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

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 4 printed pages.

HN/3

Jun04/erratum32

OXFORD CAMBRIDGE AND RSA EXAMINATIONS
Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2617

Statistics 5

Friday **11 JUNE 2004** Morning 1 hour 20 minutes

ERRATUM NOTICE

One copy to be given to each candidate

Turn to Page 2.

There is an error in Question **1(iv)** and Question **1(v).**

The Questions should read:

(iv) Suppose now that the number of females in the zoo in part **(ii)** is regarded as a random variable X having the Poisson distribution with parameter λ. This means that the pgf of S (the total number of offspring produced by these females) is now the *expectation with respect to* X of the answer to part **(ii)**. By considering the pgf of the Poisson distribution with parameter λ, namely $e^{\lambda(t-1)}$, show that

$$\mathrm{E}\left[\left(\frac{\ln(1-t\theta)}{\ln(1-\theta)}\right)^X\right]=e^{\lambda\left(\frac{\ln(1-t\theta)}{\ln(1-\theta)}-1\right)}.$$

(v) Hence, using the result

$$e^{\lambda\left(\frac{\ln(1-t\theta)}{\ln(1-\theta)}-1\right)}=\left(\frac{1-t\theta}{1-\theta}\right)^{\frac{\lambda}{\ln(1-\theta)}},$$

show that S has the negative binomial distribution with parameters $\frac{\lambda}{-\ln(1-\theta)}$ and $1-\theta$.

1 **(i)** The random variable Y takes values 1, 2, ... with

$$P(Y = y) = \frac{\theta^y}{-y\ln(1-\theta)}, \qquad y = 1, 2, \ldots,$$

where $0 < \theta < 1$. (Y is said to have the logarithmic distribution with parameter θ.) Show that the probability generating function (pgf) of Y is

$$G(t) = \frac{\ln(1-t\theta)}{\ln(1-\theta)}.$$

[You may use the expansion $\ln(1-u) = -u - \frac{1}{2}u^2 - \frac{1}{3}u^3 - \ldots$.] [6]

(ii) A zoo has x females of a particular species. The number of offspring produced by each female in a year follows the logarithmic distribution with parameter θ, independently of all other females. Write down the pgf of S, the total number of offspring produced by these females. [1]

(iii) The random variable Z takes values 0, 1, 2, ... with

$$P(Z = z) = \binom{k+z-1}{z} p^k q^z, \qquad z = 0, 1, 2, \ldots,$$

where k is a non-negative integer and $0 < p < 1$, with $q = 1 - p$. (Z is said to have the negative binomial distribution with parameters k and p.) Show that the pgf of Z is

$$H(t) = p^k(1 - tq)^{-k}.$$

[You may use the expansion $(1-u)^{-k} = 1 + ku + \frac{k(k+1)}{2!}u^2 + \frac{k(k+1)(k+2)}{3!}u^3 + \ldots$.] [6]

(iv) Suppose now that the number of females in the zoo in part **(ii)** is regarded as a random variable X having the Poisson distribution with parameter λ. This means that the pgf of S (the total number of offspring produced by these females) is now the *expectation with respect to* X of the answer to part **(ii)**. By considering the pgf of the Poisson distribution with parameter λ, namely $e^{\lambda(t-1)}$, show that

$$E\left[\left(\frac{\ln(1 \pm t\theta)}{\ln(1 \pm \theta)}\right)^X\right] = e^{\lambda\left(\frac{\ln(1 \pm t\theta)}{\ln(1 \pm \theta)} - 1\right)}.$$ [2]

(v) Hence, using the result

$$e^{\lambda\left(\frac{\ln(1 \pm t\theta)}{\ln(1 \pm \theta)} - 1\right)} = \left(\frac{1 \pm t\theta}{1 \pm \theta}\right)^{\frac{\lambda}{\ln(1-\theta)}},$$

show that S has the negative binomial distribution with parameters $\dfrac{\lambda}{-\ln(1-\theta)}$ and $1 - \theta$. [5]

2 **(i)** The probability density function of the random variable X having the standard Normal (N(0,1)) distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty.$$

Derive the moment generating function of X. [6]

(ii) The random variable Y is defined by $Y = \sigma X + \mu$; this means that $Y \sim N(\mu, \sigma^2)$. Use the moment generating function of X to obtain the moment generating function of Y. [3]

(iii) Use the moment generating function of Y to confirm that μ and σ^2 are the mean and variance of the distribution of Y. [6]

(iv) $Y_1, Y_2, \ldots, Y_n$ are independent random variables having the distributions $N(\mu_1, \sigma^2_1), N(\mu_2, \sigma^2_2), \ldots, N(\mu_n, \sigma^2_n)$ respectively. Show that the distribution of $Y_1 + Y_2 + \ldots + Y_n$ is Normal, and state its mean and variance. [5]

3 A big company needs to hire large numbers of temporary employees and has a contract with an agency that provides them. The contract includes agreed standards by which a temporary employee can be deemed satisfactory when working at the company. It is stipulated that, over an extended period, at least 80% of the employees should be satisfactory, otherwise the company will withhold its fee to the agency.

(i) Inspection of records for a random sample of 120 temporary employees shows that only 91 were satisfactory. The company informs the agency that it will withhold the fee. The agency claims that this is unreasonable as there is bound to be some natural variation. Undertake an appropriate 5% significance test and also provide a two-sided 99% confidence interval to help adjudicate in this dispute, in each case explaining your conclusions carefully. [13]

(ii) Another agency offers an improved service with a greater proportion of satisfactory employees. The company tries this agency and finds from the records that, out of a random sample of 80 temporary employees, 72 are satisfactory. Use this information and the data from the sample in part **(i)** to provide a two-sided 95% confidence interval for the true difference in proportions of satisfactory employees. Explain what you conclude from this interval. [7]

[Turn over

4 An engineer is investigating a new production process for components made in a factory.

(a) As part of the investigation, the engineer needs to compare the variability of the strengths of the components made by the new process with that of those made by the standard process.

Each component of a random sample of 8 made by the standard process is tested to destruction. The breaking strengths, in a convenient unit, are as follows.

87 108 98 100 84 93 109 96

Similarly, the breaking strengths of a random sample of 12 components made by the new process are as follows.

100 89 104 105 92 105 97 86 98 97 86 94

Examine at the 5% level of significance whether the underlying variances may be assumed equal. State the assumptions necessary for the test to be valid, and state also the distributional result on which the test is based. [10]

(b) Assume now that the *true* underlying standard deviation of strengths for components made by the new process is 6.5. A 5% test of the null hypothesis $\mu = 100$ against the alternative hypothesis $\mu < 100$ is required, where μ is the true mean breaking strength, using a random sample of size 12. Find the critical region for this test. Hence find the power of the test for $\mu = 100, 99, 98, 97, 96, 95$. If there were a *perfect* test for these hypotheses (i.e. one which *never* reached the wrong conclusion), what would be the values of its power function at these values of μ? [10]

Mark Scheme

General 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 (‘explanation’) 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.

All queries about the marking should have been resolved at the standardising 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 (i) [Logarithmic: $P(Y = y) = \dfrac{\theta^y}{-y\ln(1-\theta)}$ y = 1, 2, …]

$$\text{Pgf } G(t) = E[t^Y] = \sum_{y=1}^{\infty} \frac{-1}{\ln(1-\theta)} \frac{(t\theta)^y}{y} = \frac{1}{\ln(1-\theta)}\left\{-t\theta - \frac{(t\theta)^2}{2} - \ldots\right\} = \frac{\ln(1-t\theta)}{\ln(1-\theta)}$$

M1 **M1** **1** **M1** **1**

attempt at expansion correct **1**; beware printed answer — **6**

(ii) $S = Y_1 + Y_2 + \ldots + Y_x$ Pgf of S is $\left\{\dfrac{\ln(1-t\theta)}{\ln(1-\theta)}\right\}^x$ **1** — **1**

(iii) [negative binomial: $P(Z = z) = \dbinom{k+z-1}{z} p^k q^z$ z = 0, 1, 2, …]

$$\text{Pgf } H(t) = E[t^z] = \sum_{z=0}^{\infty} p^k \binom{k+z-1}{z} (tq)^z \quad \textbf{M1}$$

$$= p^k \left\{1 + k(tq) + \binom{k+1}{2}(tq)^2 + \binom{k+2}{3}(tq)^3 + \ldots\right\} = p^k(1 - tq)^{-k}$$

M1 **1** **1** **1**

attempt at expansion correct **1**; $= \dfrac{k(k+1)}{2!}$; $= \dfrac{k(k+1)(k+2)}{3!}$; beware printed answer — **6**

(iv) X ~ Poisson (λ) We have (pgf of X) $E[t^x] = e^{\lambda(t-1)}$. Thus

$$E\left[\left(\frac{\ln(1-t\theta)}{\ln(1-\theta)}\right)^X\right] = e^{\lambda\left\{\frac{\ln(1-t\theta)}{\ln(1-\theta)}-1\right\}}$$

M1, 1 beware printed answer — **2**

(v) we now have

pgf of S = this $= \left(\dfrac{1-t\theta}{1-\theta}\right)^{\frac{\lambda}{\ln(1-\theta)}}$ by the given result.

So we need to show that this = H(t) [answer to (iii)]
with $k = \dfrac{\lambda}{-\ln(1-\theta)}$ and $p = 1 - \theta$. **M1**

$H(t) = p^k(1 - tq)^{-k}$ which becomes $(1-\theta)^{\frac{\lambda}{-\ln(1-\theta)}} (1 - t\theta)^{\frac{\lambda}{\ln(1-\theta)}}$

A1 **A1**

$$= \left(\frac{1-t\theta}{1-\theta}\right)^{\frac{\lambda}{\ln(1-\theta)}} \quad \textbf{A1}$$

– hence result **1** (beware printed result) — **5**

Q2 (i) $M_X(\theta) = E(e^{\theta x}) = \int_{-\infty}^{\infty} e^{\theta x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ **M1**

$= e^{\frac{\theta^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} dx$

completing the square **M1**

done correctly **A1**

$e^{\frac{\theta^2}{2}}$ taken outside **M1**

$= e^{\frac{\theta^2}{2}} \int_{-\infty}^{\infty}$ pdf of $N(\theta, 1)$ **M1**

$= e^{\frac{\theta^2}{2}}$ **A1** **6**

(ii) $Y = \sigma X = \mu$

$\therefore M_Y(\theta) = e^{\mu\theta} M_X(\sigma\theta) = e^{\mu\theta} e^{\frac{(\sigma\theta)^2}{2}} = e^{\mu\theta + \frac{\sigma^2\theta^2}{2}}$

M1 **M1** **1**

(beware answer in formulae book – must be convincing) **3**

(iii) mean = $M'(0)$ $M'(0) = e^{\mu\theta+\frac{\sigma^2\theta^2}{2}}(\mu + \sigma^2\theta)$ **1**

$M'(0) = e^0(\mu + 0) = \mu$ ie mean = μ **1**

(depends on previous mark. If NEITHER awarded, but a reasonable attempt at $M'(0)$ has been made award **M1**)

variance = $M''(0) - \text{mean}^2$

$M''(\theta) = e^{\mu\theta+\frac{\sigma^2\theta^2}{2}}(\sigma^2) + (\mu + \sigma^2\theta)e^{\mu\theta+\frac{\sigma^2\theta^2}{2}}(\mu + \sigma^2\theta)$

1 **1**

$M''(0) = e^0(\sigma^2) + (\mu + 0)e^0(\mu + 0) = \sigma^2 + \mu^2$ **1**

depends on both previous marks. If NONE awarded, but a reasonable attempt at $M''(0)$ has been made, award **M1**

$\therefore$ variance = σ^2 **1** **6**

(iv) Convolution theorem **M1**

Mgf of $Y_1 + \ldots + Y_n = e^{\mu_1\theta+\frac{\sigma_1^2\theta^2}{2}} \ldots e^{\mu_n\theta+\frac{\sigma_n^2\theta^2}{2}}$

$= e^{(\mu_1+\ldots+\mu_n)\theta+\frac{(\sigma_1^2+\ldots+\sigma_n^2)\theta^2}{2}}$ **1**

This is the functional form for mgf of **Normal** distribution **E1**

with mean $\mu_1 + \ldots + \mu_n$ **1**

and variance $\sigma_1^2 + \ldots + \sigma_n^2$ **1**

allow as **B1, B1** for direct quotations **5**

Q3 (i) Test of H0:p = 0.8 against H1:p < 0.8, where p is the probability that a temporary employee is satisfactory. **1**

Test statistic is

M1 for use of cty corr (FT if no cty corr)

$$\frac{91\tfrac{1}{2}-96}{\sqrt{120\times0.8\times0.2}} = -1.02(698)$$

M1 **A1** [if no cty corr, ie 91 or $0.758\dot{3}$ used, value is -1.14(109)]

or

$$\frac{0.7625-0.8}{\sqrt{\frac{0.8\times0.2}{120}}}$$

Refer to N(0, 1), lower 5% point is -1.645. **1**

Not significant – can conclude that there is no real evidence against p = 0.8, i.e. no real case for withholding fee. **E2**

99% CI is $\frac{91}{120} \pm 2.576\sqrt{\frac{\frac{91}{120}\cdot\frac{29}{120}}{120}}$ = 0.7583 ± 2.576 √0.001527

M1 **B1** **M1** = 0.7583 ± 2.576 × 0.039

= 0.7583 ± 0.100(67)

= (0.658, 0.859) **A1**

We are "99% confident" that this interval contains the true value of p – 0.8 is in it, giving some support to the agency's claim. **E2** **13**

(ii) We have 91 out of 120 and 72 out of 80. 99% CI for true $p_2 - p_1$ is

$$\frac{72}{80}-\frac{91}{120}\pm1.96\sqrt{\frac{\frac{72}{80}\cdot\frac{8}{80}}{80}+\frac{\frac{91}{120}\cdot\frac{29}{120}}{120}} = 0.141\dot{6}\pm1.96\sqrt{0.001125+0.0015271(=0.0026521)}$$

M1 **B1** two terms **M1** $= 0.141\dot{6}\pm1.96\times0.0515$

both correct **M1** $= 0.141\dot{6}\pm0.1009 = \left(0.040\left(7\dot{6}\right), 0.242\left(5\dot{6}\right)\right)$ **A1**

The lower end of this interval is > 0, which <u>suggests</u> that the new agency is better. **E2** **7**

Q4 (a) $s_1^2 = 80.125$ **B1** [70.109 with divisor n]
$s_2^2 = 46.992$ [43.076 with divisor n]

Test statistic is $\frac{80.125}{46.992}$ = 1.70(51) **1** FT from candidate's values

Refer to $F_{7,11}$ F dist **1**, dF **1** No FT if wrong

Upper 2½% pt is 3.76 **1** no FT if wrong
Not significant **1**
Seems underlying variances are equal **1**
Requires Normality of both populations **1**

$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$ $\left(\text{allow LHS as } \frac{s_1^2}{s_2^2}\right)$ **10**

(b) Critical region is $\bar{x} < \mu_0 - 1.645\frac{\sigma}{\sqrt{n}}$ **M1** i.e. $\bar{x} < 100 - \frac{1.645 \times 6.5}{\sqrt{12}}$ **A1**

i.e. $\bar{x} < 96.913341$ **A1**
[candidates need not display this level of accuracy]

Power function is P(reject H_0 | μ) **M1**
$= P\left(\bar{X} < 96.913341 \middle/ \bar{X} \sim N\left(\mu, \frac{6.5^2}{12}\right)\right)$ **M1**
$= P\left(N(0,1) < \frac{(96.913341 - \mu)\sqrt{12}}{6.5} = 51.648871 - 0.532939\mu\right)$ **1**

μ =	z-value =	prob =	
100	-1.645	1 – 0.95	= 0.05
99	-1.112	1 – 0.8669	= 0.1331
98	-0.5791	1 – 0.7188	= 0.2812
97	-0.0462	1 – 0.5185	= 0.4815
96	0.4868		0.6868
95	1.0197		0.8460

A2 if all correct [**A1** if any four correct]

Power function for perfect test is 0 at μ = 100 **1**, 1 for the other μ values **1** **10**

Examiner’s Report

2617 Statistics 5

General Comments

There were 19 candidates from 7 centres, a substantial reduction compared with the already small entry last year. Happily, however, much of the work was good.

Comments on Individual Questions

Q.1 This question no doubt looked long and formidable but, as is usually the case with such appearances, the many intermediate results gave signposts by which candidates could work steadily and carefully through the several steps. And several candidates did precisely this, meeting with considerable success. The opening probability generating function (for a logarithmic distribution) was found by setting up the required expansion and comparing it with that of $\ln(1 - u)$ that was quoted in the question; most candidates were able to obtain the given answer in a convincing way. Nearly all candidates then knew that part (ii) was a simple application of the convolution theorem. Part (iii) was a similar sort of exercise to part (i), this time with a negative binomial distribution; again it required the expansion for the probability generating function to be compared with a result given in the question, and again most candidates did it convincingly. Parts (iv) and (v) moved into less familiar waters, dealing with the sum of a random number of random variables, put into the context of this question. The question guided candidates in what to do, and most made some progress here; indeed, some were fully successful. Unsurprisingly, though, this was found more difficult than the earlier parts of the question.

Q.2 This question was based on what ought to be standard bookwork for the Normal distribution. Most candidates clearly knew what to do and could do it well; it was, however, rather surprising to find a few attempts with clearly very little idea of how to proceed, despite the standard nature of the tasks. That said, most candidates could obtain the moment generating function of the N(0, 1) distribution by completing the square in the required integral, and then knew how to use the linear transformation result to obtain the moment generating function for $N(0, \sigma^2)$. They could then differentiate this the required number of times to confirm the mean and variance. They were also able to use and interpret the convolution theorem to do the last part.

Q.3 The Normal approximation test for p was usually well done, though there was reluctance to use a continuity correction. The confidence interval was less well done, some candidates using the null hypothesis value of p instead of the sample proportion as the centre of the interval and/or when estimating the standard deviation. Interpretations of the test and the interval in the context of the question were generally reasonably good. The confidence interval for comparing two parameters also gave some problems. A few candidates appeared not to be familiar with this method, providing instead some form of an interval based on just the second proportion but including also the null hypothesis value of the first. Other candidates, while knowing near-enough what to do, could not form the standard deviation properly. However, it should also be said that there were many correct intervals. Again the interpretations in context were usually reasonably good.

Q.4 The *F* test was usually correctly done, though there were some incorrect numbers of degrees of freedom and, independently, some cases where the upper 5% point was given instead of the upper 2½% point (this is a *two-sided* test). Also, some candidates were unable to give the distributional result on which the test is based. In part (b), candidates had first to find the critical region for the test; this was usually done correctly, though there were occasional cases of the inequality being the wrong way round or of a two-sided critical region being given. Such cases were followed through as far as was practicable, though the candidates often made independent mistakes later. Having found the critical region, the power function should have been straightforward to derive and evaluate, being simply the probability of getting a result in the critical region as a function of the parameter μ. Most candidates were indeed able to do this, but various mistakes arose. Knowledge of what a perfect power function would look like was mixed. Most candidates, but not all, knew that it should have two possible values, 0 and 1, but there was some confusion as to when it should be 0 and when it should be 1.