OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE UNIT **4769/01**

MATHEMATICS (MEI)

Statistics 4

TUESDAY 5 JUNE 2007

Afternoon

Time: 1 hour 30 minutes

Additional Materials:
Answer booklet (8 pages)
Graph paper
MEI Examination Formulae and Tables (MF2)

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

ADVICE TO CANDIDATES

- Read each question carefully and make sure you know what you have to do before starting your answer.
- 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.

This document consists of **4** printed pages.

[Turn over

Option 1: Estimation

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

$$\mathrm{f}(x) = \frac{1}{\theta}, \qquad 0 \leqslant x \leqslant \theta,$$

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

A random sample of n observations from X is denoted by $X_1, X_2, \ldots, X_n$, with sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(i) Show that $2\overline{X}$ is an unbiased estimator of θ. [4]

(ii) Evaluate $2\overline{X}$ for a case where, with $n = 5$, the observed values of the random sample are 0.4, 0.2, 1.0, 0.1, 0.6. Hence comment on a disadvantage of $2\overline{X}$ as an estimator of θ. [4]

For a general random sample of size n, let Y represent the sample maximum, $Y = \max(X_1, X_2, \ldots, X_n)$. You are given that the probability density function of Y is

$$\mathrm{g}(y) = \frac{ny^{n-1}}{\theta^n}, \qquad 0 \leqslant y \leqslant \theta.$$

(iii) An estimator kY is to be used to estimate θ, where k is a constant to be chosen. Show that the mean square error of kY is

$$k^2\mathrm{E}(Y^2) - 2k\theta\mathrm{E}(Y) + \theta^2$$

and hence find the value of k for which the mean square error is minimised. [12]

(iv) Comment on whether kY with the value of k found in part **(iii)** suffers from the disadvantage identified in part **(ii)**. [4]

Option 2: Generating Functions

2 The random variable X has the binomial distribution with parameters n and p, i.e. $X \sim \mathrm{B}(n, p)$.

(i) Show that the probability generating function of X is $\mathrm{G}(t) = (q + pt)^n$, where $q = 1 - p$. [4]

(ii) Hence obtain the mean μ and variance σ^2 of X. [6]

(iii) Write down the mean and variance of the random variable $Z = \dfrac{X - \mu}{\sigma}$. [1]

(iv) Write down the moment generating function of X and use the linear transformation result to show that the moment generating function of Z is

$$\mathrm{M}_Z(\theta) = \left(q\mathrm{e}^{-\frac{p\theta}{\sqrt{npq}}} + p\mathrm{e}^{\frac{q\theta}{\sqrt{npq}}}\right)^n. \qquad [5]$$

(v) By expanding the exponential terms in $\mathrm{M}_Z(\theta)$, show that the limit of $\mathrm{M}_Z(\theta)$ as $n \to \infty$ is $\mathrm{e}^{\theta^2/2}$.

You may use the result $\lim\limits_{n\to\infty}\left(1 + \dfrac{y + \mathrm{f}(n)}{n}\right)^n = \mathrm{e}^y$ provided $\mathrm{f}(n) \to 0$ as $n \to \infty$. [4]

(vi) What does the result in part **(v)** imply about the distribution of Z as $n \to \infty$? Explain your reasoning briefly. [3]

(vii) What does the result in part **(vi)** imply about the distribution of X as $n \to \infty$? [1]

Option 3: Inference

3 An engineering company buys a certain type of component from two suppliers, A and B. It is important that, on the whole, the strengths of these components are the same from both suppliers. The company can measure the strengths in its laboratory. Random samples of seven components from supplier A and five from supplier B give the following strengths, in a convenient unit.

Supplier A	25.8	27.4	26.2	23.5	28.3	26.4	27.2
Supplier B	25.6	24.9	23.7	25.8	26.9		

The underlying distributions of strengths are assumed to be Normal for both suppliers, with variances 2.45 for supplier A and 1.40 for supplier B.

(i) Test at the 5% level of significance whether it is reasonable to assume that the mean strengths from the two suppliers are equal. [10]

(ii) Provide a two-sided 90% confidence interval for the true mean difference. [4]

(iii) Show that the test procedure used in part **(i)**, with samples of sizes 7 and 5 and a 5% significance level, leads to acceptance of the null hypothesis of equal means if $-1.556 < \bar{x} - \bar{y} < 1.556$, where $\bar{x}$ and $\bar{y}$ are the observed sample means from suppliers A and B. Hence find the probability of a Type II error for this test procedure if in fact the true mean strength from supplier A is 2.0 units more than that from supplier B. [7]

(iv) A manager suggests that the Wilcoxon rank sum test should be used instead, comparing the median strengths for the samples of sizes 7 and 5. Give one reason why this suggestion might be sensible and two why it might not. [3]

Option 4: Design and Analysis of Experiments

4 An agricultural company conducts a trial of five fertilisers (A, B, C, D, E) in an experimental field at its research station. The fertilisers are applied to plots of the field according to a completely randomised design. The yields of the crop from the plots, measured in a standard unit, are analysed by the one-way analysis of variance, from which it appears that there are no real differences among the effects of the fertilisers.

A statistician notes that the residual mean square in the analysis of variance is considerably larger than had been anticipated from knowledge of the general behaviour of the crop, and therefore suspects that there is some inadequacy in the design of the trial.

(i) Explain briefly why the statistician should be suspicious of the design. [2]

(ii) Explain briefly why an inflated residual leads to difficulty in interpreting the results of the analysis of variance, in particular that the null hypothesis is more likely to be accepted erroneously. [3]

Further investigation indicates that the soil at the west side of the experimental field is naturally more fertile than that at the east side, with a consistent 'fertility gradient' from west to east.

(iii) What experimental design can accommodate this feature? Provide a simple diagram of the experimental field indicating a suitable layout. [4]

The company decides to conduct a new trial in its glasshouse, where experimental conditions can be controlled so that a completely randomised design is appropriate. The yields are as follows.

Fertiliser A	Fertiliser B	Fertiliser C	Fertiliser D	Fertiliser E
23.6	26.0	18.8	29.0	17.7
18.2	35.3	16.7	37.2	16.5
32.4	30.5	23.0	32.6	12.8
20.8	31.4	28.3	31.4	20.4

[The sum of these data items is 502.6 and the sum of their squares is 13 610.22.]

(iv) 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]

(v) State the assumptions about the distribution of the experimental error that underlie your analysis in part **(iv)**. [3]

Mark Scheme 4769
June 2007

1) (i)	$\text{f}(x) = \dfrac{1}{\theta} \qquad 0 \le x \le \theta$ $\text{E}[X] = \dfrac{\theta}{2}$ $\text{E}[2\overline{X}] = 2\text{E}[\overline{X}] = 2\text{E}[X]$ $= \theta$ $\therefore$ unbiased	 B1 M1 A1 E1	 Write-down, or by symmetry, or by integration.	4
(ii)	$\sum x = 2.3 \qquad \therefore\ \bar{x} = \dfrac{2.3}{5} = 0.46 \quad \therefore 2\,\bar{x} = 0.92$ But we know $\theta \ge 1$ $\therefore$ estimator can give nonsense answers, i.e. essentially useless	B1 E1 E2	 (E1, E1)	4
(iii)	$Y = \max\{X_i\}$, $\text{g}(y) = \dfrac{ny^{n-1}}{\theta^n} \qquad 0 \le y \le \theta$ MSE $(kY) = \text{E}[(kY-\theta)^2] =$ $\text{E}[k^2Y^2 - 2k\theta\,Y + \theta^2] =$ $k^2\text{E}[Y^2] - 2k\theta\,\text{E}[Y] + \theta^2$ $\dfrac{d\text{MSE}}{dk} =$ $2k\text{E}[Y^2] - 2\theta\,\text{E}[Y] = 0$ for $k = \dfrac{\theta\,\text{E}[Y]}{\text{E}[Y^2]}$ $\dfrac{d^2\text{MSE}}{dk^2} = 2\text{E}[Y^2] > 0 \ \therefore$ this is a minimum $\text{E}[Y] = \displaystyle\int_0^\theta \frac{ny^n}{\theta^n}\,\text{d}y = \frac{n}{\theta^n}\frac{\theta^{n+1}}{n+1} = \frac{n\theta}{n+1}$ $\text{E}[Y^2] = \displaystyle\int_0^\theta \frac{ny^{n+1}}{\theta^n}\,\text{d}y = \frac{n}{\theta^n}\frac{\theta^{n+2}}{n+2} = \frac{n\theta^2}{n+2}$ $\therefore$ minimising $k = \theta\,\dfrac{n\theta}{n+1}\,\dfrac{n+2}{n\theta^2} = \dfrac{n+2}{n+1}$	 M1 1 M1 M1 A1 M1 M1 A1 M1 A1 M1 A1	 BEWARE PRINTED ANSWER	12
(iv)	With this k, kY is always greater than the sample maximum So it does not suffer from the disadvantage in part (ii)	E2 E2	(E1 E1) (E1 E1)	4

2(i)	$\mathrm{G}(t)=\mathrm{E}[t^X]=\sum_{x=0}^{n}\binom{n}{x}(pt)^x(1-p)^{n-x}$ $=[(1-p)+pt]^n$ $=(q+pt)^n$	M1 2 1	Available as B2 for write-down or as 1+1 for algebra	4
(ii)	$\mu=\mathrm{G}'(1)\quad \mathrm{G}'(t)=np(q+pt)^{n-1}$ $\mathrm{G}'(1)=np\times 1=np$ $\sigma^2=\mathrm{G}''(1)+\mu-\mu^2$ $\mathrm{G}''(t)=n(n-1)p^2(q+pt)^{n-2}$ $\mathrm{G}''(1)=n(n-1)p^2$ $\therefore \sigma^2=n^2p^2-np^2+np-n^2p^2$ $=-np^2+np=npq$	1 1 1 1 M1 1		6
(iii)	$Z=\frac{X-\mu}{\sigma}$ Mean 0, Variance 1	B1	For BOTH	1
(iv)	$\mathrm{M}(\theta)=\mathrm{G}(e^\theta)=(q+pe^\theta)^n$ $Z=aX+b$ with: $a=\frac{1}{\sigma}=\frac{1}{\sqrt{npq}}$ and $b=-\frac{\mu}{\sigma}=-\sqrt{\frac{np}{q}}$ $\mathrm{M}_Z(\theta)=e^{b\theta}\mathrm{M}_X(a\theta)$ $\therefore \mathrm{M}_Z(\theta)=e^{-\sqrt{\frac{np}{q}}\theta}\left(q+pe^{\frac{1}{\sqrt{npq}}\theta}\right)^n=$ - - - - - $\left(qe^{-\frac{p\theta}{\sqrt{npq}}}+pe^{\frac{1-p}{\sqrt{npq}}\theta}\right)^n$	1 M1 1 - - - 1 1	BEWARE PRINTED ANSWER	5
(v)	$\mathrm{M}_Z(\theta)=(q-\frac{qp\theta}{\sqrt{npq}}+\frac{qp^2\theta^2}{2npq}+$ terms in $n^{-3/2}$, n^{-2}, ………….. + $p+\frac{pq\theta}{\sqrt{npq}}+\frac{pq^2\theta^2}{2npq}+\cdots\cdots)^n=$ $(1+\frac{\theta^2}{2n}+\cdots\cdots)^n\to$ $e^{\theta^2/2}$	M1 M1 1 1	For expansion of exponential terms For indication that these can be neglected as $n\to\infty$. Use of result given in question	4

(vi)	N(0,1) Because $e^{\theta^2/2}$ is the mgf of N(0,1) and the relationship between distributions and their mgfs is unique	1 E1 E1		3
(vii)	"Unstandardising", $\mathrm{N}(\mu,\sigma^2)$ ie $\mathrm{N}(np, npq)$	1	Parameters need to be given.	1

3(i)	$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A \neq \mu_B$ Where μ_A, μ_B are the population means Test statistic $\frac{26.4 - 25.38}{\sqrt{\frac{2.45}{7} + \frac{1.40}{5}}} =$ $\frac{1.02}{\sqrt{0.63} = 0.7937} = 1.285$ Refer to N(0,1) Double-tailed 5% point is 1.96 Not significant No evidence that the population means differ	1 1 M1 M1 M1 A1 1 1 1 1	Do NOT allow $\overline{X} = \overline{Y}$ or similar Accept absence of "population" if correct notation μ is used. Hypotheses stated verbally must include the word "population". Numerator Denominator two separate terms correct No FT if wrong No FT if wrong	10
(ii)	CI (for $\mu_A - \mu_B$) is $1.02 \pm$ $1.645 \times$ $0.7937 =$ $1.02 \pm 1.3056 =$ (– 0.2856, 2.3256)	M1 B1 M1 A1 cao	Zero out of 4 if not N(0,1)	4
(iii)	H_0 is accepted if –1.96< test statistic < 1.96 i.e. if $-1.96 < \frac{\bar{x} - \bar{y}}{0.7937} < 1.96$ i.e. if $-1.556 < \bar{x} - \bar{y} < 1.556$ In fact, $\overline{X} - \overline{Y} \sim \mathrm{N}(2, 0.7937^2)$ So we want $\mathrm{P}(-1.556 < \mathrm{N}(2, 0.7937^2) < 1.556) =$ $\mathrm{P}\left(\frac{-1.556 - 2}{0.7937} < \mathrm{N}(0,1) < \frac{1.556 - 2}{0.7937}\right) =$ $\mathrm{P}(-4.48 < \mathrm{N}(0,1) < -0.5594) = 0.2879$	M1 M1 A1 M1 M1 M1 A1 cao	SC1 Same wrong test can get M1,M1,A0. SC2 Use of 1.645 gets 2 out of 3. BEWARE PRINTED ANSWER Standardising	7
(iv)	Wilcoxon would give protection if assumption of Normality is wrong. Wilcoxon could not really be applied if underlying variances are indeed different. Wilcoxon would be less powerful (worse Type II error behaviour) with such small samples if Normality is correct.	E1 E1 E1		3

<table>
<tr><td>4 (i)</td><td>There might be some consistent source of plot-to-plot variation that has inflated the residual and which the design has failed to cater for.</td><td>E2</td><td>E1 – Some reference to extra variation.
E1 – Some indication of a reason.</td><td>2</td></tr>
<tr><td>(ii)</td><td>Variation between the fertilisers should be compared with experimental error.

If the residual is inflated so that it measures more than experimental error, the comparison of between - fertilisers variation with it is less likely to reach significance.</td><td>E1

E2</td><td>

(E1, E1)</td><td>3</td></tr>
<tr><td>(iii)</td><td>Randomised blocks
<table><tr><td>C
B
A
D
E</td><td>.
.
.</td><td>.
.
.</td></tr></table>
SPECIAL CASE: Latin Square $\frac{2}{4}$ (1, E1)</td><td>1

E1

E1

E1</td><td>

Blocks (strips) clearly correctly oriented w.r.t. fertiliser gradient.

All fertilisers appear in a block.

Different (random) arrangements in the blocks.</td><td>4</td></tr>
<tr><td>(iv)</td><td>Totals are: 95.0 123.2 86.8 130.2 67.4
(each from sample of size 4)
Grand total 502.6
“Correction factor” CF = $\frac{502.6^2}{20} = 12630.338$
Total SS = 13610.22 – CF = 979.882

Between fertilisers SS =
$\frac{95.0^2}{4} + ... + \frac{67.4^2}{4}$ – CF =
13308.07 – CF = 677.732

Residual SS (by subtraction) =
979.882 – 677.732=302.15
<table><tr><td>Source of variation</td><td>SS</td><td>df</td><td>MS</td><td>MS Ratio</td></tr><tr><td>Between fertiliser
Residual</td><td>677.732
302.15</td><td>4
15</td><td>169.433
20.143</td><td>8.41</td></tr><tr><td>Total</td><td>979.882</td><td>19</td><td></td><td></td></tr></table>Refer to $F_{4,\,15}$
-upper 5% point is 3.06
Significant
- seems effects of fertilisers are not all the same</td><td>M1

M1

A1

M1
M1
1,A1
1

1
1
1
1</td><td>

For correct method for any two

If each calculated SS is correct

No FT if wrong
No FT if wrong</td><td>12</td></tr>
<tr><td>(vii)</td><td>Independent N (0, σ^2 [constant])</td><td>1
1
1</td><td></td><td>3</td></tr>
</table>

4769: Statistics 4

General Comments

This is the second time that the new-specification Statistics 4 module has been sat. Although the entry is small, it is pleasing that the opportunity to proceed to high levels in the applied mathematics strands is still available.

There was some extremely good work, and only a little very poor work.

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.

Sadly there were again cases of "faking" of answers that were given within the questions. This was discussed at some length in last year's report. This year, I will merely reiterate that it is *entirely unacceptable*.

Comments on Individual Questions

1 This was on the "estimation" option. It consisted of comparison of two estimators of θ for the uniform distribution on (0, θ).

The first of those estimators was $2\bar{X}$. Most candidates showed quickly enough that it was unbiased, but surprisingly many did not spot that, with the sample data given, its value was 0.92 even though we *knew* that θ must be at least 1, thus making it a fairly useless estimator! Candidates tended instead to struggle in making unconvincing comments about its variance. The question then moved on to a new estimator whose mean square error was to be found; this was usually done fairly successfully, some candidates being much more efficient in their work than others, and some not really being able to cope at all. Candidates who had spotted the key disadvantage of the first estimator were usually able to see that the new estimator could not possibly suffer from it, but others struggled to find anything sensible to say.

2 This was on the "generating functions" option. It led candidates through the steps of proving that the limiting distribution of the B(n, p) random variable as $n \to \infty$ is N(np, npq).

Most candidates proceeded thoroughly and carefully through the technical mathematical work, much of which should have been standard bookwork. However, surprisingly many could not simply write down 0 and 1 for the mean and variance in part (iii). In part (iv), several candidates were rescued, with greater or less legitimacy, by the provision in the question of the answer. In part (v), some candidates did not realise that the first step towards the limiting result was to expand the exponential terms from part (iv). Most, however, did this quite well, sometimes not being entirely convincing in their use of the result given in the question (simply averring that their version of the f(n) in that result was actually *equal to* zero rather missed the point).

3 This question was on the "inference" option. It was based on an unpaired Normal test, proceeding to consideration of Type II error.

Mostly the test and confidence interval (parts (i) and (ii)) were well done, though some of the usual errors did appear from time to time. Part (iii) met with mixed success; it was done very well by some candidates, whereas others fell by the wayside en route. Surprisingly many failed to consider both "tails" in finding the last probability; though one of them turns out to be negligible in the extreme, this cannot be known until there has been some investigation of it! A variety of suggestions in favour of and against the Wilcoxon alternative came forward in part (iv).

4 This was on the "design and analysis of experiments" option.

It opened with some important considerations of experimental design. Some candidates showed good appreciation of the points here; others did not. In part (iii), the required design was randomised blocks (correctly oriented with respect to the fertility gradient); some credit was allowed for suggestions of Latin squares, though that design is not really appropriate here as it is too complicated for the situation.

The analysis in the last part was usually done well. However, the point must *yet again* be made that many candidates were very inefficient in their calculations. This is definitely getting worse. 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. It is poor practice. The "squared totals" method (as exhibited, somewhat in summary form, in the published mark scheme) is *very* much better for hand calculation. It is appreciated that the "s_b^2/s_w^2" method is that by which the analysis of variance is first approached in the MEI textbook that supports this module, but the book does go on to mention the "squared totals" method. Candidates should be sure to understand the "squared totals" method and to use it routinely when carrying out these calculations by hand.

Finally, it was encouraging that many candidates were able to state the assumptions about the distribution of the experimental error correctly.