OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE

MATHEMATICS (MEI) 4769

Statistics 4

Candidates answer on the Answer Booklet

OCR Supplied Materials:
- 8 page Answer Booklet
- Graph paper
- MEI Examination Formulae and Tables (MF2)

Other Materials Required:
None

Monday 15 June 2009
Afternoon

Duration: 1 hour 30 minutes

* 4 7 6 9 *

INSTRUCTIONS TO CANDIDATES

- Write your name clearly in capital letters, your Centre Number and Candidate Number in the spaces provided on the Answer Booklet.
- Use black ink. Pencil may be used for graphs and diagrams only.
- Read each question carefully and make sure that you know what you have to do before starting your answer.
- Answer any **three** questions.
- Do **not** write in the bar codes.
- 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 document consists of **4** pages. Any blank pages are indicated.

Option 1: Estimation

1 An industrial process produces components. Some of the components contain faults. The number of faults in a component is modelled by the random variable X with probability function

$$\mathrm{P}(X = x) = \theta(1-\theta)^x \quad \text{for } x = 0,\ 1,\ 2,\ \ldots$$

where θ is a parameter with $0 < \theta < 1$. The numbers of faults in different components are independent.

A random sample of n components is inspected. n_0 are found to have no faults, n_1 to have one fault and the remainder $(n - n_0 - n_1)$ to have two or more faults.

(i) Find $\mathrm{P}(X \geqslant 2)$ and hence show that the likelihood is

$$\mathrm{L}(\theta) = \theta^{n_0+n_1}(1-\theta)^{2n-2n_0-n_1}.$$ **[5]**

(ii) Find the maximum likelihood estimator $\hat{\theta}$ of θ. You are not required to verify that any turning point you locate is a maximum. **[6]**

(iii) Show that $\mathrm{E}(X) = \dfrac{1-\theta}{\theta}$. Deduce that another plausible estimator of θ is $\tilde{\theta} = \dfrac{1}{1+\overline{X}}$ where $\overline{X}$ is the sample mean. What additional information is needed in order to calculate the value of this estimator? **[6]**

(iv) You are given that, in large samples, $\tilde{\theta}$ may be taken as Normally distributed with mean θ and variance $\theta^2(1-\theta)/n$. Use this to obtain a 95% confidence interval for θ for the case when 100 components are inspected and it is found that 92 have no faults, 6 have one fault and the remaining 2 have exactly four faults each. **[7]**

Option 2: Generating Functions

2 **(i)** The random variable Z has the standard Normal distribution with probability density function

$$\mathrm{f}(z) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}, \quad -\infty < z < \infty.$$

Obtain the moment generating function of Z. **[8]**

(ii) Let $\mathrm{M}_Y(t)$ denote the moment generating function of the random variable Y. Show that the moment generating function of the random variable $aY + b$, where a and b are constants, is $\mathrm{e}^{bt}\mathrm{M}_Y(at)$. **[4]**

(iii) Use the results in parts **(i)** and **(ii)** to obtain the moment generating function $\mathrm{M}_X(t)$ of the random variable X having the Normal distribution with parameters μ and σ^2. **[4]**

(iv) If $W = \mathrm{e}^X$ where X is as in part **(iii)**, W is said to have a lognormal distribution. Show that, for any positive integer k, the expected value of W^k is $\mathrm{M}_X(k)$. Use this result to find the expected value and variance of the lognormal distribution. **[8]**

Option 3: Inference

3 **(i)** At a waste disposal station, two methods for incinerating some of the rubbish are being compared. Of interest is the amount of particulates in the exhaust, which can be measured over the working day in a convenient unit of concentration. It is assumed that the underlying distributions of concentrations of particulates are Normal. It is also assumed that the underlying variances are equal. During a period of several months, measurements are made for method A on a random sample of 10 working days and for method B on a separate random sample of 7 working days, with results, in the convenient unit, as follows.

Method A	124.8	136.4	116.6	129.1	140.7	120.2	124.6	127.5	111.8	130.3
Method B	130.4	136.2	119.8	150.6	143.5	126.1	130.7			

Use a t test at the 10% level of significance to examine whether either method is better in resulting, on the whole, in a lower concentration of particulates. State the null and alternative hypotheses under test. **[10]**

(ii) The company's statistician criticises the design of the trial in part **(i)** on the grounds that it is not paired. Summarise the arguments the statistician will have used. A new trial is set up with a paired design, measuring the concentrations of particulates on a random sample of 9 paired occasions. The results are as follows.

Pair	I	II	III	IV	V	VI	VII	VIII	IX
Method A	119.6	127.6	141.3	139.5	141.3	124.1	116.6	136.2	128.8
Method B	112.2	128.8	130.2	134.0	135.1	120.4	116.9	134.4	125.2

Use a t test at the 5% level of significance to examine the same hypotheses as in part **(i)**. State the underlying distributional assumption that is needed in this case. **[10]**

(iii) State the names of procedures that could be used in the situations of parts **(i)** and **(ii)** if the underlying distributional assumptions could not be made. What hypotheses would be under test? **[4]**

[Question 4 is printed overleaf.]

Option 4: Design and Analysis of Experiments

4 **(i)** Describe, with the aid of a specific example, an experimental situation for which a Latin square design is appropriate, indicating carefully the features which show that a completely randomised or randomised blocks design would be inappropriate. **[9]**

(ii) The model for the one-way analysis of variance may be written, in a customary notation, as

$$x_{ij} = \mu + \alpha_i + e_{ij}.$$

State the distributional assumptions underlying e_{ij} in this model. What is the interpretation of the term α_i? **[5]**

(iii) An experiment for comparing 5 treatments is carried out, with a total of 20 observations. A partial one-way analysis of variance table for the analysis of the results is as follows.

Source of variation	Sums of squares	Degrees of freedom	Mean squares	Mean square ratio
Between treatments				
Residual	68.76			
Total	161.06			

Copy and complete the table, and carry out the appropriate test using a 1% significance level. **[10]**

OCR
RECOGNISING ACHIEVEMENT

4769 Statistics 4 June 2009

Q1	Follow-through all intermediate results in this question, unless obvious nonsense.			
(i)	$P(X \geq 2) = 1 - \theta - \theta(1-\theta) = (1-\theta)^2$ [o.e.]	M1 A1		
	$L = [\theta]^{n_0}[\theta(1-\theta)]^{n_1}[(1-\theta)^2]^{n-n_0-n_1}$	M1 A1	Product form Fully correct	
	$= \theta^{n_0+n_1}(1-\theta)^{2n-2n_0-n_1}$	A1	BEWARE PRINTED ANSWER	5
(ii)	$\ln L = (n_0+n_1)\ln\theta + (2n-2n_0-n_1)\ln(1-\theta)$	M1 A1		
	$\dfrac{d\ln L}{d\theta}$	M1		
	$= \dfrac{n_0+n_1}{\theta} - \dfrac{2n-2n_0-n_1}{1-\theta}$	A1		
	$= 0$	M1		
	$\Rightarrow (1-\hat{\theta})(n_0+n_1) = \hat{\theta}(2n-2n_0-n_1)$			
	$\Rightarrow \hat{\theta} = \dfrac{n_0+n_1}{2n-n_0}$	A1		6
(iii)	$E(X) = \sum_{x=0}^{\infty} x\theta(1-\theta)^x$	M1		
	$= \theta\{0 + (1-\theta) + 2(1-\theta)^2 + 3(1-\theta)^3 + \ldots\}$ $= \dfrac{1-\theta}{\theta}$	A2	Divisible, for algebra; e.g. by "GP of GPs" BEWARE PRINTED ANSWER	
	So could sensibly use (method of moments)			
	$\tilde{\theta}$ given by $\dfrac{1-\tilde{\theta}}{\tilde{\theta}} = \bar{X}$	M1		
	$\Rightarrow \tilde{\theta} = \dfrac{1}{1+\bar{X}}$	A1	BEWARE PRINTED ANSWER	
	To use this, we need to know the exact numbers of faults for components with "two or more".	E1		6
(iv)	$\bar{x} = \dfrac{14}{100} = 0 \cdot 14$	B1		
	$\tilde{\theta} = \dfrac{1}{1+0.14} = 0{\cdot}8772$	B1		
	Also, from expression given in question, $\text{Var}(\tilde{\theta}) = \dfrac{0\cdot 8772^2(1-0\cdot 8772)}{100}$ $= 0{\cdot}000945$	B1		
	CI is given by $0{\cdot}8772 \pm 1{\cdot}96 \times \sqrt{0\cdot 000945}$ = (0·817, 0·937)	M1 B1 M1 A1	For 0.8772 For 1.96 For $\sqrt{0\cdot 000945}$	7

Q2				
(i)	Mgf of Z = E$(e^{tZ}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tz-\frac{z^2}{2}} dz$	M1		
	Complete the square	M1		
	$tz - \frac{z^2}{2} = -\frac{1}{2}(z-t)^2 + \frac{1}{2}t^2$	A1 A1		
	$= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-t)^2}{2}} dt = e^{\frac{t^2}{2}}$ Pdf of N(t,1) $\therefore \int = 1$	M1 M1 M1 A1	For taking out factor $e^{\frac{t^2}{2}}$ For use of pdf of N(t,1) For $\int$ pdf = 1 For final answer $e^{\frac{t^2}{2}}$	8
(ii)	*Y* has mgf $M_Y(t)$ Mgf of *aY* + *b* is $E[e^{t(aY+b)}]$ $= e^{bt} E[e^{(at)Y}] = e^{bt} M_Y(at)$	M1 1 1 1	For factor e^{bt} For factor $E[e^{(at)Y}]$ For final answer	4
(iii)	$Z = \frac{X-\mu}{\sigma}$, so $X = \sigma Z + \mu$ $\therefore M_X(t) = e^{\mu t}.e^{\frac{(\sigma t)^2}{2}} = e^{\mu t + \frac{\sigma^2 t^2}{2}}$	M1 1 1 1	For factor $e^{\mu t}$ For factor $e^{\frac{(\sigma t)^2}{2}}$ For final answer	4
(iv)	$W = e^X$ $E(W^k) = E[(e^X)^k] = E(e^{kX}) = M_X(k)$	M1 A1 A1	For $E[(e^X)^k]$ For $E(e^{kX})$ For $M_X(k)$	
	$\therefore E(W) = M_X(1) = e^{\mu + \frac{\sigma^2}{2}}$	M1 A1		
	$E(W^2) = M_X(2) = e^{2\mu + 2\sigma^2}$ $\therefore \mathrm{Var}(W) = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} [= e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)]$	M1 A1 A1		8

Q3

(i)	$\bar{x} = 126\cdot 2$ s = 8·7002 $s^2 = 75\cdot 69\dot{3}$ $\bar{y} = 133\cdot 9$ s = 10·4760 $s^2 = 109\cdot 74\dot{6}$	A1	A1 if all correct. [No mark for use of s_n, which are 8·2537 and 9·6989 respectively.]
	$H_0 : \mu_A = \mu_B$ $H_0 : \mu_A \neq \mu_B$ Where μ_A, μ_B are the population means.	1 1	<u>Do not</u> accept $\overline{X} = \overline{Y}$ or similar.
	Pooled s^2 $= \dfrac{9 \times 75\cdot 69\dot{3} + 6 \times 109\cdot 47\dot{6}}{15} = \dfrac{681\cdot 24 + 658\cdot 48}{15}$ $= 89\cdot 314\dot{6}$ [√ = 9·4506]	B1	
	Test statistic is $\dfrac{126\cdot 2 - 133\cdot 9}{\sqrt{89\cdot 314\dot{6}}\sqrt{\dfrac{1}{10}+\dfrac{1}{7}}} = -\dfrac{7\cdot 7}{4\cdot 6573} = -1\cdot 653$	M1 A1	
	Refer to t_{15}	1	No FT if wrong
	Double-tailed 10% point is 1·753 Not significant No evidence that population mean concentrations differ.	1 1 1	No FT if wrong 10
(ii)	There may be consistent differences between days (days of week, types of rubbish, ambient conditions,…) which should be allowed for.	E1 E1	
	Assumption: Normality of population of <u>differences.</u>	1	
	Differences are 7·4 -1·2 11·1 5·5 6·2 3·7 −0·3 1·8 3·6 [$\bar{d}$ = 4·2, s = 3·862 (s^2 =14·915)] Use of $s_n (= 3\cdot 641)$ is <u>not</u> acceptable, even in a denominator of $s_n / \sqrt{n-1}$]	M1	A1 Can be awarded here if NOT awarded in part (i)
	Test statistic is $\dfrac{4\cdot 2 - 0}{3\cdot 862/\sqrt{9}}$ = 3·26	M1 A1	
	Refer to t_8 Double-tailed 5% point is 2·306 Significant Seems population means differ	1 1 1 1	No FT if wrong No FT if wrong 10

(iii)	Wilcoxon rank sum test Wilcoxon signed rank test H_0: $\text{median}_A = \text{median}_B$ H_1: $\text{median}_A \neq \text{median}_B$	B1 B1 1 1	 [Or more formal statements]	 4
Q4				
(i)	Description must be in <u>context</u>. If no context given, mark according to scheme and then give half-marks, rounded down. Clear description of "rows". And "columns" As extraneous factors to be taken account of in the design, with "treatments" to be compared. Need same numbers of each Clear contrast with situations for completely randomised design and randomised trends.	 E1 E1 E1 E1 E1 E1 E1 E1 E1		9
(ii)	$e_{ij} \sim$ ind N $(0, \sigma^2)$ α_i is population mean effect by which ith treatment differs from overall mean	1 1 1 1 1	Allow uncorrelated For 0 For σ^2	5
(iii)	See ANOVA table below	1 1 1 1 1 1 1		
	Refer to $F_{4,15}$	1	No FT if wrong	
	Upper 1% point is 4·89	1	No FT if wrong	
	Significant, seems treatments are not all the same	1		10

Source of Variation	SS	df	MS	MS ratio
Between Treatments	92·30	4	23·075	5·034
Residual	68·76	15	4·584	
Total	161·06	19		

4769 Statistics 4

General Comments

There were 35 candidates from 17 centres (plus one more centre whose candidate was absent). While obviously a small entry, it is a noticeable and welcome increase from last year. Many centres entered just one candidate, but that is unsurprising for this advanced module at the "top" of the statistics strand. Indeed, it is pleasing that centres are able to support single candidates. Perhaps the Further Mathematics Network is making an important contribution too. A particularly pleasing feature was that there were some centres which had had no candidates for this module (or its predecessors) for many years, and one or two centres that, it is thought, were entering for the first time.

As usual, the paper consisted 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. Overall, there was some very good work, but also some distinctly poorer work.

We are seeing too many cases of unsupported numerical answers that are clearly taken straight from calculators. Candidates must be made to realise that this is a high-risk strategy. If the numerical value is wrong (beyond whatever latitude is allowed for say the second or third decimal place), then *no marks at all* can be awarded for that section of the work, because there is no evidence that a correct method is being used. A particular illustration of this was provided in question 3, where the value of a pooled estimator of variance had to be found, and where there were a number of cases of unsupported numerically incorrect answers (often quite substantially incorrect). Was there an attempt to use the right method with just a keying error, or did the candidate not know what to do? With no evidence, it cannot be assumed that the correct method was being used.

There were many cases where the conclusions in context for hypothesis tests were too assertive. This was disappointing as it had appeared that this point had been successfully made over recent years.

Comments on Individual Questions

1) This was on the "estimation" option. It was based on maximum likelihood estimation and method of moments estimation. The latter term was of course not used by name. The general idea of "moments" estimation has appeared in many previous papers.

First, there was some good work. Some candidates were able to complete the question, or at least very nearly do so, in a careful, efficient and insightful way.

However, there were some candidates who clearly had no idea what a likelihood is. This is very poor as it is an explicit and central item in this section of the syllabus.

Maximisation of the given expression for the likelihood was usually reasonably well done, but some candidates did it without first taking logarithms, which again indicates lack of understanding of the usual procedures in this work.

The work to find E(X) in part (iii) was commonly very poorly done. The random variable is obviously discrete, so the expected value is a sum; whyever did some candidates think it was an integral? The sum is *not* that of a GP. More subtle methods are required to find it. "More subtle methods" do not include simply writing down the given answer – faking was especially prevalent here. The given answer, as for the likelihood itself, is there so that candidates may *use* it in subsequent work, and of course it is entirely legitimate to do that.

The "moments" estimation in part (iii) was also commonly poorly done. There was bad confusion between estimators and parameters (poor notation was often a particular drawback here), for example in claims such as $\bar{X} = (1-\theta)/\theta$.

Finally, the confidence interval in part (iv) was sometimes done well, but often the work here was very confused. Silly nonsenses of "$s/\sqrt{n}$" for the standard deviation turned up far too often.

After all the above criticisms, it is well to reiterate that there was some very good work throughout this question.

2) This was on the "generating functions" option and was mostly based on standard work for the Normal distribution.

Many candidates knew that "completing the square" (in the exponent) is the right method for obtaining the moment generating function of the N(0, 1) distribution, but not all could do it. The step that follows, where the integral of the pdf of N(t, 1) is created and used, was not always convincing. Other candidates got themselves into various severe difficulties (it is hopeless to try to do this integral by parts) and often faked the result.

The linear combination work in part (ii) was usually done well.

The "unstandardising" in part (iii) was also usually done well, though some faking also occurred here.

In part (iv), the previous results were applied to finding the mean and variance of the lognormal distribution. Only some of the candidates got the (actually rather easy) point here.

3) This question was on the "inference" option, exploring unpaired and paired tests. It was often done very well. The usual errors (e.g. wrong number of degrees of freedom, wrong critical point) sometimes appeared. Wrong critical points were strangely more common in part (ii), often despite previous success in part (i). As mentioned in the "general comments" section above, over-assertive conclusions were seen too often.

Many candidates simply failed to discuss the arguments for pairing that are asked for in part (ii).

There were even some candidates who did part (ii) as another unpaired test, an especially disappointing error.

Solutions to part (iii) were often somewhat muddled, not clearly distinguishing the cases of parts (i) and (ii). Several candidates appeared to think that the Wilcoxon rank sum test and the Mann-Whitney test are different!

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

The examples to demonstrate a Latin square were generally fairly good. The contexts chosen by the candidates were remarkably uniform. Several contexts appeared several times (not including the classical "stream down two sides of a field", either); perhaps these are discussed in popular text books. The contrast with a randomised blocks design was not always grasped, and many candidates simply omitted the comparison with a completely randomised design.

The modelling work in part (ii) and the analysis of variance in part (iii) were usually done well.