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

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

Option 1: Estimation

1 The random variable X has the exponential distribution with probability density function

$$f(x) = \lambda e^{-\lambda x}$$

for $x > 0$, where λ is a parameter ($\lambda > 0$). $X_1, X_2, \ldots, X_n$ represent a sample of n independent observations from this distribution.

(i) Find $\hat{\lambda}$, the maximum likelihood estimator of λ. [11]

(ii) Let L denote the likelihood function, and suppose that n is large. You are given that the approximate distribution of $\hat{\lambda}$ is

$$\hat{\lambda} \sim \mathrm{N}\left(\lambda, \left[-\frac{d^2(\ln L)}{d\lambda^2}\right]^{-1}\right).$$

Find the variance of this distribution. Hence, estimating λ by $\hat{\lambda}$ in the variance, obtain an expression for an approximate 95% confidence interval for λ. [7]

(iii) A random sample of size 30 has $\bar{x} = 3.2$. Evaluate $\hat{\lambda}$, and find the limits of the approximate 95% confidence interval. [2]

Option 2: Bivariate distributions

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

X and Y are discrete random variables whose joint distribution is given in the table.

		values of Y		
		1	2	3
	1	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{24}$
values of X	2	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{12}$
	3	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{24}$

(i) Find the marginal distribution of X and its mean and variance. [4]

(ii) Find the marginal distribution of Y and its mean and variance. [3]

(iii) Show that X and Y are independent. [2]

(iv) Draw up a table showing the values XY can take and the probability of taking each value. Hence determine $\mathrm{E}(XY)$ and $\mathrm{Var}(XY)$. [4]

Let T, U be independent random variables with means μ_T, μ_U and variances σ_T^2, σ_U^2 respectively.

(v) Starting from the definition

$$\mathrm{Var}(TU) = \mathrm{E}\left[(TU - \mathrm{E}(TU))^2\right],$$

and using *without proof* the results $\mathrm{E}(TU) = \mu_T\mu_U$ and $\mathrm{E}(T^2U^2) = \mathrm{E}(T^2)\mathrm{E}(U^2)$, show that

$$\mathrm{Var}(TU) = \sigma_T^2\sigma_U^2 + \mu_T^2\sigma_U^2 + \mu_U^2\sigma_T^2.$$ [6]

(vi) Verify that the result in part **(v)** holds for the random variables X and Y used in parts **(i)** to **(iv)**. [1]

[Turn over

Option 3: Markov chains

3 A financial correspondent has to write a background article each day in the 'personal finances' section of a newspaper. She writes either about insurance or about pensions or about investments. She arranges her writing schedule as follows.

She never writes about insurance on two consecutive days; after a day writing about insurance, she writes about pensions the next day with probability $\frac{3}{5}$. After a day writing about pensions, the next day she writes about pensions again with probability $\frac{1}{3}$ and about investments with probability $\frac{1}{2}$. After a day writing about investments, she writes about insurance the next day with probability $\frac{1}{2}$ and is otherwise equally likely to write again about investments or about pensions.

(i) Write down the transition matrix of the Markov chain model of this situation. [4]

(ii) On the Wednesday of a certain week, she writes about insurance. Find the probabilities of writing about each topic on the Friday of that week. [4]

(iii) Find the long-run proportions of days on which she writes about each topic. [6]

(iv) On a certain day, she writes about investments. Find the expected number of consecutive further days on which she will write about investments. [6]

Option 4: Analysis of variance

4 State the usual model for the one-way analysis of variance for a situation having k treatments with n_i observations on the ith treatment, with x_{ij} denoting the jth observation on the ith treatment ($i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n_i$). Interpret the parameters in the model. State the usual assumptions about the term representing experimental error. [4]

State carefully the null and alternative hypotheses that are customarily tested in the analysis of variance. [2]

At a process development laboratory, engineers are investigating five methods for igniting gas in a cylinder. The percentage of the gas that remains unburnt is measured four times for each method, with the following results.

Method A	11.2	10.8	10.7	10.1
Method B	9.4	9.9	9.6	9.1
Method C	9.2	8.6	8.8	8.4
Method D	12.1	12.3	12.7	11.9
Method E	13.6	12.4	13.1	12.9

[The sum of these data items is 216.8 and the sum of their squares is 2403.86.]

Draw up the usual analysis of variance table and report your conclusions. [9]

Provide a two-sided 95% confidence interval for the true mean difference between method A and method C. [5]

Option 5: Regression

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

$$Y_i = \alpha + \beta x_i + \gamma z_i + e_i.$$

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 $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$. [4]

Summary measures for a data-set of 8 independent observations are as follows.

$n = 8$	$\Sigma x_i = 472$	$\Sigma z_i = 480$	$\Sigma y_i = 440$
	$\Sigma x_i^2 = 28\,028$	$\Sigma z_i^2 = 29\,700$	$\Sigma y_i^2 = 29\,600$
	$\Sigma x_i z_i = 28\,680$	$\Sigma x_i y_i = 26\,860$	$\Sigma z_i y_i = 28\,560$

(iii) Verify that the values $\hat{\alpha} = -124, \hat{\beta} = 1, \hat{\gamma} = 2$ satisfy the normal equations. [2]

(iv) You are given that the residual sum of squares is 180. State the value of $\hat{\sigma}^2$, the usual estimate of the variance of the error terms. [1]

(v) You are also given the following results.

$$\operatorname{Var}(\hat{\beta}) = \frac{\sigma^2}{S_{xx} - \dfrac{S_{xz}^2}{S_{zz}}}$$

$$\operatorname{Var}(\hat{\gamma}) = \frac{\sigma^2}{S_{zz} - \dfrac{S_{xz}^2}{S_{xx}}}$$

where

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n},$$

$$S_{zz} = \sum (z_i - \bar{z})^2 = \sum z_i^2 - \frac{(\sum z_i)^2}{n},$$

$$S_{xz} = \sum (x_i - \bar{x})(z_i - \bar{z}) = \sum x_i z_i - \frac{(\sum x_i)(\sum z_i)}{n},$$

and σ^2 is the variance of the error terms.

Test each of the hypotheses $\beta = 0$ and $\gamma = 0$ at the 5% level of significance, stating your conclusions clearly. [11]

Mark Scheme

(Qu 1) $f(x) = \lambda e^{-\lambda x}$

(i) $L = \lambda e^{-\lambda x_1} . \lambda e^{-\lambda x_2} \ldots . \lambda e^{-\lambda x_n}$ (M1) for product form (1) if correct

$= \lambda^n e^{-\lambda \Sigma x_i}$ (1) might be implicit in sequel

(M1) $\ln L = n \ln \lambda - \lambda \Sigma x_i$ (1)

(M1) $\frac{d \ln L}{d\lambda} = \frac{n}{\lambda} - \Sigma x_i$ (1) $= 0$ for max (etc) (1)

$\Rightarrow \hat{\lambda} = \frac{n}{\Sigma x_i} = \frac{1}{\bar{x}}$ (1) (allow full marks for solution via $\frac{dL}{d\lambda}$)

(M1) $\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0$, $\therefore$ this is a maximum (1).

11

(ii) We have $\hat{\lambda} \sim$ approx $N(\lambda, +\frac{\lambda^2}{n})$ (1) and we take $\hat{\lambda} \sim$ approx $N(\lambda, \frac{\hat{\lambda}^2}{n})$

↖ FT candidate's expression, unless absurd

So $\frac{\hat{\lambda} - \lambda}{\hat{\lambda}/\sqrt{n}} \sim$ approx $N(0,1)$ (M1)

$\therefore 0.95 \approx P\left(-1.96 < \frac{\hat{\lambda} - \lambda}{\hat{\lambda}/\sqrt{n}} < 1.96\right)$ (M1) for basic idea (B1) for 1.96

$\Rightarrow$ CI is $\hat{\lambda} - 1.96\frac{\hat{\lambda}}{\sqrt{n}} < \lambda < \hat{\lambda} + 1.96\frac{\hat{\lambda}}{\sqrt{n}}$ (M1) double inequality with λ (only) in middle

(1) (1)

7

(iii) $n = 30$, $\bar{x} = 3.2$ $\hat{\lambda} = 0.3125$ (A1)

CI is $0.3125 \pm \frac{1.96 \times 0.3125}{\sqrt{30}} = 0.3125 \pm 0.1118 = (0.2007, 0.4243)$ (A1)

2

Qu 2

(i)

		Y 1	2	3	
X	1	1/12	1/8	1/24	1/4
	2	1/6	1/4	1/12	1/2
	3	1/12	1/8	1/24	1/4
		1/3	1/2	1/6	

M1 A1

FT throughout, but A0 for negative variances or for prob dists for which $\Sigma \neq 1$.

Accept fractions not in lowest terms, but **DEDUCT 1 FROM TOTAL** if this has been done.

$E(X) = 1\times\frac{1}{4} + 2\times\frac{1}{2} + 3\times\frac{1}{4}$ [or by symmetry] $= 2$ A1

$E(X^2) = 1\times\frac{1}{4} + 4\times\frac{1}{2} + 9\times\frac{1}{4} = \frac{9}{2}$ $\quad \therefore Var(X) = \frac{9}{2} - 4 = \frac{1}{2}$ A1

[4]

(ii) A1 for marginal dist.

$E(Y) = 1\times\frac{1}{3} + 2\times\frac{1}{2} + 3\times\frac{1}{6} = \frac{11}{6}$ A1

$E(Y^2) = 1\times\frac{1}{3} + 4\times\frac{1}{2} + 9\times\frac{1}{6} = \frac{23}{6}$ $\quad \therefore Var(Y) = \frac{23}{6} - \frac{121}{36} = \frac{138-121}{36} = \frac{17}{36}$ A1

[3]

(iii) Demonstration that $P(x,y) = P(x)P(y)$ 1

for **all** x, y 1

[2]

(iv)

XY	1	2	3	4	6	9
prob	$\frac{1}{12}$	$\frac{7}{24}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{5}{24}$	$\frac{1}{24}$

1 1

$E(XY) = \frac{1}{24}\{2+14+9+24+30+9\} = \frac{88}{24} = \frac{11}{3}$ A1

$E((XY)^2) = \frac{1}{24}\{2+28+27+96+180+81\} = \frac{414}{24} = \frac{69}{4}$

$\therefore Var(XY) = \frac{69}{4} - \frac{121}{9} = \frac{621-484}{36} = \frac{137}{36}$ A1

[4]

(v) $Var(TU) = E[(TU - E(TU))^2]$

$= E[(TU - \mu_T\mu_U)^2]$ 1

$= E[T^2U^2 - 2\mu_T\mu_U TU + \mu_T^2\mu_U^2]$ 1

$= E(T^2)E(U^2) - \mu_T^2\mu_U^2$ 1 1

$= (\sigma_T^2 + \mu_T^2)(\sigma_U^2 + \mu_U^2) - \mu_T^2\mu_U^2$ 1

$= \sigma_T^2\sigma_U^2 + \mu_T^2\sigma_U^2 + \mu_U^2\sigma_T^2$ 1

[6]

(vi) Inserting numerical values for X, Y gives

$\frac{1}{2}\times\frac{17}{36} + 4\times\frac{17}{36} + \frac{121}{36}\times\frac{1}{2} = \frac{34+272+242}{4\times 36}$

$= \frac{548}{4\times 36} = \frac{137}{36}$ as required 1

[1]

Qu 3

(i)

	Ins	Pen	Inv
Insurance	0	3/5	2/5
Pensions	1/6	1/3	1/2
Investments	1/2	1/4	1/4

(B4)

4

(ii) $p_0 = [1 \ 0 \ 0]$

$p_1 = p_0 \mathbb{P} = [0 \quad 3/5 \quad 2/5]$

$p_2 = p_1 \mathbb{P} = \left[\frac{3}{5}\times\frac{1}{6} + \frac{2}{5}\times\frac{1}{2} = \underline{\underline{\frac{3}{10}}} \quad \frac{3}{5}\times\frac{1}{3} + \frac{2}{5}\times\frac{1}{4} = \underline{\underline{\frac{3}{10}}} \quad \frac{3}{5}\times\frac{1}{2} + \frac{2}{5}\times\frac{1}{4} = \underline{\underline{\frac{2}{5}}}\right]$

(M1, A1, A1, A1) (or use $\mathbb{P}^2$)

4

(iii) $\pi = \pi \mathbb{P}$ (M2)

with $\Sigma \pi_i = 1$ (M1)

$$\begin{aligned}
\pi_1 &= \tfrac{1}{6}\pi_2 + \tfrac{1}{2}\pi_3 \\
\pi_2 &= \tfrac{3}{5}\pi_1 + \tfrac{1}{3}\pi_2 + \tfrac{1}{4}\pi_3 \\
\pi_3 &= \tfrac{2}{5}\pi_1 + \tfrac{1}{2}\pi_2 + \tfrac{1}{4}\pi_3 \\
\pi_1 &+ \pi_2 + \pi_3 = 1
\end{aligned}$$

Solutions are

$\pi_1 = \frac{45}{179} = 0.251397$ (A1)

$\pi_2 = \frac{66}{179} = 0.368715$ (A1)

$\pi_3 = \frac{68}{179} = 0.379888$ (A1)

6

(iv) We have P(investments again) $= \frac{1}{4}$, P(not) $= \frac{3}{4}$. (M1)

∴ we want (M1) (M1)

$0 \times \frac{3}{4} + 1 \times \frac{1}{4} \times \frac{3}{4} + 2 \times \left(\frac{1}{4}\right)^2 \times \frac{3}{4} + 3 \times \left(\frac{1}{4}\right)^3 \times \frac{3}{4} + \ldots$

$= \frac{3}{4}\left\{\frac{1}{4} + 2\left(\frac{1}{4}\right)^2 + 3\left(\frac{1}{4}\right)^3 + \ldots\right\}$ (M2)

METHOD for summing series must be clear (eg "GP of GPs")

$= \frac{1}{3}$ (A1)

6

(Qu 4)

$x_{ij} = \mu + \alpha_i + e_{ij}$ (1)

μ is population grand mean for whole experiment (1)

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

} Must be clear reference to population

$e_{ij} \sim$ ind $N(0, \sigma^2)$ (1) Allow "uncorrelated" for "ind N". Condone absence of mean = 0.

[4]

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_k$ (1)

H_1: The α_i are not all equal (1)

Verbal statements acceptable

[2]

				TOTALS
11.2	10.8	10.7	10.1	42.8
9.4	9.9	9.6	9.1	38.0
9.2	8.6	8.8	8.4	35.0
12.1	12.3	12.7	11.9	49.0
13.6	12.4	13.1	12.9	52.0
				216.8

"Correction factor" $= \frac{(216.8)^2}{20} = 2350.112$

Total SS $= 2403.86 - 2350.112 = 53.748$ with 19 df

Between methods SS

$= \frac{42.8^2}{4} + \ldots + \frac{52.0^2}{4} - 2350.112 = 2401.46 - 2350.112 = 51.348$ with 4 df

Residual SS (by subtraction) $= 53.748 - 51.348 = 2.4$ with 15 df

Source of variation	SS (M1)	df (1)	MS (M1)	MS ratio (M1 A1)
Between methods	51.348	4	12.837	80.23
Residual	2.4	15	0.16	
Total	53.748	19		

Refer to $F_{4,15}$ (1) – overwhelming (1) evidence (1) that methods are not all the same. (1) verbal statement

[9]

CI for $\mu_A - \mu_C$ is

$10.7 - 8.75 \pm 2.131 \sqrt{0.16}\sqrt{\frac{1}{4} + \frac{1}{4}}$ $= 1.95 \pm 0.60(27)$

$= (1.34(73), 2.55(27))$

↑ t_{15} (d.t. 5%) ↑ resid MS

(M1) (B1) (M1) (M1) (A1)

[5]

(Qu 5) $Y_i = \alpha + \beta x_i + \gamma z_i + e_i$

(i) $e_i \sim \text{ind } N(0, \sigma^2)$ (2) May subdivide. Allow "uncorr" for "ind N". **[2]**

(ii) $\Omega = \sum e_i^2$

$= \sum (Y_i - \alpha - \beta x_i - \gamma z_i)^2$ (M1)

(M1)

$$\frac{\partial \Omega}{\partial \alpha} = -2\sum(Y_i - \alpha - \beta x_i - \gamma z_i) = 0$$

(M1)

$$\frac{\partial \Omega}{\partial \beta} = -2\sum x_i(Y_i - \alpha - \beta x_i - \gamma z_i) = 0$$

$$\frac{\partial \Omega}{\partial \gamma} = -2\sum z_i(Y_i - \alpha - \beta x_i - \gamma z_i) = 0$$

$\rightarrow$

$\sum Y_i - n\hat{\alpha} - \hat{\beta}\sum x_i - \hat{\gamma}\sum z_i = 0$

$\sum x_i Y_i - \hat{\alpha}\sum x_i - \hat{\beta}\sum x_i^2 - \hat{\gamma}\sum x_i z_i = 0$

(1) $\sum z_i Y_i - \hat{\alpha}\sum z_i - \hat{\beta}\sum x_i z_i - \hat{\gamma}\sum z_i^2 = 0$ **[4]**

(iii) Verification: $440 - 8(-124) - 472 - 2\times 480 = 0$ ✓ (M1) attempt to verify

$26860 - 472(-124) - 28028 - 2\times 28680 = 0$ ✓ (A1) all correct

$28560 - 480(-124) - 28680 - 2\times 29700 = 0$ ✓ **[2]**

(iv) $\hat{\sigma}^2 = \frac{RSS}{8-3} = \frac{180}{5} = 36$ (1) **[1]**

(v) $S_{xx} = 28028 - \frac{(472)^2}{8} = 180$

$S_{zz} = 29700 - \frac{(480)^2}{8} = 900$

$S_{xz} = 28680 - \frac{(472)(480)}{8} = 360$ (B1) all these correct

$\therefore \text{Var}(\hat{\beta}) = \frac{\sigma^2}{180 - 360^2/900} = \frac{\sigma^2}{36}$ (1)

$\therefore \text{Est Var}(\hat{\beta}) = \frac{36}{36} = 1$ (1)

also, $\text{Var}(\hat{\gamma}) = \frac{\sigma^2}{900 - 360^2/180} = \frac{\sigma^2}{180}$ (1)

$\therefore \text{Est Var}(\hat{\gamma}) = \frac{36}{180} = 0.2$ (1)

$\therefore$ Test statistic for $\beta = 0$ is $\frac{1-0}{\sqrt{1}} = 1$ (M1)

and for $\gamma = 0$ is $\frac{2-0}{\sqrt{0.2}} = 4.472$. (A1)

Compare with t_5 (1). Dt. 5% point is 2.571 (1).

$\therefore$ no evidence against $\beta = 0$, but (separately) there is evidence against $\gamma = 0$. (1, 1) **[11]**

Examiner’s Report

2618 Statistics 6

General Comments

There were only 14 candidates, from 6 centres. Generally the work was of good quality, as is of course to be expected from candidates who enter for this module.

Comments on Individual Questions

Q.1 This question was generally answered well. The maximum likelihood estimator, which turns out to be $1/\bar{x}$, was usually successfully found, though some candidates did not check that the turning point of the likelihood function (or its logarithm) is indeed a maximum. The approximate confidence interval in the next part of the question was sometimes worked out very well, but it was disturbing to find candidates at this level who seemed to think that confidence intervals are always of the form $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ – the unthinking use of $\left(\frac{\sigma}{\sqrt{n}}\right)$ being arguably even worse that the automatic use of $\bar{x}$ as the centre. The small numerical exercise at the end caused little difficulty.

Interval is (0.2007, 0.4243).

Q.2 As usual, the arithmetical parts of this question were usually done with great facility. In part (iii), verification of independence requires *all* the joint probabilities to be checked (or some other equivalent approach – and some candidates were ingenious in finding other correct methods); it is not enough to check only one (or even a few). It was another disappointment at this level to find occasional candidates who averred that zero covariance implied independence. In part (v), some candidates did the algebra very efficiently, but others needed varying degrees of desperate struggles.

There are many numerical answers in the question, too many to quote here. Please see the published mark scheme for details.

Q.3 This question also was usually done well. Only one candidate made a mistake with the transition matrix (this candidate's work then "followed-through" correctly). Nearly all candidates worked out the two-step distribution correctly [3/10, 3/10, 2/5], the only exception being one candidate who went one step too far. The limiting distribution was also carefully and correctly obtained [45/179, 66/179, 68/179]. The majority of candidates were likewise careful and correct in finding the expected number of consecutive further days in the last part, giving good indication of their methods. Often a generating function approach was used to sum the series, sometimes a binomial expansion was used instead. But, as seems always to be the case, there were some candidates who showed no working at all. To quote from last year's report: "These reports have repeatedly stated that unsupported writing-down of answers is regarded as in contravention of the global rubric that sufficient details of the working must be shown to indicate that a correct method is being used". Yet one still saw candidates who showed no method, or perhaps merely quoted "$\alpha/(1-\alpha)$", often without even saying what α is. Quoting again from last year: "these candidates lost marks. And similar candidates will continue to lose marks in future years".

Answer for last part is 1/3.

Q.4 This question opened with a requirement to state the model and interpret the terms in it. This is asked in many years, but there are still some candidates who do not know it thoroughly. This is disappointing. Candidates should not see this topic as merely a "recipe" for carrying out an important test procedure. They should look for a deeper understanding of what is going on.

Similar remarks apply to the next part of the question, where the customary null and alternative hypotheses were to be stated. Here, however, candidates in general do seem now to have a better grasp of the fundamentals.

An analysis of variance was then to be undertaken. All candidates knew what to do and there were very few errors in doing it. [Value of test statistic is 80.23, refer to F with 4 and 15 degrees of freedom, *overwhelming* evidence against the null hypothesis, i.e. that the methods are not all equivalent.]

The last part of the question required a confidence interval for the true mean difference between methods A and C. This was usually done well, except that some candidates used the customary "pooled estimator" of variance from just these two methods and proceeded with an interval based on t_6. However, by assumption in the modelling, the underlying variance is the same for *all* the methods, and the best estimate of it, pooling the information from *all* of them, is simply the residual mean square that has already been calculated. This leads directly to an interval based on t_{15}.

The interval is (1.347, 2.553).

Q.5 There were no attempts on this question this year. Please see the published mark scheme for details of the solution.