y^2	xy
2		119	118	14161	13924	14042
3		112	116	12544	13456	12992
4		116	108	13456	11664	12528
61		115	118	13225	13924	13570
62	SUM	7105	7132	846357	853620	846371

Figure 4.22

(i) Explain how you can tell that the value of n is 60.

(ii) State the formula in cell B62.

(iii) Find the values of S_{xx}, S_{yy} and S_{xy}.

(iv) Find the value of r.

(v) State suitable null and alternative hypotheses for a test to investigate the manager's belief.

(vi) Carry out the hypothesis test at the 5% significance level.

(vii) Comment on the effect size.

(viii) Do you think that this information can be of any use to the company?

⑮ The values of x and y in the table are the marks obtained in an intelligence test and a university examination, respectively, by 20 medical students. The data are plotted in the scatter diagram.

Table 4.29

x	98	51	71	57	44	59	75	47	39	58
y	85	40	30	25	50	40	50	35	25	90
x	77	65	58	66	79	72	45	40	49	76
y	65	25	70	45	70	50	40	20	30	60

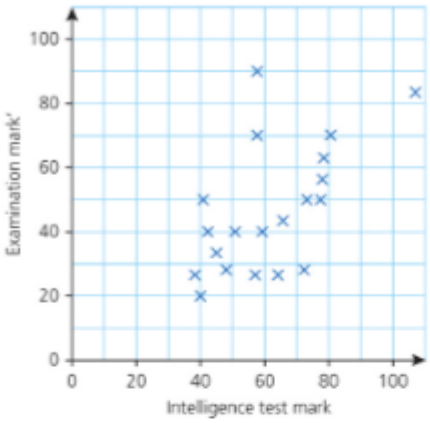

Figure 4.23

Given that $\Sigma x = 1226$, $\Sigma y = 945$, $\Sigma x^2 = 79\,732$, $\Sigma y^2 = 52\,575$ and $\Sigma xy = 61\,495$, calculate the product moment correlation coefficient, r, to 2 decimal places.

Referring to the evidence provided by the diagram and the value of r, comment briefly on the correlation between the two sets of marks.

Now eliminate from consideration those ten students whose values of x are less than 50 or more than 75. Calculate the new value of r for the marks of the remaining students. What does the comparison with the earlier value of r seem to indicate? [MEI]

⑯ A random sample of students who are shortly to sit an examination are asked to keep a record of how long they spend revising, in order to investigate whether more revision time is associated with a higher mark. The data are given below, with x hours being the revision time (correct to the nearest half hour) and y % being the mark scored in the examination.

Table 4.30

x	0	3	4.5	3.5	7	5.5	5	6.5	6	10.5	2
y	36	52	52	57	60	61	63	63	64	70	89

(i) Obtain the value of the product moment correlation coefficient for the data.

(ii) Specify appropriate null and alternative hypotheses, and carry out a suitable test at the 5% level of significance.

(iii) Without further calculation, state the effect of the data $x = 2$, $y = 89$ on the value of the product moment correlation coefficient. Explain whether or not this point should be excluded when carrying out the hypothesis test. [MEI]

⑰ In order to investigate the strength of the correlation between the value of a house and the value of the householder's car, a random sample of householders was questioned. The resulting data are shown in the table, the units being thousands of pounds.

Table 4.31

x	220	212	102	188	132	52	144	102	106	266
y	24	19	4.8	8.4	8.2	0.6	6.4	12	15.6	30

(i) Represent the data graphically.

(ii) Calculate the product moment correlation coefficient.

(iii) Carry out a suitable hypothesis test, at a suitable level of significance, to determine whether or not it is reasonable to suppose that the value of a house is positively correlated with the value of the householder's car.

(iv) A student argues that when two variables are correlated one must be the cause of the other. Briefly discuss this view with regard to the data in this question. [MEI]

18 The table below gives the heights, h, of six male Olympic 100m sprint winners together with the times, t, they took.

Table 4.32

h	1.80	1.83	1.87	1.88	1.85	1.76
t	10.00	9.95	10.06	9.99	9.84	9.87

(i) Draw a scatter diagram to illustrate the data.

(ii) Calculate the product moment correlation coefficient.

(iii) Carry out a suitable hypothesis test, at a suitable level of significance, to determine whether or not it is reasonable to suppose that the heights and times are positively correlated.

(iv) Rewrite the table giving just the rank of each data value, using a rank of 1 for the lowest value and 6 for the highest value. For example the rank associated with a height of 1.88 m would be 6 since it is the height of the tallest person. The rank associated with a time of 9.87 s would be 2 since it is the second lowest time.

(v) Calculate the product moment correlation coefficient of the ranked data.

(vi) Comment on the difference between the two correlation coefficients.

19 (i) Prove algebraically that these two formulae for S_{xx} are equivalent

$$S_{xx} = \Sigma(x_i - \bar{x})^2 \quad S_{xx} = \Sigma x^2 - n\bar{x}^2$$

(ii) Prove algebraically these two formulae for S_{xy} are equivalent

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) \quad S_{xy} = \Sigma xy - n\bar{x}\bar{y}$$

(iii) Hence prove that these two formulae for r are equivalent

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \times \Sigma(y_i - \bar{y})^2}} \quad r = \frac{\Sigma x_i y_i - n(\bar{x})(\bar{y})}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)(\Sigma y_i^2 - n\bar{y}^2)}}$$