OCR
RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 4769

Statistics 4

Wednesday **24 MAY 2006** Afternoon 1 hour 30 minutes

Additional materials:
8 page answer booklet
Graph paper
MEI Examination Formulae and Tables (MF2)

TIME 1 hour 30 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.
- Final answers should be given to a degree of accuracy appropriate to the context.

INFORMATION FOR CANDIDATES

- The number 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.
- The total number of marks for this paper is 72.

This question paper consists of 5 printed pages and 3 blank pages.

Option 1: Estimation

1 A parcel is weighed, independently, on two scales. The weights are given by the random variables W_1 and W_2 which have underlying Normal distributions as follows.

$$W_1 \sim \mathrm{N}(\mu, \sigma_1^2), \qquad W_2 \sim \mathrm{N}(\mu, \sigma_2^2),$$

where μ is an unknown parameter and σ_1^2 and σ_2^2 are taken as known.

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

$$\hat{\mu} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} W_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} W_2. \qquad [11]$$

[You may quote the probability density function of the general Normal distribution from page 9 in the MEI Examination Formulae and Tables Booklet (MF2).]

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

(iii) Obtain the variance of $\hat{\mu}$. [2]

(iv) A simpler estimator $T = \frac{1}{2}(W_1 + W_2)$ is proposed. Write down the variance of T and hence show that the relative efficiency of T with respect to $\hat{\mu}$ is

$$y = \left(\frac{2\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2} \right)^2 . \qquad [5]$$

(v) Show that $y \leqslant 1$ for all values of σ_1^2 and σ_2^2. Explain why this means that $\hat{\mu}$ is preferable to T as an estimator of μ. [4]

Option 2: Generating Functions

2 [In this question, you may use the result $\int_0^\infty u^m e^{-u}\,du = m!$ for any non-negative integer m.]

The random variable X has probability density function

$$f(x) = \begin{cases} \dfrac{\lambda^{k+1} x^k e^{-\lambda x}}{k!}, & x > 0, \\ 0, & \text{elsewhere,} \end{cases}$$

where $\lambda > 0$ and k is a non-negative integer.

(i) Show that the moment generating function of X is $\left(\dfrac{\lambda}{\lambda - \theta}\right)^{k+1}$. [7]

(ii) The random variable Y is the sum of n independent random variables each distributed as X. Find the moment generating function of Y and hence obtain the mean and variance of Y. [8]

(iii) State the probability density function of Y. [3]

(iv) For the case $\lambda = 1$, $k = 2$ and $n = 5$, it may be shown that the definite integral of the probability density function of Y between limits 10 and ∞ is 0.9165. Calculate the corresponding probability that would be given by a Normal approximation and comment briefly. [6]

Option 3: Inference

3 The human resources department of a large company is investigating two methods, A and B, for training employees to carry out a certain complicated and intricate task.

(i) Two separate random samples of employees who have not previously performed the task are taken. The first sample is of size 10; each of the employees in it is trained by method A. The second sample is of size 12; each of the employees in it is trained by method B. After completing the training, the time for each employee to carry out the task is measured, in controlled conditions. The times are as follows, in minutes.

Employees trained by method A:	35.2	47.8	25.8	38.0	53.6	31.0	33.9
	35.4	21.6	42.5				
Employees trained by method B:	43.0	57.5	68.6	20.9	31.4	44.9	62.8
	27.6	41.8	46.1	39.8	61.6		

Stating appropriate assumptions concerning the underlying populations, use a t test at the 5% significance level to examine whether either training method is better in respect of leading, on the whole, to a lower time to carry out the task. [12]

(ii) A further trial of method B is carried out to see if the performance of experienced and skilled workers can be improved by re-training them. A random sample of 8 such workers is taken. The times in minutes, under controlled conditions, for each worker to carry out the task before and after re-training are as follows.

Worker	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
Time before	32.6	28.5	22.9	27.6	34.9	28.8	34.2	31.3
Time after	26.2	24.1	19.0	28.6	29.3	20.0	36.0	19.2

Stating an appropriate assumption, use a t test at the 5% significance level to examine whether the re-training appears, on the whole, to lead to a lower time to carry out the task. [10]

(iii) Explain how the test procedure in part **(ii)** is enhanced by designing it as a paired comparison. [2]

Option 4: Design and Analysis of Experiments

4 An experiment is carried out to compare five industrial paints, A, B, C, D, E, that are intended to be used to protect exterior surfaces in polluted urban environments. Five different types of surface (I, II, III, IV, V) are to be used in the experiment, and five specimens of each type of surface are available. Five different external locations (1, 2, 3, 4, 5) are used in the experiment.

The paints are applied to the specimens of the surfaces which are then left in the locations for a period of six months. At the end of this period, a "score" is given to indicate how effective the paint has been in protecting the surface.

(i) Name a suitable experimental design for this trial and give an example of an experimental layout. [3]

Initial analysis of the data indicates that any differences between the types of surface are negligible, as also are any differences between the locations. It is therefore decided to analyse the data by one-way analysis of variance.

(ii) State the usual model, including the accompanying distributional assumptions, for the one-way analysis of variance. Interpret the terms in the model. [9]

(iii) The data for analysis are as follows. Higher scores indicate better performance.

Paint A	Paint B	Paint C	Paint D	Paint E
64	66	59	65	64
58	68	56	78	52
73	76	69	69	56
60	70	60	72	61
67	71	63	71	58

[The sum of these data items is 1626 and the sum of their squares is 106 838.]

Construct the usual one-way analysis of variance table. Carry out the appropriate test, using a 5% significance level. Report briefly on your conclusions. [12]

Mark Scheme 4769
June 2006

Q1				
(i)	$L = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(W_1-\mu)^2}{2\sigma_1^2}} \cdot \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(W_2-\mu)^2}{2\sigma_2^2}}$	M1 M1 A1	Product form. Two Normal terms. Fully correct.	
	$\ln L = \text{const} - \frac{1}{2\sigma_1^2}(W_1-\mu)^2 - \frac{1}{2\sigma_2^2}(W_2-\mu)^2$	M1 A1		
	$\frac{d \ln L}{d\mu} = \frac{2}{2\sigma_1^2}(W_1-\mu) + \frac{2}{2\sigma_2^2}(W_2-\mu)$	M1 A1	Differentiate w.r.t. μ.	
	$= 0 \Rightarrow \sigma_2^2 W_1 - \sigma_2^2\mu + \sigma_1^2 W_2 - \sigma_1^2\mu = 0$	A1		
	$\Rightarrow \hat{\mu} = \frac{\sigma_2^2 W_1 + \sigma_1^2 W_2}{\sigma_1^2 + \sigma_2^2}$	A1	BEWARE PRINTED ANSWER.	
	Check this is a maximum.	M1		
	E.g. $\frac{d^2 \ln L}{d\mu^2} = -\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} < 0$	A1		11
(ii)	$E(\hat{\mu}) = \frac{\sigma_2^2\mu + \sigma_1^2\mu}{\sigma_1^2 + \sigma_2^2} = \mu$	M1		
	$\therefore$ unbiased.	A1		2
(iii)	$\text{Var}(\hat{\mu}) = \left(\frac{1}{\sigma_1^2 + \sigma_1^2}\right)^2 \cdot (\sigma_2^4\sigma_1^2 + \sigma_1^4\sigma_2^2)$	B1 B1	First factor. Second factor.	
	$= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$		Simplification not required at this point.	2
(iv)	$T = \frac{1}{2}(W_1 + W_2)$			
	$\text{Var}(T) = \frac{1}{4}(\sigma_1^2 + \sigma_2^2)$	B1		
	Relative efficiency (y) $= \frac{\text{Var}(\hat{\mu})}{\text{Var}(T)}$	M1 M1	Any attempt to compare variances. If correct.	
	$= \frac{\sigma_2^4\sigma_1^2 + \sigma_1^4\sigma_2^2}{(\sigma_1^2 + \sigma_2^2)^2} \cdot \frac{4}{\sigma_1^2 + \sigma_2^2}$	A1		
	$= \frac{4\sigma_1^2\sigma_2^2}{(\sigma_1^2 + \sigma_2^2)^2}$	A1	BEWARE PRINTED ANSWER.	5
(v)	E.g. consider $\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2 = (\sigma_1 - \sigma_2)^2 \geq 0$	M1		
	$\therefore$ Denominator ≥ numerator, $\therefore$ fraction ≤ 1	E1		
	[Both $\hat{\mu}$ and T are unbiased,] $\hat{\mu}$ has smaller variance than T and is therefore better.	E1 E1		4
				24

Q2	$f(x)=\frac{\lambda^{k+1}x^k e^{-\lambda x}}{k!}$, $[x>0 \quad (\lambda>0,\ k \text{ integer} \geq 0)]$ Given: $\int_0^\infty u^m e^{-u}du = m!$			
(i)	$M_X(\theta)=E[e^{\theta x}]$ $=\int_0^\infty \frac{\lambda^{k+1}}{k!}x^k e^{-(\lambda-\theta)x}dx$ Put $(\lambda-\theta)x=u$ $=\frac{\lambda^{k+1}}{k!(\lambda-\theta)^{k+1}}\int_0^\infty u^k e^{-u}du$ $=\left(\frac{\lambda}{\lambda-\theta}\right)^{k+1}$	M1 M1 M1 A1 A1 A1 A1	For obtaining this expression after substitution. Take out constants. (Dep on subst.) Apply "given": integral = k! (Dep on subst.) BEWARE PRINTED ANSWER.	7
(ii)	$Y = X_1 + X_2 + \ldots + X_n$ By convolution theorem:- mgf of Y is $\{M_X(\theta)\}^n$ i.e. $\left(\frac{\lambda}{\lambda-\theta}\right)^{nk+n}$ $\mu=M'(0)$ $M'(\theta)=\lambda^{nk+n}(-nk-n)(\lambda-\theta)^{-nk-n-1}(-1)$ $\therefore \mu=\frac{nk+n}{\lambda}$ $\sigma^2=M''(0)-\mu^2$ $M''(\theta)=(nk+n)\lambda^{nk+n}(-nk-n-1)(\lambda-\theta)^{-nk-n-2}(-1)$ $\therefore M''(0)=(nk+n)(nk+n+1)/\lambda^2$ $\therefore \sigma^2=\frac{(nk+n)(nk+n+1)}{\lambda^2}-\frac{(nk+n)^2}{\lambda^2}$ $=\frac{nk+n}{\lambda^2}$	B1 M1 A1 A1 M1 A1 M1 A1		8
(iii)	[Note that $M_Y(t)$ is of the same functional form as $M_X(t)$ with k + 1 replaced by nk + n, i.e. k replaced by nk + n –1. This must also be true of the pdf.] Pdf of Y is $\frac{\lambda^{nk+n}}{(nk+n-1)!}\times y^{nk+n-1}\times e^{-\lambda y}$ [for y > 0]	B1 B1 B1	One mark for each factor of the expression. Mark for third factor shown here depends on at least one of the other two earned.	3
(iv)	λ = 1, k = 2, n = 5, Exact P(Y > 10) = 0·9165 Use of N(15, 15)	M1 M1	Mean. ft (ii). Variance. ft (ii).	

	P(this > 10) = $P\left(N(0,1) > \frac{10-15}{\sqrt{15}} = -1\cdot291\right)$ = 0·9017 Reasonably good agreement – CLT working for only small *n*.	A1 A1 E2	c.a.o. c.a.o. (E1, E1) [Or other sensible comments.]	 6
				24

Q3				
(i)	$\bar{x} = 36.48$ $s = 9.6307$ $s^2 = 92.7507$ $\bar{y} = 45.5$ $s = 14.8129$ $s^2 = 219.4218$	B1	If all correct. [No marks for use of s_n which are 9.1365 and 14.1823 respectively.]	
	Assumptions: Normality of <u>both</u> populations equal variances	B1 B1		
	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A \neq \mu_B$	B1	Do <u>NOT</u> accept $\bar{X} = \bar{Y}$ or similar.	
	Where μ_A, μ_B are the population means.	B1		
	Pooled $s^2 = \dfrac{9 \times 92.7507 + 11 \times 219.4218}{20}$ $= \dfrac{834.756 + 24136.64}{20} = 162.4198$	B1	$= (12.7444)^2$	
	Test statistic is $\dfrac{36.48 - 45.5}{\sqrt{162.4198}\sqrt{\frac{1}{10} + \frac{1}{12}}}$	M1		
	$= \dfrac{-9.02}{5.4568} = -1.653$	A1		
	Refer to t_{20}.	M1	No ft from here if wrong.	
	Double tailed 5% point is 2·086.	A1	No ft from here if wrong.	
	Not significant.	A1	ft only c's test statistic.	
	No evidence that population mean times differ.	A1	ft only c's test statistic.	12
(ii)	Assumption: Normality of underlying population of <u>differences</u>.	B1		
	$H_0 : \mu_D = 0$ $H_1 : \mu_D > 0$	B1	Do <u>NOT</u> accept $\bar{D} = 0$ or similar.	
	Where μ_D is the population mean of "before – after" differences.	B1	The "<u>direction</u>" of D must be CLEAR. Allow $\mu_A = \mu_B$ etc.	
	Differences are 6.4, 4.4, 3.9, -1.0, 5.6, 8.8, -1.8, 12.1	M1		
	($\bar{x} = 4.8$ $s = 4.6393$)		[A1 can be awarded here if NOT awarded in part (i)]. Use of s_n (=4.3396) is <u>NOT</u> acceptable, even in a denominator of $\dfrac{s_n}{\sqrt{n-1}}$	
	Test statistic is $\dfrac{4.8 - 0}{4.6393/\sqrt{8}}$	M1		
	=2.92(64)	A1		
	Refer to t_7.	M1	No ft from here if wrong.	
	Single tailed 5% point is 1.895.	A1	No ft from here if wrong.	
	Significant.	A1	ft only c's test statistic.	
	Seems mean is lowered.	A1	ft only c's test statistic.	10
(iii)	The paired comparison in part (ii) eliminates the variability between workers.	E2	(E1, E1)	2
				24

<table>
<tr><td>Q4</td><td></td><td></td><td></td><td></td></tr>
<tr><td>(i)</td><td>Latin square.

Layout such as:

Locations: 1 2 3 4 5
Surfaces I: A B C D E
II: B C D E A
III: C D E A B
IV: D E A B C
V: E A B C D</td><td>B1

B1
B1</td><td>(letters = paints)
Correct rows and columns.
A correct arrangement of letters.
SC. For a description instead of an example allow max 1 out of 2.</td><td>3</td></tr>
<tr><td>(ii)</td><td>$X_{ij} = \mu + \alpha_i + e_{ij}$

μ = population
grand mean for whole experiment.

α_i = population
mean amount by which the i^{th} treatment differs from μ.

e_{ij} are experimental errors
~ ind
$N(0, \sigma^2)$.</td><td>B1

B1
B1

B1
B1

B1
B1
B1
B1</td><td>

Allow "uncorrelated".
Mean.
Variance.</td><td>9</td></tr>
<tr><td>(iii)</td><td>Totals are: 322, 351, 307, 355, 291
(each from sample of size 5)
Grand total: 1626
"Correction factor" CF = $\frac{1626^2}{25} = 105755.04$

Total SS = 106838 – CF = 1082.96
Between paints SS = $\frac{322^2}{5} + ... + \frac{291^2}{5} - \text{CF}$
= 106368 – CF =612.96
Residual SS (by subtraction) = 1082.96 – 612.96
= 470.00

Source of variation | SS | df | MS
Between paints | 612.96 | 4 | 153.24
Residual | 470.00 | 20 | 23.5
Total | 1082.96 | 24 |

MS ratio = $\frac{153.24}{23.5} = 6.52$</td><td>M1
M1

A1

B1
B1
M1

M1
A1</td><td>
For correct methods for any two SS.
If each calculated SS is correct.

Degrees of freedom "between paints".
Degrees of freedom "residual".
MS column.

Independent of previous M1.
Dep only on this M1.</td><td></td></tr>
</table>

	Refer to $F_{4,\ 20}$ Upper 5% point is 2.87 Significant. Seems performances of paints are not all the same.	M1 A1 A1 A1	No ft if wrong. But allow ft of wrong d.o.f. above. No ft if wrong. ft only c's test statistic and d.o.f.'s. ft only c's test statistic and d.o.f.'s.	12
				24

4769 - Statistics 4

General Comments

This is the first time that the new-specification Statistics 4 module has been sat. It is now the highest module in the statistics strand of the MEI specification; its content is a selection of material from the higher modules of the old specification. The new rules under which the present specification must operate mean that opportunities to proceed to high levels in the applied mathematics strands are very limited; so it is good to see that numbers proceeding to the highest level in statistics are holding up well.

There was some very good work, but also some candidates who were perhaps not quite ready to take the examination.

The paper consists of four questions, each within a defined "option" area of the specification. The rubric requires that three be attempted. All four questions received many attempts – another encouraging feature, as it indicates that centres and candidates are spreading their work over all the options.

There were in fact several candidates who attempted all four questions. Sometimes they deleted one but, whether they did that or not, all four were marked and the best three counted. It needs to be said, and not for the first time, that in general it is not a wise policy to attempt all the questions. Of course it sometimes happens that a solution "goes wrong" and the candidate decides to give it up and proceed to another question. As a tactic, that is acceptable. But candidates should not set out with the *strategy* of expecting to attempt more questions than are required. It is far better to try to produce the required number of nearly-complete answers than a surfeit of fragments.

Another general point must again be made. There were *again* several instances of "fudging" of answers that were provided in the question. Such answers are provided partly as a reassurance and check on work so far, but mainly so that they can be used in the rest of the question even by candidates who could not derive them. It is no shame whatever to *use* a given answer in this way. A number of candidates did so in an honest and open way. In some cases, it appeared that they did not know how to derive the given answer at all, and in other cases they were let down by algebraic errors, commonly noting (sometimes in a humorous way) that something must have gone wrong somewhere. *This is entirely acceptable as examination technique*. What is *not* acceptable is faking the answer: commonly done using the magic disappearing minus sign, or by arriving at an incorrect algebraic statement (often *grossly* incorrect) and then merely stating that it equals the given answer as though it is hoped that the examiner won't notice.

Comments on Individual Questions

1) This was on the "estimation" option. The first part required a maximum likelihood estimator to be found; the remaining parts sought its mean and variance, followed by comparison with another estimator using the efficiency criterion.

   There were many excellent solutions to the first part, but a substantial minority of candidates had problems here. Some clearly did not know what to do at all; others had some idea but ran into problems right from the beginning; and others made a good start but were then let down by poor technique. Maximum likelihood was of course in the sixth statistics module of the old specification; candidates who wish to offer a solution to this option now must ensure that they are adequately prepared and practised in the work. Maximum likelihood is a fairly advanced concept but not particularly difficult technically provided one is careful and thorough in one's work.

The middle parts of the question were usually well done, even by candidates who could not derive the maximum likelihood estimator in part (i) [with reference to remarks above, note that this is a case where the answer was deliberately given so that it could be used]. Most, but not all, candidates knew how to work out the relative efficiency in part (iv), but there was insecurity in part (v). It was not enough just to *aver* that $y \leq 1$; the question says "show that", and some sort of convincing explanation was required. Many candidates understood that this result meant that the maximum likelihood estimator was preferable due to having smaller variance, but in some scripts the explanation was not fully complete.

2) This was on the "generating functions" option. It was primarily about obtaining and using a moment generating function.

Many candidates carefully and thoroughly obtained the given answer in part (i), and other candidates made a good start but then made mistakes so that the answer did not come out. However, this was one of the places where there was quite a lot of faking.

In part (ii), most candidates knew that the convolution theorem gave the moment generating function of the Y variable straight away. Obtaining the mean and variance of Y from its moment generating function was also usually done well, though many candidates were clumsy in their technique for differentiation – disappointingly so, in what must be a "Further Mathematics" module. Note that the question includes the explicit word "hence"; other methods of finding the mean and variance were not acceptable.

Very few candidates were able to answer part (iii). The result (see the published mark scheme), which is very simple, seemed not to be known.

In part (iv), candidates' commentaries on the accuracy of the Normal approximation were often very insightful. Some reference to the value of n was expected if full marks were to be obtained. The published mark scheme is based on "remarkably good agreement"; some candidates, having made earlier errors, did not get good agreement here, but their work was followed through.

3) This question was on the "inference" option. It included unpaired and paired t tests.

There were a few candidates who used some form of Wilcoxon test in one part of the question or the other, despite the explicit instruction in each part to use a t test. Unfortunately these candidates lost marks quite heavily. This also applied to candidates who used a wrong type of t test (e.g. an unpaired test in part (ii)).

Usually the work was well done. There was some insecurity in stating assumptions and hypotheses. In part (iii), the point being looked for in the discussion was that the pairing eliminated variability between workers; many candidates made that point, but others lost their way in statements about the nature of the estimated standard deviations.

4) This was on the "design and analysis of experiments" option. Most candidates realised that the required design was a Latin square and produced an example of one; a few candidates however were besotted with randomised blocks. In the next part, the formal statement of the model was sometimes very carefully set out, but many candidates were not quite complete in this.

The analysis in the last part was usually done well. However, another point that can *again* be made is that there were many candidates who were very inefficient in their calculations. This appeared to have been getting better over the last few years of the old specification, but this year has taken a turn for the worse again. What might be called the "s_b^2/s_w^2" method is *extremely* cumbersome for hand calculation. It is intricate, takes a great deal of time, and is liable to produce errors. The "squared totals" method (as exhibited, somewhat in summary form, in the published mark scheme) is *very* much better for hand calculation.

[Incidentally, it also appeared that there were candidates who were able to read the required values directly from their calculators. These candidates must be careful to get the values *right* (i.e. no keying errors), for no method marks can be given if there is only an unsupported numerical answer that happens to be wrong.]