RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2614/1

Statistics 2

Friday **17 JANUARY 2003** Afternoon 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.

Registered Charity 1066969

[Turn over

2 A pharmaceutical company has developed a new drug which is effective for 94% of patients. The drug is administered to various groups of patients.

For any group of patients, let X represent the number for whom the drug is effective and Y represent the number for whom the drug is not effective.

(i) In one week, 10 patients are given the drug. State the exact distribution of X. Hence calculate the probability that the drug is effective for at least 9 of the patients. [4]

(ii) In one month, 50 patients are given the drug. Use a Poisson approximation for Y to calculate the probability that the drug is not effective for 5 or fewer patients. [3]

(iii) In one year, 600 patients are given the drug. Use a suitable approximating distribution to

(A) calculate the probability that the drug is effective for at least 560 patients, [5]

(B) find the greatest value of k such that $\mathrm{P}(X \geqslant k)$ exceeds 99%. [3]

[Total 15]

3 Every day I check the number of emails on my computer at home. The numbers of emails, x, received per day for a random sample of 100 days are summarised by

$$\sum x = 184, \qquad \sum x^2 = 514.$$

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

(ii) Give two reasons why the Poisson distribution might be thought to be a suitable model for the number of emails received per day. [2]

(iii) Using the mean as found in part (i), calculate the expected number of days, in a period of 100 days, on which I will receive exactly 2 emails. [3]

On a working day, I also receive emails at the office. The number of emails received per day at the office follows a Poisson distribution with mean λ. On 1.5% of working days I receive no emails at the office.

(iv) Show that $\lambda = 4.2$, correct to 2 significant figures. Hence find the probability that on a working day I receive at least 5 emails at the office. [3]

(v) Find the probability that on a working day I receive a total of 10 emails (at home and at the office). [2]

(vi) Let Y be the total number of emails received at home and at the office over a period of 20 working days. Using a suitable approximating distribution, estimate values for a and b such that

$$\mathrm{P}(a \leqslant Y \leqslant b) = 0.95.$$ [4]

[Total 16]

[Turn over

1 For a random sample of 20 towns, a smoking index (x) and a cancer index (y) are constructed. These data are illustrated in Fig. 1. The associated summary statistics are also given below.

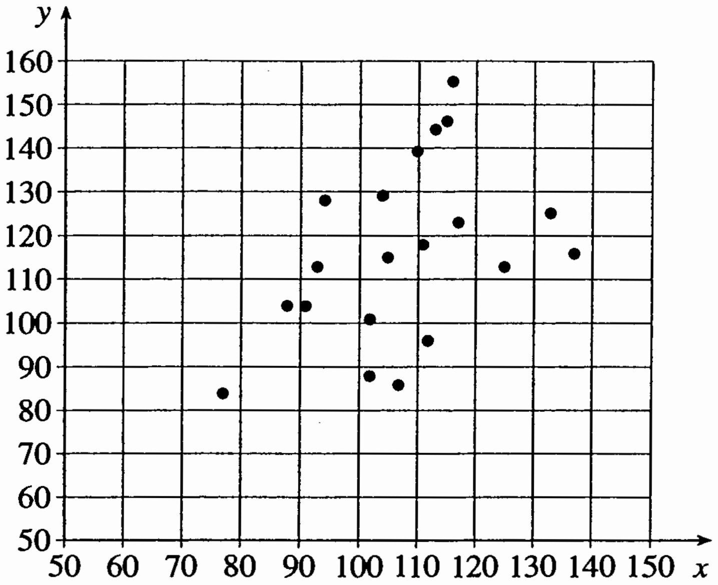


Fig. 1

$n = 20$ $\quad\sum x = 2152$ $\quad\sum y = 2327$

$\sum x^2 = 235\,724$ $\quad\sum y^2 = 278\,565$ $\quad\sum xy = 252\,811$

A community health officer claims that these data show that there is a positive connection between smoking and cancer.

(i) Show that the product moment correlation coefficient for the data is 0.425, correct to 3 significant figures. Carry out a suitable hypothesis test at the 5% significance level to check the officer's claim, stating your hypotheses and conclusion carefully. Comment on the validity of the test in relation to the scatter diagram. [10]

(ii) A spokesman for a tobacco firm claims that the data do not show that there is a connection between smoking and cancer. Discuss briefly whether or not his claim can be justified statistically. [2]

(iii) Explain the meaning of the term 'significance level', relating your answer to the test carried out in part **(i)**. [2]

[Total 14]

4 A company which hires out equipment by the day has three mowers. The number, X, of mowers which are hired on any one day has the following probability distribution.

$$\mathrm{P}(X = r) = k\left(\tfrac{1}{2}\right)^r \qquad \text{for } r = 0, 1, 2 \text{ and } 3.$$

(i) Show that $k = \frac{8}{15}$. [2]

(ii) Sketch the probability distribution of X. [2]

(iii) Calculate the expectation and variance of X. [4]

(iv) The income from hiring a mower is £25 per day. Deduce the mean and variance of the daily income from hiring out mowers. [2]

(v) Find the probability that, from Monday to Friday inclusive in one week, the total income from hiring mowers is £50. [5]

[Total 15]

Mark Scheme

Question 1

(i)	**EITHER:** $S_{xy} = \Sigma xy - n\overline{xy} = 252811 - 20 \times 107.6 \times 116.35 = 2425.8$ $S_{xx} = \Sigma x^2 - n\overline{x}^2 = 235724 - 20 \times 107.6^2 = 4168.8$ $S_{yy} = \Sigma y^2 - n\overline{y}^2 = 278565 - 20 \times 116.35^2 = 7818.55$	B1 for S_{xy} B1 for at least one of S_{xx} or S_{yy}	
	$r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{2425.8}{\sqrt{4168.8 \times 7818.55}} = 0.425$ (3 s.f.)	M1 for structure of r A1	
	OR:	**NB: ANSWER GIVEN**	
	$\text{Cov}(x,y) = \dfrac{\sum xy}{n} - \overline{xy} = 252811/20 - 107.6 \times 116.35 = 121.29$	B1 for Cov (x,y)	
	$\text{sd}(x) = \sqrt{\dfrac{\sum x^2}{n} - (\overline{x})^2} = \sqrt{(235724/20 - 107.6^2)} = \sqrt{208.44} = 14.437$ $\text{sd}(y) = \sqrt{\dfrac{\sum y^2}{n} - (\overline{y})^2} = \sqrt{(278565/20 - 116.35^2)} = \sqrt{390.93} = 19.772$	B1 for at least one sd or variance	
	$r = \dfrac{\text{Cov}(x,y)}{sd(x)sd(y)} = \dfrac{121.29}{14.437 \times 19.772} = 0.425$ (3 s.f.)	M1 for structure of r A1	**4**
	H_0: $\rho = 0$, H_1: $\rho > 0$ where ρ is the population correlation coefficient	B1 for hypotheses	
	For $n = 20$, 5% critical value = 0.3783	B1 for defining ρ B1 for critical value	
	Since $0.3783 < 0.425$ we reject H_0: There is sufficient evidence at the 5% significance level to suggest there is a positive correlation between the smoking index and the cancer index.	M1 for comparison A1FT for conclusion in words in context	
	Suitable comment on the shape of the scatter; *allow complete argument for or against appropriateness of the test depending on whether shape is thought to be roughly elliptical or not.*	E1 for explanation	**6**
(ii)	From tables for $n = 20$, the critical value at the 2.5% significance level is 0.4438, so you would come to the opposite conclusion. OR: cv at the 1% significance level is 0.5155 Hence the spokesman's claim could be justified statistically when the test is conducted at, say, 2.5% level or below. *Accept alternative sensible arguments relating to eg Type I error*	B1 for quoting a lower significance level and the relevant critical value B1 for comment explaining that at lower level the claim can be justified statistically.	**2**
(iii)	The significance level is the probability of rejecting the null hypothesis when it is in fact true. If indeed $\rho = 0$, then 5% of random bivariate samples of size 20 will produce an r value exceeding 0.3783.	E1 for definition of significance level E1 for comment related to the context	**2**
			14

Question 2

(i)	$X \sim \mathrm{B}(10, 0.94)$ P(drug effective on at least 9 patients) $= \mathrm{P}(X=9) + \mathrm{P}(X=10)$ $= 10 \times 0.94^9 \times 0.06 + 0.94^{10}$ $= 0.3438 + 0.5386 = 0.882$ (allow 0.88) *SC1 for use of tables and 1-P(X≤8)*	B1 for distribution in symbols or words M1 for the Binomial probability $\mathrm{P}(X=9)$ M1 for $\mathrm{P}(X=9) + 0.94^{10}$ A1 CAO	4
(ii)	Using $n = 50, p = 0.06$: $\lambda = 50 \times (1-0.94)$ $= 3$ P(not effective for 5 or fewer patients) = 0.9161 *from tables*	M1 for 50×(1-0.94) A1 B1FT for probability	3
(iii)	(*A*) Suitable approximating distribution: $X \sim \mathrm{N}(564, 33.84)$ Probability drug effective for at least 560 patients ≈ $\mathrm{P}(X > 559.5) = \mathrm{P}(Z > -0.7736)$ $= \mathrm{P}(Z < 0.7736)$ $= 0.7805$ (allow 0.780 to 0.781) *Alternative solutions which can gain the final A1:* *-omitted continuity correction* P(X>560) = 0.7542 NB (z = -0.6876), *-wrong continuity correction* P(X>560.5) = 0.7262 NB (z = -0.6017),	B1 for approx. dist. SOI but not if var =12×3 =36 B1 for cont. correction M1 for standardisation M1 for probability calculation including attempt at use of correct tables A1 **CAO** *NB M0M0 for use of variance*	5
	(*B*) $\mathrm{P}(Z > -2.326) = 0.99$ $\Rightarrow x = 564 - 2.326 \times 5.817 = 550.5$ hence required number is 550 (allow 551) *Allow answer with continuity correction, but it is not expected.* *Accept trial and improvement method only if supported by correct probabilities*	B1 for ±2.326 seen M1 for calculation based on a negative z value. *NB NOT (1 – z value)* *Allow use of variance for M1 if penalised in (A)* A1 CAO	3
			15

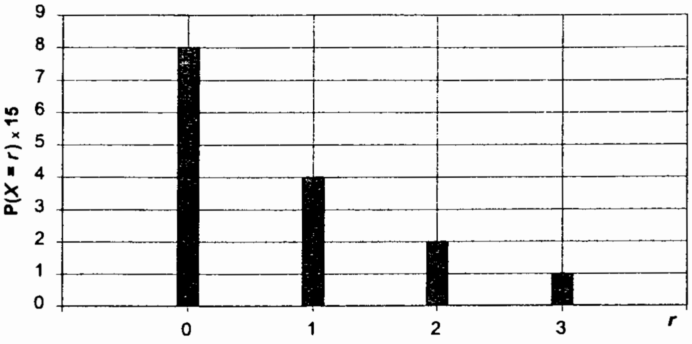
Question 3

(i)	Mean $= \frac{\Sigma x}{n} = \frac{184}{100} = 1.84$ Variance $= \frac{\Sigma x^2}{n} - \bar{x}^2 = \frac{514}{100} - 1.84^2 = 1.75$ (to 3 s.f.)	B1 for mean as fraction or decimal B1 for variance FT their mean	2
(ii)	Any two reasons why Poisson might be a suitable model, such as • Independence of arrival and random distribution through time • Uniform average rate of occurrence • Mean and variance approximately equal (*provided that this is the case in part (i)*) • Suitable "Large n and small p" argument – must be in context	E1 for one reason E1 for second reason	2
(iii)	Using $\lambda = 1.84$: $P(X=2) = e^{-1.84} \times \frac{1.84^2}{2} = 0.269$ (to 3 s.f.) Expected number of days = 26.9 (allow 27 days) *Allow correct interpolation from tables leading to P(X=2) = 0.2687*	M1 for Poisson probability P(X=2) A1 SOI - FT value of λ A1 for expected no. FT P(X=2) × 100	3
(iv)	$P(X=0) = e^{-\lambda} = 0.015 \Rightarrow \lambda = -\ln(0.015) \approx 4.2$ **OR:** $P(X=0) = e^{-4.2} \approx 0.015$ P(at least 5 emails at the office) $= 1 - 0.5898 = 0.410$ (to 3 s.f.)	B1 for finding λ OR for finding probability **NB: ANSWER GIVEN** M1 for use of tables to find 1 – P(X≤4) (*or 1 - sum of point probabilities)* A1	3

(v)	Using $\lambda = 6.04$ (i.e. 1.84 + 4.2):
	P(a total of 10 emails)
	$= e^{-6.04} \times \frac{6.04^{10}}{10!} = 0.0424$ (allow 0.042.)
		M1 for adding up their
		1.84 + 4.2
		A1 FT
			2

(vi)	$\lambda = 20 \times 6.04 = 120.8 \Rightarrow Y \sim N(120.8, 120.8)$ $a = 120.8 - 1.96 \times \sqrt{120.8} = 99.26 \quad (\approx 99)$ $b = 120.8 + 1.96 \times \sqrt{120.8} = 142.34 \quad (\approx 142)$	B1 for Normal approximation SOI (FT their 6.04) B1 for 1.96 seen M1 for an equation in a or b, with their μ, σ and suitable z-value A1 for both answers FT their μ, σ	4
			16

Question 4

(i)	$\begin{array}{|c|c|c|c|c|}\hline r & 0 & 1 & 2 & 3 \\ \hline P(X=r) & k & \frac{1}{2}k & \frac{1}{4}k & \frac{1}{8}k \\ \hline \end{array}$ $k\left(1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}\right) = 1 \Rightarrow \frac{15}{8}k = 1 \Rightarrow k = \frac{8}{15}$	M1 for forming equation with the sum of four probabilities A1 for solution **NB: ANSWER GIVEN**	**2**
(ii)		G1 for lines in proportion G1 (*dependent*) for both axes scaled	**2**
(iii)	$E(X) = \Sigma r\,P(X=r)$ $= 0 \times \frac{8}{15} + 1 \times \frac{4}{15} + 2 \times \frac{2}{15} + 3 \times \frac{1}{15} = \frac{11}{15} = 0.73$ $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$ $= 0 \times \frac{8}{15} + 1 \times \frac{4}{15} + 4 \times \frac{2}{15} + 9 \times \frac{1}{15} - \left(\frac{11}{15}\right)^2$ $= \frac{7}{5} - \frac{121}{225}$ $= \frac{194}{225} = 0.862$ (allow 0.86)	M1 for sum of four products A1 FT their probabilities provided that $\Sigma p = 1$ M1 for $E(X^2)$ A1 CAO	**4**
(iv)	Mean daily income $= 25 \times \frac{11}{15} =$ £18.33 *(at least 3 sig fig)* Variance of daily income $= 625 \times 0.862 =$ £^{2} 538.75 *(at least 3 sig fig)*	B1 for mean income (FT their $\frac{11}{15}$) B1 for variance (FT their 0.862)	**2**
(v)	Total income from hiring mowers = £50 if either (*A*) two on one day and none on the other days $P(A) = 5 \times \frac{2}{15} \times \left(\frac{8}{15}\right)^4 = 0.0539$ or (*B*) one on each of two days and none on the other days $P(B) = 10 \times \left(\frac{4}{15}\right)^2 \times \left(\frac{8}{15}\right)^3 = 0.1079$ Hence total probability $= 0.0539 + 0.1079$ $= 0.162$ (3 s.f.)	M1 for complete expression for P(*A*) M1 for $\left(\frac{4}{15}\right)^2 \times \left(\frac{8}{15}\right)^3$ M1 for 10 × fractional expression M1 for sum of two probabilities dep on at least one M1 A1 **CAO**	**5**
			15

Examiner’s Report

2614 Statistics 2

General Comments

The overall standard was similar to previous sessions, although very few candidates achieved close to full marks. Most candidates appeared to be well prepared and the paper allowed them to demonstrate their level of understanding and achievement. The earlier parts of Question 1 were well answered by almost all candidates, although good responses to the discussion and explanation required in parts (ii) and (iii) were rarely seen. Questions 2 and 3 were generally tackled well, other than the inverse normal calculations in Q2(iii) and 3(vi). Most candidates found Question 4 to be very accessible, with only part (v) causing much difficulty. There were sufficient opportunities throughout the paper for weaker candidates to gain a good deal of marks by the application of standard techniques, but equally there was enough demanding material in the latter parts of each question to test the most able candidates. On the whole, candidates' work was clear with sufficient detail of their working. Normal distribution calculations were often set out very well.

Comments on Individual Questions

Question 1

(i) Almost all candidates obtained the given value of 0.425 correctly; the formula based on covariance and standard deviations was more frequently seen than that based on sums of squares.

Candidates usually quoted suitable null and alternative hypotheses, but despite a comment in a previous examiner's report, very few made reference to the population in stating their hypotheses and so most lost an easy mark. The null and alternative hypotheses should be given in terms of ρ, and candidates must define ρ as the population correlation coefficient. Most candidates set out their test clearly, quoting the correct critical value and writing down an explicit comparison with the sample correlation coefficient. However a number of candidates did not interpret their conclusion in the context of smoking and cancer indices, thus losing a mark.

In making a comment on the validity of the test, it was expected that candidates would make a judgement on the ellipticity of the scatter diagram and thus draw a conclusion as to the validity, ideally mentioning that the data did or did not appear to come from a Bivariate-Normal parent population. Although suitable answers were often seen, many candidates spuriously based their answer on whether or not the points appeared to lie on a straight line.

(ii) Arguments such as 'correlation does not imply causation' were irrelevant here, since the question referred to a 'connection', not a causal link. Few candidates made a suitable comment related to the use of a lower significance level, and even amongst these few, it was rare to see this followed up by the quotation of the critical value for a lower significance level to justify their comment.

(iii) A clear definition of the term 'significance level' was required (ie 'the probability of rejecting H_0 when it is in fact true'). Many candidates discussed the probability of H_0 being true or of H_0 being false, which is not of course the meaning of the term 'significance level'. Others mentioned the 'accuracy' or 'reliability' of the test. Only a few candidates were able to give an adequate definition and even less were then able to relate their answer to the test in part (i).

(i) 0.425 (answer given), 0.3783, reject $H_{0;}$ (ii) comment; (iii) comment.

Question 2

(i) It was pleasing to see that most candidates were able to state the distribution as B(10,0.94) and to do so using standard notation. Many of these were able to find the correct Binomial probabilities, although there were attempts to use tables, usually based on p=0.95. Another fairly common error was 1 – P(X=8). Some candidates thought that a Poisson distribution was required, usually using tables. No credit was given in this case.

(ii) All but a few candidates found the value of $\lambda = 3$, but some then went on to calculate the point probability $P(X=5)$ rather than using tables to find $P(X \leq 5)$.

(iii)(a) Most candidates realised that a Normal approximating distribution was appropriate, but often the variance was quoted as 564 (np) rather than 33.84 (npq). Only a few attempted to use a Poisson approximation. The continuity correction was often omitted, but incorrect attempts at a continuity correction were rare. Weaker candidates often tried to standardise by dividing by the variance rather than the standard deviation – candidates should be reminded of the importance of avoiding this error. Many candidates were able to handle the evaluation of the area to the right of a negative z-value, but often the wrong tail was found, leading to a probability less than 0.5.

(iii)(b) Many candidates gained credit for using the inverse table to find $\Phi^{-1}(0.99) = 2.326$. However most did not realise that in fact the required z-value was –2.326 and so their value of k was greater than the mean! Many who did use a negative z-value failed to round their answer down to the next integer and thus lost the final mark.

(i) B(10,0.94), 0.882; (ii) 0.9161; (iii) A) 0.7805; B) 550.

Question 3

(i) Most candidates scored the full two marks here although some made errors with the variance, such as omitting the division by 100, or dividing by 100 after having subtracted the square of the mean.

(ii) Many correct responses were seen, although the suggestion of large n and small p was fairly popular, despite no indication of n and p in the question. When commenting on the rate of occurrence, candidates should discuss a uniform average rate, rather than just a uniform or constant rate. Some candidates referred to the randomness of the sample, rather than the randomness and independence of e-mail arrivals.

(iii) This part was answered very well, with many fully correct solutions seen, although a number of candidates forgot to multiply their probability by 100 to find the expected number of days.

(iv) Often candidates gave a very well presented solution using natural logarithms, or occasionally using tables or the given value of λ to derive a probability of 0.015. Relatively few candidates were unable to make an attempt at the justification. The calculation of $P(X > 5) = 1 - P(X < 4)$, was handled rather better than in previous sessions, although common errors were $P(X > 5) = 1 - P(X < 5)$ or simply the calculation of $P(X < 5)$.

(v) Many fully correct responses were seen, and it was pleasing to see that very few candidates lost the accuracy mark due to premature approximation of $\lambda = 6.04$ to $\lambda = 6$. Some candidates did not appreciate that they needed to add together the two means to find the overall mean rate.

(vi) Although some candidates did not appear to know how to begin this part, most realised that a Normal approximating distribution was required. There were a variety of incorrect values for the mean and variance, with several candidates forgetting to multiply 6.04 by 20, but many did quote the correct distribution N(120.8, 120.8) . Almost all candidates realised that a symmetrical interval was the preferred option (although credit was given for a non-symmetrical interval), but many used $z = 1.645$ rather than $z = 1.96$. Some candidates formed a single equation in a and b and were unable to proceed further. For those candidates who did use a correct method, a continuity correction was not required, but equally the use of one was not penalised.

(i) 1.84, 1.75; (ii) two reasons; (iii) 0.269, 26.9 (iv) 4.2 (answer given), 0.410
(v) 0.0424; (vi) a=99.3, b=142.3

Question 4

(i) This was answered well by most candidates, whether they formed an equation in k and solved this to find the value of k or whether they substituted the given value of k to show that the sum of the probabilities was 1. Occasionally candidates did not provide evidence that they were summing probabilities and thus forfeited the marks.

(ii) Most candidates drew a correct vertical line diagram. It is good to report that very few candidates now seem unaware that this is the correct diagram for a discrete probability distribution. However there are still some candidates who appear to think that the term 'sketch' in this context has the same meaning as in sketching a graph in pure mathematics. In statistics, vertical line diagrams should have linear scales, with lines whose length is in proportion to the probabilities. Some candidate did not take into account the value of k in their diagram.

(iii) Once again most candidates scored at least three marks here, with both E(X) and E(X^2) evaluated correctly. Some lost the fourth mark due to omission of subtraction of $E(X)^2$, or due to premature approximation of the value of E(X).

(iv) Calculation of the mean income was usually fully correct, and the required accuracy of 3 significant figures usually allowed enough leeway for those who prematurely approximated their answers to part (iii). However many candidates simply multiplied their variance from part (iii) by 25 rather than by the square of 25. A few candidates found the probability distribution of the daily income and then used this to evaluate the mean and variance, usually scoring at least one mark.

(v) Most candidates found this very difficult, with a wide variety of incorrect responses seen, including attempts at various Poisson and Normal probabilities. Up to two marks were available for those numerous candidates who thought that they were dealing with a Binomial Distribution and whose fractional probabilities added up to one. Only a small proportion of candidates were able to arrive at a fully correct solution.

(i) k=8/15 (answer given) (ii) graph (iii) 0.733, 0.862 (iv) £18.33, £538.89 (v) 0.162