OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE **4769/01**

MATHEMATICS (MEI)

Statistics 4

FRIDAY 6 JUNE 2008 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.
- Read each question carefully and make sure you know what you have to do before starting your answer.
- 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.
- 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.

© OCR 2008 [Y/102/2659] OCR is an exempt Charity

[Turn over

Option 1: Estimation

1 The random variable X has the Poisson distribution with parameter θ so that its probability function is

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

where θ ($\theta > 0$) is unknown. A random sample of n observations from X is denoted by $X_1, X_2, \ldots , X_n$.

(i) Find $\hat{\theta}$, the maximum likelihood estimator of θ. [9]

The value of $\mathrm{P}(X = 0)$ is denoted by λ.

(ii) Write down an expression for λ in terms of θ. [1]

(iii) Let R denote the number of observations in the sample with value zero. By considering the binomial distribution with parameters n and $\mathrm{e}^{-\theta}$, write down $\mathrm{E}(R)$ and $\mathrm{Var}(R)$. Deduce that the observed *proportion* of observations in the sample with value zero, denoted by $\tilde{\lambda}$, is an unbiased estimator of λ with variance $\dfrac{\mathrm{e}^{-\theta}(1 - \mathrm{e}^{-\theta})}{n}$. [7]

(iv) In large samples, the variance of the maximum likelihood estimator of λ may be taken as $\dfrac{\theta\mathrm{e}^{-2\theta}}{n}$. Use this and the appropriate result from part **(iii)** to show that the relative efficiency of $\tilde{\lambda}$ with respect to the maximum likelihood estimator is $\dfrac{\theta}{\mathrm{e}^{\theta} - 1}$. Show that this expression is always less than 1. Show also that it is near 1 if θ is small and near 0 if θ is large. [7]

© OCR 2008

Option 2: Generating Functions

2 Independent trials, on each of which the probability of a 'success' is p $(0 < p < 1)$, are being carried out. The random variable X counts the number of trials up to and including that on which the first success is obtained. The random variable Y counts the number of trials up to and including that on which the nth success is obtained.

(i) Write down an expression for $\mathrm{P}(X = x)$ for $x = 1, 2, \ldots$. Show that the probability generating function of X is

$$\mathrm{G}(t) = pt(1 - qt)^{-1}$$

where $q = 1 - p$, and hence that the mean and variance of X are

$$\mu = \frac{1}{p} \qquad \text{and} \qquad \sigma^2 = \frac{q}{p^2}$$

respectively. [11]

(ii) Explain why the random variable Y can be written as

$$Y = X_1 + X_2 + \ldots + X_n$$

where the X_i are independent random variables each distributed as X. Hence write down the probability generating function, the mean and the variance of Y. [5]

(iii) State an approximation to the distribution of Y for large n. [1]

(iv) The aeroplane used on a certain flight seats 140 passengers. The airline seeks to fill the plane, but its experience is that not all the passengers who buy tickets will turn up for the flight. It uses the random variable Y to model the situation, with $p = 0.8$ as the probability that a passenger turns up. Find the probability that it needs to sell at least 160 tickets to get 140 passengers who turn up.

Suggest a reason why the model might not be appropriate. [7]

Option 3: Inference

3 **(i)** Explain the meaning of the following terms in the context of hypothesis testing: Type I error, Type II error, operating characteristic. [6]

A machine fills salt containers that will be sold in shops. The containers are supposed to contain 750 g of salt. The machine operates in such a way that the amount of salt delivered to each container is a Normally distributed random variable with standard deviation 20 g. The machine should be calibrated in such a way that the mean amount delivered, μ, is 750 g.

Each hour, a random sample of 9 containers is taken from the previous hour's output and the sample mean amount of salt is determined. If this is between 735 g and 765 g, the previous hour's output is accepted. If not, the previous hour's output is rejected and the machine is recalibrated.

(ii) Find the probability of rejecting the previous hour's output if the machine is properly calibrated. Comment on your result. [6]

(iii) Find the probability of accepting the previous hour's output if $\mu = 725$ g. Comment on your result. [6]

(iv) Obtain an expression for the operating characteristic of this testing procedure in terms of the cumulative distribution function $\Phi(z)$ of the standard Normal distribution. Evaluate the operating characteristic for the following values (in g) of μ: 720, 730, 740, 750, 760, 770, 780. [6]

Option 4: Design and Analysis of Experiments

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

(ii) An examinations authority is considering using an external contractor for the typesetting and printing of its examination papers. Four contractors are being investigated. A random sample of 20 examination papers over the entire range covered by the authority is selected and 5 are allocated at random to each contractor for preparation. The authority carefully checks the printed papers for errors and assigns a score to each to indicate the overall quality (higher scores represent better quality). The scores are as follows.

Contractor A	Contractor B	Contractor C	Contractor D
41	54	56	41
49	45	45	36
50	50	54	46
44	50	50	38
56	47	49	35

[The sum of these data items is 936 and the sum of their squares is 44 544.]

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]

(iii) The authority thinks that there might be differences in the ways the contractors cope with the preparation of examination papers in different subject areas. For this purpose, the subject areas are broadly divided into mathematics, sciences, languages, humanities, and others. The authority wishes to design a further investigation, ensuring that each of these subject areas is covered by each contractor. Name the experimental design that should be used and describe briefly the layout of the investigation. [3]

Permission to reproduce items where third-party owned material protected by copyright is included has been sought and cleared where possible. Every reasonable effort has been made by the publisher (OCR) to trace copyright holders, but if any items requiring clearance have unwittingly been included, the publisher will be pleased to make amends at the earliest possible opportunity.

OCR is part of the Cambridge Assessment Group. Cambridge Assessment is the brand name of University of Cambridge Local Examinations Syndicate (UCLES), which is itself a department of the University of Cambridge.

© OCR 2008

4769 Statistics 4

Q1				
(i)	$\mathrm{L} = \frac{e^{-\theta}\theta^{x_1}}{x_1!}\cdots\frac{e^{-\theta}\theta^{x_n}}{x_n!}\left[= \frac{e^{-n\theta}\theta^{\sum x_i}}{x_1!x_2!\cdots x_n!}\right]$	M1 A1	product form fully correct	
	ln L = const − $n\theta$ + $\sum x_i \ln\theta$	M1 A1		
	$\frac{d\ln \mathrm{L}}{d\theta} = -n + \frac{\sum x_i}{\theta} = 0$	M1 A1		
	$\Rightarrow \hat{\theta} = \frac{\sum x_i}{n} (= \bar{x})$	A1	CAO	
	Check this is a maximum	M1		
	e.g. $\frac{d^2 \ln \mathrm{L}}{\mathrm{d}\theta^2} = -\frac{\sum x_i}{\theta^2} < 0$	A1		9
(ii)	$\lambda = \mathrm{P}(X = 0) = e^{-\theta}$	B1		1
(iii)	We have $R \sim \mathrm{B}(n, e^{-\theta})$,	M1		
	so $\mathrm{E}(R) = ne^{-\theta}$	B1		
	$\mathrm{Var}(R) = ne^{-\theta}(1-e^{-\theta})$	B1		
	$\tilde{\lambda} = \frac{R}{n}$	M1		
	$\therefore \mathrm{E}(\tilde{\lambda}) = e^{-\theta}$	A1		
	i.e. unbiased	A1		
	$\mathrm{Var}(\tilde{\lambda}) = \frac{e^{-\theta}(1-e^{-\theta})}{n}$	A1	BEWARE PRINTED ANSWER	7

(iv)	Relative efficiency of $\tilde{\lambda}$ wrt ML est $= \dfrac{\text{Var(ML Est)}}{\text{Var}(\tilde{\lambda})}$	M1	any attempt to compare variances	
		M1	if correct	
	$= \dfrac{\theta\, e^{-2\theta}}{n} \cdot \dfrac{n}{e^{-\theta}(1-e^{-\theta})} = \dfrac{\theta}{e^{\theta}-1}$	A1	BEWARE PRINTED ANSWER	
	Eg:- Expression is $\dfrac{\theta}{\theta + \dfrac{\theta^2}{2!} + \ldots}$	M1		
	always < 1	E1		
	and this is ≈ 1 if θ is small ≈ 0 if θ is large	E1 E1	Allow statement that $\dfrac{\theta}{e^{\theta}-1} \to 0 \text{ as } \theta \to \infty$	
				7

Q2				
(i)	$\mathrm{P}(X=x)=q^{x-1}p$	B1	FT into pgf only	
	Pgf $\mathrm{G}(t)=\mathrm{E}(t^X)=\sum_{x=1}^{\infty}pt^xq^{x-1}$	M1		
	$=pt(1+qt+q^2t^2+\ldots)$	A1		
	$=\underline{\underline{pt(1-qt)^{-1}}}$	A1	BEWARE PRINTED ANSWER [consideration of |qt| < 1 not required]	
	$\mu=\mathrm{G}'(1)\quad\sigma^2=\mathrm{G}''(1)+\mu-\mu^2$	M1	for attempt to find G'(*t*) and/or G"(*t*)	
	$G'(t)=pt(-1)(1-qt)^{-2}(-q)+p(1-qt)^{-1}$ $=pqt(1-qt)^{-2}+p(1-qt)^{-1}$	A1		
	$\therefore\mathrm{G}'(1)=pq(1-q)^{-2}+p(1-q)^{-1}=\dfrac{q}{p}+1=\dfrac{1}{\underline{\underline{p}}}$	A1	BEWARE PRINTED ANSWER	
	$\mathrm{G}''(t)=pqt(-2)(1-qt)^{-3}(-q)+pq(1-qt)^{-2}+$ $p(-1)(1-qt)^{-2}(-q)$	A1		
	$\therefore\mathrm{G}''(1)=2pq^2(1-q)^{-3}+pq(1-q)^{-2}+pq(1-q)^{-2}$ $=\dfrac{2q^2}{p^2}+\dfrac{2q}{p}$	A1		
	$\therefore\sigma^2=\dfrac{2q^2}{p^2}+\dfrac{2q}{p}+\dfrac{1}{p}-\dfrac{1}{p^2}=\dfrac{2q^2+2pq+p-1}{p^2}$	M1	For inserting their values	
	$=\dfrac{q}{p^2}(2q+2p-1)=\underline{\underline{\dfrac{q}{p^2}}}$	A1	BEWARE PRINTED ANSWER	
				11

(ii)	X_1=number of trials to first success X_2= “ “ “ “ “ next “ . . . . X_n= “ “ “ “ “ *n*th “ $\therefore Y=X_1+X_2+\ldots X_n$ = total no of trials to the *n*th success $\therefore$ pgf of $Y=(\text{pgf of } X)^n=\underline{\underline{p^n t^n(1-qt)^{-n}}}$ $\mu_Y=n\mu_X=\underline{\underline{\frac{n}{p}}}$ $\sigma_Y^2=n\sigma_X^2=\underline{\underline{\frac{nq}{p^2}}}$	E1 E1 1 1 1		5
(iii)	N(candidate's μ_Y, candidate's σ_Y^2)	1		1
(iv)	Y = no of tickets to be sold ~ random variable as in (ii) with n = 140 and p = 0.8 ~ Approx N($\frac{140}{0.8}=175, \frac{140\times 0.2}{(0.8)^2}=43.75$) $\text{P}(Y\geq 160)\approx \text{P}(\text{N}(175,43.75)>159\frac{1}{2})$ = P(N(0,1)>-2.343) = 0.9905 For any sensible discussion <u>in context</u> (eg groups of passengers ⇒ not indep.)	E1 1 M1 A1 A1 E1 E1	Do not award if cty corr absent or wrong, but FT if 160 used → -2.268, 0.9884 CAO	7
Q3	X = amount of salt ~ N($\mu[750],\sigma^2[20^2]$) Sample of n=9			
(i)	Type I error: rejecting null hypothesis … … when it is true. Type II error: accepting null hypothesis … … when it is false. OC: P (accepting null hypothesis … … as a function of the parameter under investigation)	B1 B1 B1 B1 B1 B1	Allow B1 for P(rej H_0 when true) Allow B1 for P(acc H_0 when false) [P(type II error | the true value of the parameter) scores B1+B1]	6
(ii)	Reject if $\bar{x}<735$ *or* $\bar{x}>765$ $\alpha=\text{P}(\bar{X}<735 \text{ or } \bar{X}>765 \mid \bar{X}\sim \text{N}(750,\frac{20^2}{9}))$ $=\text{P}(Z<\frac{(735-750)3}{20}=-2.25$ $\text{or } Z>\frac{(765-750)3}{20}=2.25)$ = 2(1-0.9878) = 2 × 0.0122 = 0.0244 This is the probability of rejecting good output and unnecessarily re-calibrating the machine – seems small [but not very small?]	M1 A1 A1 A1 E1 E1	Might be implicit CAO Accept any sensible comments	6

(iii)	Accept if $735 < \bar{x} < 765$, and now $\mu = 725$. $\beta = \mathrm{P}(735 < \bar{X} < 765 \mid \bar{X} \sim \mathrm{N}(725, {20^2}/{9}))$ = P(1.5 < Z< 6) = 1 – 0.9332 = $\underline{\underline{0.0668}}$ This is the probability of accepting output and carrying on when in fact μ has slipped to 725 – small[-ish?]	M1 A1 A1 A1 E1 E1	might be implicit CAO If upper limit 765 not considered, maximum 2 of these 4 marks. If Φ(6) not considered, maximum 3 out of 4. accept sensible comments	6
(iv)	$\left.\begin{array}{l}\text{OC} = \mathrm{P}(\ 735 < \bar{X} < 765 \mid \bar{X} \sim \mathrm{N}(\mu, {20^2}/{9}\ ))\\ = \Phi\left(\frac{(765\ -\ \mu)3}{20}\right) - \Phi\left(\frac{(735\ -\ \mu)3}{20}\right)\end{array}\right\}$ $''\ \Phi - \Phi''$ μ=720:Φ(6.75) – Φ(2.25)=1– 0.9878 =0.0122 730: 5.25 0.75 =1– 0.7734 =0.2266 740: 3.75 −0.75 =1– (1–0.7734)= 0.7734 750: similarly or by write-down from part (ii) [FT] : 0.9756 760, 770, 780 by symmetry [FT]: 0.7734, 0.2266, 0.0122	M1 M1 A1 1 1 1	 both correct if any two correct	6
Q4				
(i)	$x_{ij} = \mu + \alpha_i + e_{ij}$ μ = population … .. grand mean for whole experiment α_i = population … .. mean by which *i* th treatment differs from μ e_{ij} are experimental errors… ~ ind N $(0, \sigma^2)$	1 1 1 1 1 1 3	Allow “uncorrelated” 1 for ind N; 1 for 0; 1 for σ^2.	9
(ii)	Totals are 240, 246, 254, 264, 196 each from sample of size 5 Grand total 936 “Correction factor” CF = $\frac{936^2}{20}$ = 43804.8 Total SS = 44544 - CF = 739.2			

<table>
<tr><td rowspan="7"></td><td>Between contractors SS = $\frac{240^2}{5}+\ldots+\frac{196^2}{5}-\text{CF}$ = 44209.6 – CF = 404.8</td><td>M1
M1</td><td>For correct methods for any two,</td><td rowspan="7">12</td></tr>
<tr><td>Residual SS (by subtraction) = 739.2 – 404.8 = 334.4</td><td>A1</td><td>if each calculated SS is correct.</td></tr>
<tr><td><table>
<tr><th>Source of Variation</th><th>SS</th><th>df</th><th>MS</th><th>MS ratio</th></tr>
<tr><td>Between Contractors</td><td>404.8</td><td>3</td><td>$134.9\dot{3}$</td><td>6.456</td></tr>
<tr><td>Residual</td><td>334.4</td><td>16</td><td>20.9</td><td></td></tr>
<tr><td>Total</td><td>739.2</td><td>19</td><td></td><td></td></tr>
</table></td><td>M1
M1
1
A1
1</td><td>CAO</td></tr>
<tr><td>Refer to $F_{3,16}$</td><td>1</td><td>NO FT IF WRONG</td></tr>
<tr><td>Upper 5% point is 3.24</td><td>1</td><td>NO FT IF WRONG</td></tr>
<tr><td>Significant</td><td>1</td><td></td></tr>
<tr><td>Seems performances of contractors are not all the same</td><td>1</td><td></td></tr>
<tr><td rowspan="2">(iii)</td><td>Randomised blocks</td><td>B1</td><td></td><td rowspan="2">3</td></tr>
<tr><td>Description</td><td>E1
E1</td><td>Take the subject areas as “blocks”, ensure each contractor is used at least once in each block</td></tr>
</table>

4769 Statistics 4

General Comments

This is the third occasion on which the new-specification Statistics 4 module has been sat. There were 24 candidates from 10 centres. This rather small number is a disappointing reduction from the previous two years. There were several more candidates who had registered for the examination but in the event were absent.

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, which is encouraging as it indicates that centres and candidates are spreading their work over all the options. The least popular of the questions was Q.1, on estimation theory, but even this received several attempts, some of which were highly successful. Overall, there was some extremely good work, but it has also to be reported that there was some work distinctly at the poorer end of the spectrum.

Comments on Individual Questions

1) This was on the "estimation" option. It was based on maximum likelihood estimation for a Poisson distribution.

   Most of the candidates who attempted this question did well. Unfortunately there were a few who simply did not know how to form the likelihood, though these were usually able to recover from part (ii) onwards, where another estimator was introduced for comparison. It was good that most of the candidates who dealt well with part (i) were aware that it needed to be checked that the turning point found was indeed a maximum.

   Moving onwards, the new estimator was usually dealt with successfully, and candidates knew how to find its relative efficiency with respect to the maximum likelihood estimator. Showing that this expression was always less than 1 and considering its limiting behaviour taxed the mathematical ingenuity of some candidates, but most had a reasonable idea of what to do.

2) This was on the "generating functions" option and was based on the geometric distribution.

   Many candidates proceeded thoroughly and carefully through the technical mathematical work, though this was one of the places where faking of answers was too common. The explanation in part (ii) was often somewhat sketchy, though usually sufficient to indicate that the candidate understood near enough what was happening. Most candidates knew the convolution theorem result at the end of part (ii) and could also straightway write down the mean and variance of the sum, but it was *very* disappointing, at Statistics 4 level, to find quite a few thinking that the variance of the sum had a factor of n^2 rather than n.

   In part (iii), not quite everybody realised that the approximation was simply a Normal distribution and, of those that did, some took it to be N(0, 1) even though they had explicitly obtained the mean and variance immediately before. Further difficulties arose in part (iv) where some candidates did not really know what to do and where use of a continuity correction was rare. It was however pleasing that many candidates gave a sensible reason why the model might not be appropriate, usually based on lack of independence for groups of passengers travelling together.

As some of the remarks above are somewhat critical of candidates' work, it is fitting to add that there were many very good attempts at this question.

3) This question was on the "inference" option, exploring ideas of Type I and Type II errors and the Operating Characteristic.

The opening explanations were usually correct, though perhaps it was inevitable that some candidates would get things the wrong way round. The technical work following in parts (ii) and (iii) was usually done well, though again it was perhaps inevitable that there would be a few errors in setting up the probabilities and/or in reading the Normal tables (the latter really should not occur in a Statistics 4 paper!). In part (iii), consideration of the upper limit 765 was sometimes forgotten; $\Phi(6)$ is indeed extremely near to 1, but it did need to be brought into account. Part (iv) was perhaps less successful; perhaps candidates were not too familiar with obtaining an algebraic expression for the Operating Characteristic. Nevertheless, there were some good solutions here.

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

It is extremely pleasing that there was some very good work here. Candidates seemed well prepared and to know the material thoroughly. Statements of the model and interpretation of its terms were often impeccable, though some candidates were still not careful enough in including words such as "population" and "independent". The analysis in part (ii) was also usually done well, and it is with much pleasure that I can report that nearly all candidates used the efficient "squared totals" method of calculation rather than the extremely cumbersome and error-prone "s_b^2/s_w^2" method. This is a real improvement from previous years and in particular from last year, when this had become much worse.

Part (iii) introduced a short discussion of "design" features for the experimentation. Most candidates appreciated that the design being introduced was that of randomised blocks and were able to give a good description of the layout.