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

**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 $\theta$ 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 $\theta$. 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 $\theta$ 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 $\theta$ and variance $\theta^2(1 - \theta)/n$. Use this to obtain a 95% confidence interval for $\theta$ 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 $\mu$ and $\sigma^2$. **[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 $\alpha_i$? **[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]**