RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2615

Statistics 3

Monday **21 JUNE 2004** Morning 1 hour 20 minutes

Additional materials:
Answer booklet
Graph paper
MEI Examination Formulae and Tables (MF12)

TIME 1 hour 20 minutes

INSTRUCTIONS TO CANDIDATES

- Write your Name, Centre Number and Candidate Number in the spaces provided on the answer booklet.
- Answer **all** questions.
- You are permitted to use a graphical calculator in this paper.

INFORMATION FOR CANDIDATES

- The allocation 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.
- Final answers should be given to a degree of accuracy appropriate to the context.
- The total number of marks for this paper is 60.

This question paper consists of 3 printed pages and 1 blank page.

Registered Charity 1066969

[Turn over

1 The amount of dust in an air-conditioned room is measured at frequent intervals. The amount is recorded on a scale from 0 to 1, and is found to be well modelled by the continuous random variable X having probability density function given by

$$f(x) = k(5x^3 - 13x^2 + 8x), \qquad 0 \leqslant x \leqslant 1.$$

(i) Show that $k = \frac{12}{11}$. [2]

(ii) Find $E(X)$ and show that $\text{Var}(X) = \frac{29}{605}$. [6]

(iii) Find the cumulative distribution function of X. Deduce that the median of X is a root of

$$30x^4 - 104x^3 + 96x^2 - 11 = 0. \qquad [4]$$

[You are *not* required to solve this equation.]

(iv) State the name and parameters of the distribution that may be used to model the average amount of dust over a random sample of 50 measurements. [3]

2 A garden centre sells four fertilisers, A, B, C and D. The weekly demands for these are modelled by independent Normal distributions with means and standard deviations, in kilograms, as follows.

	Mean	Standard deviation
A	65	7
B	45	3
C	30	8
D	50	4

(i) Find the probability that, in any week, the demand for A is less than 70 kg. [1]

(ii) Find the probability that, in any week, the total demand for these fertilisers is more than 200 kg. [3]

(iii) Find the probability that, in any week, the demand for A exceeds the combined demands for B and C. [4]

(iv) The profits (in pence per kilogram) for fertilisers A, B, C, D are 35, 40, 60, 25 respectively. Find the mean and variance of the weekly overall profit. [5]

(v) Find the probability that, in any week, the overall profit exceeds £70. [2]

3 A metal alloy is specified as containing, on average, 4.5% of a particular element. An inspector selects a random sample of 10 specimens of the alloy and undertakes a detailed analysis of the composition of each specimen. The percentages of this element are found to be as follows.

4.12 4.31 4.60 4.53 4.88 4.47 4.76 4.51 4.50 4.62

(i) State the appropriate null and alternative hypotheses for the usual t test that the inspector would wish to carry out. [2]

(ii) What condition is necessary for the correct use of this test? [1]

(iii) Carry out the test, using a 5% significance level. [7]

The inspector later analyses a random sample of 80 specimens of the alloy.

(iv) The inspector finds that the percentages of the element in these 80 specimens are summarised by $\Sigma x = 357.6$ and $\Sigma x^2 = 1605.20$. Use this information to obtain a 95% confidence interval for the mean percentage of the element. [5]

4 A researcher is studying the distribution of a particular species of tree over a large area of countryside.

(a) The researcher divides part of the area into 100 regions of equal size and counts the number, x, of these trees in each region. The observed frequencies, f, are as follows.

x	0	1	2	3	4	5	>5
f	30	26	19	15	9	1	0

You are given that the sample mean $\bar{x}$ is 1.5.

Initially the researcher investigates whether trees of this species are scattered at random over the whole area so that the underlying random variable X would have a Poisson distribution.

Use a suitable statistical procedure and a 10% significance level to assess the goodness of fit of a Poisson distribution, commenting briefly on any discrepancies. [10]

(b) The researcher notices that none of these trees is less than 16 m or more than 180 m from its nearest neighbour. The researcher assumes that a continuous uniform distribution describes the distances between each of these trees and its nearest neighbour. State the mean and use calculus to find the standard deviation of this distribution. [5]

Mark Scheme

Q1	f(x) = $k(5x^3 - 13x^2 + 8x)$, $0 \le x \le 1$.			
(i)	$1=\int_0^1 \mathrm{f}(x)dx$ $=k\left[\frac{5x^4}{4}-\frac{13x^3}{3}+\frac{8x^2}{2}\right]_0^1=k\left(\frac{5}{4}-\frac{13}{3}+\frac{8}{2}\right)$	M1	Set up requirement: integral, including limits (which may appear later), "= 1". Use of c.d.f. requires sight of F(1) = 1 o.e.	
	$=k\frac{15-52+48}{12}=k\frac{11}{12}$ $\therefore k=\frac{12}{11}$	A1	Exact answer shown convincingly. (e.g. evidence of use of upper limit.) BEWARE PRINTED ANSWER.	2
(ii)	$\mathrm{E}(X)=\int_0^1 x\mathrm{f}(x)dx$	M1	Definition of mean, including limits (which may appear later).	
	$=\frac{12}{11}\left[\frac{5x^5}{5}-\frac{13x^4}{4}+\frac{8x^3}{3}\right]_0^1$	A1	Successfully integrated.	
	$=\frac{12}{11}\left(\frac{5}{5}-\frac{13}{4}+\frac{8}{3}\right)=\frac{12}{11}\times\frac{12-39+32}{12}=\frac{5}{11}$	A1	Correct use of limits leading to final answer. C.a.o.; accept decimal 0·455 o.b.	
	$\mathrm{E}(X^2)=\int_0^1 x^2\mathrm{f}(x)dx$ $=\frac{12}{11}\left(\frac{5}{6}-\frac{13}{5}+\frac{8}{4}\right)=\frac{12}{11}\times\frac{50-156+120}{60}$	M1	Integral for E(X^2) including limits (which may appear later).	
	$=\frac{12\times 14}{11\times 60}=\frac{14}{55}$	A1	Accept unsimplified and/or decimal 0·255 o.b. Allow in terms of k.	
	$\mathrm{Var}(X)=\mathrm{E}(X^2)-(\mathrm{E}(X))^2=\frac{14}{55}-\left(\frac{5}{11}\right)^2$ $=\frac{14\times 11-5\times 5\times 5}{55\times 11}=\frac{29}{605}$	A1	Method must be clear, and exact answer shown convincingly. BEWARE PRINTED ANSWER.	6
(iii)	$\mathrm{F}(x)=\int_0^x \mathrm{f}(t)dt$	M1	Definition of c.d.f., including limits or use of "+c" (which may appear later).	
	$=\frac{12}{11}\left(\frac{5x^4}{4}-\frac{13x^3}{3}+\frac{8x^2}{2}\right)$	A1	… or equivalent expression; condone absence of domain [0,1].	
	Median is a root of F(x) = ½. i.e. $\frac{12}{11}\times\frac{15x^4-52x^3+48x^2}{12}=\frac{1}{2}$	M1	Definition of median, f.t. c's c.d.f.	
	i.e. $30x^4-104x^3+96x^2-11=0$	A1	BEWARE PRINTED ANSWER.	4
(iv)	$\mathrm{N}\left(\frac{5}{11},\frac{29}{50\times 605}\left[=\frac{29}{30250}=0\cdot 000958(6)\right]\right)$	B1 B1F B1	Normal. Mean; f.t. candidate's mean. Variance. Accept sd if indicated clearly as such. If the name of the distribution is wrong or missing then allow the marks for the parameters either if they are the conventional parameters for the named distribution or they are named explicitly.	3
				15

Q2	Mean SD Profit A 65 7 35 B 45 3 40 C 30 8 60 D 50 4 25		Throughout this question do not allow continuity corrections. When a candidate's answers suggest that (s)he appears to have neglected to use the difference columns of the Normal distribution tables penalise the first occurrence only.	
(i)	$\mathrm{P}(A<70)=\mathrm{P}\left(\mathrm{N}(0,1)<\frac{70-65}{7}=0\cdot7143\right)=0\cdot7624$	B1		1
(ii)	Total demand ~ N(190, $7^2+3^2+8^2+4^2=138$) $\therefore \mathrm{P}(\text{total}>200)=$ $\mathrm{P}\left(\mathrm{N}(0,1)>\frac{200-190}{\sqrt{138}}=0\cdot8513\right)=0\cdot1973$	B1 B1 B1	Mean. Variance. Accept sd = √138 = 11·747… c.a.o. (= 1 – 0·8027).	 3
(iii)	Want $\mathrm{P}(A-(B+C)>0)$ $A-B-C\sim$ N(– 10, $7^2+3^2+8^2=122$) $\therefore \mathrm{P}(\text{this}>0)$ $=\mathrm{P}\left(\mathrm{N}(0,1)>\frac{0-(-10)}{\sqrt{122}}=0\cdot9054\right)=0\cdot1827$	M1 B1 B1 A1	Or $\mathrm{P}((B+C)-A<0)$. Mean. Or +10 for alternative method. Variance. Accept sd = √122 = 11·045… N.B. Method and mean should be consistent with each other. c.a.o. (= 1 – 0·8173). Or $\mathrm{P}\left(\mathrm{N}(0,1)<\frac{0-10}{\sqrt{122}}=-0\cdot9054\right)$.	 4
(iv)	Profit = $35A+40B+60C+25D$ Mean profit = 35×65 + 40×45 + 60×30 + 25×50 = 7125 (pence) Variance = $35^2\times7^2+40^2\times3^2+60^2\times8^2+25^2\times4^2$ = 314825 (pence squared)	 M1 A1 M1 M1 A1	 Accept 71.25 (£). For 35^2 etc. For $\times7^2$ etc. and the sum of the products. Depends on <u>both</u> M's. Accept 31·4825 (£^{2}).	 5
(v)	$\therefore \mathrm{P}(\text{profit}>7000)=\mathrm{P}(\mathrm{N}(7125,314825)>7000)$ $=\mathrm{P}\left(\mathrm{N}(0,1)>\frac{-125}{\sqrt{314825}}=-0\cdot2228\right)=0\cdot5882$	M1 A1	For use of Normal (might be implicit) with c's parameters from part (iv). c.a.o.	 2
				15

Q3				
(i)	$H_0 : \mu = 4\cdot5$ $H_1 : \mu \neq 4\cdot5$ Where μ is the (population) mean percentage of the element in the alloy.	B1 B1	*Both* must be correct. Do **not** allow any other symbol, including $\overline{X}$ or similar, unless it is clearly and explicitly stated to be a population mean. Allow statements in words (see below). μ must be defined verbally. Must indicate "mean"; condone "average". Allow absence of "population" if μ is used, otherwise insist on "population".	2
(ii)	Underlying distribution is Normal.	B1	Do not accept statements about "the data" or "the sample" etc.	1
(iii)	$\bar{x} = 4\cdot53,\ s_{n-1} = 0\cdot2134\ \ (s_{n-1}{}^2 = 0\cdot04553)$ Test statistic is $\dfrac{4\cdot53-4\cdot5}{\left(\dfrac{0\cdot2134}{\sqrt{10}}\right)}$ $= 0\cdot44(46)$ Refer to t_9. Double-tail 5% point is 2·262. Not significant. Seems mean can be taken as 4·5.	B1 M1 A1 M1 A1 E1 E1	Allow $s_n = 0\cdot2024$, ($s_n{}^2 = 0\cdot04098$) only if used correctly in sequel. Allow c's $\bar{x}$ and/or s_{n-1}. Allow alternative: 4·5 ± (c's 2·262) × $\frac{0\cdot2134}{\sqrt{10}}$ (= 4·3474, 4·6526) for subsequent comparison with $\bar{x}$. (Or $\bar{x}$ ± (c's 2·262) × $\frac{0\cdot2134}{\sqrt{10}}$ (= 4·3774, 4·6826) for comparison with 4·5.) c.a.o. (but ft from here if this is wrong.) Use of $4\cdot5 - \bar{x}$ scores M1A0, but next 4 marks still available. No ft from here if wrong. No ft from here if wrong. ft only c's test statistic. ft only c's test statistic. S.C. (t_9 and 1·833) or (t_{10} and 2·228) can score max 1 of these last 2 marks if either form of conclusion is stated, consistent with the test statistic and critical value.	7
(iv)	Now $\bar{x} = 4\cdot47$ $s_{n-1}{}^2 = \dfrac{1}{79}\left\{1605\cdot2 - \dfrac{(357\cdot6)^2}{80}\right\}$ $= \dfrac{1}{79} \times 6\cdot728 = 0\cdot0851(645)$ $s_{n-1} = 0\cdot2918$ CI is given by 4·47 ± 1·96 $\times \dfrac{0\cdot2918}{\sqrt{80}}$ = 4·47 ± 0·06(39) = (4·41, 4·53)	B1 M1 B1 M1 A1	For both $\bar{x}$ and s or s^2. Allow divisor 80 here, giving 0·29 ($s_n{}^2 = 0\cdot0841$). Must be c's $\bar{x}$ ± … (Implies correct method of use of Normal distribution.) Allow t_{79} (1·99(45)) provided $s_{n-1}/\sqrt{80}$ or $s_n/\sqrt{79}$ also used. Allow c's s_n or s_{n-1}. Allow $\sqrt{79}$ if s_n used. c.a.o. Must be written as an interval.	5
				15

Q4				
(a)	x 0 1 2 3 4 5 >5 o 30 26 19 15 9 1 0 p 0·2231 0·3347 0·2510 0·1256 0·0470 0·0141 0·0045 From Poisson (1·5)*. e 22·31 33.47 25·10 12·56 4·70 1·41 0·45 Combining last 3 cells: o = 10 e = 6·56 * These are from cumulative tables. Might differ by 1 in last d.p. if calculated directly.			
		M1	For apparently correct method for e's (>5 cell must be present, or equivalent if candidate obviously realises to group cells earlier).	
		A1	If all correct or if $\Sigma e_i = 100$. (But A0 if rounded to integers.)	
		M1	For grouping (cells where $e \leq 5$).	
	$X^2 = 2·6506(5) + 1·6672 + 1·4825 + 0·4740 + 1·8039$ $= 8·07(82)$	M1 A1	For evidence of correct method for X^2. c.a.o. (but ft from here if this is wrong.) e_i to 1 d.p. gives $X^2 = 8·16$.	
	Refer to χ_ν^2, where ν = no of cells in candidate's table – 2 (ideally, $\nu = 3$)	M1	Allow this mark if it agrees with candidate's table, and then ft as below. Accept anything that implies use of this distribution.	
	For $\nu = 3$ upper 10% point is 6·251.	A1	Allow candidate's ν if preceding M1 awarded. No ft from here if not correct point from candidate's χ^2.	
	Significant.	E1	No f.t. of above M1 or A1 if wrong, except for Special Case: $\nu + 1$ and its 10% point can get EITHER (but not both) of these 2 marks for the conclusion. ($\nu = 4$, cv = 7·779)	
	Seems Poisson does not fit.	E1	"Model does not fit data" NOT "data do not fit model".	
	Seems there are too many zeros and (4's and 5's) – possibly suggesting "clumping".	E1	Accept any reasonable descriptive comment about discrepancies.	10
(b)	$\mathrm{f}(x) = \dfrac{1}{180-16} = \dfrac{1}{164}$	B1	Condone absence of domain [16, 180].	
	Mean is $\dfrac{196}{2} = 98$	B1	Mean.	
	$\mathrm{E}(X^2) = \int_{16}^{180} \dfrac{x^2}{164} dx = \left[\dfrac{x^3}{492}\right]_{16}^{180} = 11845\cdot\dot{3}$	M1	Must be using calculus for E(X^2).	
	$\mathrm{Var}(X) = 11845\cdot\dot{3} - (98)^2$ $= 2241\cdot\dot{3}$	M1	Use of E(X^2) – E(X)^2. Depends on previous method mark.	
	SD = 47·34(27)	A1	c.a.o. Award A0 if left as variance.	5
				15

Examiner’s Report

2615 Statistics 3

General Comments

There were about 1000 candidates for this paper, which was very similar to the size of entry in June 2003. The general standard of the scripts seen was satisfactory: many candidates were clearly well prepared for the paper, although one wonders how many candidates are made aware of the recommendations of reports such as this one.

Comments and explanations continue to be a consistent weakness.

In questions about hypothesis tests, the conclusion at the end of the test should make explicit reference to the context of the question. Furthermore the wording of conclusions needs to be less assertive, retaining an element of doubt and acknowledging that hypothesis tests do not *prove* anything conclusively.

Invariably all four questions were attempted. The average marks scored on questions 1, 2 and 3 were pleasing and broadly similar to each other. However the marks for Question 4 were considerably lower: many attempts looked rushed or half-hearted or both and it appeared that laborious attempts at Questions 1 and 2 may well have resulted in a shortage of time for this question.

Comments on Individual Questions

Q.1 Throughout this question candidates made a poor impression by their lack of attention to correct notation and carefully presented working. For example in part (iii) (and again in Question 4 part (b)) the variance was often presented as $\mathrm{E}(X^2)$, with " – $\mathrm{E}(X)^2$ " only appearing, without explanation, at some later point. Furthermore candidates should be alerted to the need for full and detailed working when asked to show a given result.

(i) The correct integral was usually attempted, but sometimes the limits were missing and sometimes it was not equated to 1. Also, since the required, *exact* value of k was given in the question, the onus was firmly on the candidate to show full working details of how (s)he obtained it. Such evidence was not always seen.

(ii) The expectation was usually obtained correctly. The value of $\mathrm{E}(X^2)$ was also often found correctly. However for the variance candidates were required to obtain another *exact* printed value, and it was frequently unclear whether or not the final calculation (an addition of fractions) had genuinely been carried out.

(iii) This part was less well done. Most seemed to know that an integral was required in order to find the c.d.f. but, more often than not, the limits (or the use of a constant of integration) were omitted. Even when limits were present, the letter used was the same as the independent variable, and explicit evidence of their use was likely to be missing.

There was more success with the second half of this part: the definition of the median was often stated correctly and some algebraic manipulation was shown towards the required equation (another printed answer).

(iv) Weaker candidates tended to omit this part of the question while many others quoted “Normal” but without parameters.

Q.2 (i) This part was usually correct. The most common error was finding the wrong tail.

(ii) There were very many correct answers to this part too, but there were also quite a few careless errors: finding the wrong tail, as above, or mean = 170 and s.d. = 22 were fairly common.

(iii) This time the problem for weaker candidates was the identification of the correct variable (*A* – *B* – *C*). Candidates who were able to do this correctly were usually able to complete the part successfully.

(iv) Quite a few candidates encountered problems with this part. In particular the variance was usually incorrect. For instance 35Var(*A*) + 40Var(*B*) + 60Var(*C*) + 25Var(*D*) = 6315 was a common error. The mean and variance of {35, 40, 60, 25} were also seen on a number of occasions. Candidates who could not convert correctly from "pence squared" to "pounds squared" created further problems for themselves.

(v) Candidates usually attempted to use the appropriate method for this part but, since it depended on the previous part for the parameters of the distribution, the correct probability was relatively rare.

Q.3 (i) The required hypotheses were usually, but not always, stated correctly in symbolic form. However in questions about hypothesis tests of the mean, candidates are expected to have acquired the habit of defining the symbol, μ, as the *population* mean of the variable under consideration. In many cases it is not clear that they realise that they are testing a population mean. In addition, the conclusion at the end of the test should also make explicit reference to this, as part of a statement set in the context of the question.

(ii) Various misconceptions were fairly commonplace, e.g. "the data is (sic) Normally distributed" or "the sample is Normally distributed". The response of many other candidates implied that they were unaware of any distributional requirement for the *t* test.

(iii) In general there was some good work here – the test statistic was usually correct as was the critical value leading to a correct conclusion of "not significant". However, as mentioned above, this conclusion needs to be interpreted in context, with reference to the population mean, and in language that is non-assertive. Contextual conclusions of this nature were usually either deficient or completely missing.

(iv) A sample size of 80 should be large enough to invoke the Central Limit Theorem, and so a percentage point taken from the Normal distribution was appropriate for the calculation of this confidence interval. Even so there were some attempts to use the *t* distribution, with and without interpolation for v = 79. The most common error was the confusion of the standard deviation and the variance.

Q.4 (a) Although there were scripts with good high scoring attempts at this part, completely correct answers were very few and far between. The calculation of the test statistic tended to be reasonably well done. Some common errors included a failure to calculate an open ended class for $x > 5$ and/or forgetting to group classes where the expected frequencies were below 5. However, marks were consistently lost in conducting the hypothesis test mainly as a result of deciding on the wrong number of degrees of freedom: candidates overlooked the fact that the mean of the Poisson model had been estimated from the data. Added to this, the contextual conclusion was usually either missing or too assertive.

Candidates were also asked to comment “briefly on any discrepancies”. Almost all candidates offered no comment at all; perhaps they thought it was no more than a reminder to state the conclusion of their test.

(b) This part was usually either not attempted or done badly. Many candidates gave 82 as the mean, and many quoted the formula for the variance of a rectangular distribution, despite being told to use calculus. A few of those who did use calculus lost the final mark because they gave the variance instead of the standard deviation. As in Question 1 correct notation and careful presentation were in short supply.