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.

Registered Charity 1066969
[Turn over

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 $\int_0^\infty u^m \mathrm{e}^{-u}\,\mathrm{d}u = m!$ for any non-negative integer m.]

The random variable X has probability density function

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

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]

[Turn over

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]

OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE UNIT

MATHEMATICS (MEI)

4769/01

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

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

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

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

(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

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

(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]

[Turn over

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]

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.

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]

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]

[Turn over

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]

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 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

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

(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

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:
- Scientific or graphical calculator

Friday 18 June 2010
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 The random variable X has probability density function

$$f(x) = \frac{xe^{-x/\lambda}}{\lambda^2} \qquad (x > 0),$$

where λ is a parameter ($\lambda > 0$). $X_1, X_2, \ldots, X_n$ are n independent observations on X, and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is their mean.

(i) Obtain $E(X)$ and deduce that $\hat{\lambda} = \frac{1}{2}\overline{X}$ is an unbiased estimator of λ. **[7]**

(ii) Obtain $\text{Var}(\hat{\lambda})$. **[7]**

(iii) Explain why the results in parts **(i)** and **(ii)** indicate that $\hat{\lambda}$ is a good estimator of λ in large samples. **[2]**

(iv) Suppose that $n = 3$ and consider the alternative estimator

$$\tilde{\lambda} = \tfrac{1}{8}X_1 + \tfrac{1}{4}X_2 + \tfrac{1}{8}X_3.$$

Show that $\tilde{\lambda}$ is an unbiased estimator of λ. Find the relative efficiency of $\tilde{\lambda}$ compared with $\hat{\lambda}$. Which estimator do you prefer in this case? **[8]**

Option 2: Generating Functions

2 The random variable X has the Poisson distribution with parameter λ.

(i) Show that the probability generating function of X is $G(t) = e^{\lambda(t-1)}$. **[3]**

(ii) Hence obtain the mean μ and variance σ^2 of X. **[5]**

(iii) Write down the mean and variance of the random variable $Z = \dfrac{X - \mu}{\sigma}$. **[2]**

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

$$M_Z(\theta) = e^{f(\theta)} \quad \text{where } f(\theta) = \lambda\left(e^{\theta/\sqrt{\lambda}} - \frac{\theta}{\sqrt{\lambda}} - 1\right). \qquad \textbf{[7]}$$

(v) Show that the limit of $M_Z(\theta)$ as $\lambda \to \infty$ is $e^{\theta^2/2}$. **[4]**

(vi) Explain briefly why this implies that the distribution of Z tends to $N(0, 1)$ as $\lambda \to \infty$. What does this imply about the distribution of X as $\lambda \to \infty$? **[3]**

Option 3: Inference

3 At a factory, two production lines are in use for making steel rods. A critical dimension is the diameter of a rod. For the first production line, it is assumed from experience that the diameters are Normally distributed with standard deviation 1.2 mm. For the second production line, it is assumed from experience that the diameters are Normally distributed with standard deviation 1.4 mm. It is desired to test whether the mean diameters for the two production lines, μ_1 and μ_2, are equal. A random sample of 8 rods is taken from the first production line and, independently, a random sample of 10 rods is taken from the second production line.

(i) Find the acceptance region for the customary test based on the Normal distribution for the null hypothesis $\mu_1 = \mu_2$, against the alternative hypothesis $\mu_1 \neq \mu_2$, at the 5% level of significance. **[6]**

(ii) The sample means are found to be 25.8 mm and 24.4 mm respectively. What is the result of the test? Provide a two-sided 99% confidence interval for $\mu_1 - \mu_2$. **[7]**

The production lines are modified so that the diameters may be assumed to be of equal (but unknown) variance. However, they may no longer be Normally distributed. A two-sided test of the equality of the population medians is required, at the 5% significance level.

(iii) The diameters in independent random samples of sizes 6 and 8 are as follows, in mm.

First production line	25.9	25.8	25.3	24.7	24.4	25.4		
Second production line	23.8	25.6	24.0	23.5	24.1	24.5	24.3	25.1

Use an appropriate procedure to carry out the test. **[11]**

[Question 4 is printed overleaf.]

Option 4: Design and Analysis of Experiments

4 At an agricultural research station, a trial is made of four varieties (A, B, C, D) of a certain crop in an experimental field. The varieties are grown on plots in the field and their yields are measured in a standard unit.

(i) It is at first thought that there may be a consistent trend in the natural fertility of the soil in the field from the west side to the east, though no other trends are known. Name an experimental design that should be used in these circumstances and give an example of an experimental layout. **[5]**

Initial analysis suggests that any natural fertility trend may in fact be ignored, so the data from the trial are analysed by one-way analysis of variance.

(ii) The usual model for one-way analysis of variance of the yields y_{ij} may be written as

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

where the e_{ij} represent the experimental errors. Interpret the other terms in the model. State the usual distributional assumptions for the e_{ij}. **[7]**

(iii) The data for the yields are as follows, each variety having been used on 5 plots.

Variety			
A	B	C	D
12.3	14.2	14.1	13.6
11.9	13.1	13.2	12.8
12.8	13.1	14.6	13.3
12.2	12.5	13.7	14.3
13.5	12.7	13.4	13.8

$[\Sigma\Sigma y_{ij} = 265.1, \quad \Sigma\Sigma y_{ij}^2 = 3524.31.]$

Construct the usual one-way analysis of variance table and carry out the usual test, at the 5% significance level. Report briefly on your conclusions. **[12]**

OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE

MATHEMATICS (MEI)

4769

Statistics 4

Candidates answer on the answer booklet.

OCR supplied materials:
- 8 page answer booklet (sent with general stationery)
- MEI Examination Formulae and Tables (MF2)

Other materials required:
- Scientific or graphical calculator

Thursday 26 May 2011
Morning

Duration: 1 hour 30 minutes

INSTRUCTIONS TO CANDIDATES

- Write your name, centre number and candidate number in the spaces provided on the answer booklet. Please write clearly and in capital letters.
- Use black ink. Pencil may be used for graphs and diagrams only.
- Read each question carefully. Make sure 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 scientific or 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 The random variable X has the Normal distribution with mean 0 and variance θ, so that its probability density function is

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

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. **[14]**

(ii) Show that $\hat{\theta}$ is an unbiased estimator of θ. **[4]**

(iii) In large samples, the variance of $\hat{\theta}$ may be estimated by $\frac{2\hat{\theta}^2}{n}$. Use this and the results of parts **(i)** and **(ii)** to find an approximate 95% confidence interval for θ in the case when $n = 100$ and $\Sigma X_i^2 = 1000$. **[6]**

Option 2: Generating Functions

2 The random variable X has the χ_n^2 distribution. This distribution has moment generating function $M(\theta) = (1 - 2\theta)^{-\frac{1}{2}n}$, where $\theta < \frac{1}{2}$.

(i) Verify the expression for $M(\theta)$ quoted above for the cases $n = 2$ and $n = 4$, given that the probability density functions of X in these cases are as follows. **[10]**

$n = 2$: $\quad f(x) = \frac{1}{2}e^{-\frac{1}{2}x} \quad (x > 0)$

$n = 4$: $\quad f(x) = \frac{1}{4}xe^{-\frac{1}{2}x} \quad (x > 0)$

(ii) For the general case, use $M(\theta)$ to find the mean and variance of X in terms of n. **[7]**

(iii) $Y_1, Y_2, \ldots, Y_k$ are independent random variables, each with the χ_1^2 distribution. Show that $W = \sum_{i=1}^{k} Y_i$ has the χ_k^2 distribution. **[4]**

(iv) Use the Central Limit Theorem to find an approximation for $P(W < 118.5)$ for the case $k = 100$. **[3]**

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, power. **[8]**

(ii) A market research organisation is designing a sample survey to investigate whether expenditure on everyday food items has increased in 2011 compared with 2010. For one of the populations being studied, the random variable X is used to model weekly expenditure, in £, on these items in 2011, where X is Normally distributed with mean μ and variance σ^2. As the corresponding mean value in 2010 was 94, the hypotheses to be examined are

H_0: $\mu = 94$,

H_1: $\mu > 94$.

By comparison with the corresponding 2010 value, σ^2 is assumed to be 25.

The following criteria for the survey are laid down.

- If in fact $\mu = 94$, the probability of concluding that $\mu > 94$ must be only 2%
- If in fact $\mu = 97$, the probability of concluding that $\mu > 94$ must be 95%

A random sample of size n is to be taken and the usual Normal test based on $\overline{X}$ is to be used, with a critical value of c such that H_0 is rejected if the value of $\overline{X}$ exceeds c. Find c and the smallest value of n that is required. **[13]**

(iii) Sketch the power function of an ideal test for examining the hypotheses in part **(ii)**. **[3]**

Option 4: Design and Analysis of Experiments

4 **(a)** Provide an example of an experimental situation where there is one factor of primary interest and where a suitable experimental design would be

(i) randomised blocks,

(ii) a Latin square.

In each case, explain carefully why the design is suitable and why the other design would not be appropriate. **[12]**

(b) An industrial experiment to compare four treatments for increasing the tensile strength of steel is carried out according to a completely randomised design. For various reasons, it is not possible to use the same number of replicates for each treatment. The increases, in a suitable unit of tensile strength, are as follows.

Treatment A	Treatment B	Treatment C	Treatment D
10.1	21.1	9.2	22.6
21.2	20.3	8.8	17.4
11.6	16.0	15.2	23.1
13.6		15.0	19.2
		12.4	

[The sum of these data items is 256.8 and the sum of their squares is 4471.92.]

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

THERE ARE NO QUESTIONS PRINTED ON THIS PAGE

OCR
RECOGNISING ACHIEVEMENT

Thursday 24 May 2012 – Morning

A2 GCE MATHEMATICS (MEI)

4769 Statistics 4

QUESTION PAPER

Candidates answer on the Printed Answer Book.

OCR supplied materials:
- Printed Answer Book 4769
- MEI Examination Formulae and Tables (MF2)

Other materials required:
- Scientific or graphical calculator

Duration: 1 hour 30 minutes

INSTRUCTIONS TO CANDIDATES

These instructions are the same on the Printed Answer Book and the Question Paper.

- The Question Paper will be found in the centre of the Printed Answer Book.
- Write your name, centre number and candidate number in the spaces provided on the Printed Answer Book. Please write clearly and in capital letters.
- **Write your answer to each question in the space provided in the Printed Answer Book**. Additional paper may be used if necessary but you must clearly show your candidate number, centre number and question number(s).
- Use black ink. HB pencil may be used for graphs and diagrams only.
- Read each question carefully. Make sure 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 scientific or graphical calculator in this paper.
- Final answers should be given to a degree of accuracy appropriate to the context.

INFORMATION FOR CANDIDATES

This information is the same on the Printed Answer Book and the Question Paper.

- The number of marks is given in brackets **[ ]** at the end of each question or part question on the Question Paper.
- 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**.
- The Printed Answer Book consists of **16** pages. The Question Paper consists of **8** pages. Any blank pages are indicated.

INSTRUCTION TO EXAMS OFFICER/INVIGILATOR

- Do not send this Question Paper for marking; it should be retained in the centre or destroyed.

Option 1: Estimation

1 In a certain country, any baby born is equally likely to be a boy or a girl, independently for all births. The birthweight of a baby boy is given by the continuous random variable X_B with probability density function (pdf) $f_B(x)$ and cumulative distribution function (cdf) $F_B(x)$. The birthweight of a baby girl is given by the continuous random variable X_G with pdf $f_G(x)$ and cdf $F_G(x)$.

The continuous random variable X denotes the birthweight of a baby selected at random.

(i) By considering

$$P(X \leqslant x) = P(X \leqslant x \,|\, \text{boy})\ P(\text{boy}) + P(X \leqslant x \,|\, \text{girl})\ P(\text{girl}),$$

find the cdf of X in terms of $F_B(x)$ and $F_G(x)$, and deduce that the pdf of X is

$$f(x) = \tfrac{1}{2}\{f_B(x) + f_G(x)\}. \quad \textbf{[3]}$$

(ii) The birthweights of baby boys and girls have means μ_B and μ_G respectively. Deduce that

$$E(X) = \tfrac{1}{2}(\mu_B + \mu_G). \quad \textbf{[1]}$$

(iii) The birthweights of baby boys and girls have common variance σ^2. Find an expression for $E(X^2)$ in terms of μ_B, μ_G and σ^2, and deduce that

$$\text{Var}(X) = \sigma^2 + \tfrac{1}{4}(\mu_B - \mu_G)^2. \quad \textbf{[7]}$$

(iv) A random sample of size $2n$ is taken from all the babies born in a certain period. The mean birthweight of the babies in this sample is $\overline{X}$. Write down an approximation to the sampling distribution of $\overline{X}$ if n is large. **[4]**

(v) Suppose instead that a stratified sample of size $2n$ is taken by selecting n baby boys at random and, independently, n baby girls at random. The mean birthweight of the $2n$ babies in this sample is $\overline{X}_{st}$. Write down the expected value of $\overline{X}_{st}$ and find the variance of $\overline{X}_{st}$. **[4]**

(vi) Deduce that both $\overline{X}$ and $\overline{X}_{st}$ are unbiased estimators of the population mean birthweight. Find which is the more efficient. **[5]**

Option 2: Generating Functions

2 The random variable X ($X = 1, 2, 3, 4, 5, 6$) denotes the score when a fair six-sided die is rolled.

(i) Write down the mean of X and show that $\text{Var}(X) = \frac{35}{12}$. **[3]**

(ii) Show that $\text{G}(t)$, the probability generating function (pgf) of X, is given by

$$\text{G}(t) = \frac{t(1 - t^6)}{6(1 - t)}.$$ **[2]**

The random variable N ($N = 0, 1, 2, \ldots$) denotes the number of heads obtained when an unbiased coin is tossed repeatedly until a tail is first obtained.

(iii) Show that $\text{P}(N = r) = \left(\frac{1}{2}\right)^{r+1}$ for $r = 0, 1, 2, \ldots$. **[1]**

(iv) Hence show that $\text{H}(t)$, the pgf of N, is given by $\text{H}(t) = (2 - t)^{-1}$. **[2]**

(v) Use $\text{H}(t)$ to find the mean and variance of N. **[4]**

A game consists of tossing an unbiased coin repeatedly until a tail is first obtained and, each time a head is obtained in this sequence of tosses, rolling a fair six-sided die. The die is not rolled on the first occasion that a tail is obtained and the game ends at that point. The random variable Q ($Q = 0, 1, 2, \ldots$) denotes the total score on all the rolls of the die. Thus, in the notation above, $Q = X_1 + X_2 + \ldots + X_N$ where the X_i are independent random variables each distributed as X, with $Q = 0$ if $N = 0$. The pgf of Q is denoted by $\text{K}(t)$. The familiar result that the pgf of a sum of independent random variables is the product of their pgfs does **not** apply to $\text{K}(t)$ because N is a random variable and not a fixed number; you should instead **use without proof** the result that $\text{K}(t) = \text{H}(\text{G}(t))$.

(vi) Show that $\text{K}(t) = 6(12 - t - t^2 - \ldots - t^6)^{-1}$. **[4]**

[Hint. $(1 - t^6) = (1 - t)(1 + t + t^2 + \ldots + t^5)$.]

(vii) Use $\text{K}(t)$ to find the mean and variance of Q. **[6]**

(viii) Using your results from parts **(i)**, **(v)** and **(vii)**, verify the result that (in the usual notation for means and variances)

$$\sigma_Q^2 = \sigma_N^2 \mu_X^2 + \mu_N \sigma_X^2.$$ **[2]**

Option 3: Inference

3 At an agricultural research station, trials are being made of two fertilisers, A and B, to see whether they differ in their effects on the yield of a crop. Preliminary investigations have established that the underlying variances of the distributions of yields using the two fertilisers may be assumed equal. Scientific analysis of the fertilisers has suggested that fertiliser A may be inferior in that it leads, on the whole, to lower yield. A statistical analysis is being carried out to investigate this.

The crop is grown in carefully controlled conditions in 14 experimental plots, 6 with fertiliser A and 8 with fertiliser B. The yields, in kg per plot, are as follows, arranged in ascending order for each fertiliser.

Fertiliser A	9.8	10.2	10.9	11.5	12.7	13.3		
Fertiliser B	10.8	11.9	12.0	12.2	12.9	13.5	13.6	13.7

(i) Carry out a Wilcoxon rank sum test at the 5% significance level to examine appropriate hypotheses. **[9]**

(ii) Carry out a *t* test at the 5% significance level to examine appropriate hypotheses. **[11]**

(iii) Goodness of fit tests based on more extensive data sets from other trials with these fertilisers have failed to reject hypotheses of underlying Normal distributions. Discuss the relative merits of the analyses in parts **(i)** and **(ii)**. **[4]**

Option 4: Design and Analysis of Experiments

4 **(i)** In an engineering research laboratory, a study is being made of the strength of steel girders supplied by four different manufacturers. Four techniques for casting the girders are to be used, as are four slightly different chemical compositions of the steel. Sixteen girders are to be supplied for testing purposes, four by each manufacturer.

Name an experimental design that should be used for allocating the work to the manufacturers in such a way that any differences in strength of girders between the different manufacturers can be studied, whether or not there are consistent differences resulting from the casting techniques or from the chemical compositions. Give an example of a suitable layout of the experiment. **[5]**

(ii) After initial investigation, it is decided that differences in strength resulting from the casting techniques or the chemical compositions can be ignored. A one-way analysis of variance is therefore carried out on the results, which are as follows, measured in a convenient unit.

Strength of girder

Manufacturer			
A	B	C	D
109.4	114.4	114.8	115.1
110.0	113.1	113.7	114.0
110.9	113.5	115.4	114.7
110.3	112.5	114.3	115.6

[The sum of these data items is 1811.7 and the sum of their squares is 205 202.57.]

Construct the usual one-way analysis of variance table. Carry out the appropriate test and report your conclusion. **[12]**

(iii) Using the customary notation, write down the usual model underlying the one-way analysis of variance. Carefully interpret the terms in this model. State the assumptions that are usually made for the error term in the model. **[7]**

BLANK PAGE

BLANK PAGE

OCR
RECOGNISING ACHIEVEMENT

Friday 24 May 2013 – Morning

A2 GCE MATHEMATICS (MEI)

4769/01 Statistics 4

QUESTION PAPER

Candidates answer on the Printed Answer Book.

OCR supplied materials:
- Printed Answer Book 4769/01
- MEI Examination Formulae and Tables (MF2)

Other materials required:
- Scientific or graphical calculator

Duration: 1 hour 30 minutes

INSTRUCTIONS TO CANDIDATES

These instructions are the same on the Printed Answer Book and the Question Paper.

- The Question Paper will be found in the centre of the Printed Answer Book.
- Write your name, centre number and candidate number in the spaces provided on the Printed Answer Book. Please write clearly and in capital letters.
- **Write your answer to each question in the space provided in the Printed Answer Book**. Additional paper may be used if necessary but you must clearly show your candidate number, centre number and question number(s).
- Use black ink. HB pencil may be used for graphs and diagrams only.
- Read each question carefully. Make sure 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 scientific or graphical calculator in this paper.
- Final answers should be given to a degree of accuracy appropriate to the context.

INFORMATION FOR CANDIDATES

This information is the same on the Printed Answer Book and the Question Paper.

- The number of marks is given in brackets **[ ]** at the end of each question or part question on the Question Paper.
- 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**.
- The Printed Answer Book consists of **16** pages. The Question Paper consists of **4** pages. Any blank pages are indicated.

INSTRUCTION TO EXAMS OFFICER / INVIGILATOR

- Do not send this Question Paper for marking; it should be retained in the centre or recycled. Please contact OCR Copyright should you wish to re-use this document.

Option 1: Estimation

1 Traffic engineers are studying the flow of vehicles along a road. At an initial stage of the investigation, they assume that the average flow remains the same throughout the working day. An automatic counter records the number of vehicles passing a certain point per minute during the working day. A random sample of these records is selected; the sample values are denoted by $x_1, x_2, \ldots, x_n$.

(i) The engineers model the underlying random variable X by a Poisson distribution with unknown parameter θ. Obtain the likelihood of $x_1, x_2, \ldots, x_n$ and hence find the maximum likelihood estimate of θ. **[10]**

(ii) Write down the maximum likelihood estimate of the probability that no vehicles pass during a minute. **[3]**

(iii) The engineers note that, in a sample of size 1000 with sample mean $\bar{x} = 5$, there are no observations of zero. Suggest why this might cast some doubt on the investigation. **[3]**

(iv) On checking the automatic counter, the engineers find that, due to a fault, no record at all is made if no vehicle passes in a minute. They therefore model X as a Poisson random variable, again with an unknown parameter θ, except that the value $x = 0$ cannot occur. Show that, under this model,

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

and hence show that the maximum likelihood estimate of θ satisfies the equation

$$\frac{\theta \mathrm{e}^\theta}{\mathrm{e}^\theta - 1} = \bar{x}.$$ **[8]**

Option 2: Generating Functions

2 The random variable X takes values -2, 0 and 2, each with probability $\frac{1}{3}$.

(i) Write down the values of

(A) μ, the mean of X,

(B) $\mathrm{E}(X^2)$,

(C) σ^2, the variance of X. **[3]**

(ii) Obtain the moment generating function (mgf) of X. **[2]**

A random sample of n independent observations on X has sample mean $\overline{X}$, and the standardised mean is denoted by Z where

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

(iii) Stating carefully the required general results for mgfs of sums and of linear transformations, show that the mgf of Z is

$$\mathrm{M}_Z(\theta) = \left\{\frac{1}{3}\left(1 + \mathrm{e}^{\frac{\theta\sqrt{3}}{\sqrt{2n}}} + \mathrm{e}^{-\frac{\theta\sqrt{3}}{\sqrt{2n}}}\right)\right\}^n. \quad \textbf{[8]}$$

(iv) By expanding the exponential functions in $\mathrm{M}_Z(\theta)$, show that, for large n,

$$\mathrm{M}_Z(\theta) \approx \left(1 + \frac{\theta^2}{2n}\right)^n. \quad \textbf{[7]}$$

(v) Use the result $\mathrm{e}^y = \lim_{n\to\infty}\left(1 + \frac{y}{n}\right)^n$ to find the limit of $\mathrm{M}_Z(\theta)$ as $n \to \infty$, and deduce the approximate distribution of Z for large n. **[4]**

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, power. **[8]**

(ii) A test is to be carried out concerning a parameter θ. The null hypothesis is that θ has the particular value θ_0. The alternative hypothesis is $\theta \neq \theta_0$. Draw a sketch of the operating characteristic for a perfect test that never makes an error. **[3]**

(iii) The random variable X is distributed as $\mathrm{N}(\mu, 9)$. A random sample of size 25 is available. The null hypothesis $\mu = 0$ is to be tested against the alternative hypothesis $\mu \neq 0$. The null hypothesis will be accepted if $-1 < \overline{x} < 1$ where $\overline{x}$ is the value of the sample mean, otherwise it will be rejected. Calculate the probability of a Type I error. Calculate the probability of a Type II error if in fact $\mu = 0.5$; comment on the value of this probability. **[9]**

(iv) Without carrying out any further calculations, draw a sketch of the operating characteristic for the test in part **(iii)**. **[4]**

Option 4: Design and Analysis of Experiments

4 **(i)** Explain the advantages of randomisation and replication in a statistically designed experiment. **[6]**

(ii) The usual statistical model underlying the one-way analysis of variance is given, in the usual notation, by

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

where x_{ij} denotes the jth observation on the ith treatment. Define carefully all the terms in this model and state the properties of the term that represents experimental error. **[7]**

(iii) A trial of five fertilisers is carried out at an agricultural research station according to a completely randomised design in which each fertiliser is applied to four experimental plots of a crop (so that there are 20 experimental units altogether). The sums of squares in a one-way analysis of variance of the resulting data on yields of the crop are as follows.

Source of variation	Sum of squares
Between fertilisers	219.2
Residual	304.5
Total	523.7

State the customary null and alternative hypotheses that are tested. Provide the degrees of freedom for each sum of squares. Hence copy and complete the analysis of variance table and carry out the test at the 5% level. **[11]**

OCR

Wednesday 18 June 2014 – Afternoon

A2 GCE MATHEMATICS (MEI)

4769/01 Statistics 4

QUESTION PAPER

Candidates answer on the Printed Answer Book.

OCR supplied materials:
- Printed Answer Book 4769/01
- MEI Examination Formulae and Tables (MF2)

Other materials required:
- Scientific or graphical calculator

Duration: 1 hour 30 minutes

INSTRUCTIONS TO CANDIDATES

These instructions are the same on the Printed Answer Book and the Question Paper.

- The Question Paper will be found inside the Printed Answer Book.
- Write your name, centre number and candidate number in the spaces provided on the Printed Answer Book. Please write clearly and in capital letters.
- **Write your answer to each question in the space provided in the Printed Answer Book**. Additional paper may be used if necessary but you must clearly show your candidate number, centre number and question number(s).
- Use black ink. HB pencil may be used for graphs and diagrams only.
- Read each question carefully. Make sure 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 scientific or graphical calculator in this paper.
- Final answers should be given to a degree of accuracy appropriate to the context.

INFORMATION FOR CANDIDATES

This information is the same on the Printed Answer Book and the Question Paper.

- The number of marks is given in brackets **[ ]** at the end of each question or part question on the Question Paper.
- 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**.
- The Printed Answer Book consists of **16** pages. The Question Paper consists of **8** pages. Any blank pages are indicated.

INSTRUCTION TO EXAMS OFFICER / INVIGILATOR

- Do not send this Question Paper for marking; it should be retained in the centre or recycled. Please contact OCR Copyright should you wish to re-use this document.

Option 1: Estimation

1 $X_1, X_2, \ldots, X_n$ represent n independent observations on the random variable X with probability density function

$$f(x) = \frac{\theta^3 x^2 e^{-\theta x}}{2}, \quad x > 0,$$

where θ is an unknown parameter ($\theta > 0$). $\overline{X}$ denotes the sample mean of $X_1, X_2, \ldots, X_n$, ie $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(i) Show that the maximum likelihood estimator of θ is $\hat{\theta} = \frac{3}{\overline{X}}$. **[9]**

(ii) Show that, in the case $n = 1$, $\hat{\theta}$ is a biased estimator of θ. **[8]**

(iii) For large n, the distribution of $\hat{\theta}$ is well approximated by $N(\theta, H(\theta))$ where

$$H(\theta) = \frac{1}{E\left(-\frac{d^2 \ln L}{d\theta^2}\right)}$$

where L is the likelihood. Show that $H(\theta) = \frac{\theta^2}{3n}$. For the case where $n = 100$ and the value of $\overline{X}$ is 5.0, evaluate $\hat{\theta}$ and $H(\hat{\theta})$, and use these values to find an approximate 95% confidence interval for θ. **[7]**

Option 2: Generating Functions

2 **(i)** The probability density function of the random variable X is

$$\mathrm{f}(x) = \frac{x^{k-1}\mathrm{e}^{-x/\phi}}{\phi^k (k-1)!}, \quad x > 0,$$

where k is a known positive integer and ϕ is an unknown parameter ($\phi > 0$). Show that the moment generating function (mgf) of X is

$$\mathrm{M}_X(\theta) = (1-\phi\theta)^{-k}$$

for $\theta < \frac{1}{\phi}$. **[12]**

(ii) Write down the mgf of the random variable $W = \sum_{i=1}^{n} X_i$ where $X_1, X_2, \ldots, X_n$ are independent random variables each with the same distribution as X. **[1]**

(iii) Write down the mgf of the random variable $Y = \frac{2W}{\phi}$. Given that the mgf of the random variable V having the χ^2_m distribution is $\mathrm{M}_V(\theta) = (1-2\theta)^{-m/2}$ (for $\theta < \frac{1}{2}$), deduce the distribution of Y. **[3]**

(iv) Deduce that $\mathrm{P}\left(l < \frac{2W}{\phi} < u\right) = 0.95$ where l and u are the lower and upper $2\frac{1}{2}\%$ points of the χ^2_{2nk} distribution. Hence deduce that a 95% confidence interval for ϕ is given by $\left(\frac{2w}{u}, \frac{2w}{l}\right)$ where w is an observation on the random variable W. **[2]**

(v) For the case $k = 2$ and $n = 10$, use percentage points of the χ^2 distribution to write down, in terms of w, an expression for a 95% confidence interval for ϕ. By considering the mgf of W, find in terms of ϕ the expected length of this interval. **[6]**

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, power. **[8]**

(ii) A chemical manufacturer is endeavouring to reduce the amount of a certain impurity in one of its bulk products by improving the production process. The amount of impurity is measured in a convenient unit of concentration, and this is modelled by the Normally distributed random variable X. In the old production process, the mean of X, denoted by μ, was 63 and the standard deviation of X was 3.7. Experimental batches of the product are to be made using the new process, and it is desired to examine the hypotheses $H_0: \mu = 63$ and $H_1: \mu < 63$ for the new process. Investigation of the variability in the new process has established that the standard deviation may be assumed unchanged.

The usual Normal test based on $\overline{X}$ is to be used, where $\overline{X}$ is the mean of X over n experimental batches (regarded as a random sample), with a critical value c such that H_0 is rejected if the value of $\overline{X}$ is less than c. The following criteria are set out.

- If in fact $\mu = 63$, the probability of concluding that $\mu < 63$ must be only 1%.
- If in fact $\mu = 60$, the probability of concluding that $\mu < 63$ must be 90%.

Find c and the smallest value of n that is required. With these values, what is the power of the test if in fact $\mu = 58.5$? **[16]**

Option 4: Design and Analysis of Experiments

4 A trial is being made of four experimental methods, A, B, C and D, for carrying out an industrial process. These are being compared with each other and with the standard method M. The trial is conducted according to a completely randomised design. The results, x, are as follows, in a suitable unit.

Method	Results x	Total	Mean
M	25.0 23.0 30.1 27.5 28.8 25.6 29.2 31.6	220.8	27.6
A	37.3 34.9 30.8 40.2	143.2	35.8
B	36.4 36.6 29.2 44.0 34.8	181.0	36.2
C	32.0 40.1 33.0 36.5	141.6	35.4
D	35.0 31.8 39.0 38.2	144.0	36.0
	Grand total	830.6	

You are also given that $\Sigma x^2 = 28\,260.18$.

(i) The usual statistical model underlying a one-way analysis of variance is given, in the usual notation, by

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

where x_{ij} denotes the jth observation on the ith treatment. State the properties that are assumed for the term that represents experimental error. **[3]**

(ii) Construct the usual analysis of variance table for these data. Stating your hypotheses carefully, test whether there is evidence of differences among the means for the five methods, using a 5% significance level. **[12]**

(iii) In each case using the residual mean square as the estimate of the variance of the experimental error, find a 95% confidence interval for the population mean for method M and a 95% confidence interval for the population mean for method A. What do these confidence intervals suggest about these population means? **[5]**

(iv) The residuals, calculated by subtracting the corresponding method mean from each observation, are given in the table below. For example the first residual for method M is 25.0 – 27.6 = –2.6. Each residual gives a measure of experimental error.

Method	Residuals
M	–2.6 –4.6 2.5 –0.1 1.2 –2.0 1.6 4.0
A	1.5 –0.9 –5.0 4.4
B	0.2 0.4 –7.0 7.8 –1.4
C	–3.4 4.7 –2.4 1.1
D	–1.0 –4.2 3.0 2.2

The diagram in the printed answer book shows a dotplot of the residuals for method M. Complete the diagram by adding the dotplots for the other methods.

Use these dotplots to comment briefly on the assumptions you have stated in part **(i)**. **[4]**

END OF QUESTION PAPER

BLANK PAGE

BLANK PAGE

OCR
Oxford Cambridge and RSA

Tuesday 16 June 2015 – Afternoon

A2 GCE MATHEMATICS (MEI)

4769/01 Statistics 4

QUESTION PAPER

Candidates answer on the Printed Answer Book.

OCR supplied materials:
- Printed Answer Book 4769/01
- MEI Examination Formulae and Tables (MF2)

Other materials required:
- Scientific or graphical calculator

Duration: 1 hour 30 minutes

INSTRUCTIONS TO CANDIDATES

These instructions are the same on the Printed Answer Book and the Question Paper.

- The Question Paper will be found inside the Printed Answer Book.
- Write your name, centre number and candidate number in the spaces provided on the Printed Answer Book. Please write clearly and in capital letters.
- **Write your answer to each question in the space provided in the Printed Answer Book**. Additional paper may be used if necessary but you must clearly show your candidate number, centre number and question number(s).
- Use black ink. HB pencil may be used for graphs and diagrams only.
- Read each question carefully. Make sure 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 scientific or graphical calculator in this paper.
- Final answers should be given to a degree of accuracy appropriate to the context.

INFORMATION FOR CANDIDATES

This information is the same on the Printed Answer Book and the Question Paper.

- The number of marks is given in brackets **[ ]** at the end of each question or part question on the Question Paper.
- 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**.
- The Printed Answer Book consists of **16** pages. The Question Paper consists of **8** pages. Any blank pages are indicated.

INSTRUCTION TO EXAMS OFFICER / INVIGILATOR

- Do not send this Question Paper for marking; it should be retained in the centre or recycled. Please contact OCR Copyright should you wish to re-use this document.

Option 1: Estimation

1 The random variable X has the following probability density function, in which a is a (positive) parameter.

$$\mathrm{f}(x) = \frac{2}{a} x \mathrm{e}^{-x^2/a}, \quad x \geqslant 0.$$

(i) Verify that $\int_0^\infty \mathrm{f}(x)\,\mathrm{d}x = 1$. **[1]**

(ii) Show that $\mathrm{E}(X^2) = a$ and $\mathrm{E}(X^4) = 2a^2$. **[7]**

The parameter a is to be estimated by maximum likelihood based on an independent random sample from the distribution, $X_1, X_2, \ldots, X_n$.

(iii) Show that the logarithm of the likelihood function is

$$n \ln 2 - n \ln a + \sum_{i=1}^{n} \ln X_i - \frac{1}{a} \sum_{i=1}^{n} X_i^2.$$

Hence obtain the maximum likelihood estimator, $\hat{a}$, for a.

[You are not required to verify that any turning point you find is a maximum.] **[7]**

(iv) Using the results from part **(ii)**, show that $\hat{a}$ is unbiased for a and find the variance of $\hat{a}$. **[5]**

(v) In a particular random sample from this distribution, $n = 100$ and $\sum x_i^2 = 147.1$. Obtain an approximate 95% confidence interval for a. (You may assume that the Central Limit Theorem holds in this case.) **[4]**

Option 2: Generating Functions

2 The random variable Z has the standard Normal distribution. The random variable Y is defined by $Y = Z^2$.

You are given that Y has the following probability density function.

$$f(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}, \quad y > 0.$$

(i) Show that the moment generating function (mgf) of Y is given by

$$M_Y(\theta) = (1-2\theta)^{-\frac{1}{2}}.$$ **[6]**

(ii) Use the mgf to obtain $E(Y)$ and $Var(Y)$. **[5]**

The random variable U is defined by

$$U = Z_1^2 + Z_2^2 + \ldots + Z_n^2,$$

where $Z_1, Z_2, \ldots, Z_n$ are independent standard Normal random variables.

(iii) State an appropriate general theorem for mgfs and hence write down the mgf of U. State the values of $E(U)$ and $Var(U)$. **[4]**

The random variable W is defined by

$$W = \frac{U-n}{\sqrt{2n}}.$$

(iv) Show that the logarithm of the mgf of W is

$$-\sqrt{\frac{n}{2}}\theta - \frac{n}{2}\ln\left(1-\sqrt{\frac{2}{n}}\theta\right).$$

Use the series expansion of $\ln(1-t)$ to show that, as $n \to \infty$, this expression tends to $\frac{1}{2}\theta^2$.

State what this implies about the distribution of W for large n. **[9]**

Option 3: Inference

3 At an agricultural research station, trials are being carried out to compare a standard variety of tomato with one that has been genetically modified (GM). The trials are concerned with the mean weight of the tomatoes and also with the aesthetic appearance of the tomatoes.

(a) **(i)** Tomatoes of the standard and GM varieties are grown under similar conditions. The tomatoes are weighed and the data are summarised as follows.

Variety	Sample size	Sum of weights (g)	Sum of squares of weights (g^2)
Standard	30	3218.3	349257
GM	26	2954.1	338691

Carry out a test, using the Normal distribution, to investigate whether there is evidence, at the 5% level of significance, that the two varieties of tomato differ in mean weight.

State one assumption required for this test to be valid. **[10]**

(ii) The data in part **(i)** could have been used to carry out a test for the equality of means based on the *t* distribution. State **two** additional assumptions required for this test to be valid.

Discuss briefly which test would be preferable in this case. **[4]**

(b) In order to judge whether, on the whole, GM tomatoes have a better aesthetic appearance than standard tomatoes, a trial is carried out as follows. 10 of each variety are chosen and a consumer panel is asked to arrange the 20 tomatoes in order according to their appearance.

(i) State **two** important features of the way in which this trial should be designed.

Comment briefly on how reliable the evidence from the trial is likely to be. **[3]**

(ii) The order in which the consumer panel arranges the tomatoes is as follows. The tomato with best appearance is listed first. *G* and *S* denote GM and standard tomatoes respectively.

G G G S G G G S G S S S G G S G S S S S

Carry out an appropriate test at the 1% level of significance. **[7]**

Option 4: Design and Analysis of Experiments

4 **(a)** The standard one-way Analysis of Variance (ANOVA) model is expressed in the usual notation as follows.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

(i) Explain what the terms Y_{ij}, μ, α_i and ε_{ij} represent. **[4]**

(ii) State a distributional assumption about ε_{ij} and explain briefly why this assumption is required. **[4]**

(iii) State the null and alternative hypotheses for the usual one-way ANOVA test. Explain clearly how to interpret the two possible outcomes of the ANOVA test. **[4]**

(b) I drive frequently between two cities, A and B. There are k different routes that I can take. On each of these routes the journey time varies according to time of day, traffic conditions and so on.

In order to test whether or not there are any differences between the mean journey times on the k routes, I chose a route at random for each of N journeys. I recorded the time for each journey, entered the data into a spreadsheet, and carried out an ANOVA analysis. Part of the output was as follows.

Source of variation	Sum of squares	Degrees of freedom
Between groups	333.77	
Within groups		15
Total	752.96	18

(i) State the values of k and N. **[2]**

(ii) Complete the analysis using a 5% significance level. **[10]**

END OF QUESTION PAPER

BLANK PAGE

BLANK PAGE