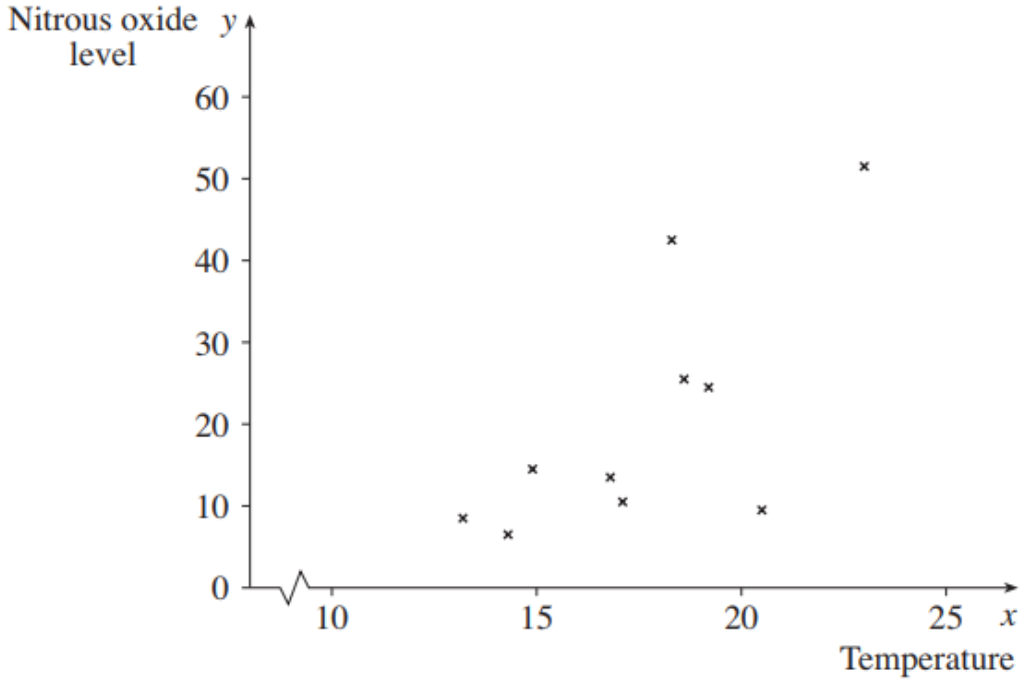
Hypothesis Tests for Correlation (From OCR 4767)

Q1, (Jan 2006, Q3)

A researcher is investigating the relationship between temperature and levels of the air pollutant nitrous oxide at a particular site. The researcher believes that there will be a positive correlation between the daily maximum temperature, x, and nitrous oxide level, y. Data are collected for 10 randomly selected days. The data, measured in suitable units, are given in the table and illustrated on the scatter diagram.

x	13.3	17.2	16.9	18.7	18.4	19.3	23.1	15.0	20.6	14.4
y	9	11	14	26	43	25	52	15	10	7



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

(ii) Perform a hypothesis test at the 5% level to check the researcher's belief, stating your hypotheses clearly. [5]

(iii) It is suggested that it would be preferable to carry out a test based on the product moment correlation coefficient. State the distributional assumption required for such a test to be valid. Explain how a scatter diagram can be used to check whether the distributional assumption is likely to be valid and comment on the validity in this case. [3]

(iv) A statistician investigates data over a much longer period and finds that the assumptions for the use of the product moment correlation coefficient are in fact valid. Give the critical region for the test at the 1% level, based on a sample of 60 days. [2]

(v) In a different research project, into the correlation between daily temperature and ozone pollution levels, a positive correlation is found. It is argued that this shows that high temperatures cause increased ozone levels. Comment on this claim. [3]

Q2, (Jun 2006, Q3)

A student is investigating the relationship between the length x mm and circumference y mm of plums from a large crop. The student measures the dimensions of a random sample of 10 plums from this crop. Summary statistics for these dimensions are as follows.

$$\Sigma x = 4715 \quad \Sigma y = 13\,175 \quad \Sigma x^2 = 2\,237\,725$$

$$\Sigma y^2 = 17\,455\,825 \quad \Sigma xy = 6\,235\,575 \quad n = 10$$

(i) Calculate the sample product moment correlation coefficient. [5]

(ii) Carry out a hypothesis test at the 5% significance level to determine whether there is any correlation between length and circumference of plums from this crop. State your hypotheses clearly, defining any symbols which you use. [6]

(iii) (*A*) Explain the meaning of a 5% significance level. [2]

(*B*) State one advantage and one disadvantage of using a 1% significance level rather than a 5% significance level in a hypothesis test. [2]

The student decides to take another random sample of 10 plums. Using the same hypotheses as in part **(ii)**, the correlation coefficient for this second sample is significant at the 5% level. The student decides to ignore the first result and concludes that there is correlation between the length and circumference of plums in the crop.

(iv) Comment on the student's decision to ignore the first result. Suggest a better way in which the student could proceed. [3]

Q3, (Jun 2007, Q2)

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.

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

(i) Calculate the value of Spearman's rank correlation coefficient. [5]

(ii) Carry out a hypothesis test at the 5% level to determine whether there is positive association between the student's estimates and the actual birth weights of babies in the underlying population. [5]

(iii) Calculate the value of the product moment correlation coefficient of the sample. You may use the following summary statistics in your calculations:

$$\Sigma x = 31.63, \quad \Sigma y = 33.1, \quad \Sigma x^2 = 101.92, \quad \Sigma y^2 = 112.61, \quad \Sigma xy = 106.51.$$ [5]

(iv) Explain why, if the underlying population has a bivariate Normal distribution, it would be preferable to carry out a hypothesis test based on the product moment correlation coefficient.

Comment briefly on the significance of the product moment correlation coefficient in relation to that of Spearman's rank correlation coefficient. [4]

Q4, (Jan 2009, Q1i,ii)

A researcher is investigating whether there is a relationship between the population size of cities and the average walking speed of pedestrians in the city centres. Data for the population size, x thousands, and the average walking speed of pedestrians, $y\,\mathrm{m\,s^{-1}}$, of eight randomly selected cities are given in the table below.

x	18	43	52	94	98	206	784	1530
y	1.15	0.97	1.26	1.35	1.28	1.42	1.32	1.64

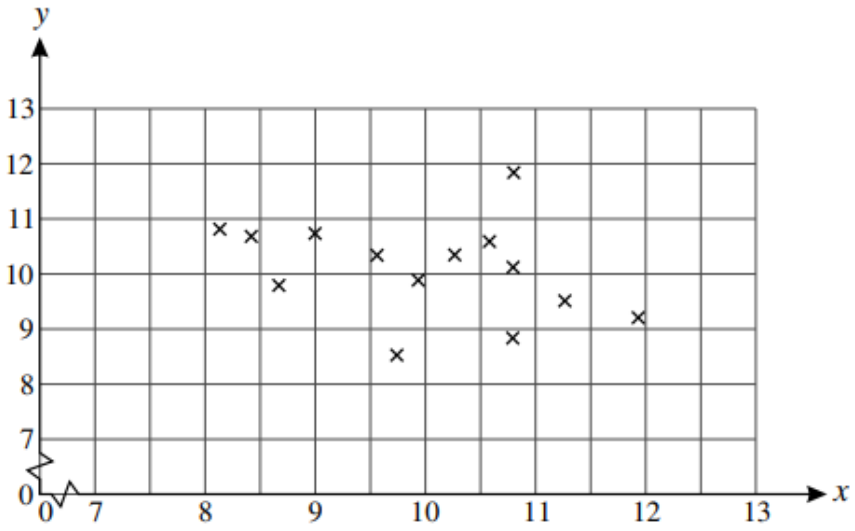
(i) Calculate the value of Spearman's rank correlation coefficient. **[5]**

(ii) Carry out a hypothesis test at the 5% significance level to determine whether there is any association between population size and average walking speed. **[6]**

Q5, (Jan 2011, Q1)

The scatter diagram below shows the birth rates x, and death rates y, measured in standard units, in a random sample of 14 countries in a particular year. Summary statistics for the data are as follows.

$$\Sigma x = 139.8 \quad \Sigma y = 140.4 \quad \Sigma x^2 = 1411.66 \quad \Sigma y^2 = 1417.88 \quad \Sigma xy = 1398.56 \quad n = 14$$



(i) Calculate the sample product moment correlation coefficient. **[5]**

(ii) Carry out a hypothesis test at the 5% significance level to determine whether there is any correlation between birth rates and death rates. **[6]**

(iii) State the distributional assumption which is necessary for this test to be valid. Explain briefly in the light of the scatter diagram why it appears that the assumption may be valid. **[2]**

(iv) The values of x and y for another country in that year are 14.4 and 7.8 respectively. If these values are included, the value of the sample product moment correlation coefficient is −0.5694. Explain why this one observation causes such a large change to the value of the sample product moment correlation coefficient. Discuss whether this brings the validity of the test into question. **[4]**

Q6, (Jun 2012, Q1)

The times, in seconds, taken by ten randomly selected competitors for the first and last sections of an Olympic bobsleigh run are denoted by x and y respectively. Summary statistics for these data are as follows.

$$\Sigma x = 113.69 \quad \Sigma y = 52.81 \quad \Sigma x^2 = 1292.56 \quad \Sigma y^2 = 278.91 \quad \Sigma xy = 600.41 \quad n = 10$$

(i) Calculate the sample product moment correlation coefficient. **[5]**

(ii) Carry out a hypothesis test at the 10% significance level to investigate whether there is any correlation between times taken for the first and last sections of the bobsleigh run. **[6]**

(iii) State the distributional assumption which is necessary for this test to be valid. Explain briefly how a scatter diagram may be used to check whether this assumption is likely to be valid. **[2]**

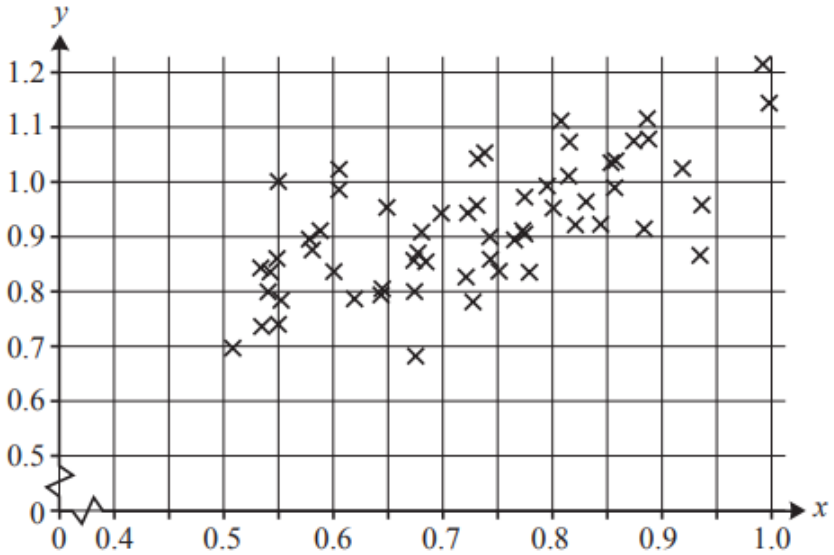
(iv) A commentator says that in order to have a fast time on the last section, you must have a fast time on the first section. Comment briefly on this suggestion. **[2]**

(v) (*A*) Would your conclusion in part **(ii)** have been different if you had carried out the hypothesis test at the 1% level rather than the 10% level? Explain your answer. **[2]**

(*B*) State one advantage and one disadvantage of using a 1% significance level rather than a 10% significance level in a hypothesis test. **[2]**

Q7, (Jun 2013, Q1)

Salbutamol is a drug used to improve lung function. In a medical trial, a random sample of 60 people with impaired lung function was selected. The forced expiratory volume in one second (FEV1) was measured for each person, both before being given salbutamol and again after a two-week course of the drug. The variables x and y, measured in suitable units, represent FEV1 before and after the two-week course respectively. The data are illustrated in the scatter diagram below, together with the summary statistics for these data.



Summary statistics:

$$n = 60, \quad \Sigma x = 43.62, \quad \Sigma y = 55.15, \quad \Sigma x^2 = 32.68, \quad \Sigma y^2 = 51.44, \quad \Sigma xy = 40.66$$

(i) Calculate the sample product moment correlation coefficient. **[5]**

(ii) Carry out a hypothesis test at the 5% significance level to investigate whether there is positive correlation between FEV1 before and after the course. **[6]**

(iii) State the distributional assumption which is necessary for this test to be valid. State, with a reason, whether the assumption appears to be valid. **[2]**

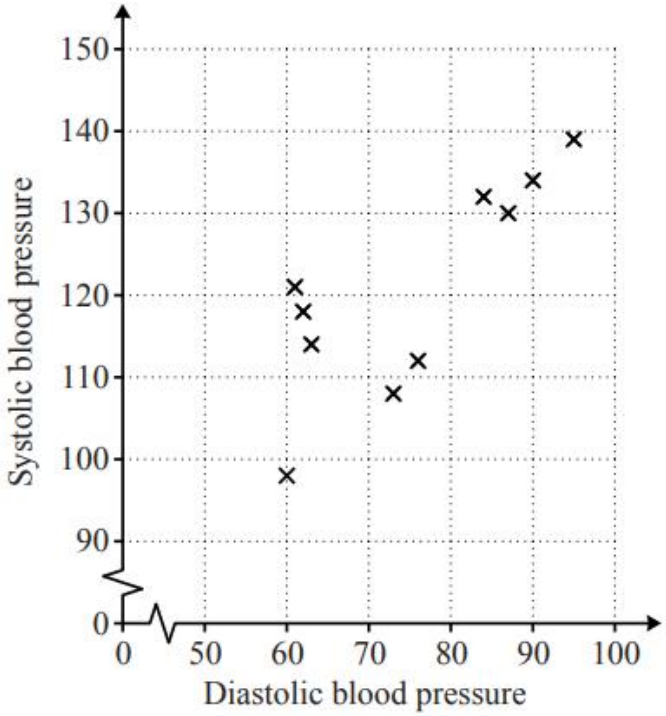
(iv) Explain the meaning of the term 'significance level'. **[2]**

(v) Calculate the values of the summary statistics if the data point $x - 0.55$, $y - 1.00$ had been incorrectly recorded as $x = 1.00$, $y = 0.55$. **[3]**

Q8, (Jun 2014, Q1)

A medical student is investigating the claim that young adults with high diastolic blood pressure tend to have high systolic blood pressure. The student measures the diastolic and systolic blood pressures of a random sample of ten young adults. The data are shown in the table and illustrated in the scatter diagram.

Diastolic blood pressure	60	61	62	63	73	76	84	87	90	95
Systolic blood pressure	98	121	118	114	108	112	132	130	134	139



(i) Calculate the value of Spearman's rank correlation coefficient for these data. **[5]**

(ii) Carry out a hypothesis test at the 5% significance level to examine whether there is positive association between diastolic blood pressure and systolic blood pressure in the population of young adults. **[6]**

(iii) Explain why, in the light of the scatter diagram, it might not be valid to carry out a test based on the product moment correlation coefficient. **[2]**

The product moment correlation coefficient between the diastolic and systolic blood pressures of a random sample of 10 athletes is 0.707.

(iv) Carry out a hypothesis test at the 1% significance level to investigate whether there appears to be positive correlation between these two variables in the population of athletes. You may assume that in this case such a test is valid. **[5]**

Q9, (Jun 2016, Q1)

A researcher believes that there may be negative association between the quantity of fertiliser used and the percentage of the population who live in rural areas in different countries. The data below show the percentage of the population who live in rural areas and the fertiliser use measured in kg per hectare, for a random sample of 11 countries.

Percentage of population	33	6	58	35	81	69	61	7	74	71	17
Fertiliser use	76	44	6	68	3	10	7	176	5	137	157

(i) Draw a scatter diagram to illustrate the data. **[3]**

(ii) Explain why it might not be valid to carry out a test based on the product moment correlation coefficient in this case. **[2]**

(iii) Calculate the value of Spearman's rank correlation coefficient. **[5]**

(iv) Carry out a hypothesis test at the 1% significance level to investigate the researcher's belief. **[6]**

(v) Explain the meaning of '1% significance level'. **[1]**

(vi) In order to carry out a test based on Spearman's rank correlation coefficient, what modelling assumptions, if any, are required about the underlying distribution? **[1]**