OCR
RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2617

Statistics 5

Thursday **9 JUNE 2005** 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 5 printed pages and 3 blank pages.

HN/3

[Turn over

1 The random variable X has the Poisson distribution with parameter λ, so that

$$\mathrm{P}(X = x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \dots .$$

(i) Derive the probability generating function of X. [3]

(ii) Hence obtain the mean and variance of X. [5]

(iii) The random variable Y is independent of X and has the Poisson distribution with parameter μ. Find the distribution of $X + Y$. [6]

(iv) Use the distributions of X, Y and $X + Y$ to find the conditional probability that $X = x$ given that $X + Y = n$, where n is a non-negative integer. Deduce that the conditional distribution of X given that $X + Y = n$ is binomial with parameters n and $\frac{\lambda}{\lambda + \mu}$. [6]

2 **(i)** The moment generating function (mgf) of a random variable U is defined by $\mathrm{M}_U(\theta) = \mathrm{E}(\mathrm{e}^{\theta U})$. Prove from this definition that, if $V = aU + b$, where a and b are constants, the mgf of V is $\mathrm{e}^{b\theta}\mathrm{M}_U(a\theta)$. [3]

(ii) The non-negative random variable X has probability density function $\mathrm{f}(x) = \mathrm{e}^{-x}$, $x \geqslant 0$. Show that the mgf of X is $\mathrm{M}_X(\theta) = (1 - \theta)^{-1}$. Deduce that $\mathrm{E}(X) = 1$ and $\mathrm{Var}(X) = 1$. [6]

(iii) The random variable Y is defined by

$$Y = X_1 + X_2 + \dots + X_n,$$

where $X_1, X_2, \dots, X_n$ are independent random variables each distributed as X. Write down the mgf of Y. [1]

(iv) The random variable $\bar{X}$ is defined by

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{Y}{n}.$$

Use the result of part **(i)**, taking $a = \frac{1}{n}$ and $b = 0$, to show that the mgf of $\bar{X}$ is $\left(1 - \frac{\theta}{n}\right)^{-n}$. Write down the mean and variance of $\bar{X}$. [3]

(v) The random variable Z is defined by

$$Z = \frac{\bar{X} - 1}{\frac{1}{\sqrt{n}}};$$

this has mean 0 and variance 1 for any n, and is called the *standardised mean* of n independent realisations of X. Again using the result of part **(i)**, show that the mgf of Z is

$$\mathrm{M}_Z(\theta) = \mathrm{e}^{-\theta\sqrt{n}}\left(1 - \frac{\theta}{\sqrt{n}}\right)^{-n}.$$ [2]

(vi) Use the expansion $\ln(1 - s) = -s - \frac{s^2}{2} - \frac{s^3}{3} - \dots$ to show that

$$\ln \mathrm{M}_Z(\theta) = \frac{\theta^2}{2} + \frac{\theta^3}{3\sqrt{n}} + \frac{\theta^4}{4n} + \dots .$$

Given that the mgf of the standard Normal random variable is $\mathrm{e}^{\theta^2/2}$, what can you deduce about the distribution of Z as n becomes large? [5]

[Turn over

3 (i) The following data are a random sample of size 5 from a Normal distribution with unknown variance σ^2.

12.6 8.4 14.4 10.2 14.6

Test the null hypothesis H_0: $\sigma^2 = 2$ against the alternative hypothesis H_1: $\sigma^2 > 2$ at the 5% significance level. [5]

(ii) The probability density function of the random variable Y having the χ_4^2 distribution is

$$f(y) = \tfrac{1}{4} y e^{-\frac{1}{2}y}$$

(for $y \geqslant 0$). Show that

$$P(Y < y) = 1 - e^{-\frac{1}{2}y}\left(1 + \tfrac{1}{2}y\right)$$

(for $y \geqslant 0$). [3]

(iii) Hence show that the level of significance of the data in part (i) is 0.00577. [2]

(iv) Explain how the result in part (iii) is related to appropriate entries in the table of percentage points of the χ_4^2 distribution. [2]

(v) For a test of H_0: $\sigma^2 = 2$ against H_1: $\sigma^2 > 2$, given a random sample of size 5 from a Normal distribution with variance σ^2, show that the null hypothesis is accepted if $Y < \frac{2}{\sigma^2} \times 9.488$. [3]

(vi) Hence find the value of the operating characteristic of the test for $\sigma^2 = 3$ and for $\sigma^2 = 6$. Given also the values 0.685 and 0.245 of the operating characteristic for $\sigma^2 = 4$ and $\sigma^2 = 10$ respectively, comment briefly on the test. [5]

4 A market gardener is testing a new insect-repellent spray.

(a) 60 tomato seedlings, regarded as a random sample, are sprayed with the new spray. Another 100 tomato seedlings, also regarded as a random sample, are sprayed with the existing spray. After five weeks, the numbers of the seedlings that have still not been attacked by insects are 54 and 76 respectively. Provide a two-sided 90% confidence interval for the difference in the proportions of seedlings not attacked by insects after five weeks in the corresponding underlying populations. Explain what you conclude from this interval. [8]

(b) The gardener is concerned that the new spray, even if effective in reducing attacks by insects, might as a side-effect decrease the average yield of the crop or increase the variability in yields. To examine this, tomatoes in 10 experimental plots are sprayed with the new spray and those in 8 other experimental plots (of the same size) are sprayed with the existing spray. Each set of experimental plots is regarded as a random sample. All other conditions on the experimental plots are carefully controlled. The yields, in kg, from the experimental plots are found to be as follows.

New spray	26.8	28.1	30.8	29.1	32.6	30.2	25.8	30.6	29.5	28.0
Existing spray	31.2	32.9	29.6	28.2	29.9	30.3	30.6	29.6		

Test at the 5% level of significance whether the underlying variances may be assumed equal. State the distributional result on which the test is based.

State the assumptions necessary for the test to be valid. Assuming they are satisfied, and in the light of the result of your test for the variances, explain briefly whether you would be prepared to proceed to a t test for comparing the population means. [12]

Mark Scheme 2617
June 2005

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:

Question 1

(i)	X ~ Poisson (λ) Pgf of X is $\mathrm{G}_X(t) = \mathrm{E}\left\lfloor t^X \right\rfloor$ $= \sum_{x=o}^{\infty} t^x \frac{e^{-\lambda}\lambda^x}{x!}$ $= \mathrm{e}^{-\lambda}\sum_{x=o}^{\infty} \frac{(\lambda t)^x}{x!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda t}$	**M1** **M1, A1**	3
(ii)	$\mu = \mathrm{G}'(1) = \mathrm{e}^{-\lambda}(\lambda e^{\lambda t})\big\vert_{t=1} = \lambda$ $\sigma^2 = \mathrm{G}''(1) + \mu(1-\mu) = \lambda e^{-\lambda}\cdot\lambda e^{\lambda t}\big\vert_{t=1} + \lambda - \lambda^2$ $= \lambda^2 + \lambda - \lambda^2 = \lambda$	**M1, A1** **M1** **A1**	
			5
(iii)	Y(~ Poisson (μ)) has pgf $\mathrm{e}^{-\mu}\ \mathrm{e}^{\mu t}$ $\therefore X + Y$ has pgf $(\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda t}).(\mathrm{e}^{-\mu}\mathrm{e}^{\mu t})$ $= \mathrm{e}^{-(\lambda+\mu)}\mathrm{e}^{(\lambda+\mu)t}$ which is pgf of Poisson (λ + μ) and as pgfs are unique it follows that X + Y ~ Poisson (λ+ μ) [or, <u>much</u> longer, by convolution of probabilities]	**M1** **M1** product of pgfs **A1** **M1** **E1** **1**	6

(iv)	$\mathrm{P}(X=x \mid X+Y=n) = \dfrac{\mathrm{P}(X=x \;\cap\; X+Y=n)}{\mathrm{P}(X+Y=n)}$	**M1**	
	$= \dfrac{\mathrm{P}(X=x \;\cap\; Y=n-x)}{\mathrm{P}(X+Y=n)}$	**M1**	
		o.e. **1**	
	$= \dfrac{e^{-\lambda}\lambda^{x}}{x!}\cdot\dfrac{e^{-\mu}\mu^{n-x}}{(n-x)!}\cdot\dfrac{n!}{e^{-(\lambda+\mu)}(\lambda+\mu)^{n}}$ [for x=0,1,…,n]		
	$= \binom{n}{x}\Big(\dfrac{\lambda}{\lambda+\mu)}\Big)^{x}\Big(\dfrac{\mu}{\lambda+\mu} \quad = \boxed{1-\dfrac{\lambda}{\lambda+\mu}}\Big)^{n-x}$	**1** for algebraic terms **1** for $\binom{n}{x}$	
	i.e. $\mathrm{B}(n, \dfrac{\lambda}{\lambda+\mu})$	**1**	6

Question 2

(i)	$\mathrm{M}_V(\theta) = \mathrm{E}(\mathrm{e}^{\theta V})$ $= \mathrm{E}(\mathrm{e}^{\theta(aU+b)})$ $= \mathrm{e}^{b\theta}\mathrm{E}(\mathrm{e}^{(a\theta)U})\ \mathrm{M}_X(\theta) = \int_0^\infty \mathrm{e}^{\theta x}\mathrm{e}^{-x}\mathrm{d}x = \int_0^\infty \mathrm{e}^{-x(1-\theta)}\mathrm{d}x$ $= \mathrm{e}^{b\theta}\mathrm{M}_U(a\theta)$	**M1** **1** **1**	3
(ii)	$\mathrm{M}_X(\theta) = \int_0^\infty \mathrm{e}^{\theta x}\mathrm{e}^{-x}\mathrm{d}x = \int_0^\infty \mathrm{e}^{-x(1-\theta)}\mathrm{d}x$ $= \left[\frac{\mathrm{e}^{-x(1-\theta)}}{-(1-\theta)}\right]_0^\infty \quad = 0 + \frac{1}{1-\theta} = (1-\theta)^{-1}$ (OK for $\theta < 1$: candidates are not expected to discuss this) $\mathrm{E}(X) = \mathrm{M}_x{}'(0) = -1(1-\theta)^{-2}(-1)\Big\rvert_{\theta=0} = 1$ $\mathrm{E}(X^2) = \mathrm{M}_x{}''(0) = -2(1-\theta)^{-3}(-1)\Big\rvert_{\theta=0} = 2$ $\therefore$ Var (X) = 2 – 1 = 1 [or by series expansion]	**M1** **A1** beware printed answer **M1** Note – Answer is given **M1, A1** **1** Answer given	6
(iii)	$\mathrm{M}_Y(\theta) = \{\mathrm{M}_X(\theta)\}^n = (1-\theta)^{-n}$	**1**	1
(iv)	$\mathrm{M}_{\overline{X}}(\theta) = \mathrm{e}^{0\theta}\mathrm{M}_Y(\frac{1}{n}\theta) = (1-\frac{\theta}{n})^{-n}$ Mean 1 Variance $\frac{1}{n}$	**1** Answer given **1** **1**	3
(v)	$Z = \sqrt{n}\ \overline{X} - \sqrt{n}$ $\therefore \mathrm{M}_Z(\theta) = \mathrm{e}^{-\sqrt{n}\theta}\ \mathrm{M}_{\overline{X}}(\ \sqrt{n}\theta\ ) = \mathrm{e}^{-\sqrt{n}\,\theta}\left(\ 1 - \frac{\theta}{\sqrt{n}}\ \right)^{-n}$	**M1** **1** Answer given	2

(vi)	$\ln \mathrm{M}_Z(\theta) = -\sqrt{n}\theta - n\ln\left(1 - \frac{\theta}{\sqrt{n}}\right)$		
	$= -\sqrt{n}\ \theta + n\left\{\frac{\theta}{\sqrt{n}} + \frac{\theta^2}{2n} + \frac{\theta^3}{3n^{3/2}} + \frac{\theta^4}{4n^2} + \cdots\right\}$	**M1**	
	$= \frac{\theta^2}{2} + \frac{\theta^3}{3}n^{-1/2} + \frac{\theta^4}{4}n^{-1} + \cdots$	**1** Answer given	
	As $n \to \infty$, this $\to \frac{\theta^2}{2}$, so $M_z(\theta) \to e^{\theta^2/2}$ Which is mgf of N(0,1)	**1**	
	and, as mgfs are unique	**E1**	
	this implies Z tends to N(0,1)	**1**	
			5

Question 3

(i)	$s_{n-1}{}^2 = 7.268 \quad \left[s_n^2 = 5.8144, \text{allow only if correctly used}\right]$ Test statistic is $\dfrac{(n-1)\,S^2}{\sigma_o{}^2}\left[n = 5, \sigma_o{}^2 = 2\right] = 14.536$ Compare with χ_4^2 Upper 5% point is 9.488 Significant/reject H_O	**M1, A1** **1** No FT if wrong **1** No FT if wrong **1**	5
(ii)	$P(Y < y) = \int_0^y \frac{1}{4} te^{-t/2} dt$ $= \frac{1}{4}\left\{\left[-2te^{-t/2}\right]_0^y + 2\int_0^y e^{-t/2} dt\right\}$ $= \frac{1}{4}\left\{-2ye^{-y/2} + (-4)\left[e^{-t/2}\right]_0^y\right\}$ $= -\frac{1}{2} ye^{-y/2} - e^{-y/2} + 1 \quad = 1 - e^{-y/2}\,(1 + \frac{y}{2})$	**M1** for attempt to integrate by parts **2,** divisible, for algebra BEWARE PRINTED ANSWER	3
(iii)	Want $P(Y > 14.536) = 1 - \left\{1 - e^{-7.268}\,(8.268)\right\} = 0.0057\ (669)$	**M1, A1** BEWARE PRINTED ANSWER	2
(iv)		**E2**	2

	14.536 is between upper 0.01 point (13.28) and upper 0.005 point (14.86)	(E0,E1,E2)	

(v)	H_0 is accepted if $\dfrac{(n-1)S^2}{\sigma_o^2}\left[i.e.\dfrac{4S^2}{2},\ i.e.\ 2S^2\right]\text{is} < 9.488$ But we now have $\dfrac{(n-1)S^2}{\sigma^2}\left[i.e.\dfrac{4S^2}{\sigma^2}\right]\sim \chi_4^2$ So $2\,S^2 \sim \dfrac{\sigma^2}{2}\chi_4^2$, ie accept H_0 if $\dfrac{\sigma^2}{2}\chi_4^2 < 9.488$ ie if $\chi_4^2 < \dfrac{2}{\sigma^2}\times 9.488$	M1 1 E1	3
(vi)	$\sigma^2 = 3:$ Want $P(\chi_4^2 < \dfrac{2}{3}\times 9.488 = 6.325\dot{3})$ $= 1 - e^{-3.162\dot{6}}\ (4.162\dot{6})$ $= 0.823(866)$ $\sigma^2 = 6:$ Want $P(\chi_4^2 < \dfrac{2}{6}\times 9.488 = 3.162\dot{6})$ $= 1 - e^{-1.581\dot{3}}\ (2.581\dot{3})$ $= 0.469(018)$ Also given $\sigma^2 = 4\ :\ 0.685$ $\sigma^2 = 10\ :\ 0.245$	M1 ; A1, A1 E2 (E0,E1,E2)	5

	These are quite high probabilities of accepting H_O when it is false [or other sensible comments]		

Question 4

(a)	We have 54 out of 60 and 76 out of 100 90% CI for $p_1 - p_2$ is $\frac{54}{60}-\frac{76}{100}\pm 1.645\sqrt{\frac{(\frac{54}{60}\cdot\frac{6}{60})}{60}+\frac{(\frac{76}{100}\cdot\frac{24}{100})}{100}}$ $= 0.14 \pm 1.645\sqrt{0.0015+0.001824(=0.003324)}$ $= 0.14 \pm 1.645\times 0.0576\,(54)$ $= 0.14 \pm 0.094\,(84)$ $= (0.045\,(16)\,,\,0.234(84))$ The lower end of this interval is >0, which suggests that the new spray is better.	**M1** for $\frac{54}{60}-\frac{76}{100}$ **B1** for 1.645 **M1** Two Terms **M1** Both Correct **A1** **A1 CAO** **E2** (E0,E1,E2)	8
(b)	$s_1^2 = 4.125$ (3.7125 with divisor n) $s_2^2 = 1.887$ (1.651 with divisor n) Test statistic is $\frac{4.125}{1.887} = 2.186$	**B1** **1** [FT from candidates' values] Refer to $F_{9,7}$	

		1 for *F* **1** for df No FT if wrong	

	$F_{9,7}$ is not in tables. Upper 2 ½ % point of $F_{8,7}$ is 4.90, of $F_{10,7}$ is 4.76. Accept any convincing explanation that result is not significant. Seems underlying variances can be assumed equal $\dfrac{\left(\dfrac{S_1^2}{\sigma_1^2}\right)}{\left(\dfrac{S_2^2}{\sigma_2^2}\right)} \sim F_{n_1-1,\, n_2-1}$ Require Normality of both populations This is also required for *t* test. *t* test needs same population variances – we have no evidence against this. So *t* test seems OK.	**E2** [Only **1** if based on 5% points : 3.73 and 3.64] **1** **1, 1** Accept $\left[\text{LHS as } \frac{S_1^2}{S_2^2}\right]$ **1** **E2** (E0,E1,E2)	12

2617 - Statistics 5

General Comments

There were only 13 candidates, from 7 centres – including some unfamiliar ones, which it was nice to see among such a small entry for the last regular sitting of this module.

In view of the small number of candidates, this report is couched in very general terms so as to avoid any possibility that individuals are identifiable.

Comments on Individual Questions

1) This was on probability generating functions, based on the Poisson distribution and the sum of Poisson distributions. Candidates were able to do the technical work in the first three parts of the question, as far as and including using the pgf to find the distribution of the sum. However, in the last part there was considerable insecurity in use of conditional probability, so that several candidates were left with a struggle to try to manipulate incorrect expressions so as to achieve the given result.

2) This question was based on moment generating functions, leading to a proof of the central limit theorem for the case of an exponential distribution. Most candidates met with considerable success here, perhaps helped by the substantial number of intermediate steps given in the question.

3) This question was based on the chi-squared test for variance. The initial test was usually done correctly. Most candidates could then integrate the given pdf of the chi-squared distribution with 4 degrees of freedom so as to obtain the cdf, and most candidates then knew how to use this to obtain the level of significance of the data. Not all, however, grasped the point about the relation of this to entries in the chi-squared table. The next part of the question was concerned with setting up the acceptance region for the test in a general way. Most candidates seemed to know what to try to do, but there were some difficulties in doing it. However, the given result was used well in the final part in deriving values of the operating characteristic of the test, usually with sensible interpretations of the rather poor nature of the test in this (very small sample) case.

4) This was a composite question covering a confidence interval for a difference between two proportions and a test for the equality of two variances. Mostly it was done quite well. The *F* distribution with 9 and 7 degrees of freedom is not tabulated in the MEI tables; candidates were expected to overcome this and did so in a variety of ways. A fairly common error was to work with upper-tail 5% points whereas, as the test is two-sided, upper-tail 2½% points should have been used.