OCR Statistics 1

Past Paper Questions

- Product moment correlation coefficient
- Regression
- Spearman's rank correlation coefficient

Edited by K V Kumaran

Email: kvkumaran@gmail.com

Phone: 07961319548

Correlation

- The Product Moment Correlation Coefficient is a number (r) calculated on a set of bivariate data that tells us how correlated two data sets are.
- The value of r is such that $-1 < r < 1$. If $r = 1$ you have perfect positive linear correlation. If $r = -1$ you have perfect negative linear correlation. If $r = 0$ then there exists no correlation between the data sets.
- It is defined

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where we define the individual components as

$$\begin{aligned} S_{xx} &= \sum x^2 - \tfrac{1}{n}\left(\sum x\right)^2, \\ S_{yy} &= \sum y^2 - \tfrac{1}{n}\left(\sum y\right)^2, \\ S_{xy} &= \sum xy - \tfrac{1}{n}\sum x \sum y. \end{aligned}$$

- So to calculate r for the data set

x	14	12	16	18	21	13	15	17
y	1	2	4	5	2	8	5	6

we write the data in columns and add extra ones. We then sum the columns and calculate from these sums. Note that in the above example $n = 8$ (i.e. the number of pairs, not the number of individual data pieces).

x	y	x^2	y^2	xy
14	1	196	1	14
12	2	144	4	24
16	4	256	16	64
18	5	324	25	90
21	2	441	4	42
13	8	169	64	104
15	5	225	25	75
17	6	289	36	102
126	**33**	**2044**	**175**	**515**

Therefore

$$\begin{aligned} S_{xx} &= \sum x^2 - \tfrac{1}{n}\left(\sum x\right)^2 = 2044 - \tfrac{126^2}{8} = 59.5, \\ S_{yy} &= \sum y^2 - \tfrac{1}{n}\left(\sum y\right)^2 = 175 - \tfrac{33^2}{8} = 38.875, \\ S_{xy} &= \sum xy - \tfrac{1}{n}\sum x \sum y = 515 - \tfrac{126\times 33}{8} = -4.75. \end{aligned}$$

Therefore

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-4.75}{\sqrt{59.5 \times 38.875}} = -0.09876\ldots$$

Therefore the data has very, very weak negative correlation. Basically it has no *meaningful* correlation.

- It can be shown that if one (or both) of the variables are transformed in a linear fashion i.e. if we replace the x values by, say, $\frac{x-4}{3}$ (or any transformation formed by $+$, $-$, $\div$ or $\times$ with constants) then the value of r will be unchanged.

- You need to be able to calculate Spearman's rank correlation coefficient (r_s). You will be given a table and you will need to (in the next 2 columns) rank the data. If two data points are tied then you (e.g. the 2nd and 3rd are tied) then you rank them both 2.5.

%	IQ	Rank %	Rank IQ	d	d^2
89	143	2.5	1	1.5	2.25
55	89	7	8	−1	1
72	102	5	6	−1	1
91	136	1	2	−1	1
89	126	2.5	3	−0.5	0.25
30	60	9	9	0	0
71	115	6	4	2	4
53	100	8	7	1	1
78	103	4	5	−1	1

Now $\boxed{r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}}$. $\sum d^2$ is just the sum of the d^2 column in the table and n is the number of pairs of data; here $n = 9$. We therefore find $r_s = 1 - \frac{6\times 11.5}{9(81-1)} = 0.90416\dot{}$. Therefore we see a strong degree of positive association.

- If r_s is close to -1 then strong negative association. If close to zero then no meaningful association/agreement.

Regression

- For any set of bivariate data (x_i, y_i) there exist two possible regression lines; 'y on x' and 'x on y'.
- If neither is controlled (see below) then if you want to predict y from a given value of x, you use the 'y on x' line. If you want to predict x from a given value of y, you use the 'x on y' line.
- The 'y on x' line is defined

$$y = a + bx \quad \text{where} \quad b = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

- The 'x on y' line is defined

$$x = a' + b'y \quad \text{where} \quad b' = \frac{S_{xy}}{S_{yy}} \quad \text{and} \quad a' = \bar{x} - b'\bar{y}.$$

- Both regression lines pass through the average point $(\bar{x}, \bar{y})$.
- In the example in the book (P180) the height of the tree is the dependent variable and the circumference of the tree is the independent variable. This is because the experiment has been constructed to see how the height of the tree depends on its circumference.
- If one variable is being controlled by the experimenter (e.g. x), it is called a controlled variable. If x is controlled you would never use the 'x on y' regression line. Only use the 'y on x' line. You would use this to predict y from x (expected) *and* x from y (not-expected)

1.

(i) Calculate the value of Spearman's rank correlation coefficient between the two sets of rankings, A and B, shown in Table 1. [4]

A	1	2	3	4	5
B	4	1	3	2	5

Table 1

(ii) The value of Spearman's rank correlation coefficient between the set of rankings B and a third set of rankings, C, is known to be -1. Copy and complete Table 2 showing the set of rankings C. [2]

B	4	1	3	2	5
C					

Table 2

Q1 June 2005

2.

A chemical solution was gradually heated. At five-minute intervals the time, x minutes, and the temperature, $y\,^{\circ}\mathrm{C}$, were noted.

x	0	5	10	15	20	25	30	35
y	0.8	3.0	6.8	10.9	15.6	19.6	23.4	26.7

$[n = 8,\ \Sigma x = 140,\ \Sigma y = 106.8,\ \Sigma x^2 = 3500,\ \Sigma y^2 = 2062.66,\ \Sigma xy = 2685.0.]$

(i) Calculate the equation of the regression line of y on x. [4]

(ii) Use your equation to estimate the temperature after 12 minutes. [2]

(iii) It is given that the value of the product moment correlation coefficient is close to +1. Comment on the reliability of using your equation to estimate y when

(a) $x = 17$,

(b) $x = 57$.

[2]

Q5 Jan 2007

3.

The table shows the population, x million, of each of nine countries in Western Europe together with the population, y million, of its capital city.

	Germany	United Kingdom	France	Italy	Spain	The Netherlands	Portugal	Austria	Switzerland
x	82.1	59.2	59.1	56.7	39.2	15.9	9.9	8.1	7.3
y	3.5	7.0	9.0	2.7	2.9	0.8	0.7	1.6	0.1

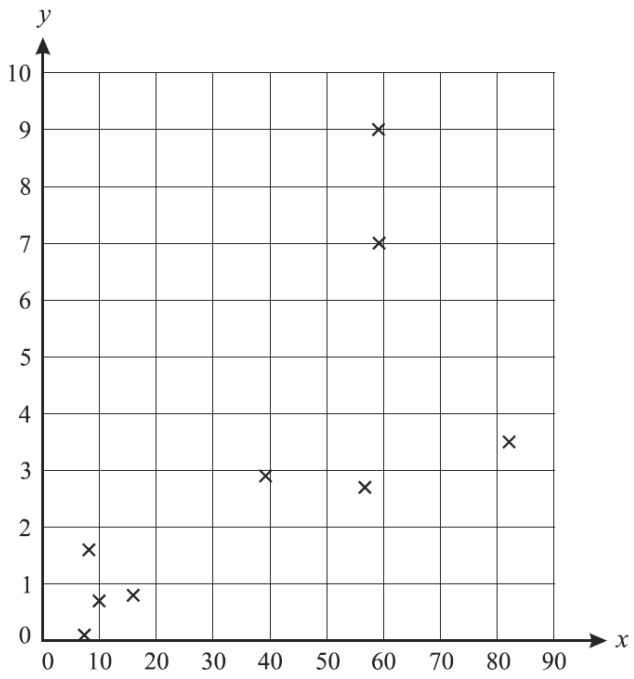
$[n = 9,\ \Sigma x = 337.5,\ \Sigma x^2 = 18\,959.11,\ \Sigma y = 28.3,\ \Sigma y^2 = 161.65,\ \Sigma xy = 1533.76.]$

(i) (a) Calculate Spearman's rank correlation coefficient, r_s. [5]

(b) Explain what your answer indicates about the populations of these countries and their capital cities. [1]

(ii) Calculate the product moment correlation coefficient, r. [2]

The data are illustrated in the scatter diagram.



(iii) By considering the diagram, state the effect on the value of the product moment correlation coefficient, r, if the data for France and the United Kingdom were removed from the calculation. [1]

(iv) In a certain country in Africa, most people live in remote areas and hence the population of the country is unknown. However, the population of the capital city is known to be approximately 1 million. An official suggests that the population of this country could be estimated by using a regression line drawn on the above scatter diagram.

(a) State, with a reason, whether the regression line of y on x or the regression line of x on y would need to be used. [2]

(b) Comment on the reliability of such an estimate in this situation. [2]

Q8 Jan 2006

4.

The table shows the total distance travelled, in thousands of miles, and the amount of commission earned, in thousands of pounds, by each of seven sales agents in 2005.

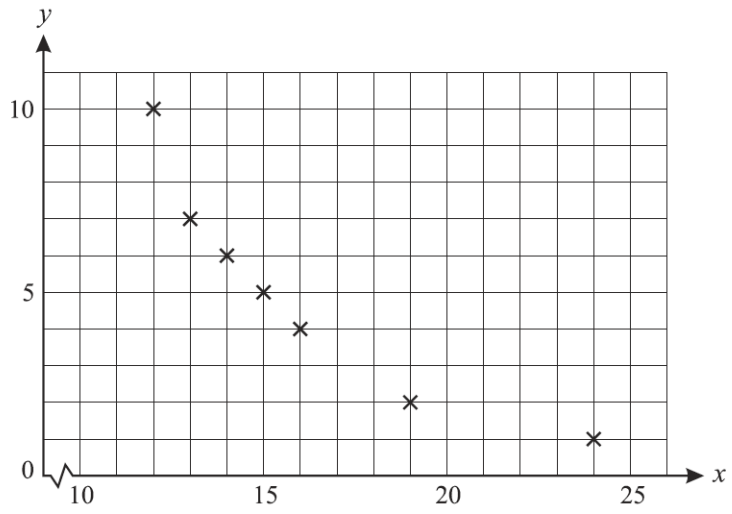
Agent	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Distance travelled	18	15	12	14	16	24	13
Commission earned	18	45	19	24	27	22	23

(i) (a) Calculate Spearman's rank correlation coefficient, r_s, for these data. [5]

(b) Comment briefly on your value of r_s with reference to this context. [1]

(c) After these data were collected, agent *A* found that he had made a mistake. He had actually travelled 19 000 miles in 2005. State, with a reason, but without further calculation, whether the value of Spearman's rank correlation coefficient will increase, decrease or stay the same. [2]

The agents were asked to indicate their level of job satisfaction during 2005. A score of 0 represented no job satisfaction, and a score of 10 represented high job satisfaction. Their scores, y, together with the data for distance travelled, x, are illustrated in the scatter diagram below.



(ii) For this scatter diagram, what can you say about the value of

(a) Spearman's rank correlation coefficient, [1]

(b) the product moment correlation coefficient? [1]

Q6 June 2006

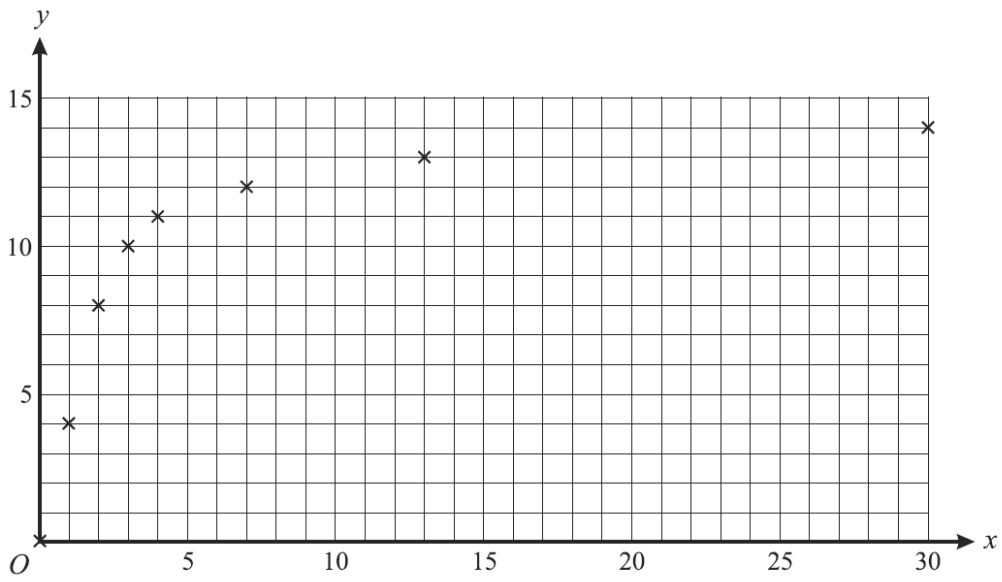
5.

A machine with artificial intelligence is designed to improve its efficiency rating with practice. The table shows the values of the efficiency rating, y, after the machine has carried out its task various numbers of times, x.

x	0	1	2	3	4	7	13	30
y	0	4	8	10	11	12	13	14

$[n = 8,\ \Sigma x = 60,\ \Sigma y = 72,\ \Sigma x^2 = 1148,\ \Sigma y^2 = 810,\ \Sigma xy = 767.]$

These data are illustrated in the scatter diagram.



(i) (a) Calculate the value of r, the product moment correlation coefficient. [3]

(b) Without calculation, state with a reason the value of r_s, Spearman's rank correlation coefficient. [2]

(ii) A researcher suggests that the data for $x = 0$ and $x = 1$ should be ignored. Without calculation, state with a reason what effect this would have on the value of

(a) r, [2]

(b) r_s. [2]

(iii) Use the diagram to estimate the value of y when $x = 29$. [1]

(iv) Jack finds the equation of the regression line of y on x for all the data, and uses it to estimate the value of y when $x = 29$. Without calculation, state with a reason whether this estimate or the one found in part **(iii)** will be the more reliable. [2]

Q6 June 2007

6.

A sample of bivariate data was taken and the results were summarised as follows.

$$n = 5 \qquad \Sigma x = 24 \qquad \Sigma x^2 = 130 \qquad \Sigma y = 39 \qquad \Sigma y^2 = 361 \qquad \Sigma xy = 212$$

(i) Show that the value of the product moment correlation coefficient r is 0.855, correct to 3 significant figures. [2]

(ii) The ranks of the data were found. One student calculated Spearman's rank correlation coefficient r_s, and found that $r_s = 0.7$. Another student calculated the product moment coefficient, R, of these ranks. State which one of the following statements is true, and explain your answer briefly.

(A) $R = 0.855$

(B) $R = 0.7$

(C) It is impossible to give the value of R without carrying out a calculation using the original data.

[2]

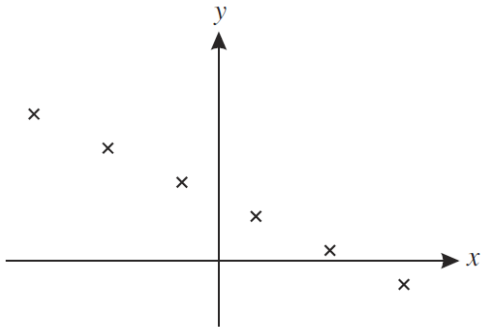
(iii) All the values of x are now multiplied by a scaling factor of 2. State the new values of r and r_s. [2]

Q3 Jan 2008

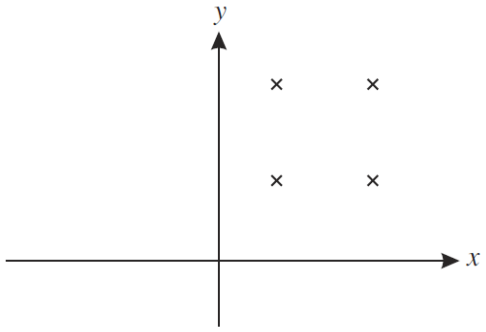
7.

(i) State the value of the product moment correlation coefficient for each of the following scatter diagrams. [2]

(a)



(b)



(ii) Calculate the value of Spearman's rank correlation coefficient for the following data. [5]

x	3.8	4.1	4.5	5.3
y	1.4	0.8	0.7	1.2

Q3 June 2008

8.

It is thought that the pH value of sand (a measure of the sand's acidity) may affect the extent to which a particular species of plant will grow in that sand. A botanist wished to determine whether there was any correlation between the pH value of the sand on certain sand dunes, and the amount of each of two plant species growing there. She chose random sections of equal area on each of eight sand dunes and measured the pH values. She then measured the area within each section that was covered by each of the two species. The results were as follows.

	Dune	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
	pH value, x	8.5	8.5	9.5	8.5	6.5	7.5	8.5	9.0
Area, y cm^2, covered	Species *P*	150	150	575	330	45	15	340	330
	Species *Q*	170	15	80	230	75	25	0	0

The results for species *P* can be summarised by

$$n = 8, \quad \Sigma x = 66.5, \quad \Sigma x^2 = 558.75, \quad \Sigma y = 1935, \quad \Sigma y^2 = 711\,275, \quad \Sigma xy = 17\,082.5.$$

(i) Give a reason why it might be appropriate to calculate the equation of the regression line of y on x rather than x on y in this situation. [1]

(ii) Calculate the equation of the regression line of y on x for species *P*, in the form $y = a + bx$, giving the values of a and b correct to 3 significant figures. [4]

(iii) Estimate the value of y for species *P* on sand where the pH value is 7.0. [2]

The values of the product moment correlation coefficient between x and y for species *P* and *Q* are $r_P = 0.828$ and $r_Q = 0.0302$.

(iv) Describe the relationship between the area covered by species *Q* and the pH value. [1]

(v) State, with a reason, whether the regression line of y on x for species *P* will provide a reliable estimate of the value of y when the pH value is

(a) 8, [1]

(b) 4. [1]

(vi) Assume that the equation of the regression line of y on x for species *Q* is also known. State, with a reason, whether this line will provide a reliable estimate of the value of y when the pH value is 8. [1]

Q9 Jan 2008

9.

A city council attempted to reduce traffic congestion by introducing a congestion charge. The charge was set at £4.00 for the first year and was then increased by £2.00 each year. For each of the first eight years, the council recorded the average number of vehicles entering the city centre per day. The results are shown in the table.

Charge, £x	4	6	8	10	12	14	16	18
Average number of vehicles per day, y million	2.4	2.5	2.2	2.3	2.0	1.8	1.7	1.5

[$n = 8$, $\Sigma x = 88$, $\Sigma y = 16.4$, $\Sigma x^2 = 1136$, $\Sigma y^2 = 34.52$, $\Sigma xy = 168.6$.]

(i) Calculate the product moment correlation coefficient for these data. [3]

(ii) Explain why x is the independent variable. [1]

(iii) Calculate the equation of the regression line of y on x. [4]

(iv) **(a)** Use your equation to estimate the average number of vehicles which will enter the city centre per day when the congestion charge is raised to £20.00. [2]

(b) Comment on the reliability of your estimate. [2]

(v) The council wishes to estimate the congestion charge required to reduce the average number of vehicles entering the city per day to 1.0 million. Assuming that a reliable estimate can be made by extrapolation, state whether they should use the regression line of y on x or the regression line of x on y. Give a reason for your answer. [2]

Q8 June 2008

10.

The table shows the age, x years, and the mean diameter, y cm, of the trunk of each of seven randomly selected trees of a certain species.

Age (x years)	11	12	20	28	35	45	51
Mean trunk diameter (y cm)	12.2	16.0	26.4	39.2	39.6	51.3	60.6

[$n = 7$, $\Sigma x = 202$, $\Sigma y = 245.3$, $\Sigma x^2 = 7300$, $\Sigma y^2 = 10\,510.65$, $\Sigma xy = 8736.9$.]

(i) **(a)** Use an appropriate formula to show that the gradient of the regression line of y on x is 1.13, correct to 2 decimal places. **[2]**

(b) Find the equation of the regression line of y on x. **[2]**

(ii) Use your equation to estimate the mean trunk diameter of a tree of this species with age

(a) 30 years, **[1]**

(b) 100 years. **[1]**

It is given that the value of the product moment correlation coefficient for the data in the table is 0.988, correct to 3 decimal places.

(iii) Comment on the reliability of each of your two estimates. **[2]**

Q2 Jan 2008

11.

Three tutors each marked the coursework of five students. The marks are given in the table.

Student	*A*	*B*	*C*	*D*	*E*
Tutor 1	73	67	60	48	39
Tutor 2	62	50	61	76	65
Tutor 3	42	50	63	54	71

(i) Calculate Spearman's rank correlation coefficient, r_s, between the marks for tutors 1 and 2. **[5]**

(ii) The values of r_s for the other pairs of tutors, are as follows.

Tutors 1 and 3: $r_s = -0.9$

Tutors 2 and 3: $r_s = 0.3$

State which two tutors differ most widely in their judgements. Give your reason. **[2]**

Q4 Jan 2008

12.

Two judges placed 7 dancers in rank order. Both judges placed dancers *A* and *B* in the first two places, but in opposite orders. The judges agreed about the ranks for all the other 5 dancers. Calculate the value of Spearman's rank correlation coefficient. **[4]**

Q2 June 2009

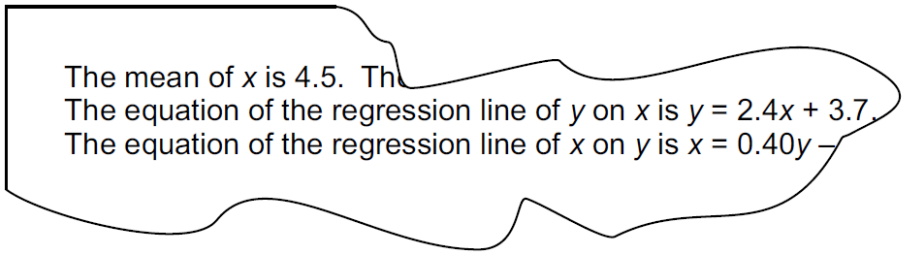
(a) A student calculated the values of the product moment correlation coefficient, r, and Spearman's rank correlation coefficient, r_s, for two sets of bivariate data, *A* and *B*. His results are given below.

A: $r = 0.9$ and $r_s = 1$

B: $r = 1$ and $r_s = 0.9$

With the aid of a diagram where appropriate, explain why the student's results for *A* could both be correct but his results for *B* cannot both be correct. **[3]**

(b) An old research paper has been partially destroyed. The surviving part of the paper contains the following incomplete information about some bivariate data from an experiment.



Calculate the missing constant at the end of the equation of the second regression line. **[4]**

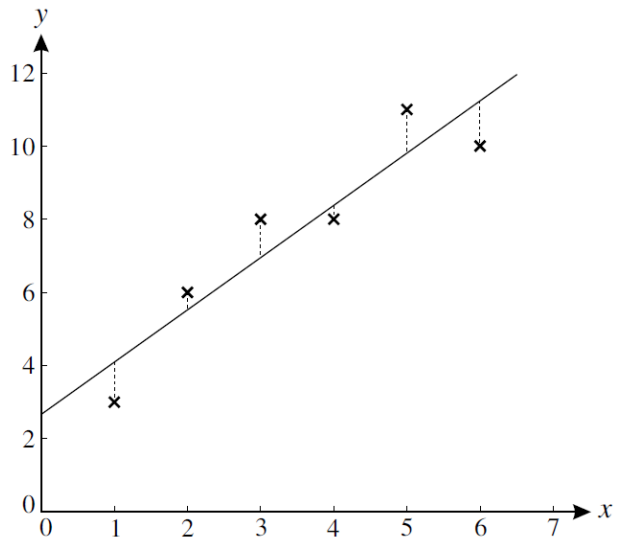
Q6 Jan 2010

13.

In an agricultural experiment, the relationship between the amount of water supplied, x units, and the yield, y units, was investigated. Six values of x were chosen and for each value of x the corresponding value of y was measured. The results are shown in the table.

x	1	2	3	4	5	6
y	3	6	8	8	11	10

These results, together with the regression line of y on x, are plotted on the graph.



(i) Give a reason why the regression line of x on y is not suitable in this context. **[1]**

(ii) Explain the significance, for the regression line of y on x, of the distances shown by the vertical dotted lines in the diagram. **[2]**

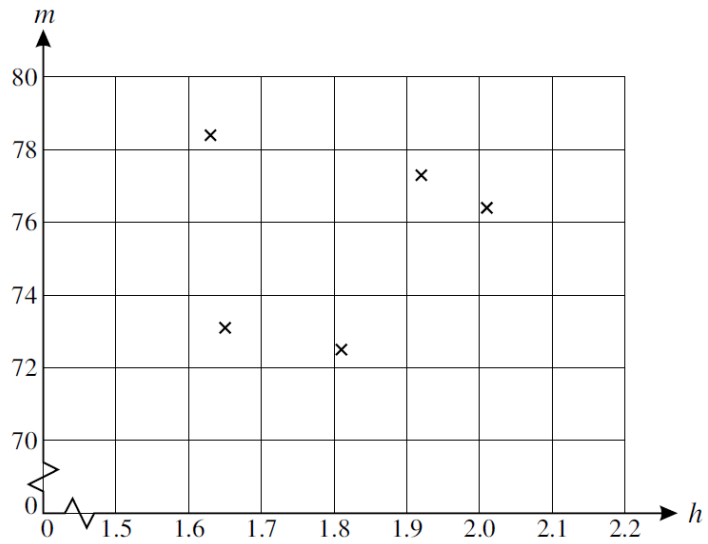
(iii) Calculate the value of the product moment correlation coefficient, r. **[3]**

(iv) Comment on your value of r in relation to the diagram. **[2]**

Q3 June 2009

14.

The heights, h m, and weights, m kg, of five men were measured. The results are plotted on the diagram.



The results are summarised as follows.

$$n = 5 \quad \Sigma h = 9.02 \quad \Sigma m = 377.7 \quad \Sigma h^2 = 16.382 \quad \Sigma m^2 = 28\,558.67 \quad \Sigma hm = 681.612$$

(i) Use the summarised data to calculate the value of the product moment correlation coefficient, r. **[3]**

(ii) Comment on your value of r in relation to the diagram. **[2]**

(iii) It was decided to re-calculate the value of r after converting the heights to feet and the masses to pounds. State what effect, if any, this will have on the value of r. **[1]**

(iv) One of the men had height 1.63 m and mass 78.4 kg. The data for this man were removed and the value of r was re-calculated using the original data for the remaining four men. State in general terms what effect, if any, this will have on the value of r. **[1]**

Q3 Jan 2010

15.

The orders in which 4 contestants, P, Q, R and S, were placed in two competitions are shown in the table.

Position	1st	2nd	3rd	4th
Competition 1	Q	R	S	P
Competition 2	Q	P	R	S

Calculate Spearman's rank correlation coefficient between these two orders. **[5]**

Q2 June 2011

16.

Three skaters, A, B and C, are placed in rank order by four judges. Judge P ranks skater A in 1st place, skater B in 2nd place and skater C in 3rd place.

(i) Without carrying out any calculation, state the value of Spearman's rank correlation coefficient for the following ranks. Give a reason for your answer. **[1]**

Skater	A	B	C
Judge P	1	2	3
Judge Q	3	2	1

(ii) Calculate the value of Spearman's rank correlation coefficient for the following ranks. **[3]**

Skater	A	B	C
Judge P	1	2	3
Judge R	3	1	2

(iii) Judge S ranks the skaters at random. Find the probability that the value of Spearman's rank correlation coefficient between the ranks of judge P and judge S is 1. **[3]**

Q2 June 2010

17.

(i) Some values, (x, y), of a bivariate distribution are plotted on a scatter diagram and a regression line is to be drawn. Explain how to decide whether the regression line of y on x or the regression line of x on y is appropriate. **[2]**

(ii) In an experiment the temperature, $x\,^{\circ}\text{C}$, of a rod was gradually increased from $0\,^{\circ}\text{C}$, and the extension, y mm, was measured nine times at $50\,^{\circ}\text{C}$ intervals. The results are summarised below.

$$n = 9 \quad \Sigma x = 1800 \quad \Sigma y = 14.4 \quad \Sigma x^2 = 510\,000 \quad \Sigma y^2 = 32.6416 \quad \Sigma xy = 4080$$

(a) Show that the gradient of the regression line of y on x is 0.008 and find the equation of this line. **[4]**

(b) Use your equation to estimate the temperature when the extension is 2.5 mm. **[1]**

(c) Use your equation to estimate the extension for a temperature of $-50\,^{\circ}\text{C}$. **[1]**

(d) Comment on the meaning and the reliability of your estimate in part **(c)**. **[2]**

Q3 June 2010

18.

A firm wishes to assess whether there is a linear relationship between the annual amount spent on advertising, £x thousand, and the annual profit, £y thousand. A summary of the figures for 12 years is as follows.

$$n = 12 \quad \Sigma x = 86.6 \quad \Sigma y = 943.8 \quad \Sigma x^2 = 658.76 \quad \Sigma y^2 = 83\,663.00 \quad \Sigma xy = 7351.12$$

(i) Calculate the product moment correlation coefficient, showing that it is greater than 0.9. **[3]**

(ii) Comment briefly on this value in this context. **[1]**

(iii) A manager claims that this result shows that spending more money on advertising in the future will result in greater profits. Make two criticisms of this claim. **[2]**

(iv) Calculate the equation of the regression line of y on x. **[4]**

(v) Estimate the annual profit during a year when £7400 was spent on advertising. **[2]**

Q3 Jan 2011

19.

Five dogs, A, B, C, D and E, took part in three races. The order in which they finished the first race was $ABCDE$.

(i) Spearman's rank correlation coefficient between the orders for the 5 dogs in the first two races was found to be -1. Write down the order in which the dogs finished the second race. **[1]**

(ii) Spearman's rank correlation coefficient between the orders for the 5 dogs in the first race and the third race was found to be 0.9.

(a) Show that, in the usual notation (as in the List of Formulae), $\Sigma d^2 = 2$. **[2]**

(b) Hence or otherwise find a possible order in which the dogs could have finished the third race. **[2]**

Q8 Jan 2011

20.

Five salesmen from a certain firm were selected at random for a survey. For each salesman, the annual income, x thousand pounds, and the distance driven last year, y thousand miles, were recorded. The results were summarised as follows.

$$n = 5 \quad \Sigma x = 251 \quad \Sigma x^2 = 14\,323 \quad \Sigma y = 65 \quad \Sigma y^2 = 855 \quad \Sigma xy = 3247$$

(i) **(a)** Show that the product moment correlation coefficient, r, between x and y is -0.122, correct to 3 significant figures. **[3]**

(b) State what this value of r shows about the relationship between annual income and distance driven last year for these five salesmen. **[1]**

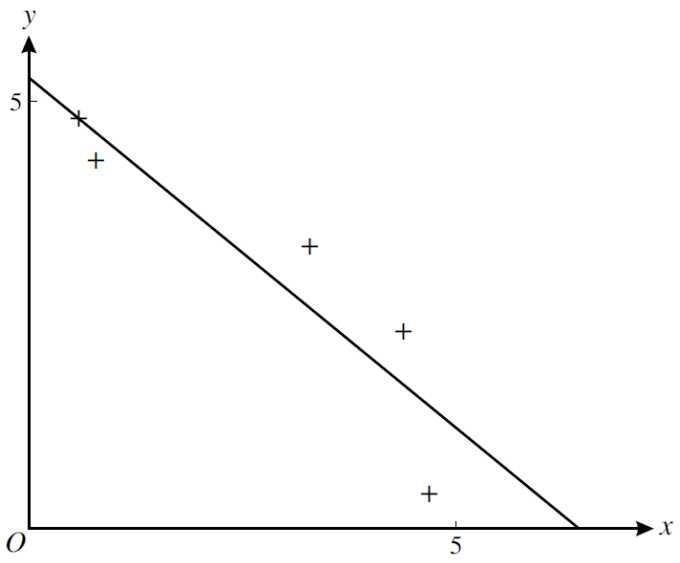
(c) It was decided to recalculate r with the distances measured in kilometres instead of miles. State what effect, if any, this would have on the value of r. **[1]**

(ii) Another salesman from the firm is selected at random. His annual income is known to be £52 000, but the distance that he drove last year is unknown. In order to estimate this distance, a regression line based on the above data is used. Comment on the reliability of such an estimate. **[2]**

Q1 June 2011

21.

The diagram shows the results of an experiment involving some bivariate data. The least squares regression line of y on x for these results is also shown.



(i) Given that the least squares regression line of y on x is used for an estimation, state which of x or y is treated as the independent variable. **[1]**

(ii) Use the diagram to explain what is meant by 'least squares'. **[2]**

(iii) State, with a reason, the value of Spearman's rank correlation coefficient for these data. **[2]**

(iv) What can be said about the value of the product moment correlation coefficient for these data? **[1]**

Q7 June 2011

22.

In an experiment, the percentage sand content, y, of soil in a given region was measured at nine different depths, x cm, taken at intervals of 6 cm from 0 cm to 48 cm. The results are summarised below.

$$n = 9 \quad \Sigma x = 216 \quad \Sigma x^2 = 7344 \quad \Sigma y = 512.4 \quad \Sigma y^2 = 30\,595 \quad \Sigma xy = 10\,674$$

(i) State, with a reason, which variable is the independent variable. **[1]**

(ii) Calculate the product moment correlation coefficient between x and y. **[3]**

(iii) (a) Calculate the equation of the appropriate regression line. **[3]**

(b) This regression line is used to estimate the percentage sand content at depths of 25 cm and 100 cm. Comment on the reliability of each of these estimates. You are not asked to find the estimates. **[3]**

Q2 Jan 2012

23.

(a) The table gives the heights and masses of 5 people.

Person	*A*	*B*	*C*	*D*	*E*
Height (m)	1.72	1.63	1.77	1.68	1.74
Mass (kg)	75	62	64	60	70

Calculate Spearman's rank correlation coefficient. **[5]**

(b) In an art competition the value of Spearman's rank correlation coefficient, r_s, calculated from two judges' rankings was 0.75. A late entry for the competition was received and both judges ranked this entry lower than all the others. By considering the formula for r_s, explain whether the new value of r_s will be less than 0.75, equal to 0.75, or greater than 0.75. **[3]**

Q4 Jan 2012

24.

For each of the last five years the number of tourists, x thousands, visiting Sackton, and the average weekly sales, £ y thousands, in Sackton Stores were noted. The table shows the results.

Year	2007	2008	2009	2010	2011
x	250	270	264	290	292
y	4.2	3.7	3.2	3.5	3.0

(i) Calculate the product moment correlation coefficient r between x and y. **[4]**

(ii) It is required to estimate the average weekly sales at Sackton Stores in a year when the number of tourists is 280 000. Calculate the equation of an appropriate regression line, and use it to find this estimate. **[4]**

(iii) Over a longer period the value of r is –0.8. The mayor says, "This shows that having more tourists causes sales at Sackton Stores to decrease." Give a reason why this statement is not correct. **[1]**

Q1 June 2012

25.

(i) Write down the value of Spearman's rank correlation coefficent, r_s, for the following sets of ranks.

(a)

Judge *A* ranks	1	2	3	4
Judge *B* ranks	1	2	3	4

[1]

(b)

Judge *A* ranks	1	2	3	4
Judge *C* ranks	4	3	2	1

[1]

(ii) Calculate the value of r_s for the following ranks.

Judge *A* ranks	1	2	3	4
Judge *D* ranks	2	4	1	3

[3]

(iii) For each of parts **(i)(a)**, **(i)(b)** and **(ii)**, describe in everyday terms the relationship between the two judges' opinions. **[3]**

Q5 June 2012

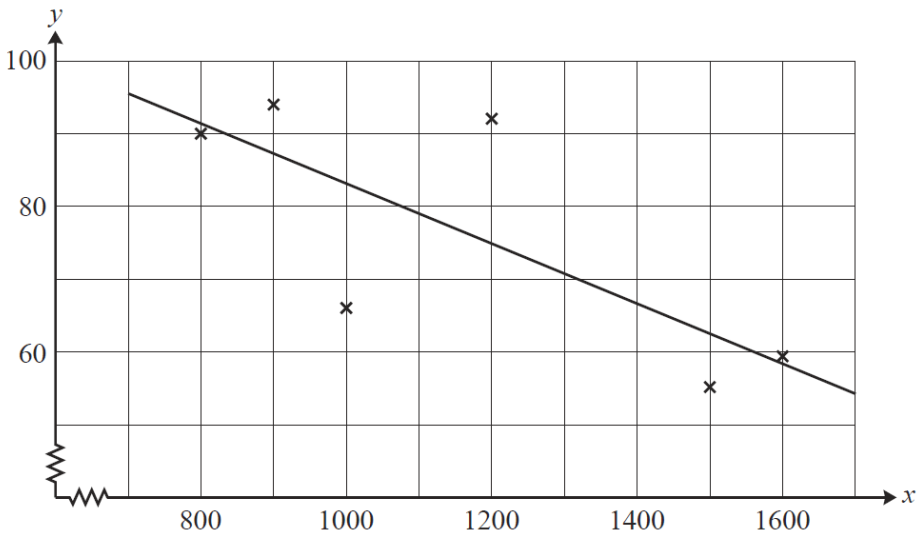
26.

The Gross Domestic Product per Capita (GDP), x dollars, and the Infant Mortality Rate per thousand (IMR), y, of 6 African countries were recorded and summarised as follows.

$$n = 6 \quad \Sigma x = 7000 \quad \Sigma x^2 = 8\,700\,000 \quad \Sigma y = 456 \quad \Sigma y^2 = 36\,262 \quad \Sigma xy = 509\,900$$

(i) Calculate the equation of the regression line of y on x for these 6 countries. **[4]**

The original data were plotted on a scatter diagram and the regression line of y on x was drawn, as shown below.



(ii) The GDP for another country, Tanzania, is 1300 dollars. Use the regression line in the diagram to estimate the IMR of Tanzania. **[1]**

(iii) The GDP for Nigeria is 2400 dollars. Give two reasons why the regression line is unlikely to give a reliable estimate for the IMR for Nigeria. **[2]**

(iv) The actual value of the IMR for Tanzania is 96. The data for Tanzania ($x = 1300$, $y = 96$) is now included with the original 6 countries. Calculate the value of the product moment correlation coefficient, r, for all 7 countries. **[4]**

(v) The IMR is now redefined as the infant mortality rate per hundred instead of per thousand, and the value of r is recalculated for all 7 countries. Without calculation state what effect, if any, this would have on the value of r found in part **(iv)**. **[1]**

Q3 Jan 2013

27.

(i) Two judges rank n competitors, where n is an even number. Judge 2 reverses each consecutive pair of ranks given by Judge 1, as shown.

Competitor	C_1	C_2	C_3	C_4	C_5	C_6	.....	C_{n-1}	C_n
Judge 1 rank	1	2	3	4	5	6	.....	$n-1$	n
Judge 2 rank	2	1	4	3	6	5	.....	n	$n-1$

Given that the value of Spearman's coefficient of rank correlation is $\frac{63}{65}$, find n. **[4]**

(ii) An experiment produced some data from a bivariate distribution. The product moment correlation coefficient is denoted by r, and Spearman's rank correlation coefficient is denoted by r_s.

(a) Explain whether the statement

$$r = 1 \Rightarrow r_s = 1$$

is true or false. **[1]**

(b) Use a diagram to explain whether the statement

$$r \neq 1 \Rightarrow r_s \neq 1$$

is true or false. **[2]**

Q7 Jan 2013

28.

(i) The table shows the times, in minutes, spent by five students revising for a test, and the grades that they achieved in the test.

Student	Ann	Bill	Caz	Den	Ed
Time revising	0	60	35	100	45
Grade	C	D	E	B	A

Calculate Spearman's rank correlation coefficient. **[5]**

(ii) The table below shows the ranks given by two judges to four competitors.

Competitor	P	Q	R	S
Judge 1 rank	1	2	3	4
Judge 2 rank	3	2	1	4

Spearman's rank correlation coefficient for these ranks is denoted by r_s. With the same set of ranks for Judge 1, write down a different set of ranks for Judge 2 which gives the same value of r_s. There is no need to find the value of r_s. **[2]**

Q2 June 2013

29.

5 The table shows some of the values of the seasonally adjusted Unemployment Rate (UR), x%, and the Consumer Price Index (CPI), y%, in the United Kingdom from April 2008 to July 2010.

Date	April 2008	July 2008	October 2008	January 2009	April 2009	July 2009	October 2009	January 2010	April 2010	July 2010
UR, x%	5.2	5.7	6.1	6.8	7.5	7.8	7.8	7.9	7.8	7.7
CPI, y%	3.0	4.4	4.5	3.0	2.3	1.8	1.5	3.5	3.7	3.1

These data are summarised below.

$$n = 10 \quad \Sigma x = 70.3 \quad \Sigma x^2 = 503.45 \quad \Sigma y = 30.8 \quad \Sigma y^2 = 103.94 \quad \Sigma xy = 211.9$$

(i) Calculate the product moment correlation coefficient, r, for the data, showing that $-0.6 < r < -0.5$. **[3]**

(ii) Karen says "The negative value of r shows that when the Unemployment Rate increases, it causes the Consumer Price Index to decrease." Give a criticism of this statement. **[1]**

(iii) **(a)** Calculate the equation of the regression line of x on y. **[3]**

(b) Use your equation to estimate the value of the Unemployment Rate in a month when the Consumer Price Index is 4.0%. **[2]**

Q5 June 2013

30.

Tariq collected information about typical prices, £y million, of four-bedroomed houses at varying distances, x miles, from a large city. He chose houses at 10-mile intervals from the city. His results are shown below.

x	10	20	30	40	50	60	70	80
y	1.2	1.4	1.2	0.9	0.8	0.5	0.5	0.3

$n = 8$ $\quad \Sigma x = 360$ $\quad \Sigma x^2 = 20\,400$ $\quad \Sigma y = 6.8$ $\quad \Sigma y^2 = 6.88$ $\quad \Sigma xy = 241$

(i) Use an appropriate formula to calculate the product moment correlation coefficient, r, showing that $-1.0 < r < -0.9$. **[3]**

(ii) State what this value of r shows in this context. **[1]**

(iii) Tariq decides to recalculate the value of r with the house prices measured in hundreds of thousands of pounds, instead of millions of pounds. State what effect, if any, this will have on the value of r. **[1]**

(iv) Calculate the equation of the regression line of y on x. **[3]**

(v) Explain why the regression line of y on x, rather than x on y, should be used for estimating a value of x from a given value of y. **[1]**

Q5 June 2014

31.

Fiona and James collected the results for six hockey teams at the end of the season. They then carried out various calculations using Spearman's rank correlation coefficient, r_s.

(i) Fiona calculated the value of r_s between the number of goals scored FOR each team and the number of goals scored AGAINST each team. She found that $r_s = -1$. Complete the table in the answer book showing the ranks.

Team	A	B	C	D	E	F
Number of goals FOR (rank)	1	2	3	4	5	6
Number of goals AGAINST (rank)						

[1]

(ii) James calculated the value of r_s between the number of goals scored and the number of points gained by the 6 teams. He found the value of r_s to be 1. He then decided to include the results of another two teams in the calculation of r_s. The table shows the ranks for these two teams.

Team	G	H
Number of goals scored (rank)	7	8
Number of points gained (rank)	8	7

Calculate the value of r_s for all 8 teams. **[4]**

Q6 June 2014

32.

For the top 6 clubs in the 2010/11 season of the English Premier League, the table shows the annual salary, £x million, of the highest paid player and the number of points scored, y.

Club	Manchester United	Manchester City	Chelsea	Arsenal	Tottenham	Liverpool
x	5.6	7.4	6.5	4.1	3.6	6.5
y	80	71	71	68	62	58

$n = 6$ $\quad \Sigma x = 33.7$ $\quad \Sigma x^2 = 200.39$ $\quad \Sigma y = 410$ $\quad \Sigma y^2 = 28\,314$ $\quad \Sigma xy = 2313.9$

(i) Use a suitable formula to calculate the product moment correlation coefficient, r, between x and y, showing that $0 < r < 0.2$. **[3]**

(ii) State what this value of r shows in this context. **[1]**

(iii) A fan suggests that the data should be used to draw a regression line in order to estimate the number of points that would be scored by another Premier League club, whose highest paid player's salary is £1.7 million. Give two reasons why such an estimate would be unlikely to be reliable. **[2]**

Q1 June 2015

33.

An expert tested the quality of the wines produced by a vineyard in 9 particular years. He placed them in the following order, starting with the best.

1980 1983 1981 1982 1984 1985 1987 1986 1988

(i) Calculate Spearman's rank correlation coefficient, r_s, between the year of production and the quality of these wines. The years should be ranked from the earliest (1) to the latest (9). **[5]**

(ii) State what this value of r_s shows in this context. **[1]**

Q3 June 2015

34.

The table shows the load a lorry was carrying, x tonnes, and the fuel economy, y km per litre, for 8 different journeys. You should assume that neither variable is controlled.

Load (x tonnes)	5.1	5.8	6.5	7.1	7.6	8.4	9.5	10.5
Fuel economy (y km per litre)	6.2	6.1	5.9	5.6	5.3	5.4	5.3	5.1

$n = 8$ $\Sigma x = 60.5$ $\Sigma y = 44.9$ $\Sigma x^2 = 481.13$ $\Sigma y^2 = 253.17$ $\Sigma xy = 334.65$

(i) Calculate the equation of the regression line of y on x. **[4]**

(ii) Estimate the fuel economy for a load of 9.2 tonnes. **[2]**

(iii) An analyst calculated the equation of the regression line of x on y. Without calculating this equation, state the coordinates of the point where the two regression lines intersect. **[1]**

(iv) Describe briefly the method required to estimate the load when the fuel economy is 5.8 km per litre. **[2]**

Q4 June 2015