OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2614/1

Statistics 2

Friday **11 JUNE 2004** Morning 1 hour 20 minutes

Additional materials:
Answer booklet
Graph paper
MEI Examination Formulae and Tables (MF12)

TIME 1 hour 20 minutes

INSTRUCTIONS TO CANDIDATES

- Write your Name, Centre Number and Candidate Number in the spaces provided on the answer booklet.
- Answer **all** questions.
- You are permitted to use a graphical calculator in this paper.

INFORMATION FOR CANDIDATES

- The allocation of marks is given in brackets [ ] at the end of each question or part question.
- You are advised that an answer may receive no marks unless you show sufficient detail of the working to indicate that a correct method is being used.
- Final answers should be given to a degree of accuracy appropriate to the context.
- The total number of marks for this paper is 60.

This question paper consists of 4 printed pages.

HN/3

Registered Charity 1066969

[Turn over

1 A meteorologist wishes to test whether there is any correlation between temperature, x°C, and wind speed, y mph. For a selection of places in the United Kingdom on a particular day, he collects data which are illustrated in Fig. 1 and summarised as follows.

$$n = 25 \qquad \sum x = 307 \qquad \sum y = 250$$

$$\sum x^2 = 3853 \qquad \sum y^2 = 3008 \qquad \sum xy = 3143$$

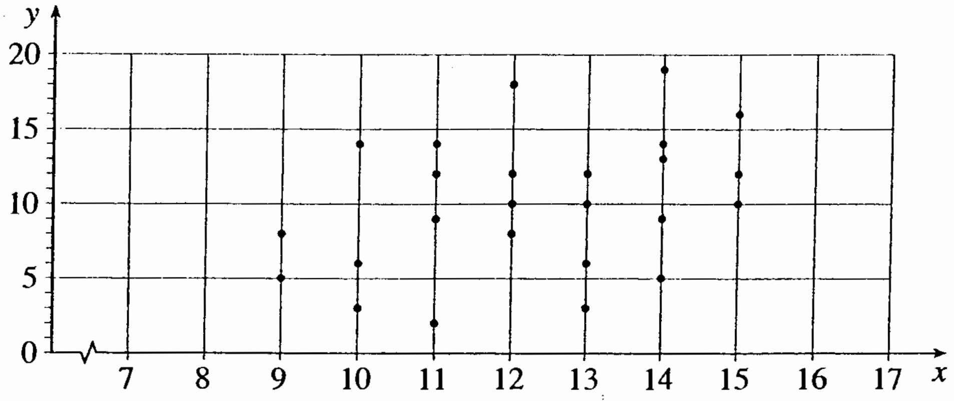

Fig. 1

(i) Calculate the product moment correlation coefficient for the data. Carry out a suitable hypothesis test at the 5% significance level, using the null hypothesis H_0: $\rho = 0$. Define ρ and state your alternative hypothesis and conclusion carefully. [9]

(ii) What must be assumed about the underlying distribution for the test to be valid? Discuss, with reference to Fig. 1, whether this assumption is reasonable in this case. [3]

(iii) Another data point, $x = 6, y = 29$, was omitted from the data set. Explain briefly the effect its inclusion would have on the product moment correlation coefficient. Comment on the validity of the test if this point were included. [3]

2 The Royal Automobile Association defines *peak-time* as 6 am to 6 pm, Monday to Friday. It records the number of vehicle breakdowns reported per hour. The figures for a random sample of 40 peak-time hours in a certain area are as follows.

Number of breakdowns, x	0	1	2	3	4 or more
Number of hours, f	25	11	3	1	0

(i) Find the mean and variance of the data. [2]

(ii) Give *two* distinct reasons why the Poisson distribution might be thought to be a suitable model for the number of breakdowns reported per peak-time hour. [2]

(iii) Using the mean as calculated in part **(i)**, calculate the expected frequencies corresponding to the given frequencies in the table. Comment further on the suitability of the Poisson model. [6]

(iv) For a peak-time 60-hour week, state a suitable approximating distribution for T, the total number of reported breakdowns. Hence estimate an interval which covers 95% of the values of T, in the form $a < T < b$. [5]

3 In a certain hospital's outpatients department, 3% of patients miss an appointment.

(i) For a random sample of n patients, state the distribution of the number who miss an appointment. [2]

(ii) For a random sample of 20 patients, find the probability that at least one of them misses an appointment. [2]

(iii) On a particular day, 320 patients have appointments at the hospital. Using a Poisson approximating distribution, find the probability that between 8 and 12 patients (inclusive) miss their appointments. [3]

Each week, there are 2000 outpatient appointments at the hospital.

(iv) Use a suitable approximating distribution to find the probability that at least 50 patients miss their appointments in a particular week. [5]

(v) The hospital managers wish to reduce the probability that at least 50 patients miss their appointments in a week to 0.5. What percentage of patients missing their appointments is consistent with this? [3]

[Turn over

4 Ben has five coins in his pocket, consisting of two 10p coins, two 20p coins and one 50p coin. He takes two of these coins from his pocket at random to put into a collecting tin.

The probability distribution of the value, X pence, of the coins he takes from his pocket is given in the following table.

r	20	30	40	60	70
$\mathrm{P}(X=r)$	0.1	0.4	0.1	0.2	0.2

(i) Use a probability argument to show that $\mathrm{P}(X = 30) = 0.4$. [2]

(ii) Find the mean and standard deviation of X. [4]

Ellie says she will contribute the same amount as Ben, and Charlotte says she will contribute 50p.

(iii) Deduce the mean and standard deviation of the total contribution from the three people. [4]

Ben takes a third coin from his pocket to add to the other two coins he gave.

(iv) Find the probability that he gives a *total* of 50p. [3]

(v) Show that the expectation of the total amount he gives is 66p. [2]

Mark Scheme

GENERAL INSTRUCTIONS

Marks in the mark scheme are explicitly designated as **M**, **A**, **B**, **E** or **G**.

M marks ("method") are for an attempt to use a correct method (not merely for stating the method).

A marks ("accuracy") are for accurate answers and can only be earned if corresponding **M** mark(s) have been earned. Candidates are expected to give answers to a sensible level of accuracy in the context of the problem in hand. The level of accuracy quoted in the mark scheme will sometimes deliberately be greater than is required, when this facilitates marking.

B marks are independent of all others. They are usually awarded for a single correct answer. Typically they are available for correct quotation of points such as 1.96 from tables.

E marks ("explanation") are for explanation and/or interpretation. These will frequently be sub divisible depending on the thoroughness of the candidate's answer.

G marks ("graph") are for completing a graph or diagram correctly.

- Insert part marks in **right-hand** margin in line with the mark scheme. For fully correct parts tick the answer. For partially complete parts indicate clearly in the body of the script where the marks have been gained or lost, in line with the mark scheme.
- Please indicate incorrect working by ringing or underlining as appropriate.
- Insert total in **right-hand** margin, ringed, at end of question, in line with the mark scheme.
- Numerical answers which are not exact should be given to at least the accuracy shown. Approximate answers to a greater accuracy *may* be condoned.
- Probabilities should be given as fractions, decimals or percentages.
- FOLLOW-THROUGH MARKING SHOULD NORMALLY BE USED WHEREVER POSSIBLE. There will, however, be an occasional designation of '**c.a.o.**' for "correct answer only".
- Full credit MUST be given when correct alternative methods of solution are used. If errors occur in such methods, the marks awarded should correspond as nearly as possible to equivalent work using the method in the mark scheme.
- The following notation should be used where applicable:

FT	Follow-through marking
BOD	Benefit of doubt
ISW	Ignore subsequent working

Question 1

(i)	**EITHER:** $S_{xy} = \Sigma xy - n\overline{xy} = 3143 - 25 \times 12.28 \times 10 = 73$ $S_{xx} = \Sigma x^2 - n\overline{x}^2 = 3853 - 25 \times 12.28^2 = 83.04$ $S_{yy} = \Sigma y^2 - n\overline{y}^2 = 3008 - 25 \times 10^2 = 508$ $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{73}{\sqrt{83.04 \times 508}} = 0.355$ **OR:** Cov (x,y) = $\dfrac{\sum xy}{n} - \overline{xy}$ = 3143/25 – 12.28×10 = 2.92 sd(x) = $\sqrt{\dfrac{\sum x^2}{n} - (\overline{x})^2}$ = √(3853/25 – 12.28^2) =√3.3216 = 1.823 sd(y) = $\sqrt{\dfrac{\sum y^2}{n} - (\overline{y})^2}$ = √(3008/25 – 10^2) =√20.32 = 4.508 $r = \dfrac{\text{Cov}(x,y)}{sd(x)sd(y)} = \dfrac{2.92}{1.823 \times 4.508} = 0.355$	B1 for S_{xy} B1 for at least one of S_{xx} or S_{yy} M1 for structure of r A1 (0.35 to 0.36) B1 for Cov (x,y) B1 for at least one sd or variance M1 for structure of r A1 (0.35 to 0.36)	**4**
	H_1: $\rho \neq 0$ (two-tailed test) where ρ is the population correlation coefficient For n = 25, 5% critical value = 0.3961 Since 0.355 < 0.3961 we cannot reject H_0: There is not sufficient evidence at the 5% significance level to suggest there is a correlation between temperature and wind speed.	B1 for H_1 in symbols B1 for defining ρ B1FT for critical value M1 for comparison leading to a conclusion A1 for conclusion in words FT their values	**5**
(ii)	Underlying distribution must be bivariate normal. The elliptical shape of the scatter would seem to indicate that it is reasonable in this case.	B1 CAO for bivariate normal B1 indep for elliptical shape E1 dep for conclusion	**3**
(iii)	With the addition of the pair of data, x = 6, y = 29, the product moment correlation coefficient will be reduced, since it is influenced by extreme values. The test is less likely to be valid since the “elliptical” shape of the scatter has been distorted.	B1 for reduced or negative E1 indep for reason for change B1 for validity	**3**
			15

Question 2

<table>
<tr>
<td><b>(i)</b></td>
<td>Mean $= \dfrac{\Sigma xf}{\Sigma f} = \dfrac{20}{40} = 0.5$
Variance $= \dfrac{\Sigma x^2 f}{\Sigma f} - \bar{x}^2 = \dfrac{32}{40} - 0.5^2 = 0.55$</td>
<td>B1
B1 FT</td>
<td><b>2</b></td>
</tr>
<tr>
<td><b>(ii)</b></td>
<td>Any two from the following:
• Mean ≈ variance
• Breakdowns random and independent
• Uniform (mean) rate of occurrence</td>
<td>B1
B1</td>
<td><b>2</b></td>
</tr>
<tr>
<td><b>(iii)</b></td>
<td>Use $\lambda = 0.5$ and Poisson tables or formula to calculate probabilities
<table>
<tr><td>number of breakdowns, <i>x</i></td><td>0</td><td>1</td><td>2</td><td>3</td><td>4+</td></tr>
<tr><td>Probability P(<i>X</i> = <i>x</i>)</td><td>0.6065</td><td>0.3033</td><td>0.0758</td><td>0.0126</td><td>0.0018</td></tr>
<tr><td>expected frequency, <i>f</i></td><td>24.3</td><td>12.1</td><td>3.03</td><td>0.505</td><td>0.072</td></tr>
</table>
Multiply by 40 to obtain expected frequencies
Since observed and expected frequencies are approximately the same, this is further evidence that a Poisson distribution with $\lambda = 0.5$ is suitable.</td>
<td>M1
A2 FT first four probabilities (A1 if at least one correct)
A1 FT for final probability using 1 – P(X ≤ 3)
B1 for multiplication by 40
E1 for explanation based on sensible exp. freq.</td>
<td><b>6</b></td>
</tr>
<tr>
<td><b>(iv)</b></td>
<td>For a peak-time 60 hour working week, a suitable approximating distribution for the number of reported breakdowns is
$X \sim \mathrm{N}(30, 30)$
<i>Using a symmetrical interval:</i>
Require <i>a</i> and <i>b</i> such that $\mathrm{P}(a < T < b) = 0.95$
But $\mathrm{P}(-1.96 < Z < 1.96) = 0.95$
Hence $a = 30 - 1.96 \times \sqrt{30} = 19.26$
and $b = 30 + 1.96 \times \sqrt{30} = 40.74$
Must round to $a = 19$ and $b = 41$
NB allow non-symmetrical intervals</td>
<td>B1 for Normal
B1dep for parameters
B1 for 1.96 seen
M1 for at least one equation in any form
A1 CAO for both values</td>
<td><b>5</b></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><b>15</b></td>
</tr>
</table>

Question 3

(i)	Distribution of number of people who miss their appointment is given by $X \sim \mathrm{B}(n, 0.03)$	B1 for binomial B1 dep for parameters	**2**
(ii)	For $n = 20$: P(at least one misses an appointment) $= 1 - 0.97^{20} = 0.4562$	M1 for complete method A1 CAO for 0.46 or better	**2**
(iii)	For a suitable Poisson distribution, $\lambda = 320 \times 0.03 = 9.6$ $\mathrm{P}(8 \le X \le 12) = 0.8279 - 0.2584 = 0.5695$	B1 for $\lambda = 9.6$ M1 for tables using 7 and 12 and subtraction A1 CAO (0.568 to 0.570)	**3**
(iv)	A suitable approximating distribution is N(60, 58.2) P(at least 50 outpatients miss their appointment) $= \mathrm{P}(X > 49.5) = \mathrm{P}\left(Z > \frac{49.5 - 60}{\sqrt{58.2}}\right)$ $= \mathrm{P}(Z > -1.376) = 0.9157$ Alternatives: No CC $\mathrm{P}(Z > -1.311) = 0.9051$ B1B0M1M1A1 max Wrong CC $\mathrm{P}(Z > -1.245) = 0.8934$ B1B0M1M1A1 max N(60,60) with CC $\mathrm{P}(Z > -1.356) = 0.9124$ B1B1M1M1A1 max N(60,60) no CC $\mathrm{P}(Z > -1.291) = 0.9017$ B1B0M1M1A1 max N(60,60) wrong CC $\mathrm{P}(Z > -1.226) = 0.8899$ B1B0M1M1A1 max	B1 for distribution B1 CAO for contin. corr. M1 for standardizing M1 for prob. calc.with correct tail, and p between 0.5 and 1 A1 4dp required	**5**
(v)	For $n = 2000$ and p = probability outpatient does not turn up, we require P(at least 50 outpatients miss their appointment) = 0.5 $\Rightarrow \mathrm{P}(X > 49.5) = \mathrm{P}\left(Z > \frac{49.5 - 2000p}{\sqrt{2000(1-p)}}\right) = 0.5$ $\Rightarrow 49.5 - 2000p = 0$ $\Rightarrow p = \frac{49.5}{2000} = 0.02475$ Hence, reduce p to 2.475 %	B1 for z = 0 or μ = 50 M1 dep for equation A1 CAO Allow omission of CC leading to 2.5%	**3**
			15

Question 4

(i)	Contribution is 30 pence if he chooses one 10 pence coin and one 20 pence coin $\text{P}(X=30) = 2\times\frac{2}{5}\times\frac{2}{4} = 0.4$ *or* $\text{P}(X=30) = \dfrac{\binom{2}{1}\times\binom{2}{1}}{\binom{5}{2}} = \dfrac{2\times2}{10} = 0.4$ NB $\frac{4}{5}\times\frac{1}{2}$ *or* $\frac{1}{5}+\frac{1}{5}$ scores B1B0 unless explained	B2 B2 **NB ANSWER GIVEN**	**2**
(ii)	$\text{E}(X) = 20\times0.1+30\times0.4+40\times0.1+60\times0.2+70\times0.2$ $= 44$ pence $\text{E}(X^2)$ $= 20^2\times0.1+30^2\times0.4+40^2\times0.1+60^2\times0.2+70^2\times0.2$ $(= 2260)$ Hence $\text{Var}(X) = \text{E}(X^2)-[\text{E}(X)]^2$ $= 2260-44^2 = 324$ Hence $\text{s.d.}(X) = \sqrt{324} = 18$ pence	B1CAO for E(X) M1 for E(X^2) M1 dep for positive variance A1 **cao**	**4**
(iii)	Total amount given $= 2X+50$ Mean $= \text{E}(Y) = 2\text{E}(X) + 50 = 2\times44+50 = £1.38$ $\text{S.d.}(Y) = 2\ \text{s.d.}(X) = 2\times18 = 36$ pence	B1 for new total (may be implied by correct answers to mean or sd) B1 FT their 44 M1 for 2 * s.d.(X) or 4 * var(X) A1 FT their 18	**4**
(iv)	*Either:* P(Ben gives total of 50p) $= 0.4\times\frac{1}{3}+0.1\times\frac{2}{3} = 0.2$ *or:* $3\times\frac{2}{5}\times\frac{1}{4}\times\frac{2}{3}$ *or:* $12\times\frac{1}{5}\times\frac{1}{4}\times\frac{1}{3}$ *or:* $\dfrac{\binom{2}{2}\times\binom{2}{1}}{\binom{5}{3}}$	M1 for one of the couplets M1 for sum of both A1 CAO	**3**
(v)	*Either:* Table method 40 50 70 80 90 0.2 0.2 0.1 0.4 0.1 *or:* 110 – 44 = 66 pence *or:* $\frac{3}{5}$ * 110 = 66 pence *or:* $\frac{3}{2}$ * 44 = 66 pence *or:* 3 * 22 = 66 pence	M1 for all amounts and at least 3 probs correct A1 **NB ANSWER GIVEN** M1 for * 110, 44 or 22	**2**
			15

Examiner’s Report

2614 Statistics 2

General Comments

Candidates were generally well prepared for this paper, and most candidates appeared to be able to demonstrate their knowledge and understanding. It is pleasing to report that, as last summer, few candidates appeared to have been inappropriately entered for the paper. In questions where comments were required, notably Question 1 parts ii and iii, candidates appeared more successful than in previous papers in making reasonably convincing statements. The vast majority of candidates appeared to have adequate time to complete the paper. Most parts of Questions 1, 3 and 4 on correlation, Poisson and Normal approximations and random variables respectively, were generally answered well. Question 2 on the Poisson distribution proved to be more demanding and candidates often found difficulty, particularly with part iii, where expected frequencies had to be calculated and with part iv where in effect a 95% confidence interval was required.

Comments on Individual Questions

Q.1 (i) The vast majority of candidates correctly calculated the test statistic. The most common error was the omission of use of square root in the denominator, even when the square root was quoted in the formula.

In the hypothesis test, there was an improvement in the number of candidates who dealt well with the pmcc test, even if a correct definition of ρ as the 'population correlation coefficient' was relatively rarely seen. Rather more candidates than in previous years put the concluding statement in context, rather than simply stating that 'there is no correlation'. A surprising number of candidates thought that a one-tailed test was required. Such candidates could follow through and lose just one mark.

(ii) Most candidates failed to quote the required assumption of a bivariate normal distribution. This failing was often strongly linked to Centres, with many Centres in which no candidates gained this mark, and others where almost all did so. The fact that an elliptical scatter diagram can be used as an indication that the test is valid was better known, although by no means universally so, and again the knowledge thereof was strongly linked to Centres. Candidates who were aware of this usually did conclude that this diagram satisfied the requirement. This lack of awareness is somewhat surprising, since a proper understanding of the need for bivariate normality, and the ellipticity of the scatter diagram, is a key feature of the compulsory coursework that presently accompanies this Unit. Centres will need to give even greater emphasis to this topic when the coursework is removed for the revised "Curriculum 2004" specification.

(iii) This was generally fairly well answered, with frequent sensible comments about the effect of the anomalous point. However some candidates mentioned a change in the value of the pmcc, but did not specify that this change would be a decrease. There was also some confusion between the validity of the test and the conclusion of the test.

Q.2 (i) The majority of candidates were able to compute the mean and variance correctly, although a surprising number of errors were seen, particularly in finding the variance.

(ii) A number of candidates felt that 'random' and 'independent' were good enough to count as two separate reasons, despite the fact that previous examiner's reports and mark schemes pair them up to score just one mark. A good many candidates did correctly refer to a uniform mean rate over time, and/or note the similarity of the mean to the variance. However 'n is large and p is small', 'there is a known mean' or 'the distribution is discrete' were three of the most frequently seen spurious comments.

(iii) Candidates who knew how to begin this part usually scored most of the six marks available, although a considerable minority were completely defeated by the question. Of those who knew how to start, a number calculated P(X=4), rather than P(X≥4), and others forgot to multiply by 40 to find expected frequencies. A more surprising error was the use of each of the original frequencies rather than their sum, as the multipliers of the relevant probabilities. For candidates who had calculated reasonable expected frequencies, the final comment was usually correct.

(iv) Most candidates realised that a normal approximating distribution was required, but in calculating the parameters, many arrived at spurious values, often simply using 0.5 as the mean rather than multiplying by 60. Most candidates who made a reasonable attempt at the question tried to find a two-tailed interval, but many used 1.645 rather than 1.960 as their two-tailed z-value. Some candidates were unsure as to whether or not a continuity correction was required; the mark scheme allowed for both interpretations. Occasionally candidates calculated a single-sided interval fully correctly.

Q.3 (i) This was generally answered well, although some candidates were not able to express their answers correctly, despite demonstrating a perfectly sound understanding of the Binomial distribution in the next part.

(ii) This was again fairly well answered, although many candidates used the correct Binomial model, but calculated 1 – P(X=0) – P(X=1). This attracted no credit. Some credit was given to those who used a Poisson approximation.

(iii) Most candidates correctly recognised that Po(9.6) was required, but a very common error was to use P(X≤12) – P(X≤8), rather than P(X≤12) – P(X≤7). A number of candidates interpreted the instruction to use a 'Poisson approximating distribution' as a requirement to use a Normal approximation to the Poisson distribution. Such candidates gained no credit.

(iv) Although this part was not particularly easy, a good proportion of candidates scored highly, with the omission of a continuity correction being the most common error. Some candidates found the correct answer and then used it to find the opposite tail, by subtracting their correct answer from one, whilst others divided by the variance rather than the standard deviation. A disappointing aspect was the high number of candidates who simplified the solution by rounding their z-value to only 2 d.p., thus avoiding the need to use difference columns. Centres should be aware that this approach is penalized.

(v) Many candidates realised that z = 0, but were unable to proceed further with a comparatively straightforward calculation. The alternative approach, seen less frequently, which uses the fact that the new mean is 50 usually led immediately to a correct answer. There was some confusion between 2.5% and 0.025, with occasional answers of 0.025% being seen, despite fully correct working. Omission of a continuity correction was not penalized.

Q.4 (i) This was generally answered fairly well, although the given answer did lead to a variety of spurious attempts at justification. Some candidates thought that it was sufficient to use the fact that the sum of the probabilities is one.

(ii) This standard calculation was not answered as well as might have been expected. A common error was to divide the correct mean value of 44 by 5. Similar errors occurred in the variance calculation, and some candidates forgot to take the square root at the end.

(iii) Many candidates correctly used 2X + 50 to deduce the new mean of 138, but quite a few unfortunately went on to divide by 3. A high proportion of candidates were also able to correctly deduce that the new standard deviation was twice the original one; a minority claimed (incorrectly) that there had been no change, whilst others multiplied the variance by 2 or the standard deviation by 2^2. Some candidates calculated the new values by finding the probability distribution of the total contribution, often arriving at correct answers, although wasting time on the process.

(iv) A wide variety of correct methods was seen here, including a complete enumeration of the three coin outcomes, which could then also be used in part v). Equally many incorrect attempts were seen, one common error being the use of 5 as denominator in all fractions.

(v) Again, numerous correct approaches were seen. Many were based on multiples of 22 or 44. Others used the enumeration of the new probability distribution, from which the expectation was calculated. Some credit was earned for an incorrect attempt at this, provided that at least three correct probabilities were obtained; however those candidates who thought that all probabilities were equal to 0.2 gained no credit.

Coursework: Statistics 2

The work from 329 centres was moderated by a team of 12. This compares with 314 this time last year. The work of 52 (16%) centres was recommended to change all but 1 in the downward direction.

The majority of centres assess their work carefully and accurately and complete all the administration with efficiency.

The major problems of an over generous allocation of marks have mostly been detailed before but:

- An unconvincing aim leads to a penalty in domain 1 but also makes it difficult for the candidate to score full marks in domain 5.
- Candidates must define their population clearly and address issues of sampling, including how they tried to achieve a representative sample in order to gain full marks in domain 2.
- Scatter diagrams should be clearly labelled.

- Modelling discussions should include a statement of whether the variables are random, mention the shape of the scatter diagram and state whether the assumption of an underlying bivariate normal distribution is a valid one. Several candidates happily calculate r without this necessary assumption being stated.
- Hypothesis tests should be stated formally and the alternative hypothesis should match the aim. There were many instances of incorrect notation and a lack of clarity in defining ρ here too.
- In some cases weak comment in the interpretation domain are being given too much credit as are basic comment of correlation only.
- Accuracy and refinements are still misunderstood by some candidates who refer, for example, to their inability to use a calculator.