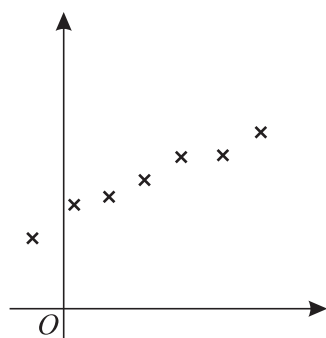OCR Maths S1
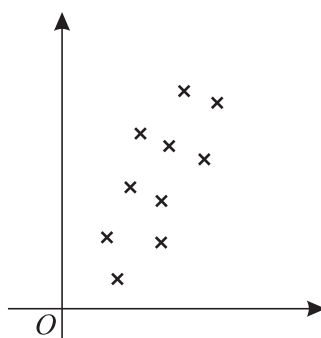
Topic Questions from Papers

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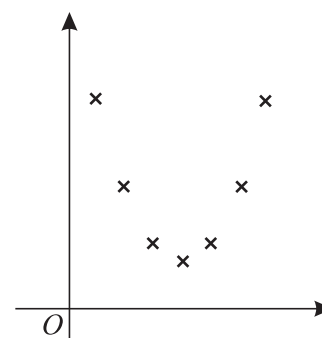
1 The scatter diagrams below illustrate three sets of bivariate data, *A*, *B* and *C*.

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]

(Q1, Jan 2005)

physicsandmathstutor.com

2 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]

(Q3, Jan 2005)

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

$$(13, 2.7) \qquad (13, 4.0) \qquad (18, 2.8) \qquad (23, 3.3) \qquad (23, 2.2)$$

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

  (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]

(Q9, Jan 2005)

4   **(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)

5   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]

(Q4, June 2005)

physicsandmathstutor.com

6   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]

(Q1, June 2006)

7   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.
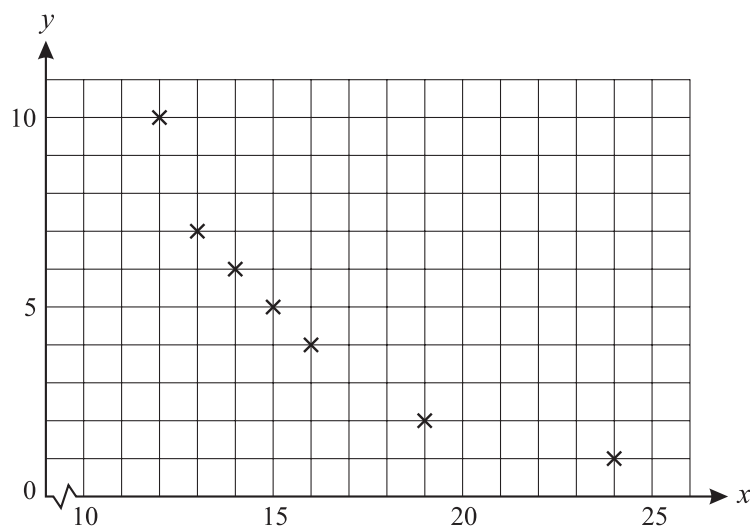
  (ii) For this scatter diagram, what can you say about the value of

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

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

(Q6, June 2006)

8 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]

(Q2, Jan 2007)

9 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]

(Q5, Jan 2007)

10 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]

(Q2, June 2007)

11   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)

12   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 2009)

13 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)

14 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]**

physicsandmathstutor.com

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

$$\text{Tutors 1 and 3:} \quad r_s = -0.9$$
$$\text{Tutors 2 and 3:} \quad r_s = 0.3$$

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

(Q4, Jan 2009)

physicsandmathstutor.com

15 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)

16 **(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.

$$A: \quad r = 0.9 \text{ and } r_s = 1$$
$$B: \quad r = 1 \quad \text{ and } r_s = 0.9$$

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.

The mean of x is 4.5. Th...
The equation of the regression line of y on x is $y = 2.4x + 3.7$.
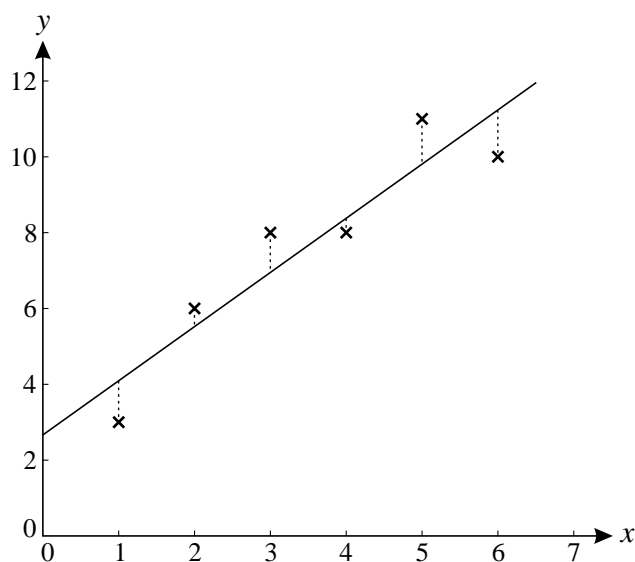The equation of the regression line of x on y is $x = 0.40y -$

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

(Q6, Jan 2010)

17 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)

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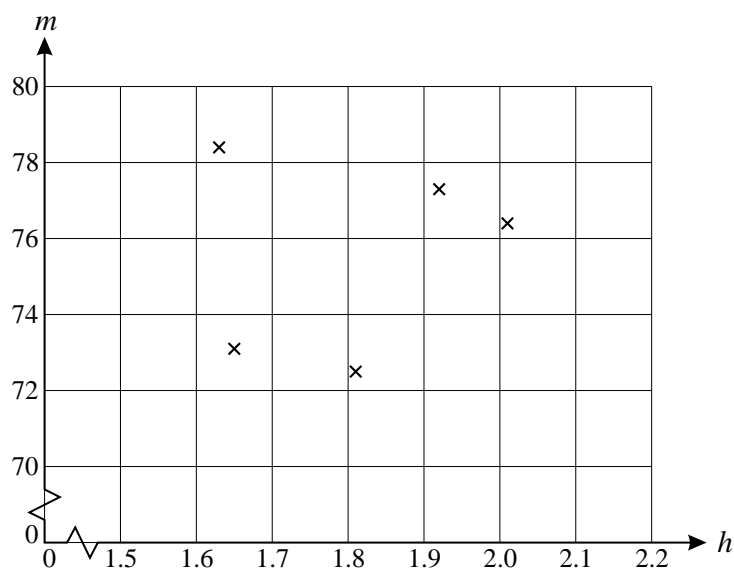
The results are summarised as follows.

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

(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)

19 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)

physicsandmathstutor.com

20 **(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, $x\,°C$, of a rod was gradually increased from $0\,°C$, and the extension, y mm, was measured nine times at $50\,°C$ 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 $-50\,°C$.   **[1]**

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

(Q3, June 2010)

21 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)

22 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)

23 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)

24 The orders in which 4 contestants, *P*, *Q*, *R* and *S*, were placed in two competitions are shown in the table.
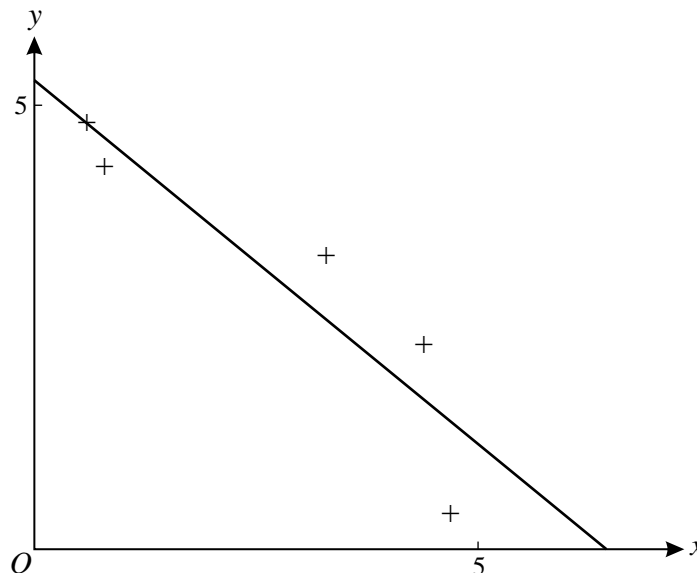
Position	1st	2nd	3rd	4th
Competition 1	*Q*	*R*	*S*	*P*
Competition 2	*Q*	*P*	*R*	*S*

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

(Q2, June 2011)

physicsandmathstutor.com

25 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)

26 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)

27 **(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)

28 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)

29   **(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)

30   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.

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

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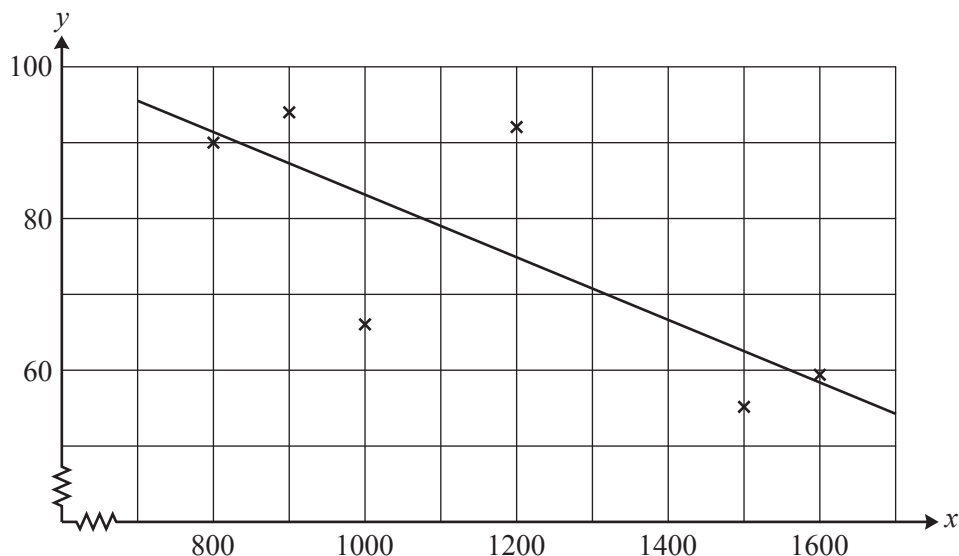
(ii) The GDP for another country, Tanzania, is 1300 dollars. Use the regression line in the diagram to estimate the IMR of Tanzania.   **[1]**

(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)

31   **(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)