RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2618

Statistics 6

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 any **three** 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 6 printed pages and 2 blank pages.

Option 1: Estimation

1 The random variable X has probability density function

$$f(x) = \frac{2x}{\theta} e^{-x^2/\theta} \qquad x > 0,$$

where θ is a parameter ($\theta > 0$). [X is said to have a Weibull distribution.] You are *given* the result that

$$E(X^{2r}) = \theta^r r! \qquad r = 1, 2, 3, \ldots$$

which you may quote without proof in parts **(ii)** and **(iii)** of this question.

$X_1, X_2, \ldots, X_n$ represent a random sample of n independent observations from this distribution.

(i) Show that the maximum likelihood estimator of θ is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$ [10]

(ii) Show that $\hat{\theta}$ is an unbiased estimator of θ. [2]

(iii) Show that

$$E(\hat{\theta}^2) = \frac{n+1}{n} \theta^2.$$

[You may *assume* the result that (for all $i, j = 1, 2, \ldots, n$) the independence of X_i and X_j implies that $E(X_i^2 X_j^2) = E(X_i^2)E(X_j^2)$.] [5]

(iv) Hence write down the variance of $\hat{\theta}$. [1]

(v) It may be shown that no unbiased estimator of θ has variance smaller than

$$-\frac{1}{E\left(\dfrac{d^2 \ln L}{d\theta^2}\right)}$$

where L is the likelihood function. Confirm that the variance of $\hat{\theta}$ attains this minimum. [2]

Option 2: Bivariate distributions

2 [Numerical answers in this question should be given as fractions in their lowest terms.]

T and U are discrete random variables whose joint distribution is given in the table.

		Values of U		
		1	2	3
Values of T	1	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{1}{20}$
	2	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{10}$
	3	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{20}$

(i) Find $\mathrm{E}(T)$ and $\mathrm{Var}(U)$. [6]

(ii) Find the three conditional probabilities that $T = 1$,

(A) given that $U = 1$,

(B) given that $U = 2$,

(C) given that $U = 3$. [4]

(iii) The random variables X and Y are defined by

$$X = \min(T, U) \quad \text{[i.e. } X = T \text{ when } T \leqslant U;\ X = U \text{ when } U < T\text{]},$$
$$Y = \max(T, U) \quad \text{[i.e. } Y = T \text{ when } T \geqslant U;\ Y = U \text{ when } U > T\text{]}.$$

Give a table showing the joint probabilities for X and Y. [4]

(iv) Obtain $\mathrm{Cov}(X, Y)$. [6]

Option 3: Markov chains

3 A man has 2 umbrellas which he uses in going to and from his office each working day. If it is raining when he leaves home in the morning, he takes an umbrella, provided there is one to be taken. Similarly, if it is raining when he leaves the office in the evening, he takes an umbrella, provided there is one to be taken. If it is not raining when he sets out on either journey, he does not take an umbrella. Independently of the weather at any other time, the probability that it is raining at the start of any journey is p. The number of umbrellas at home on the morning of day n (before setting out for the office) is denoted by X_n.

(i) Explain briefly why X_n is a Markov chain. [2]

(ii) Explain why $\mathrm{P}(X_{n+1} = 2 \mid X_n = 2) = p^2 + q$, where $q = 1 - p$. Find the other transition probabilities and write down the transition matrix of the Markov chain. [5]

(iii) On the morning of a particular working day, he has one umbrella at home. Use the transition matrix to show that the probability that there is one umbrella at home on the morning two working days later is

$$p^2q + (p^2 + q^2)^2 + p^2q^2$$

and to find the corresponding probabilities for there being no umbrellas and two umbrellas at home that morning. [4]

(iv) Find the long-run probabilities of there being no umbrellas, one umbrella or two umbrellas at home on the morning of a working day. [7]

(v) In the case $p = \frac{1}{4}$, what is the proportion of mornings on which he gets wet on leaving home? [2]

Option 4: Analysis of variance

4 **(i)** A one-way analysis of variance has k treatments and a total of n observations. State the number of degrees of freedom for each of the between-samples sum of squares, the within-samples sum of squares and the total sum of squares. [3]

(ii) State carefully the usual assumptions about the term representing experimental error in the customary model. [3]

(iii) The data below are the observations in a one-way analysis of variance situation.

Treatment 1	2.4	3.0	3.3
Treatment 2	5.2	3.8	4.0
Treatment 3	3.6	2.9	3.4
Treatment 4	2.8	4.2	
Treatment 5	4.9	4.3	

Draw up the usual analysis of variance table and carry out the test at the 5% level of significance. [9]

(iv) The cumulative distribution function of the F distribution used in the analysis of variance in part **(iii)** is

$$\mathrm{G}(y) = 1 - \frac{80y + 32}{(y+2)^5}.$$

Use this to verify the upper 5% point quoted in the tables and to find the exact level of significance of the data in part **(iii)**. [5]

Option 5: Regression

5 In the usual linear regression model, the random variable Y is related to the non-random variable x by

$$Y_i = \alpha + \beta x_i + e_i$$

where the "error" terms e_i are independent $N(0, \sigma^2)$ random variables. Given a set of n independent observations, the least squares estimators a and b of the parameters α and β respectively are

$$a = \overline{Y} - b\overline{x}$$

$$b = \frac{S_{xy}}{S_{xx}}$$

where

$$S_{xy} = \sum(x_i - \overline{x})(Y_i - \overline{Y}), \qquad S_{xx} = \sum(x_i - \overline{x})^2.$$

(i) By first writing $\sum(x_i - \overline{x})(Y_i - \overline{Y})$ as $\sum(x_i - \overline{x})Y_i - \sum(x_i - \overline{x})\overline{Y}$, show that

$$S_{xy} = \sum(x_i - \overline{x})Y_i.$$

Show similarly that

$$S_{xx} = \sum(x_i - \overline{x})x_i.$$ [3]

(ii) Show that b is an unbiased estimator of β and that its variance is $\dfrac{\sigma^2}{S_{xx}}$. [7]

(iii) The following data are for an agricultural experiment to investigate the effect of the concentration (%) of nitrogen in the fertilizer (x) on the yield (kg/plot) of the crop (Y).

x	5	7	9	11	13	15
Y	33	38	37	42	50	49

Plot the data [a simple diagram in your answer book will suffice; there is no need to use graph paper] and calculate the values of a and b for the linear regression model. Assuming that $\sigma^2 = 5$, provide a 95% confidence interval for the slope of the line. Explain briefly how this confidence interval would be obtained if σ^2 were not known. [10]

Mark Scheme

Marking Instructions

Some marks in the mark scheme are explicitly designated as ‘M’, ‘A’, ‘B’ or ‘E’.

‘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. 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-dividable depending on the thoroughness of the candidate’s answer.

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.

All queries about the mark scheme should have been resolved at the standardisation meeting. Assistant Examiners should telephone the Principal Examiner (or Team Leader if appropriate) if further queries arise during the marking.

Assistant Examiners may find it helpful to use shorthand symbols as follows:

FT	Follow-through marking
✓^	Correct work after error
✗^	Incorrect work after error
C	Condonation of a minor slip
BOD	Benefit of doubt
NOS	Not on scheme (to be used *sparingly*)
I	Work of no value

Q1		$\mathrm{f}(x)=\frac{2x}{\theta}\mathrm{e}^{-x^2/\theta}\quad x>0$ [Weibull] Given: $\mathrm{E}[X^{2r}]=\theta^r.r!\quad r=1,2,\ldots$		
	(i)	$\mathrm{L}=\frac{2x_1}{\theta}e^{-\frac{x_1^2}{\theta}}.\frac{2x_2}{\theta}e^{-\frac{x_2^2}{\theta}}\cdot\ldots\cdot\frac{2x_n}{\theta}e^{-\frac{x_n^2}{\theta}}$	1	
		$=\frac{2^n}{\theta^n}x_1x_2\ldots x_n\mathrm{e}^{-\frac{1}{\theta}\Sigma x_i^2}$ might be implicit in sequel	1	
		$\ell\mathrm{nL}=\mathrm{const}-n\ell\mathrm{n}\theta-\frac{1}{\theta}\Sigma x_i^2$	M1, 1	
		$\frac{\mathrm{d}\ell\mathrm{nL}}{\mathrm{d}\theta}=-\frac{n}{\theta}+\frac{\Sigma x_i^2}{\theta^2}$	M1, 1	
		$=0$ for max (etc)	M1	
		$\Rightarrow\hat{\theta}=\frac{1}{n}\Sigma x_i^2$ $\left[\text{or via }\frac{\mathrm{dL}}{\mathrm{d}\theta}\right]$	1	
		Need to show $\frac{\mathrm{d}^2\ell\mathrm{nL}}{\mathrm{d}\theta^2}<0$	M1	
		$\frac{\mathrm{d}^2\ell\mathrm{nL}}{\mathrm{d}\theta^2}=\frac{n}{\theta^2}-\frac{2\Sigma x_i^2}{\theta^3}$ which, at $\hat{\theta}$, equals $\frac{n^3}{(\Sigma x_i^2)^2}-\frac{2\Sigma x_i^2 n^3}{(\Sigma x_i^2)^3}<0$	1	10
	(ii)	$\mathrm{E}[\hat{\theta}]=\frac{1}{n}\Sigma\mathrm{E}[X_i^2]=\frac{1}{n}\Sigma\theta$ using given result with $r=1$	1	
		$=\theta\ \therefore$ unbiased	1	2
	(iii)	$\mathrm{E}[\hat{\theta}^2]=\mathrm{E}[\{\frac{1}{n}(X_1^2+X_2^2+\ldots+X_n^2)\}^2]$		
		$=\frac{1}{n^2}\{\mathrm{E}[X_1^4]+\mathrm{E}[X_2^4]+\ldots+\mathrm{E}[X_n^4]$ and $\mathrm{E}[X_i^4]=2\theta^2$ by given result	1	
		$+\mathrm{E}[X_1^2X_2^2]+\ldots\}$	1	
		↳ $=\mathrm{E}[X_1^2]\mathrm{E}[X_2^2]$ by result given in question	1	
		$=\theta.\theta$		
		There are $n(n-1)$ such terms	1	
		$=\frac{1}{n^2}\left\{2n\theta^2+n(n-1)\theta^2\right\}=\frac{n+1}{n}\theta^2$ beware printed answer	1	5
	(iv)	$\mathrm{Var}(\hat{\theta})=\mathrm{E}[\hat{\theta}^2]-(\mathrm{E}[\hat{\theta}])^2=\frac{\theta^2}{n}$	1	1
	(v)	We have $\mathrm{E}\left[\frac{\mathrm{d}^2\ell\mathrm{nL}}{\mathrm{d}\theta^2}\right]=\frac{n}{\theta^2}-\frac{2}{\theta^3}.n\theta=-\frac{n}{\theta^2}$	1	
		$\therefore-\frac{1}{\mathrm{E}\left[\frac{\mathrm{d}^2\ell\mathrm{nL}}{\mathrm{d}\theta^2}\right]}=\frac{\theta^2}{n}=\mathrm{Var}(\hat{\theta})$	1	2

2	(i)	Joint distribution of T (rows) and U (columns), with marginals (table below)	FT throughout, but A0 for negative variance or for prob dist for which $\Sigma \neq 1$. Accept fractions not in lowest terms, but **DEDUCT 1 FROM TOTAL** if this has been done.		
		For marginals For T margin For U margin		M1 A1 A1	
		$E(T) = 1 \times \frac{1}{5} + 2 \times \frac{2}{5} + 3 \times \frac{2}{5} = \frac{11}{5}$		A1	
		$E(U) = 1 \times \frac{2}{5} + 2 \times \frac{2}{5} + 3 \times \frac{1}{5} = \frac{9}{5}$			
		$E(U^2) = 1^2 \times \frac{2}{5} + 2^2 \times \frac{2}{5} + 3^2 \times \frac{1}{5} = \frac{19}{5}$			
		$\therefore \text{Var}(U) = \frac{19}{5} - \left(\frac{9}{5}\right)^2 = \frac{19}{5} - \frac{81}{25} = \frac{14}{25}$	for attempt at Var (U)	M1 A1	6
	(ii)	$P(T=1 \mid U=1) = \frac{\frac{1}{20}}{\frac{2}{5}} = \frac{1}{8}$	award once	M1	
		$P(T=1 \mid U=2) = \frac{\frac{1}{10}}{\frac{2}{5}} = \frac{1}{4}$ $P(T=1 \mid U=3) = \frac{\frac{1}{20}}{\frac{1}{5}} = \frac{1}{4}$		A1 A1 A1	4
	(iii)	$X = \min(T, U)$ $Y = \max(T, U)$			
		Joint distribution of X (rows) and Y (columns), with marginals (table below)	Joint probs Deduct 1 per error	A4	4
	(iv)	$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$		M1	
		$E(X) = \frac{3}{2}$		A1	
		$E(Y) = \frac{5}{2}$		A1	
		$E(XY) = \frac{1.1.1}{20} + \frac{1.2.4}{20} + \frac{1.3.6}{20} + 0 + \frac{2.2.4}{20} + \frac{2.3.4}{20} + 0 + 0 + \frac{3.3.1}{20}$		M1	
		$= \frac{1+8+18+16+24+9}{20} = \frac{76}{20} = \frac{19}{5}$		A1	
		$\therefore \text{Cov}(X, Y) = \frac{19}{5} - \frac{3}{2} \cdot \frac{5}{2} = \frac{76-75}{20} = \frac{1}{20}$		A1	6

Table for (i):

		U = 1	2	3	
T	1	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{5}$
	2	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{2}{5}$
	3	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{2}{5}$
		$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	

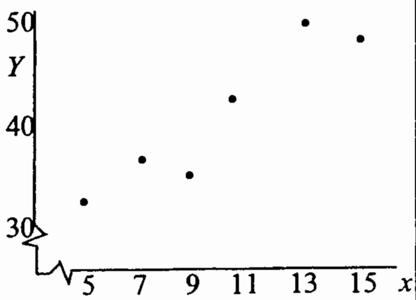
Table for (iii):

		Y = 1	2	3	
X	1	$\frac{1}{20}$	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{11}{20}$
	2	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$
	3	0	0	$\frac{1}{20}$	$\frac{1}{20}$
		$\frac{1}{20}$	$\frac{2}{5}$	$\frac{11}{20}$	

3	(i)	X_n depends only on X_{n-1} and not directly on anything previous to X_{n-1}	E2	2
	(ii)	$P(X_{n+1} = 2 \mid X_n = 2) = p^2 + q \, . \, 1$ takes one in morning, brings one back in evening (p^2) ; does not take one in morning – so none at office in evening anyway (q . 1)	E1 E1	
		Other transition probabilities as in matrix:- $$\begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 1.q & 1.p & 0 \\ pq & p^2+q^2 & qp \\ 0 & pq & p^2+q \end{bmatrix} \end{array}$$ Deduct one per error, but FT – but no A marks for any nonsense answers	A3	5
	(iii)	$p_0 = [0 \;\; 1 \;\; 0]$ $p_1 = p_0\mathbf{P} = [pq \;\; p^2+q^2 \;\; qp]$ $$p_2 = p_1\mathbf{P} = \begin{bmatrix} pq^2 + \left(p^2+q^2\right)pq \\ p^2q + \left(p^2+q^2\right)^2 + p^2q^2 \\ \left(p^2+q^2\right)qp + qp\left(p^2+q\right) \end{bmatrix}$$ or equivalent; or via $p_0\,\mathbf{P}^2$	M1 A1 A1 A1	4
	(iv)	$\pi = \pi\,\mathbf{P}$ with $\pi_0 + \pi_1 + \pi_2 = 1$	M2 M1	
		$$\left.\begin{array}{ll} \pi_0 q + \pi_1 pq & = \pi_0 \\ \pi_0 p + \pi_1\left(p^2+q^2\right) + \pi_2 pq & = \pi_1 \\ \pi_1 qp + \pi_2\left(p^2+q\right) & = \pi_2 \end{array}\right\}$$ Attempt to solve	M1	
		e.g.: from [1]: $\pi_1 \not{p} q = \pi_0 \not{p}$ $\quad \therefore \pi_1 = \frac{\pi_0}{q}$ from [3]: $\pi_1 qp = \pi_2(p - p^2)$ $\quad \therefore \pi_2 = \pi_1$ Now substitute in $\pi_0 + \pi_1 + \pi_2 = 1$		
		Answers are $\pi_0 = \frac{1-p}{3-p}$ $\quad \pi_1 = \frac{1}{3-p}$ $\quad \pi_2 = \frac{1}{3-p}$	A1 A1 A1	7
	(v)	$p = \frac{1}{4}$ gives $\pi_0 = \frac{3}{11}$	1	
		Proportion of mornings when he gets wet $= p\pi_0 = \frac{1}{4} \cdot \frac{3}{11} = \frac{3}{44}$	1	2

4	(i)	$k-1 \quad n-k \quad n-1$	1, 1, 1	3
	(ii)	Ind N (0, σ^2 [constant]) allow 'uncorrelated' for 'ind N' 1 1 1		3
	(iii)	2.4 3.0 3.3 | 8.7 5.2 3.8 4.0 | 13.0 3.6 2.9 3.4 | 9.9 2.8 4.2 | 7.0 4.9 4.3 | 9.2 $\Sigma\Sigma x_{ij} = 47.8 \quad \Sigma\Sigma x_{ij}^2 = 184.04$ 'correction factor' $= \frac{47.8^2}{13} = 175.76$ total SS = 184.04 − 175.76 = 8.28 $SS_B = \frac{8.7^2}{3} + \ldots + \frac{9.2^2}{2} - 175.76$ = 25.23 + 56.33 + 32.67 + 24.5 + 42.32 − 175.76 $= 181.05(\dot{3}) - 175.76 = 5.29$ [note calculation is ill-conditioned]		
		(table below)		
		Refer to $F_{4,8}$ — No FT if wrong	1	
		Upper 5% point 3.84 — No FT if wrong	1	
		Not significant	1	
		Seems treatments are all the same	1	9
	(iv)	cdf [of $F_{4,8}$, but ignore if candidate has df wrong] is $G(y) = 1 - \frac{80y+32}{(y+2)^5}$		
		$G(3.84) = 1 - \frac{339.2}{6793.041} = 0.95(01)$ ie 3.84 is upper 5% pt	M1 E1	
		$G(3.538(46)) = 1 - \frac{315.077}{5211.296} = 1 - 0.06(05) = 0.93(95)$	M1 A1	
		∴ level of significance of the data = 0.06(05) [6(.05)%]	1	5

	M1	1	M1	M1 A1
Source of variation	SS	df	MS	MS ratio
Between treatments	5.29	4	1.3225	3.538(46)
Residual	2.99	8	0.37375	
Total	8.28	12		

5	(i)	We have $S_{xy} = \overbrace{\Sigma(x_i - \bar{x})Y_i - \bar{Y}\,\underbrace{\Sigma(x_i - \bar{x})}_{=0}}^{\text{M1}}$ 1	M1 1	
		Similarly, $S_{xx} = \Sigma(x_i - \bar{x})x_i - \bar{x}\,\underbrace{\Sigma(x_i - \bar{x})}_{=0}$ 1	1	3
	(ii)	$\mathrm{E}(b) = \dfrac{\Sigma(x_i - \bar{x})\mathrm{E}(Y_i)}{S_{xx}}$	M1	
		$= \dfrac{\Sigma(x_i - \bar{x})(\alpha + \beta x_i)}{S_{xx}}$	M1	
		$= \dfrac{\alpha}{S_{xx}}.0$	1	
		$+ \dfrac{\beta\Sigma(x_i - \bar{x})x_i}{S_{xx}} = \beta$ (cancelled)	1	
		$\mathrm{Var}(b) = \dfrac{1}{S_{xx}^{\ 2}}\Sigma(x_i - \bar{x})^2\,\mathrm{Var}(Y_i)$	M1	
		$= \dfrac{S_{xx}}{S_{xx}^{\ 2}}\sigma^2$	1	
		$= \dfrac{\sigma^2}{S_{xx}}$	1	7
	(iii)	x: 5, 7, 9, 11, 13, 15 Y: 33, 38, 37, 42, 50, 49 $\bar{x} = 10 \quad \bar{Y} = \frac{249}{6} = 41.5$ $S_{xx} = 70$ $S_{xy}\left(= 2611 - \frac{60 \times 249}{6}\right) = 121$ 	G2	
		$b = \frac{121}{70} = 1.728(57)$	A1	
		$a = 41.5 - b \times 10 = 24.214(29)$	A1	
		95% CI for β is given by $b \pm 1.96\sqrt{\dfrac{\sigma^2}{S_{xx}}}$		
		ie by 1.728(57)	M1	
		$\pm 1.96\sqrt{\frac{5}{10}}$	B1 M1	
		= 1.728(57) ± 0.523(83)		
		= (1.205, 2.252)	A1	
		If σ^2 is unknown, estimate it using the residual mean square and use t_{n-2} (t_4 here, dt 5% pt is 2.776) instead of N(0, 1)	E2	10

Examiner’s Report

2618 Statistics 6

General Comments

There were only 8 candidates, from a total of 5 centres. This report has perforce to be written in very general terms so that accidental identification of individual candidates is avoided. Most candidates attempted questions 1 (estimation), 2 (bivariate) and 3 (Markov chain); there were a few attempts on question 4 (analysis of variance) but none on question 5 (regression). Much of the work was of a quite good standard.

In question 1, the work was generally well done. In part (iii), some candidates appeared not to realise the obvious point that X_i^2 cannot be independent of itself, so $\mathrm{E}(X_i^4)$ is not equal to $\{\mathrm{E}(X_i^2)\}^2$; instead, it is immediately available from the result quoted near the head of the question. In part (v), some candidates wrongly inserted the expression for the maximum likelihood estimator into the second derivative of the log-likelihood before attempting to find the expected value. In question 2, apart from slips, the mistake of consequence was that some candidates found the marginal probabilities of the X and Y variables and then wrongly thought that their joint probabilities were obtained by multiplication as if they were independent. Question 3 did not produce such good work as has been common in recent years. There were some very good solutions, but some candidates had difficulties forming the transition matrix and there were also difficulties in using it correctly in parts (iii) and (iv). Question 4 was also a little disappointing in some cases, with strange errors in calculations and in using the F tables.

Please see the published mark schemes for the algebraic and numerical answers and details of solutions.