SPEARMANS RANK

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

3 Two commentators gave ratings out of 100 for seven sports personalities. The ratings are shown in the table below.

Personality	A	B	C	D	E	F	G
Commentator I	73	76	78	65	86	82	91
Commentator II	77	78	79	80	86	89	95

(i) Calculate Spearman's rank correlation coefficient for these ratings. [5]

(ii) State what your answer tells you about the ratings given by the two commentators. [1]

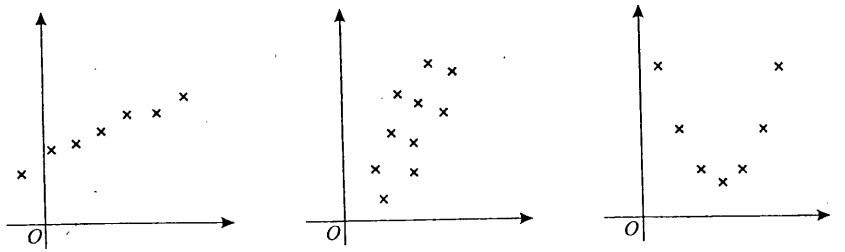
2 Two judges each placed skaters from five countries in rank order.

Position	1st	2nd	3rd	4th	5th
Judge 1	UK	France	Russia	Poland	Canada
Judge 2	Russia	Canada	France	UK	Poland

Calculate Spearman's rank correlation coefficient, r_s, for the two judges' rankings. [5]

PMCC

1 The scatter diagrams below illustrate three sets of bivariate data, A, B and C.



Set A Set B Set C

State, with an explanation in each case, which of the three sets of data has

(i) the largest,

(ii) the smallest,

value of the product moment correlation coefficient. [4]

2 The table contains data concerning five households selected at random from a certain town.

Number of people in the household	2	3	3	5	7
Number of cars belonging to people in the household	1	1	3	2	4

(i) Calculate the product moment correlation coefficient, r, for the data in the table. [5]

(ii) Give a reason why it would not be sensible to use your answer to draw a conclusion about all the households in the town. [1]

4 The table shows the latitude, x (in degrees correct to 3 significant figures), and the average rainfall y (in cm correct to 3 significant figures) of five European cities.

City	x	y
Berlin	52.5	58.2
Bucharest	44.4	58.7
Moscow	55.8	53.3
St Petersburg	60.0	47.8
Warsaw	52.3	56.6

[$n = 5$, $\Sigma x = 265.0$, $\Sigma y = 274.6$, $\Sigma x^2 = 14\,176.54$, $\Sigma y^2 = 15\,162.22$, $\Sigma xy = 14\,464.10$.]

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

(ii) The values of y in the table were in fact obtained from measurements in inches and converted into centimetres by multiplying by 2.54. State what effect it would have had on the value of the product moment correlation coefficient if it had been calculated using inches instead of centimetres. [1]

(iii) It is required to estimate the annual rainfall at Bergen, where $x = 60.4$. Calculate the equation of an appropriate line of regression, giving your answer in simplified form, and use it to find the required estimate. [5]

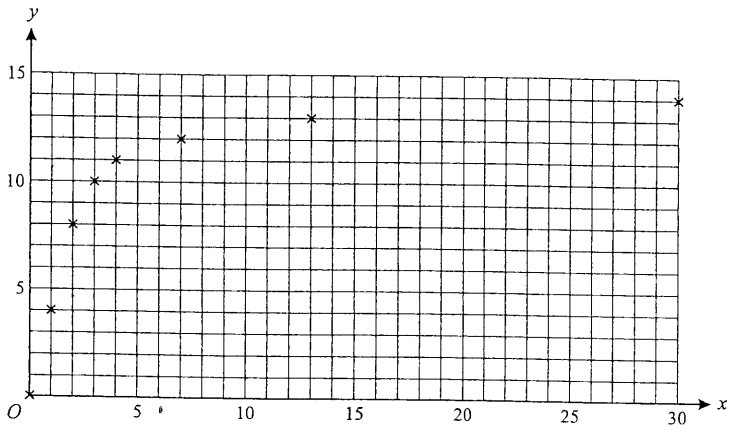
Spear mans Rank

6 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]

6 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.

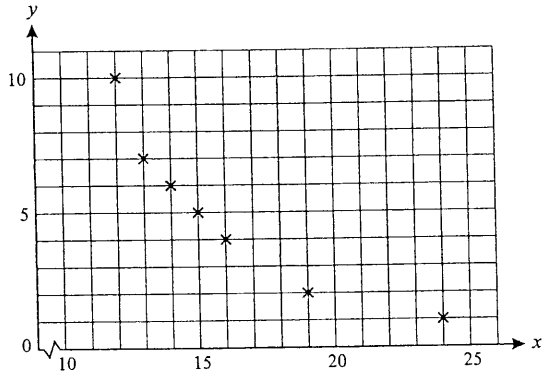
Agent	A	B	C	D	E	F	G
Distance travelled	18	15	12	14	16	24	13
Commission earned	18	45	19	24	27	22	23

(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]

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

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

(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]

Regression Lines

1 Some observations of bivariate data were made and the equations of the two regression lines were found to be as follows.

$$y \text{ on } x: \quad y = -0.6x + 13.0$$
$$x \text{ on } y: \quad x = -1.6y + 21.0$$

(i) State, with a reason, whether the correlation between x and y is negative or positive. [1]

(ii) Neither variable is controlled. Calculate an estimate of the value of x when $y = 7.0$. [2]

(iii) Find the values of $\bar{x}$ and $\bar{y}$. [3]

5 A chemical solution was gradually heated. At five-minute intervals the time, x minutes, and the temperature, y°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]

9 Five observations of bivariate data produce the following results, denoted as (x_i, y_i) for $i = 1, 2, 3, 4, 5$.

(13, 2.7) (13, 4.0) (18, 2.8) (23, 3.3) (23, 2.2)

[$\Sigma x = 90$, $\Sigma y = 15.0$, $\Sigma x^2 = 1720$, $\Sigma y^2 = 46.86$, $\Sigma xy = 264.0$.]

(i) Show that the regression line of y on x has gradient -0.06, and find its equation in the form $y = a + bx$. [4]

(ii) The regression line is used to estimate the value of y corresponding to $x = 20$, but the value $x = 20$ is accurate only to the nearest whole number. Calculate the difference between the largest and the smallest values that the estimated value of y could take. [3]

The numbers e_1, e_2, e_3, e_4, e_5 are defined by

$$e_i = a + bx_i - y_i \quad \text{for } i = 1, 2, 3, 4, 5.$$

(iii) The values of e_1, e_2 and e_3 are 0.6, −0.7 and 0.2 respectively. Calculate the values of e_4 and e_5. [2]

(iv) Calculate the value of $e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$ and explain the relevance of this quantity to the regression line found in part (i). [2]

(v) Find the mean and the variance of e_1, e_2, e_3, e_4, e_5. [4]

9 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]