OCR
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

Thursday **5 JUNE 2003** 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.

[Turn over

1 A medical screening programme predicts the remaining number of years (y) an individual aged x years has to live. The data for four individuals are shown in the table.

x	35	45	55	65
y	45.0	39.0	30.0	18.0

(i) Represent the data by a scatter diagram, drawn on graph paper. [2]

(ii) Calculate the equation of the regression line of y on x, and plot it on your scatter diagram. [5]

(iii) Use your regression equation to predict the remaining number of years of life of a person aged

(*A*) 65 years of age,

(*B*) 90 years of age,

commenting on your second prediction. [3]

(iv) Calculate the sum of the squares of the residuals. What is the relevance of this value with respect to the regression line? [5]

2 Every day Morse attempts the crossword puzzle in his newspaper. The time taken, X minutes, to complete the crossword may be modelled by a Normal distribution with mean 22 minutes and standard deviation 4.5 minutes.

(i) Calculate the probability that he takes

(*A*) more than 25 minutes,

(*B*) between 15 and 25 minutes,

to complete the crossword. [5]

(ii) What length of time would be enough for Morse to finish the crossword on 95% of days? [3]

Each day Morse takes a train to work. The journey takes 25 minutes. He starts his crossword at the beginning of his journey.

(iii) Find the probability that he completes the puzzle by the end of his journey at least twice in a five-day week. [4]

(iv) Morse changes his newspaper and finds that on 99% of occasions he completes the crossword during his morning train journey. Assuming that the time taken, Y minutes, to complete the crossword has the distribution $\mathrm{N}(18, \sigma^2)$, find the value of σ. [3]

3 A long-distance rail journey is scheduled to take 5 hours. Recent records show that trains completing this journey arrive late on 4% of occasions. Let X represent the number of times the train is late out of n journeys.

(i) State the distribution of X, giving an assumption you have to make for it to be valid. Under what conditions would a Poisson approximation be suitable? [4]

You are given that 27 such journeys occur per week.

(ii) Use a Poisson approximation to calculate the probability of exactly five late arrivals during the next 4 weeks. You are given that, using the exact distribution of X, $P(X = 5) = 0.1704$, correct to 4 significant figures. Calculate the percentage error in using the corresponding Poisson approximation. [4]

The remaining parts of this question refer to a period of 10 weeks. In these parts, use a suitable Poisson approximation.

(iii) Find the probability that between 8 and 12 (inclusive) journeys result in a late arrival. [3]

(iv) Find the smallest value of k such that $P(X \geqslant k)$ is less than 5%. How many late arrivals do you think the rail 'watchdog' would tolerate before a significant deterioration in punctuality was detected? Give reasons for your answers. [4]

TURN OVER for Q 4

4 Contestants taking part in a quiz show are asked at most five questions. A contestant who answers a question wrongly is asked no further questions. The five questions are of increasing difficulty.

For each individual question, the probability of a randomly chosen contestant answering correctly is given in Table 4.1.

Question number	1	2	3	4	5
Probability that this question is answered correctly	0.75	0.5	0.4	0.3	0.25

Table 4.1

Let X represent the number of questions that a randomly chosen contestant answers correctly. The probability distribution for X is given in Table 4.2.

r	0	1	2	3	4	5
$\mathrm{P}(X = r)$	0.25	0.375	0.225	0.105	0.033 75	0.011 25

Table 4.2

(i) Use information from Table 4.1 to show how $\mathrm{P}(X = 0)$ and $\mathrm{P}(X = 5)$ in Table 4.2 can be derived. [3]

(ii) Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. [4]

A prize of £(1000×2^r) is awarded to a contestant who answers r successive questions correctly. In one particular show there are 3 contestants.

(iii) Find the probability that

(*A*) the first contestant wins at least £8000,

(*B*) at least one of the three contestants leaves with the top prize. [4]

(iv) People who apply to take part in the quiz have to pay £1 to the quiz organisers. How many applications are required in order to provide the expected prize money to be given out in this show? [4]

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.

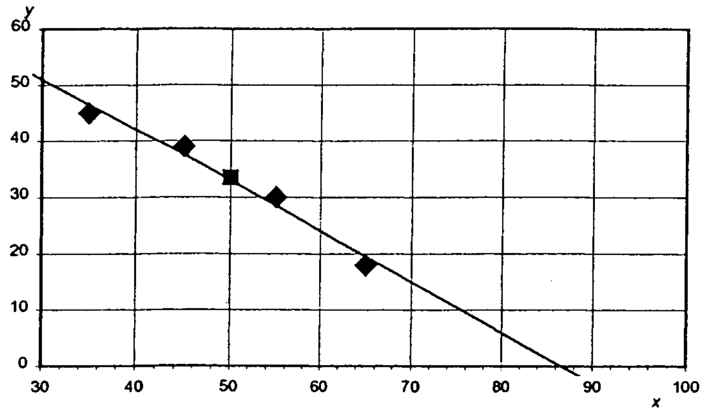
E marks ("explanation") are for explanation and/or interpretation. These will frequently be sub divisible depending on the thoroughness of the candidate's answer.

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
W	**Work worthy of credit but of no value**

Question 1

(i)		G1 for all points plotted G1 for linear scaled axes *dep on some points plotted*	**2**
(ii)	$S_{xy} = \Sigma xy - n\overline{xy} = 6150 - 4 \times 50 \times 33 = -450$ $S_{xx} = \Sigma x^2 - n\bar{x}^2 = 10500 - 4 \times 50^2 = 500$ $b = -450/500 = -0.9$ OR $b = \dfrac{6150/4 - 50 \times 33}{10500/4 - 50^2} = \dfrac{-112.5}{125} = -0.9$ $(y - \bar{y}) = b(x - \bar{x})$, where $\Rightarrow \quad y - 33 = -0.9(x - 50)$ $\Rightarrow \quad y = 78 - 0.9x$	M1 for attempt at b A1 CAO for b M1 for forming equation A1 for simplified equation (FT their b with correct $\bar{x}$ and $\bar{y}$) G1 for line on graph (FT their equation if b<0)	**5**
(iii)	(*A*) Predicted years to live for a 65-year old (let $x = 65$): $\hat{y} = 78 - 0.9 \times 65 = 19.5$ years (*B*) Predicted years to live for a 90-year old (let $x = 90$): $\hat{y} = 78 - 0.9 \times 90 = -3$ years Any valid relevant comment such as eg 'Prediction is meaningless since the person has negative years to live'	M1 for at least one prediction attempted A1 for both answers (FT their equation if b<0) NB no marks for reading predictions off the graph B1 FT for reason (allow sensible alternatives)	**3**
(iv)	Sum of the squares of the residuals: $(45 - 46.5)^2 + (39 - 37.5)^2 + (30 - 28.5)^2 + (18 - 19.5)^2$ $= 2.25 + 2.25 + 2.25 + 2.25 = 9$ This is the minimum value of the sum of the squares of residuals for any straight line (passing through $(\bar{x}, \bar{y})$).	M1 for attempt at one residual M1 for sum squares of four residuals A1 CAO E1 for minimum value E1 DEP for relating minimum values to lines	**5**
			15

Question 2

(i)	*(A)* $P(X > 25) = P\left(Z > \frac{25-22}{4.5}\right) = P(Z > \frac{2}{3})$ $= 1 - P(Z < \frac{2}{3}) = 1 - 0.7477 = 0.252$ (to 3 s.f.) *(B)* $P(15 < X < 25) = P(-1.556 < Z < 0.667)$ $= P(Z < 0.667) - P(Z < -1.556)$ $= 0.7477 - (1 - 0.9401) = 0.688$ (to 3 s. f.) NB Spurious use of CC with 25.5 and 14.5 leading to 0.218 in *A* and 0.734 in *B* can score M1M1A0M1A1 Use of √4.5 can score M0M1A0M1A1	M1 for standardising *[in either part]* M1 for probability A1 for awrt 0.252 M1 for probability A1 awrt 0.68 or 0.69, but FT their 0.252	**5**
(ii)	Require to find *k* such that $P(X < k) = 0.95$ Since $P(Z < 1.645) = 0.95$: $k = 22 + 1.645 \times 4.5$ $= 29.4$	B1 for ±1.645 M1 for equation for *k* based on positive z-value A1 FT z-value between 1.5 and 2.0.	**3**
(iii)	$P(X < 25) = 0.7477$ *[from part (i)]* Hence probability that puzzle completed by end of journey at least twice in a 5-day week is given by 1 – P(complete puzzles on time on 0 or 1 days) $= 1 - (0.2523^5 + 5 \times 0.7477 \times 0.2523^4)$ $= 0.984$ (to 3 s.f.)	B1 for 0.7477 seen FT 1 – ans to (i)A M1 for "$5 \times pq^4$" M1 for "$1 - (q^5 + 5pq^4)$" A1CAO NB Calculation of P(2)+P(3)+P(4)+P(5) scores M1 for any binomial coeff and M1 for fully correct expression	**4**
(iv)	$P(Z < 2.326) = 0.99$ $\Rightarrow \frac{25-18}{\sigma} = 2.326$ $\Rightarrow \sigma = \frac{25-18}{2.326} = 3.01$ minutes	B1 for ±2.326 M1 for equation in σ, using 25, (not y) and suitable positive z-value. A1CAO for awrt 3.0	**3**
			15

Question 3

(i)	Distribution of X: $X \sim B(n, 0.04)$ To be valid, assume independence of lateness of arrival of the trains. For a Poisson approximation to be appropriate, n must be large.	B1 for binomial B1DEP for parameters B1DEP on binomial for independence B1 for condition	4
(ii)	Using Poisson approximation with $\lambda = 108 \times 0.04 = 4.32$: $P(X=5) = e^{-4.32} \times \frac{4.32^5}{5!} = 0.167$ (3sf) Hence percentage error $= \frac{0.1704 - 0.1668}{0.1704} \times 100 = 2.1\%$	B1 CAO for λ M1 for calculation A1FT their $\lambda = n\,p$ NB Use of tables without interpolation scores M0A0 B1FT for percentage error	4
(iii)	Using Poisson approximation with $\lambda = 270 \times 0.04 = 10.8$: P(between 8 and 12 journeys inc. late) $= P(8 \le X \le 12) = P(X \le 12) - P(X \le 7)$ $= 0.7104 - 0.1566$ $= 0.554$ (accept answer between 0.55 and 0.56)	B1 FT $2.5 \times$ their 4.32 for λ M1 for probability calculation using tables and correct x values A1FT their λ NB Use of binomial or Normal scores M0A0	3
(iv)	Using tables with $\lambda = 10.8$: $P(X \ge 17) = 1 - P(X \le 16) = 1 - 0.9511 = 0.0489 < 5\%$ $P(X \ge 16) = 1 - P(X \le 15) = 1 - 0.9177 = 0.0823 > 5\%$, Hence smallest value of k is 17. "The rail watchdog would tolerate up to 16" oe If the "watchdog" found there to be, say, 17 or more late arrivals in a 10-week period, then the probability of this occurring with $p = 0.04$ is so small that a significant deterioration in punctuality could be detected.	M1 for at least one comparison A1FT their λ B1 FT their k E1DEP for "small probability" oe NB SC1 in place of B1 E1 for value above 16 with justification.	4
			15

Question 4

(i)	$P(X=0) = 1-0.75 = 0.25$ $P(X=5) = 0.75\times0.5\times0.4\times0.3\times0.25 = 0.01125$	B1 for "1 – 0.75" M1 for product of 5 probs A1 *(answer given)*	3
(ii)	$E(X) = 0\times0.25+1\times0.375+2\times0.225+3\times0.105+4\times0.03375+5\times0.01125$ $= 1.33125$ $E(X^2) = 0\times0.25+1\times0.375+4\times0.225+9\times0.105+16\times0.03375+25\times0.01125$ $= 3.04125$ $\mathrm{Var}(X) = E(X^2)-[E(X)]^2 = 3.04125-1.33125^2$ $= 1.27$ (3 s.f.)	M1 for $E(X)$ A1 CAO (to at least 3 s.f.) M1 for $E(X^2)$ A1 FT $3.04-(E(X))^2$ provided ans > 0	4
(iii)	(*A*) P(contestant wins at least £8000) $= 0.105+0.03375+0.01125 = 0.15$ OR $= 1-(0.25+0.375+0.225) = 0.15$ OR $= 0.75 = 0.75\times0.5\times0.4 = 0.15$ (*B*) P(at least one leaves with the top prize $= 1 -$ P(no one leaves with the top prize) $= 1-(1-0.01125)^3$ $= 0.0334$ (to 3 s.f.)	M1 for probability A1 CAO M1 for probability OR binomial $P(X=1)+P(X=2)+P(X=3)$ A1 CAO (to at least 2 s.f.) NB $3\times0.01125 = 0.03375$ scores M0A0 Also $P(X=1) = 0.03299$ scores M0A0	4
(iv)	Expected prize money for one contestant $= £[1000\times0.25+2000\times0.375+4000\times0.225+8000\times0.105+16000\times0.03375+32000\times0.01125]$ $= £3640$ Hence number of prospective contestants required $= 3\times3640 = 10920$	M1 for at least 3 products M1 for sum of all six products A1 CAO B1 FT expected prize money	4
			15

Examiner's Report

2614 Statistics 2

General Comments

Candidates generally performed well on this paper, with relatively few candidates scoring under 20 marks and a pleasing number over 45 marks. Almost all candidates attempted all of the questions, with very few appearing to be short of time. Both Question 1 on regression and Question 2 on the Normal distribution were answered particularly well, with only the section on residuals causing much difficulty. Candidates found Question 3 on the Poisson distribution to be more demanding, particularly part (iv), to which very few were able to produce convincing solutions. In Question 4 on discrete random variables, parts (i) and (ii) proved straightforward for most candidates but parts (iii) and (iv) were found to be much more difficult.

Comments on Individual Questions

Q.1 (i) Almost all candidates were able to gain both marks for plotting the graph. On this occasion, no penalty was applied to those candidates who used an axis break but omitted to indicate this on the axis. However, in future papers, such an omission may result in the loss of a mark.

(ii) Most candidates found the equation of the regression line successfully, using either one of the two formulae (variance and covariance or sums of squares) or the statistical functions on their calculator. Some candidates were careless in plotting their line on their graph, particularly those who had chosen an inappropriate scale or an axis break. A few drew a line of best fit on their graph and then found the equation of this line, which gained no credit.

(iii) Candidates were instructed to use their regression equation to make predictions and a few lost credit because they chose to use their graph instead. Most candidates made a sensible comment about the second prediction, although some invalid comments related to 90 years olds having exceeded their life expectancy.

(iv) Many candidates made a good attempt to find the residuals, whether from their equation or their graph. However some then simply added their residuals without squaring them, or did not appear to realise that the square of a negative number is positive. Candidates found it difficult to comment clearly on the relevance of the sum of the squares of the residuals. The requirement was for a comment to indicate that, amongst all possible straight lines, the regression line is that line which minimises the sum of the squares of the residuals. There were some centres from which no candidates appeared to have any knowledge of residuals.

(ii) $y=0.78-0.9x$; (iii)19.5, -3; (iv) 9.

Q.2 (i) Although most candidates knew how to tackle this question, scoring both method marks in part *(A)*, many did not gain the final answer mark due to premature rounding or failure to use the difference column. Some candidates used a standard deviation of $\sqrt{4.5}$, or less frequently divided by 4.5^2 to standardise. A common error was the application of a spurious continuity correction. Many candidates appear to think that a continuity correction is required whenever they use a Normal distribution. In part *(B)* many candidates were able to produce a fully correct solution, although there was some confusion with calculation of probabilities involving a negative z-value. No further penalty was applied to those who again rounded their z-value prior to use of tables or those who had used a consistent continuity correction throughout. It was pleasing to see that most candidates set out their Normal distribution calculations clearly, using Φ notation and centres are encouraged to continue to emphasize the importance of using this notation.

(ii) This part, which involved an inverse Normal calculation, was answered correctly by many candidates. Occasionally two-tailed 95% confidence intervals were given, a negative z-value was used or no z-value at all.

(iii) Most candidates realised that they needed to use a binomial distribution based on P(X<25) as evaluated in part (i) *(A)* . Many candidates then calculated the probability of exactly two puzzles completed rather than at least two completed. Some candidates omitted the binomial coefficients.

(iv) Candidates who knew how to begin this were often able to produce a fully correct solution. However, a surprising number failed to use a y-value of 25 and thus had two unknowns, y and σ in their equation; consequently they were unable to arrive at a solution. A few candidates found the correct answer, but apparently thought that this was the variance and so found the square root of their answer. Some candidates misread 2.326 from the tables as 0.2326 or 2.236.

(i)(A) 0.252, (B) 0.688; (ii) 29.4; (iii) 0.984; (iv) 3.01minutes.

Q.3 (i) Many candidates stated the correct distribution clearly and were able to quote the required assumption of independence. However, weaker candidates often did not specify the distribution, which resulted in the loss of three marks. A frequent error was to give the value of n as 5. In addition to independence a number of other 'requirements' were often quoted. The final requirement for a large value of n was stated by most candidates. However, many candidates appeared to write down every condition they could think of in the hope that one of them would be correct. Candidates are reminded that careful reading of the question and answers in full sentences are recommended.

(ii) Most candidates found the value of λ = 4.32, and used this to find the Poisson probability accurately. Occasionally credit was lost when λ was rounded and tables were used. Weaker candidates made a variety of errors, often based on a value of λ=1.08 followed by further incorrect working. The percentage error was found accurately by the majority of candidates, with only a few using an incorrect divisor of 0.1668.

(iii) Disappointingly few candidates were able to provide a correct solution to this part, even though all that was required was a simple Poisson probability calculation based on tables. Most candidates scored 1 mark for finding λ=10.8, but then the majority either found P(X≤12) – P(X≤8) rather than P(X≤12) – P(X≤7), or they used a Normal approximation, despite the clear instruction to use 'a suitable Poisson approximation'. Presumably such candidates interpreted this instruction as 'a suitable approximation to the Poisson distribution'.

(iv) Most candidates found this to be very demanding, with relatively few coming close to a correct solution. Once again a Normal distribution was very frequently used, rather than the correct Poisson distribution. Many who used the correct distribution either looked at the wrong tail or used the correct tail but made the error P(X≥k) = 1 – P(X≤k), rather than P(X≥k) = 1 – P(X≤k-1). Candidates who did find the correct value of k sometimes then offered a sensible response to the final part of the question. However many did not realise that if k=17 then the watchdog would tolerate a maximum of 16, (based on a 5% significance level). Candidates who gave answers based on other significance levels were able to achieve full credit provided that a clear explanation was offered.

(i) B(n, 0.04); (ii) 0.167,2.1%; (iii) 0.554; (iv) k = 17.

Q.4

(i) This was answered correctly by the vast majority of candidates.

(ii) It was pleasing to see that most candidates scored highly on these routine calculations, with both E(X) and E(X^2) evaluated correctly. Some lost the fourth mark due to omission of subtraction of E$(X)^2$. A few candidates used the probabilities given in Table 4.1 and a few made calculation errors.

(iii) In part (*A*) many candidates realised that r=3 was the crucial value, but then found P(X=3) rather than P(X≥3). Part (*B*) proved more difficult and whilst many candidates used the given probability of 0.01125, they often found the binomial P(X=1), or simply multiplied 0.01125 by 3, rather than the correct binomial P(X=0).

(iv) Fully correct answers were rarely seen. Many candidates scored the first method mark, for multiplication of correct prizes by probabilities, but they often omitted the smallest prize of £1000 when $r = 0$, presumably not realising that $2^0 = 1$. A common attempt was based on applying the given formula for prize money to the expectation found in part (ii); the logic appeared to be that $\mathrm{E}(1000 \times 2^X) = 1000 \times 2^{\mathrm{E}(X)}$, which is of course not true. Many candidates scored the final mark for multiplying their attempt at the expected prize money by 3 for the 3 contestants.

(i) (answer given); (ii) 1.33,1.27; (iii) (A) 0.15, (B) 0.0334; (iv) 10920.