OCR RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2618

Statistics 6

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 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

$$f(x) = \frac{2x}{\theta^2}, \qquad 0 < x \leqslant \theta,$$

where θ is a parameter ($\theta > 0$).

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

(i) Write down the likelihood function $L(\theta)$. [1]

(ii) Deduce that $L(\theta)$ is maximised by taking θ to be as small as possible. [3]

(iii) Deduce further that the maximum likelihood estimator of θ is $\hat{\theta} = X_{max}$ where X_{max} represents the largest of $X_1, X_2, \ldots, X_n$. [3]

(iv) You are *given* the result

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

Use this to find the constant k such that $\tilde{\theta} = k\hat{\theta}$ is an unbiased estimator of θ. [3]

(v) You are *given* the further result

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

Use this to find $\text{Var}(\tilde{\theta})$. [2]

(vi) Obtain the mean of X. Deduce that another plausible estimator of θ is $\theta^* = \frac{3}{2}\overline{X}$ where $\overline{X}$ represents the mean of $X_1, X_2, \ldots, X_n$. Use the result $\text{Var}(X) = \frac{1}{18}\theta^2$ to find $\text{Var}(\theta^*)$. [5]

(vii) Deduce that $\tilde{\theta}$ is a better estimator of θ than θ^*, for each fixed $n > 1$. [3]

Option 2: Bivariate distributions

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

X and Y are discrete random variables whose joint and marginal distributions are given in the table.

		Values of Y			Marginal dist. of X
		0	1	2	
Values of X	0	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{5}{12}$
	1	$\frac{1}{6}$	$\frac{1}{12}$	0	$\frac{1}{4}$
	2	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{6}$	$\frac{1}{3}$
Marginal dist. of Y		$\frac{13}{24}$	$\frac{1}{4}$	$\frac{5}{24}$	

You are *given* that $\mathrm{E}(X)=\frac{11}{12}$ and $\mathrm{Var}(X)=\frac{107}{144}$.

(i) Find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$. [5]

(ii) Find $\mathrm{E}(XY)$ and hence show that $\mathrm{Cov}(X, Y)=\frac{2}{9}$. [5]

(iii) Use the results above to find $\mathrm{E}(2X + 3Y)$ and $\mathrm{Var}(2X + 3Y)$. [4]

(iv) Construct a table showing the possible values $2X + 3Y$ can take and their respective probabilities. *Hence* find the mean and variance of $2X + 3Y$. [6]

Option 3: Markov chains

3 In a certain football league, it is a rule that a player who is sent off during a match is automatically suspended for the next 2 matches.

A particular player is always selected to play if he is not suspended but has probability p of being sent off during any match, independently for all matches.

A Markov chain model for this player's availability has three "states" for any match: that he is playing in the match (though he might get sent off during it), that he is serving the first match of a 2-match suspension, and that he is serving the second match of a 2-match suspension.

(i) Write down the transition matrix $\mathbf{P}$ of the Markov chain. [3]

(ii) Find $\mathbf{P}^4$ and hence find the probability for each "state" for this player on the fourth match after one in which he has been playing. [5]

(iii) Find the long-run probabilities for each "state" for this player. [4]

The first recurrence time, T, to any "state" is defined as follows. Suppose the player is in that "state"; then T is the number of transitions needed for the player to *first* return to the "state".

(iv) Find the possible values of T for the playing "state" and their respective probabilities. Hence find the mean recurrence time, $\mathrm{E}(T)$, for this "state". [3]

(v) Repeat part **(iv)** for the "state" of being in the first match of a 2-match suspension. Verify that the mean recurrence time in each case is the reciprocal of the corresponding long-run probability. [5]

Option 4: Analysis of variance

4 (i) Interpret the parameters μ and α_i in the one-way analysis of variance model expressed in the usual notation as $x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. [2]

(ii) A government agency is comparing the performances of high-technology companies in five regions of the country. In each region, a random sample of four such companies is taken and their financial results are carefully scrutinised. An index of performance is calculated for each company (higher values indicate better performances). The results are as follows.

	Region				
	1	2	3	4	5
	77	81	71	93	56
	107	73	27	84	50
	119	60	101	58	31
	93	106	88	66	64
Totals	396	320	287	301	201

[The sum of these data items is 1505 and the sum of their squares is 124747.]

Draw up the usual analysis of variance table and report your conclusions, using a 5% signficance level. [9]

(iii) A manager suggests that the index value of 27 in region 3 might be an error. If this index is omitted from the analysis of variance in part **(ii)**, the sum of squares between regions is found to be 5114.99 (with 4 degrees of freedom) and the residual sum of squares is found to be 3930.17 (with 14 degrees of freedom). Draw up the analysis of variance table for this revised situation, showing that the result is highly significant. Construct a table of means for the regions for the revised situation and report briefly on the conclusions from the analysis. [5]

(iv) Discuss carefully how you would report to the manager in respect of the index value of 27 and its influence on the conclusions. [4]

[Turn over

Option 5: Regression

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

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + \delta x_i^3 + e_i.$$

A set of n independent observations is available.

(i) State the usual assumptions about the "error" terms e_i. [2]

(ii) Use the method of least squares to obtain the normal equations for the parameter estimators $\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}$. [4]

In a particular case with $n = 5$, the data are as follows.

x	−2	−1	0	1	2
y	−2	4	−1	−6	1

(iii) Solve the normal equations to obtain the values of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\delta}$. [6]

(iv) Show that two of the normal equations, expressed in terms of the underlying random variables Y_i, have the forms

$$\sum Y_i - 5\hat{\alpha} - 10\hat{\gamma} = 0,$$

$$\sum x_i^2 Y_i - 10\hat{\alpha} - 34\hat{\gamma} = 0.$$

Hence obtain the expression

$$\hat{\alpha} = \tfrac{17}{35}\sum Y_i - \tfrac{1}{7}\sum x_i^2 Y_i.$$

Inserting the values $x_1 = -2, x_2 = -1, x_3 = 0, x_4 = 1, x_5 = 2$ in this expression, show that

$$\hat{\alpha} = -\tfrac{3}{35}Y_1 + \tfrac{12}{35}Y_2 + \tfrac{17}{35}Y_3 + \tfrac{12}{35}Y_4 - \tfrac{3}{35}Y_5$$

and hence that

$$\text{Var}(\hat{\alpha}) = \tfrac{17}{35}\sigma^2.$$

where σ^2 is the variance of the error terms. [5]

(v) Taking $\sigma^2 = 2$, test the hypothesis $\alpha = 0$ at the 5% level of significance. [3]

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 (‘explanatioin’) are for explanation and/or interpretation. These will frequently be subdividable 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.

The following is a list of frequently used shorthand symbols:

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*)
⌶	Work of no value

Q.1 $f(x) = \frac{2x}{\theta^2} \quad 0 < x < \theta$

(i) $L = \frac{2x_1}{\theta^2} \cdot \frac{2x_2}{\theta^2} \cdot \cdots \cdot \frac{2x_n}{\theta^2} \left(= \frac{2^n}{\theta^{2n}} x_1 x_2 \cdots x_n \right)$ 1 **1**

(ii) L is of form $\frac{\text{constant}}{\theta^{2n}}$ which is maximised by making θ as small as possible. E2 **3**

M1 might be implicit

(iii) We have $0 < x < \theta$, i.e. $\theta > x$ – so θ is > each of $x_1, x_2, \ldots, x_n$ – so the smallest θ can possibly be is x_{max}. E2 M1 might be implicit **3**

(iv) $\left[\text{Given result} : E(\hat{\theta}) = \frac{2n\theta}{2n+1} \right]$

We want $E[k\hat{\theta}] = \theta$ M1

$E[k\hat{\theta}] = \frac{2nk\theta}{2n+1}$ 1 $\therefore k = \frac{2n+1}{2n}$ 1 $\left[\text{i.e.} \tilde{\theta} = \frac{2n+1}{2n} \hat{\theta} \right]$ **3**

(v) $\left[\text{Given result} : \text{Var}(\hat{\theta}) = \frac{n\theta^2}{(n+1)(2n+1)^2} . \right]$

$\text{Var}(\tilde{\theta}) = \left(\frac{2n+1}{2n} \right)^2 \text{Var}(\hat{\theta})$ M1 $= \frac{(2n+1)^2}{(2n)^2} \cdot \frac{n\theta^2}{(n+1)(2n+1)^2} = \frac{\theta^2}{4n(n+1)}$ 1 **2**

(vi) $E(x) = \int_0^\theta \frac{2x^2}{\theta^2} dx = \frac{2}{\theta^2} \cdot \frac{\theta^3}{3} = \frac{2\theta}{3}$ 1

Reasonable to estimate E(*x*) by $\bar{x}$, so reasonable to estimate θ by $\theta^* = \frac{3}{2}\bar{x}$. E2

We are given $\text{Var}(x) = \frac{\theta^2}{18}$, so $\text{Var}(\theta^*) = \frac{9}{4} \cdot \frac{\theta^2}{18n} = \frac{\theta^2}{8n}$ **5**

1 1

(vii) Compare variances. M1

$\text{Var}(\tilde{\theta}) \quad \text{Var}(\theta^*)$

$\frac{\theta^2}{4n(n+1)} < \frac{\theta^2}{8n}$ for $n + 1 > 2$ i.e. for all $n > 1$ 1

$\therefore \tilde{\theta}$ is better E1 **3**

Q.2

		Y 0	1	2	
	0	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{5}{12}$
X	1	$\frac{1}{6}$	$\frac{1}{12}$	0	$\frac{1}{4}$
	2	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{6}$	$\frac{1}{3}$
		$\frac{13}{24}$	$\frac{1}{4}$	$\frac{5}{24}$	

$\left.\begin{array}{l} \mathrm{E}(x)=\frac{11}{12} \\ \mathrm{Var}(x)=\frac{107}{144} \end{array}\right\}$ given

FT throughout, but AO for negative variance or for probability distribution for which $\Sigma \neq 1$.
Accept fractions not in lowest terms, but **DEDUCT 1 FROM TOTAL** if this has been done.

(i) $\mathrm{E}(Y)=0\times\frac{13}{24}+1\times\frac{1}{4}+2\times\frac{5}{24}$ M1 here or elsewhere = $\frac{2}{3}$ A1

$\mathrm{E}(Y^2)=0^2\times\frac{13}{24}+1^2\times\frac{1}{4}+2^2\times\frac{5}{24}=\frac{13}{12}$ A1

$\therefore$ Var(Y) = $\frac{13}{12}-\left(\frac{2}{3}\right)^2=\frac{39-16}{36}=\frac{23}{36}$ M1 A1 **5**

(ii) E(XY) = 0 + 0 + 0 M1 bivariate expectation
+ 0 + $1.1.\frac{1}{12}$ + 0 A1 correct values
+ 0 + $2.1.\frac{1}{24}$ + $2.2.\frac{1}{6}$

$=\frac{1}{12}+\frac{2}{24}+\frac{4}{6}=\frac{5}{6}$ A1

$\therefore$ Cov(X, Y) = E(XY) – E(X)E(Y) M1 $=\frac{5}{6}-\frac{11}{12}.\frac{2}{3}=\frac{2}{9}$ A1 Beware printed answer **5**

(iii) E(2X + 3Y) = 2E(X) + 3E(Y) M1 $=2.\frac{11}{12}+3.\frac{2}{3}=3\frac{5}{6}$ A1

Var(2X + 3Y) = 4 Var(X) + 9 Var(Y) + 2 . 2 . 3 Cov (X, Y) M1

$= 4.\frac{107}{144}+9.\frac{23}{36}+12.\frac{2}{9}=\frac{107+207+96}{36}=\frac{410}{36}=\frac{205}{18}$ A1 **4**

(iv)

2X + 3Y:	0	2	3	4	5	6	7	[8]	10	←values 1
probs:	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{24}$	[0]	$\frac{1}{6}$	←M1 if method appears correct
										A1 if all correct

Hence E(2X + 3Y) = $0\times\frac{1}{4}+\ldots+10\times\frac{1}{6}=\frac{92}{24}=\frac{23}{6}$ A1

$\mathrm{E}[(2X+3Y)^2]=\ldots=\frac{626}{24}=\frac{313}{12}$ A1 $\therefore \mathrm{Var}=\frac{313}{12}-\left(\frac{23}{6}\right)^2=\frac{939-529}{36}=\frac{205}{18}$ A1 **6**

Q.3 (i)

	Play	Suspend 1	Suspend 2	
Play →	$1-p$	p	0	1 zero entries
Suspended 1	0	0	1	1 $1-p$ and p
Suspended 2	1	0	0	1 unit entries

3

(ii) $$p^2 = \begin{bmatrix}(1-p)^2 & p(1-p) & p \\ 1 & 0 & 0 \\ 1-p & p & 0\end{bmatrix} \qquad P^4 = \begin{bmatrix}(1-p)^4 + 2p(1-p) & p(1-p)^3 + p^2 & p(1-p)^2 \\ (1-p)^2 & p(1-p) & p \\ (1-p)^3 + p & p(1-p)^2 & p(1-p)\end{bmatrix}$$

M1 A1 If any 2 rows right
A1 if 3rd row right also

Want $[1 \quad 0 \quad 0]p^4$
$= [(1-p)^4 + 2p(1-p) \quad p(1-p)^3 + p^2 \quad p(1-p)^2]$
M1 A1 ← FT from candidate's p^4 unless obvious nonsense **5**

(iii) $\pi = \pi P$ M1 with $\Sigma\pi_i = 1$ M1

$$\left.\begin{aligned}(1-p)\pi_1 \qquad\qquad + \pi_3 &= \pi_1 \\ p\pi_1 &= \pi_2 \\ \pi_2 &= \pi_3\end{aligned}\right\} \therefore \pi_1 + p\pi_1 + p\pi_1 = 1 \rightarrow \pi_1 = \tfrac{1}{1+2p} \quad 1$$

$\pi_2 = \pi_3 = \frac{p}{1+2p}$ 1 **4**

(iv) For 'playing':

T	1	[2]	3	←1
Prob	$1-p$	[0]	p	←1

$\therefore E(T) = 1(1-p) + 3p$
$= 1 + 2p$ 1 **3**

(v) For 'suspended I':

T	[1]	[2]	3	4	5	… ←1
Prob	[0]	[0]	p	$(1-p)p$	$(1-p)^2p$	… ←1

$\therefore$ E(T) $= 3p + 4(1-p)p + 5p(1-p)^2 + \ldots$

$= p\left\{\frac{3}{1-(1-p)} + \frac{1-p}{1-(1-p)} + \frac{(1-p)^2}{1-(1-p)} + \ldots\right\}$ or other methods M1

$= 3 + \frac{p}{p}\cdot\frac{1-p}{1-(1-p)} = 3 + \frac{1-p}{p} = \frac{2p+1}{p}$ 1

For each case, E(T) $= \frac{1}{\pi_i}$ 1 **5**

Q.4 (i) μ is the population grand mean for the whole experiment 1

α_i is the population mean amount by which the i'th treatment differs from μ 1

must be clear reference to <u>population</u>

2

(ii) Totals are 396 320 287 301 201 (each from sample of size 4).

Grand total 1505 "correction factor" CF = $\frac{1505^2}{20}$ = 113251.25

Total SS = 124747 − CF = 11495.75

Between regions SS = $\frac{396^2}{4} + \ldots + \frac{201^2}{4}$ − CF = 118146.75 − CF = 4895.5

Residual SS (by subtraction) = 11495.75 − 4895.5 = 6600.25

Source of variation	SS M1	df M1	MS M1	MS ratio M1 A1
Between regions	4895.5	4	1223.875	2.78(14)
Residual	6600.25	15	$440.01\dot{6}$	
Total	11495.75	19		

Refer to $F_{4,\,15}$ 1 – 5% point is 3.06 1 – not significant 1

– seems performances do not differ between regions 1 **9**

(iii) [NOTE The new B really is 5114.99 to 2 dp – it is **not** (nor does it round to) 5115.]

We have now

5114.99	4	1278.75	4.55(5)	A1
3930.17	14	280.73		
9045.16				

Refer to $F_{4,\,14}$ – highly significant (5% pt 3.11, $2\frac{1}{2}$ % pt 3.89, 1% pt 5.04) 1

Means are	1	2	3	4	5	B1
	99	80	$86.\dot{6}$	75.25	50.25	

Seems regions differ 1

looks as though performance in region 5 is distinctly lower than elsewhere E1 **5**

(iv) The 27 does seem suspiciously low in context (e.g. transcription error for 77 – or even 127??)

Should try to check whether it is correct.

Conclusions are quite crucially dependent on it.

If no good reason to omit (or revise) it, arguably one should report both analyses of variance.

E4 for these or other valid points. **4**

Q.5 $Y = \alpha + \beta x + \gamma x^2 + \delta x^3 + e$

x	–2	–1	0	1	2
y	–2	4	–1	–6	1

(i) $e_i \sim \text{ind } N(0, \sigma^2)$ 2 [1 if only 1 part missing or wrong. Allow 'uncorrelated' for 'ind N'] **2**

(ii) $\Omega = \Sigma e_i^2 = \Sigma(Y_i - \alpha - \beta x_i - \gamma x_i^2 - \delta x_i^3)^2$ M1 M1 consider $\frac{\partial\Omega}{\partial\alpha}$ etc = 0 M1

$$\left.\begin{aligned}\frac{\partial\Omega}{\partial\alpha} &= -2\Sigma\left(Y_i - \alpha - \beta x_i - \gamma x_i^2 - \delta x_i^3\right) = 0\\ \frac{\partial\Omega}{\partial\beta} &= -2\Sigma x_i\,(---) = 0\\ \frac{\partial\Omega}{\partial\gamma} &= -2\Sigma x_i^2\,(---) = 0\\ \frac{\partial\Omega}{\partial\delta} &= -2\Sigma x_i^3\,(---) = 0\end{aligned}\right\}$$

$\therefore$ equations are

$\Sigma(Y_i - \alpha - \beta x_i - \gamma x_i^2 - \delta x_i^3) = 0$

$\Sigma x_i(Y_i - \alpha - \beta x_i - \gamma x_i^2 - \delta x_i^3) = 0$

$\Sigma x_i^2(Y_i - \alpha - \beta x_i - \gamma x_i^2 - \delta x_i^3) = 0$

$\Sigma x_i^3(Y_i - \alpha - \beta x_i - \gamma x_i^2 - \delta x_i^3) = 0$

[or equivalent] 1 **4**

(iii) We have $\Sigma x = 0$ $\Sigma x^2 = 10$ $\Sigma x^3 = 0$ $\Sigma x^4 = 34$ $\Sigma x^5 = 0$ $\Sigma x^6 = 130$

$\Sigma y = -4$ $\Sigma xy = -4$ $\Sigma x^2 y = -6$ $\Sigma x^3 y = 14$ ($n = 5$)

Equations become
$$\left.\begin{aligned}-4 - 5\alpha \quad - 10\gamma \quad &= 0\\ -4 \quad - 10\beta \quad - 34\delta &= 0\\ -6 - 10\alpha \quad - 34\gamma \quad &= 0\\ 14 \quad - 34\beta \quad - 130\delta &= 0\end{aligned}\right\}$$ A2 FT any errors

$$\left.\begin{aligned}&(1)\text{ and }(3)\text{ give }\hat\alpha = -\tfrac{38}{35}(=-1.0857)\text{ and }\hat\gamma = \tfrac{1}{7}(=0.142857)\\ &(2)\text{ and }(4)\text{ give }\hat\beta = -\tfrac{83}{12}(=-6.91\dot6)\text{ and }\hat\delta = \tfrac{23}{12}(=1.91\dot6)\end{aligned}\right\}$$ A1 A1 A1 A1 **6**

(iv) (1) and (3) are $\left.\begin{aligned}\Sigma Y_i - 5\hat\alpha - 10\hat\gamma = 0\\ \Sigma x_i^2 Y_i - 10\hat\alpha - 34\hat\gamma = 0\end{aligned}\right\} 1 \rightarrow \left.\begin{aligned}2\Sigma Y_i - 10\hat\alpha - 20\hat\gamma = 0\\ \Sigma x_i^2 Y_i - 10\hat\alpha - 34\hat\gamma = 0\end{aligned}\right\} \rightarrow \hat\gamma = \frac{\Sigma x^2 Y - 2\Sigma Y}{14}$ 1

$\therefore 5\hat\alpha = \Sigma Y - \frac{10}{14}\left(\Sigma x^2 Y - 2\Sigma Y\right) \rightarrow \hat\alpha = \frac{1}{5}\Sigma Y - \frac{1}{7}\Sigma x^2 Y + \frac{2}{7}\Sigma Y = \frac{17}{35}\Sigma Y - \frac{1}{7}\Sigma x^2 Y$ 1

Thus $\hat\alpha = Y_1\left(\frac{17}{35} - \frac{4}{7}\right) + Y_2\left(\frac{17}{35} - \frac{1}{7}\right) + Y_3\left(\frac{17}{35} - 0\right) + Y_4\left(\frac{17}{35} - \frac{1}{7}\right) + Y_5\left(\frac{17}{35} - \frac{4}{7}\right)$

$= -\frac{3}{35}Y_1 + \frac{12}{35}Y_2 + \frac{17}{35}Y_3 + \frac{12}{35}Y_4 - \frac{3}{35}Y_5$ 1

$\therefore \text{Var}(\hat\alpha) = \left(\frac{3}{35}\right)^2\sigma^2 + \ldots = \frac{9+144+289+144+9[=595]}{35^2[=1225]}\sigma^2 = \frac{17}{35}\sigma^2$ 1

Beware printed answer **5**

(v) Test statistic is $\frac{-\frac{38}{35} - 0}{\sqrt{\frac{17}{35}\times 2}} = -1.10(16)$ M1

Refer to N(0, 1) – d.t. 5% pt is 1.96 1; not significant, can accept $\alpha = 0$ 1 **3**

Examiner’s Report

2618 Statistics 6

General Comments

There were 12 candidates from 4 centres for this, the last sitting of this module. It is not being offered in 2005 in the "legacy" specification, while the structure of the new specification militates against proceeding as far as a sixth module in any of the applied strands. Some of the content, however, will be incorporated into other modules. This report has perforce to be written in fairly general terms so that accidental identification of individual candidates is avoided. It is pleasing to be able to start by saying that much of the work was of a quite good standard.

Comments on Individual Questions

Q.1 Surprisingly, some candidates were not quite sure how to write down the likelihood here. Others were a little unconfident in deducing from the form of the likelihood that it is maximised by taking θ to be as small as possible (which is not quite the same as saying that $\theta \to 0$, for we find in part (iii) that it can't necessarily get very close to zero). Those who embarked on a calculus path presumably realised fairly soon that it was not going to get anywhere useful. The deduction in part (iii), that because $x \le \theta$ it necessarily follows that θ can be no smaller than the largest value of X, was sometimes drawn very securely, but other candidates had some further difficulty here. The next two parts, concerned with means and variances, were usually well done. In part (vi), the straightforward instruction to "obtain the mean of X" was strangely misunderstood by some candidates, though most did the simple piece of calculus very readily. The deduction of "another plausible estimator" that followed from this was usually satisfactorily done, though most candidates worked on the basis that the given estimator is unbiased; indeed it is, but this was not quite the point being made (though it was readily allowed for full marks), but rather the usual "method of moments" idea was being looked for. Finally, in the last part there was generally a realisation that variances had to be compared, though it was not always explained why, and mostly this was correctly done, following through in some cases from earlier mistakes.

Q.2 This was a popular question and generally very well done.

Q.3 Candidates were able to write down the transition matrix readily enough, but raising it to the 4th power often caused problems. The matrix included four entries of zero and also two of 1, so there was not very much algebra to do; it is sad that some candidates could not do it. Having found (correctly or otherwise) $\mathbf{P}^2$, there was a remarkable reluctance to obtain $\mathbf{P}^4$ by simply multiplying $\mathbf{P}^2$ by itself; many candidates wasted time and effort by finding $\mathbf{P}^3$ and then multiplying by $\mathbf{P}$ yet again to get $\mathbf{P}^4$. Nearly all candidates, however, knew how to use their matrices to find the four-step probabilities and the equilibrium distribution. The last two parts of the question moved on to introduce the idea of the first recurrence time. The majority of candidates worked through this carefully and successfully, but there were some who went astray.

Q.4 The opening part of this question, on interpretation of parameters in the usual model, sought some formality and completeness, for example in being clear that *population* means were being referred to. Not all candidates were sufficiently careful about this; there was some sloppiness of expression. The test itself was usually carried out correctly. It is good to see that this method is now well known and that nearly all candidates use the "squared totals" computing formulae which are convenient and efficient for hand calculation. The third part of the question called for a repeat analysis (with most values provided in the question) with a suspicious observation omitted. Candidates here were reluctant to do what the question explicitly stated, which was (after setting out the basic analysis of variance) to construct a table of the "treatment" means for the new situation and report briefly on the conclusions from the analysis. The means now would have shown region 5 very much less than the others (which themselves remain quite variable), which might well be the reason why the basic test result now is highly significant whereas previously, with all the data included, it was not significant. Finally, candidates were asked to discuss how to report to a manager in respect of the suspicious observation and its influence on the analysis. Most discussions here were reasonably sensible.

Q.5 For the last few years, this has been a very unpopular topic, so it is pleasing to report that in this last sitting a quarter of the (admittedly few) candidates attempted it – and usually quite well. The assumptions about the error term in the model were usually known, and the normal equations could be set up and solved, including obtaining the algebraic expression for the estimator of the parameter α. However, the variance of this estimator caused some problems. Candidates usually had some idea how to proceed to the test at the very end of the question, but not always completely successfully.