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

Tuesday **28 MAY 2002** 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 approximate 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 3 printed pages and 1 blank page.

[Turn over

1 A science student took the temperature of a cup of coffee at one-minute intervals with the following results.

Time (x minutes)	0	1	2	3	4	5	6	7	8	9	10
Temperature (y °C)	96	84	73	64	55	49	44	40	36	33	31

$$n=11 \qquad \sum x=55 \qquad \sum x^2=385 \qquad \sum y=605 \qquad \sum xy=2326,$$

(i) The science student calculates the equation of the least squares regression line of y on x and uses it to predict the temperatures after 4.3 minutes and after 15 minutes. Obtain the regression line and the predictions. Show that *both* of these predictions are unsatisfactory. [8]

(ii) Plot a scatter diagram on graph paper to illustrate the data. Draw the regression line and the residuals on your diagram. By considering the pattern of the residuals, discuss whether there is a better way of modelling the relationship between temperature and time. [7]

2 The number of marks gained by candidates in a particular Statistics examination, for which the maximum mark is 60, is modelled by a Normal distribution with mean 36 and standard deviation 8. The marks are reported as integers.

(i) Find the probability that a randomly chosen candidate scores exactly 30 marks. [4]

(ii) Three candidates are chosen at random. Find the probability that just one of them gets fewer than 30 marks. [3]

(iii) It is intended that the proportion of candidates receiving a grade A should be as near as possible to 20%. What is the lowest integer mark that should be awarded a grade A? [4]

(iv) In a future Statistics examination it is intended that the top 25% of students should gain a reported mark of at least 45. Determine the required value for the mean mark, assuming the standard deviation remains at 8. [4]

3 The number of arrivals per minute at a drive-in fast food outlet is modelled by a Poisson distribution with mean λ. During a Saturday evening, $\lambda = 0.78$.

(i) Give one reason why the proposed Poisson distribution might be a suitable model. [1]

(ii) Calculate the probability of exactly two arrivals during a one-minute interval. [2]

(iii) Calculate the probability of at least four arrivals during a five-minute interval. [3]

(iv) Using a suitable approximating distribution, calculate the probability that there are more than 40 arrivals between 7 pm and 8 pm. [4]

Due to capacity constraints, the management of the fast food outlet want the probability of more than 40 arrivals in an hour to be 0.02. They ask a statistician to determine the value to which λ should be reduced.

(v) Show that λ must satisfy the equation $\frac{40.5 - 60\lambda}{\sqrt{60\lambda}} = 2.054$. [2]

(vi) Use the substitution $u = \sqrt{60\lambda}$ to formulate a quadratic equation in u. Solve this equation to show that the value of λ is just less than 0.5. [3]

4 Four children are playing a game. They each throw an ordinary fair die. Let L represent the largest score when the four dice are thrown.

You are given that the cumulative distribution function of L is

$$P(L \leqslant r) = \left(\frac{r}{6}\right)^4, \qquad r = 1, 2, 3, 4, 5, 6.$$

(i) Copy and complete the cumulative distribution of L.

r	1	2	3	4	5	6
$P(L \leqslant r)$	$\frac{1}{1296}$	$\frac{16}{1296}$				

[2]

(ii) Hence write down the probability distribution of L. [3]

(iii) Find the expected value of L. [2]

(iv) Use a probability argument to explain why $P(L \leqslant r) = \left(\frac{r}{6}\right)^4$. [3]

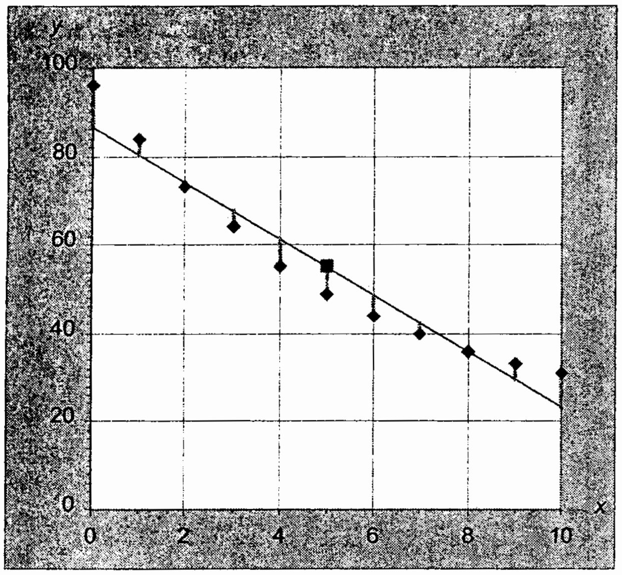
(v) Find the probability that just one child scores a six. [2]

(vi) Given that the largest score is a six, find the probability that just one child scores a six. [3]

Mark Scheme

May 2002 — 2614 MEI Statistics 2

Question 1

(i)	$\bar{x} = 5,\ \bar{y} = 55$ $b = \frac{2326 - 11\times5\times55}{385 - 11\times5^2} = \frac{-699}{110}$ *or* $\frac{2326/11 - 5\times55}{385/11 - 5^2} = \frac{-63.5}{10}$ hence least squares regression line is: $y - \bar{y} = b(x - \bar{x})$ $\Rightarrow\ y - 55 = -6.3545(x-5)$ $\Rightarrow\ y = 86.75 - 6.35x$ *or* $86.8 - 6.4x$	B1 for $\bar{x}$ and $\bar{y}$ used (SOI) M1 for attempt at gradient (b) M1 for equation of line A1 (explicit) **cao**	**4**
	$x = 4.3 \Rightarrow y = 86.75 - 6.35\times4.3 = 59.4$ $x = 15 \Rightarrow y = 86.75 - 6.35\times15 = -8.5$ First prediction is likely to be fairly inaccurate since, although it is within the data range, the fit does not look good. Second prediction is not only outside the data range, but implies that the coffee will freeze!	M1 for prediction(s) A1 for both answers E1 for first comment *allow suitable alternative* E1 for second comment *allow suitable alternative*	**4**
(ii)	 Line on scatter diagram Residuals on scatter diagram	G1 for scales *dependant on some plotted points* G1 for all points plotted G1 line through $(\bar{x}, \bar{y})$ G1 for correct gradient G1 for residuals	**5**
	Since the residuals are first negative, then positive and finally negative, a linear regression line does not seem to be appropriate. A better way might be to fit a curve.	E1 for description of pattern of residuals E1 for alternative	**2**
			15

Question 2

(i)	P(student scores exactly 30 marks) $= P(29.5 < X < 30.5)$ $= P\left(\frac{29.5-36}{8} < Z < \frac{30.5-36}{8}\right)$ $= P(-0.8125 < Z < -0.6875)$ $= (1 - 0.7542) - (1 - 0.7917)$ $= 0.2458 - 0.2083$ $= 0.0375$ (3 s.f.) *or* 0.038 (2 s.f.)	B1 for correct continuity corrections B1 for standardisation M1 for probability A1 **cao**	**4**
(ii)	P(just one gets fewer than 30 marks) $= 3 \times 0.2083 \times 0.7917^2$ $= 0.392$ (3 s.f.) *or* 0.39 (3 s.f.)	B1 for sight of their 0.2083 and 0.7917 M1 for binomial terms A1	**3**
(iii)	$P(Z > 0.8416) = 0.2$ $\Rightarrow$ corresponding value of X is given by $x = 36 + 8 \times 0.8416 = 42.7$ Hence lowest mark that should be awarded a grade A should be 43.	B1 for "± 0.8416" seen M1 for calculation of x A1 B1 **cao** for rounding as an integer	**4**
(iv)	Require $P(X > 44.5) = 0.25$ Now $P(Z > 0.6745) = 0.25$ With mean μ and standard deviation 8: $\frac{44.5 - \mu}{8} = 0.6745$ $\Rightarrow \mu = 44.5 - 8 \times 0.6745 = 39.1$ (3 s.f.) Hence target mean should be 39.1 marks	B1 for statement inc. c.c. B1 for "0.6745" M1 for attempt at setting up equation in μ A1 for solving equation	**4**
			15

Question 3

(i)	*E.g.* uniform (average) rate of occurrence; random arrivals through time *Allow* random *and* independent	B1 for suitable reason	**1**
(ii)	$\mathrm{P}(X=2) = e^{-0.78}\dfrac{0.78^2}{2!} = 0.139 \text{ (3 s.f.)} = 0.14 \text{ (2 s.f.)}$	M1 for probability calc. A1	**2**
(iii)	In a 5-minute interval mean no. arrivals $= 5 \times 0.78 = 3.9$ Using tables: $\mathrm{P}(X \geq 4) = 1 - \mathrm{P}(X \leq 3)$ $= 1 - 0.4532 = 0.5468$ $= 0.547 \text{ (3 s.f.)} = 0.55 \text{ (2 s.f.)}$	B1 for mean (SOI) M1 for probability A1	**3**
(iv)	Mean no. arrivals per hour $= 60 \times 0.78 = 46.8$ Using Normal approx. to the Poisson $X \sim \mathrm{N}(46.8, 46.8)$: $\mathrm{P}(X > 40.5) = \mathrm{P}\left(Z > \dfrac{40.5 - 46.8}{\sqrt{46.8}}\right)$ $= \mathrm{P}(Z > -0.9209) = \mathrm{P}(Z < 0.9209) = 0.821 \text{ (3 s.f.)}$	B1 for Normal approx. (SOI) B1 for continuity corr. M1 for probability A1 **cao**	**4**
(v)	Require λ so that $\mathrm{P}(X < 40.5) = 0.98$ $\Rightarrow \mathrm{P}\left(Z < \dfrac{40.5 - 60\lambda}{\sqrt{60\lambda}}\right) = 0.98$, but $\mathrm{P}(Z < 2.054) = 0.98$ hence $\dfrac{40.5 - 60\lambda}{\sqrt{60\lambda}} = 2.054$	B1 for statement with "2.054" and "0.98" B1 for using $\mu = 60\lambda$ and $\sigma = \sqrt{60\lambda}$ in formulation	**2**
(vi)	Using the substitution $u = \sqrt{60\lambda}$ results in the equation: $u^2 + 2.054u - 40.5 = 0$ $\Rightarrow u = \dfrac{-2.054 \pm \sqrt{2.054^2 - 4 \times 1 \times (-40.5)}}{2} = 5.419$ $\Rightarrow \lambda = \dfrac{5.419^2}{60} = 0.489 \text{ (3 s.f.)} = 0.49 \text{ (2 s.f.)}$	B1 for setting up quadratic equation M1 for attempt at solving equation in terms of u A1 **cao** for solution in terms of λ	**3**
			15

Question 4

(i)	<table><tr><td>r</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td></tr><tr><td>$P(L \le r)$</td><td>$\frac{1}{1296}$</td><td>$\frac{16}{1296}$</td><td>$\frac{81}{1296}$</td><td>$\frac{256}{1296}$</td><td>$\frac{625}{1296}$</td><td>1</td></tr></table>	B2 for all correct [−1 for 1 or 2 errors]	**2**
(ii)	<table><tr><td>r</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td></tr><tr><td>$P(L = r)$</td><td>$\frac{1}{1296}$</td><td>$\frac{15}{1296}$</td><td>$\frac{65}{1296}$</td><td>$\frac{175}{1296}$</td><td>$\frac{369}{1296}$</td><td>$\frac{671}{1296}$</td></tr></table>	M1 for probabilities A2 [−1 for 1 or 2 errors]	**3**
(iii)	$E(L) = \Sigma\, r\, P(L = r)$ $= 1 \times \frac{1}{1296} + 2\times\frac{15}{1296} + 3\times\frac{65}{1296} + 4\times\frac{175}{1296} + 5\times\frac{369}{1296} + 6\times\frac{671}{1296}$ $= 5.24 \text{ (3 s.f.)} = 5.2 \text{ (2 s.f.)} = \frac{6797}{1296} = 5\frac{317}{1296}$	M1 for expectation A1 [f.t. only if $\Sigma p_i = 1$]	**2**
(iv)	The probability of any child scoring $\le r = \frac{r}{6}$ If all 4 of the children score $\le r$, then the largest of the 4 scores is $\le r$. Hence (assuming independence): $P(\text{largest score} \le r) = \left(\frac{r}{6}\right)^4$	B1 for single probability B1 for "all 4 $\le r$" B1 for raising to power 4	**3**
(v)	P(just one child throws a six) $= 4\times\frac{1}{6}\times\left(\frac{5}{6}\right)^3 = \frac{500}{1296} = 0.386 \text{ (3 s.f.)} = 0.39 \text{ (2 s.f.)}$	M1 for binomial prob: $B(4, p) \to P(X = 1)$ A1 **cao**	**2**
(vi)	P(just one child throws a six | largest is 6) $= \dfrac{\frac{500}{1296}}{\frac{671}{1296}} = \frac{500}{671} = 0.745 \text{ (3 s.f.)}$	B1 for denominator M1 for ratio of probabilities A1 **cao**	**3**
			15

Examiner's Report

2614 Statistics 2

General Comments

The overall standard seemed rather better than in previous sessions with relatively few weak candidates. Most candidates appeared to be well prepared and were able to demonstrate their level of understanding and achievement. Question 1 was usually very well answered, noticeably more so than in questions on this topic in the past, with many candidates scoring 12 or more marks. The discussion and interpretation parts were tackled much more successfully than in many past examinations. Many good answers were also seen in Questions 2 and 3, although incorrect or omitted continuity corrections were common in the former and weaker candidates had problems with normal distribution calculations. Question 4 caused difficulties for many candidates, who often appeared unfamiliar with the cumulative distribution and thus lost credit. There were opportunities for all candidates to gain some easy marks in each question, although the latter parts of each question did prove testing for almost all candidates.

Comments on Individual Questions

Q.1 (i) Most candidates successfully applied the formula given in the formula book for the equation of the regression line, although a relatively small number made errors of one sort or another, often by square rooting the denominator of b. Very few attempted to calculate the equation from the original data and those who did rarely succeeded. Candidates are advised to use the given formula as it usually leads to a successful outcome.

Almost all candidates were able to calculate predictions based on their regression line and most were able to make a valid comment about the prediction at 4.3 minutes, although some felt erroneously that it was not acceptable to predict a temperature at a non-integral time value. Many candidates commented correctly on the unsuitability of the prediction at 15 minutes, which was due to the fact that the prediction was well below room temperature and would lead to the coffee freezing; however a common error was to suggest that this value was unsuitable because it was outside the range of the given data. Extrapolation has to be considered with caution, but it is does not necessarily lead to an unsatisfactory prediction – the context must be considered.

(ii) Virtually all candidates were able to draw the graph correctly and most plotted their regression line correctly; it was surprising that so many candidates who had made an error in part (i) and whose regression line was thus clearly invalid made no comment on this nor any apparent effort to find their error. Many candidates drew the residuals correctly, although a large minority appeared to have no knowledge of the concept.

In commenting on the modelling of the relationship, most candidates suggested that a curve, an exponential relationship or the use of logarithms would be more suitable. However many felt that the use of Spearman's coefficient of rank correlation would be more appropriate, failing to appreciate that this is simply a measure of correlation, and cannot be used to model a relationship. Only a few felt that the original model was appropriate. Very few candidates discussed the pattern of the residuals, or even mentioned them, in justifying their alternative model. Some candidates wasted time by measuring or calculating the residuals, which was not required.

(i) $y = 86.75 - 6.35x$; 59.4, −8.5, comments on unsuitability of estimates;
(ii) scatter diagram with regression line and residuals; comments on modelling assumption.

Q.2 (i) This part caused difficulties for many candidates. Almost all were able to standardise a value close to 30 and thus gain some credit, but relatively few realised that they should apply a continuity correction and find $P(29.5 < X < 30.5)$. A common error, which was marked generously, was to find $P(29 < X < 30)$ or $P(30 < X < 31)$.

(ii) Candidates were often more successful here than in part (i), with many scoring full marks. A number of candidates omitted a continuity correction despite having used one in part (i) and were thus penalised, usually reaching a final answer of 0.407; this penalty was not applied to those who had already lost credit in part (i) for failure to apply a continuity correction. Occasionally the binomial nature of the situation was not recognised or the binomial coefficient was omitted.

(iii) Most candidates answered this part correctly using inverse normal tables and rounding to 43. Some failed to use inverse tables, just using 0.8 or using 1 – 0.8416 rather than 0.8416.

(iv) Again many candidates were able to tackle this, although a continuity correction was very often omitted, or occasionally applied incorrectly, even by candidates who had previously used one correctly. The same errors with the z-value (in this case 0.6745) as in part (iii) were seen in a number of scripts.

(i) 0.0375; (ii) 0.392; (iii) least mark = 43; (iv) mean mark = 39.1.

Q.3 (i) It was hoped that candidates would state that the arrivals would occur randomly and independently over time, at a uniform rate and many included sufficient of this in their response to gain credit, although the uniform rate of occurrence was not often seen. Common errors included mention of rare events or large n and small p. This type of question is asked frequently and candidates should be aware of the conditions required.

(ii) This part was answered correctly by almost all candidates.

(iii) Most candidates realised that the new value of the mean λ was 3.9, but many were unable to correctly handle $P(X \geqslant 4) = 1 - P(X \leqslant 3)$, with a variety of errors here, the most common being $P(X \geqslant 4) = 1 - P(X \leqslant 4)$ or $P(X \geqslant 4) = 1 - P(X = 4)$. Very few candidates misread the tables, although a few did not use tables at all, thus doing unnecessary work although usually arriving at the correct answer. A significant number used $\lambda = 0.78$, sometimes then multiplying their answer by 5.

(iv) Candidates often recognised that a Normal approximation was needed, and most used the correct parameters. However the continuity correction was often omitted or applied in the wrong direction. Occasionally the variance was used in standardisation instead of the standard deviation. Having obtained the correct probability, many candidates then proceeded to subtract this from 1 and hence found the wrong tail.

(v) This proved demanding for many candidates who did not justify the value of 2.054, linking to $\Phi^{-1}(0.02)$ rather than $\Phi^{-1}(0.98)$. Justification of the values of 60λ as the new mean and variance was also often omitted. Explanations were often lacking in clarity.

(v) The substitution $u = \sqrt{60\lambda}$ caused difficulties for many candidates, who often did not realise that 60λ could be replaced by u^2. Others substituted $u = 60\lambda$, leading to an equation which they were unable to deal with. Nevertheless a good number of correct solutions was seen, although some candidates who obtained a correct equation were unable to solve it.

(i) reason; (ii) 0.139; (iii) 0.547; (iv) 0.821; (v) explanations; (v) 0.489.

Q.4 (i) Most candidates gave fully correct answers, although some adjusted $P(L \leq 6)$ so that their probabilities summed to 1, presumably not realising the cumulative nature of the distribution. Some converted their answers to decimals, rather than leaving them as fractions, which occasionally led to inaccuracy later on.

(ii) Candidates usually scored 3 marks or no marks in this part, with many leaving it out altogether.

(iii) The calculation of the expectation of a random variable was clearly familiar to most candidates, although many used cumulative probabilities and so gained no credit.

(iv) This proved to be challenging, with most candidates only able to justify the power of 4. A few explained the reason for the fraction $r/6$, but hardly any mentioned the essential idea that if all four children score at most r, then the largest of the four scores is at most r. Many candidates simply gave numerical examples for one or two values of r; this was not awarded any credit unless some generalisation to any value of r was clearly justified..

(v) Many binomial probabilities were seen, but the parameter p was usually wrong, often 671/1296 rather than the correct value of 1/6, leading to an answer of 0.2323. Occasionally the binomial coefficient ${}^4C_1$ was omitted.

(vi) This caused difficulties for the vast majority of candidates. Frequently the answer given was the same as that for part (v). Candidates who attempted to use the conditional probability formula often gave the product of two probabilities as the numerator and/or did not use their $P(L = 6)$ as the denominator. Only a few fully correct answers were seen.

(i) cumulative probabilities; (ii) probabilities; (iii) $E(X) = 5.24$;
(iv) explanations; (v) 0.386; (vi) 0.745.