

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

**Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education**

MEI STRUCTURED MATHEMATICS

4768

Statistics 3

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 **all** the 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.

2

- 1** Design engineers are simulating the load on a particular part of a complex structure. They intend that the simulated load, measured in a convenient unit, should be given by the random variable X having probability density function

$$f(x) = 12x^3 - 24x^2 + 12x, \quad 0 \leq x \leq 1.$$

- (i)** Find the mean and the mode of X . [6]

- (ii)** Find the cumulative distribution function $F(x)$ of X .

Verify that $F\left(\frac{1}{4}\right) = \frac{67}{256}$, $F\left(\frac{1}{2}\right) = \frac{11}{16}$ and $F\left(\frac{3}{4}\right) = \frac{243}{256}$. [3]

The engineers suspect that the process for generating simulated loads might not be working as intended. To investigate this, they generate a random sample of 512 loads. These are recorded in a frequency distribution as follows.

Load x	$0 \leq x \leq \frac{1}{4}$	$\frac{1}{4} < x \leq \frac{1}{2}$	$\frac{1}{2} < x \leq \frac{3}{4}$	$\frac{3}{4} < x \leq 1$
Frequency	126	209	131	46

- (iii)** Use a suitable statistical procedure to assess the goodness of fit of X to these data. Discuss your conclusions briefly. [9]

- 2 A bus route runs from the centre of town A through the town's urban area to a point B on its boundary and then through the country to a small town C. Because of traffic congestion and general road conditions, delays occur on both the urban and the country sections. All delays may be considered independent.

The scheduled time for the journey from A to B is 24 minutes. In fact, journey times over this section are given by the Normally distributed random variable X with mean 26 minutes and standard deviation 3 minutes.

The scheduled time for the journey from B to C is 18 minutes. In fact, journey times over this section are given by the Normally distributed random variable Y with mean 15 minutes and standard deviation 2 minutes.

Journey times on the two sections of route may be considered independent. The timetable published to the public does not show details of times at intermediate points; thus, if a bus is running early, it merely continues on its journey and is not required to wait.

- (i) Find the probability that a journey from A to B is completed in less than the scheduled time of 24 minutes. [3]
- (ii) Find the probability that a journey from A to C is completed in less than the scheduled time of 42 minutes. [3]
- (iii) It is proposed to introduce a system of bus lanes in the urban area. It is believed that this would mean that the journey time from A to B would be given by the random variable $0.85X$. Assuming this to be the case, find the probability that a journey from A to B would be completed in less than the currently scheduled time of 24 minutes. [3]
- (iv) An alternative proposal is to introduce an express service. This would leave out some bus stops on both sections of the route and its overall journey time from A to C would be given by the random variable $0.9X + 0.8Y$. The scheduled time from A to C is to be given as a whole number of minutes. Find the least possible scheduled time such that, with probability 0.75, buses would complete the journey on time or early. [6]
- (v) A programme of minor road improvements is undertaken on the country section. After their completion, it is thought that the random variable giving the journey time from B to C is still Normally distributed with standard deviation 2 minutes. A random sample of 15 journeys is found to have a sample mean journey time from B to C of 13.4 minutes. Provide a two-sided 95% confidence interval for the population mean journey time from B to C. [3]

3 An employer has commissioned an opinion polling organisation to undertake a survey of the attitudes of staff to proposed changes in the pension scheme. The staff are categorised as management, professional and administrative, and it is thought that there might be considerable differences of opinion between the categories. There are 60, 140 and 300 staff respectively in the categories. The budget for the survey allows for a sample of 40 members of staff to be selected for in-depth interviews.

- (i) Explain why it would be unwise to select a simple random sample from all the staff. [2]
- (ii) Discuss whether it would be sensible to consider systematic sampling. [3]
- (iii) What are the advantages of stratified sampling in this situation? [2]
- (iv) State the sample sizes in each category if stratified sampling with as nearly as possible proportional allocation is used. [1]

The opinion polling organisation needs to estimate the average wealth of staff in the categories, in terms of property, savings, investments and so on. In a random sample of 11 professional staff, the sample mean is £345 818 and the sample standard deviation is £69 241.

- (v) Assuming the underlying population is Normally distributed, test at the 5% level of significance the null hypothesis that the population mean is £300 000 against the alternative hypothesis that it is greater than £300 000. Provide also a two-sided 95% confidence interval for the population mean. [10]

- 4 A company has many factories. It is concerned about incidents of trespassing and, in the hope of reducing if not eliminating these, has embarked on a programme of installing new fencing.

- (i) Records for a random sample of 9 factories of the numbers of trespass incidents in typical weeks before and after installation of the new fencing are as follows.

Factory	A	B	C	D	E	F	G	H	I
Number before installation	8	12	6	4	14	22	4	13	14
Number after installation	6	11	0	1	18	10	11	5	4

Use a Wilcoxon test to examine at the 5% level of significance whether it appears that, on the whole, the number of trespass incidents per week is lower after the installation of the new fencing than before. [9]

- (ii) Records are also available of the costs of damage from typical trespass incidents before and after the introduction of the new fencing for a random sample of 7 factories, as follows (in £).

Factory	T	U	V	W	X	Y	Z
Cost before installation	1215	95	546	467	2356	236	550
Cost after installation	1268	110	578	480	2417	318	620

Stating carefully the required distributional assumption, provide a two-sided 99% confidence interval based on a t distribution for the population mean difference between costs of damage before and after installation of the new fencing.

Explain why this confidence interval should not be based on the Normal distribution. [9]

Mark Scheme 4768
June 2006

Q1	$f(x) = 12x^3 - 24x^2 + 12x, \quad 0 \leq x \leq 1$												
(i)	$E(X) = \int_0^1 xf(x)dx$ $= 12 \left[\frac{x^5}{5} - 2\frac{x^4}{4} + \frac{x^3}{3} \right]_0^1$ $= 12 \left[\frac{1}{5} - \frac{2}{4} + \frac{1}{3} \right] = 12 \times \frac{1}{30} = \frac{2}{5}$ <p>For mode, $f'(x) = 0$</p> $f'(x) = 12(3x^2 - 4x + 1) = 12(3x - 1)(x - 1)$ $\therefore f'(x) = 0 \text{ for } x = 1 \text{ and } x = \frac{1}{3}$ <p>Any convincing argument (e.g. $f''(x)$) that $\frac{1}{3}$ (and not 1) is the mode.</p>	<p>M1 Integral for E(X) including limits (which may appear later). A1 Successfully integrated.</p> <p>A1 Correct use of limits leading to final answer. C.a.o.</p> <p>M1</p> <p>A1</p> <p>A1</p>	6										
(ii)	$\text{cdf } F(x) = \int_0^x f(t)dt$ $= 12 \left(\frac{x^4}{4} - 2\frac{x^3}{3} + \frac{x^2}{2} \right)$ $= 3x^4 - 8x^3 + 6x^2$ $F\left(\frac{1}{4}\right) = \frac{3}{256} - \frac{8}{64} + \frac{6}{16} = \frac{3-32+96}{256} = \frac{67}{256}$ $F\left(\frac{1}{2}\right) = \frac{3}{16} - \frac{8}{8} + \frac{6}{4} = \frac{3-16+24}{16} = \frac{11}{16}$ $F\left(\frac{3}{4}\right) = \frac{3 \times 81}{256} - \frac{8 \times 27}{64} + \frac{6 \times 9}{16} = \frac{243}{256}$	<p>M1 Definition of cdf, including limits (or use of "+c" and attempt to evaluate it), possibly implied later. Some valid method must be seen.</p> <p>A1</p> <p>B1 For all three; answers given; must show convincing working (such as common denominator)! Use of decimals is not acceptable.</p>	3										
(iii)	<table border="1" data-bbox="252 1442 831 1576"> <tr> <td>o_i</td> <td>12 6</td> <td>209</td> <td>131</td> <td>46</td> </tr> <tr> <td>e_i</td> <td>13 4</td> <td>352 - 134 = 218</td> <td>486 - 352 = 134</td> <td>26</td> </tr> </table> $\chi^2 = 0.4776 + 0.3716 + 0.0672 + 15.3846 = 16.30(1)$ <p>Refer to χ^2_3.</p> <p>Very highly significant. Very strong evidence that the model does not fit.</p> <p>The main feature is that we observe many</p>	o_i	12 6	209	131	46	e_i	13 4	352 - 134 = 218	486 - 352 = 134	26	<p>B2 For e_i. B1 if any 2 correct, provided $\Sigma = 512$.</p> <p>M1 A1 M1 Must be some clear evidence of reference to χ^2_3, probably implicit by reference to a critical point (5% : 7.815; 1% : 11.34). No ft (to the A marks) if incorrect χ^2 used, but E marks are still available.</p> <p>A1 There must be at least one reference to "very ...", i.e. the extremeness of the test statistic.</p> <p>A1</p> <p>Or e.g. "big/small" contributions</p>	
o_i	12 6	209	131	46									
e_i	13 4	352 - 134 = 218	486 - 352 = 134	26									

	more loads at the “top end” than expected. The other observations are below expectation, but discrepancies are comparatively small.	E1 E1	to X^2 gets E1, and directions of discrepancies gets E1.	9
				18

Q2	A to B : $X \sim N(26, \sigma = 3)$ B to C : $Y \sim N(15, \sigma = 2)$		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)	$P(X < 24) = P\left(Z < \frac{24 - 26}{3} = -0.6667\right)$ $= 1 - 0.7476 = 0.2524$	M1 A1 A1	For standardising. Award once, here or elsewhere. c.a.o.	3
(ii)	$X + Y \sim N(41,$ $\sigma^2 = 9 + 4 = 13 [\sigma = 3.6056])$ $P(\text{this} < 42) =$ $P\left(Z < \frac{42 - 41}{3.6056} = 0.2774\right) = 0.6093$	B1 B1 A1	Mean. Variance. Accept sd. c.a.o.	3
(iii)	$0.85X \sim N(22.1,$ $\sigma^2 = (0.85)^2 \times 9 = 6.5025 [\sigma = 2.55])$ $P(\text{this} < 24) = P\left(Z < \frac{24 - 22.1}{2.55} = 0.7451\right)$ $= 0.7719$	B1 B1 A1	Mean. Variance. Accept sd. c.a.o.	3
(iv)	$0.9X + 0.8Y \sim N(23.4 + 12 = 35.4,$ $\sigma^2 = (0.9)^2 \times 9 + (0.8)^2 \times 4 = 9.85 [\sigma = 3.1385])$ Require t such that $0.75 = P(\text{this} < t)$ $= P\left(Z < \frac{t - 35.4}{3.1385}\right) = P(Z < 0.6745)$ $\therefore t - 35.4 = 3.1385 \times 0.6745 = 2.1169$ $\Rightarrow t = 37.52$ Must therefore take scheduled time as 38	B1 B1 M1 B1 A1 M1	Mean. Variance. Accept sd. Formulation of requirement (using c's parameters). Any use of a continuity correction scores M0 (and hence A0). 0.6745 c.a.o. Round to next integer above c's value for t .	6
(v)	CI is given by $13.4 \pm 1.96 \frac{2}{\sqrt{15}}$ $= 13.4 \pm 1.0121 = (12.38(79), 14.41(21))$	M1 B1 A1	If <u>both</u> 13.4 and $2/\sqrt{15}$ are correct. (N.B. 13.4 is given as \bar{x} in the question.) (If $3/\sqrt{15}$ used, treat as mis-read and award this M1, but not the final A1.) For 1.96 c.a.o. Must be expressed as an interval.	3
				18

Q3				
(i)	Simple random sample might not be representative - e.g. it might contain only managers.	E1 E1	Or other sensible comment.	2
(ii)	Presumably there is a list of staff, so systematic sampling would be possible. List is likely to be alphabetical, in which case systematic sampling might not be representative. But if the list is in categories, systematic sampling could work well.	E1 E1 E1	Or other sensible comments.	3
(iii)	Would cover the entire population. Can get information for each category.	E1 E1		2
(iv)	5, 11, 24	B1	(4.8, 11.2, 24)	1
(v)	$\bar{x} = 345818$, $s_{n-1} = 69241$ Underlying Normality $H_0: \mu = 300\,000$, $H_1: \mu > 300\,000$ Test statistic is $\frac{345818 - 300000}{\frac{69241}{\sqrt{11}}}$ $= 2.19(47)$. Refer to t_{10} . Upper 5% point is 1.812. Significant. Evidence that mean wealth is greater than 300 000. CI is given by $345818 \pm 2.228 \times \frac{69241}{\sqrt{11}}$ $= 345818 \pm 46513.84 = (299304(.2),$	M1 A1 M1 A1 A1 A1 M1 B1 M1 A1	All given in the question. Allow alternatives: $300000 + (c's 1.812) \times \frac{69241}{\sqrt{11}}$ (= 337829) for subsequent comparison with 345818. or $345818 - (c's 1.812) \times \frac{69241}{\sqrt{11}}$ (= 307988) for comparison with 300000. c.a.o. but ft from here in any case if wrong. Use of $\mu - \bar{d}$ scores M1A0, but ft. No ft from here if wrong. No ft from here if wrong. ft only c's test statistic. ft only c's test statistic. Special case: (t_{11} and 1.796) can score 1 of these last 2 marks if either form of conclusion is given. c.a.o. Must be expressed as an	10

	392331(.8))		interval. ZERO/4 if not same distribution as test. Same wrong distribution scores maximum M1B0M1A0. Recovery to t_{10} is OK.	
				18

Q4																							
(i)	<table border="1" data-bbox="248 331 627 719"> <thead> <tr> <th>Difference s</th> <th>Rank of diff </th> </tr> </thead> <tbody> <tr><td>-2</td><td>2</td></tr> <tr><td>-1</td><td>1</td></tr> <tr><td>-6</td><td>5</td></tr> <tr><td>-3</td><td>3</td></tr> <tr><td>4</td><td>4</td></tr> <tr><td>-12</td><td>9</td></tr> <tr><td>7</td><td>6</td></tr> <tr><td>-8</td><td>7</td></tr> <tr><td>-10</td><td>8</td></tr> </tbody> </table> <p data-bbox="248 752 799 786">$T = 4 + 6 = 10$ (or $1+2+3+5+7+8+9 = 35$)</p> <p data-bbox="248 819 799 1055">Refer to tables of Wilcoxon paired (/single sample) statistic. Lower (or upper if 35 used) 5% tail is needed. Value for $n = 9$ is 8 (or 37 if 35 used). Result is not significant. No evidence to suggest a real change.</p>	Difference s	Rank of diff	-2	2	-1	1	-6	5	-3	3	4	4	-12	9	7	6	-8	7	-10	8	<p data-bbox="858 371 1233 472">M1 For differences. ZERO in this section if differences not used.</p> <p data-bbox="858 607 1302 707">M1 A1 For ranks. FT from here if ranks wrong</p> <p data-bbox="858 752 898 786">B1</p> <p data-bbox="858 819 1254 853">M1 No ft from here if wrong.</p> <p data-bbox="858 887 1370 954">M1 i.e. a 1-tail test. No ft from here if wrong.</p> <p data-bbox="858 954 1254 987">A1 No ft from here if wrong.</p> <p data-bbox="858 987 1233 1021">A1 ft only c's test statistic.</p> <p data-bbox="858 1021 1233 1055">A1 ft only c's test statistic.</p>	9
Difference s	Rank of diff																						
-2	2																						
-1	1																						
-6	5																						
-3	3																						
4	4																						
-12	9																						
7	6																						
-8	7																						
-10	8																						
(ii)	<p data-bbox="248 1093 715 1126">Normality of <u>differences</u> is required.</p> <p data-bbox="248 1160 767 1193">CI MUST be based on DIFFERENCES.</p> <p data-bbox="248 1227 778 1294">Differences are 53, 15, 32, 13, 61, 82, 70</p> <p data-bbox="248 1305 655 1339">$\bar{d} = 46.5714$ $s_{n-1} = 27.0485$</p> <p data-bbox="248 1395 587 1496">CI is given by 46.5714 ± 3.707</p> <p data-bbox="288 1653 818 1753">$= 46.5714 \pm 37.8980 = (8.67(34), 84.47)$</p> <p data-bbox="248 1854 751 1989">Cannot base CI on Normal distribution because sample is small population s.d. is not known</p>	<p data-bbox="858 1093 1353 1261">B1 ZERO/6 for the CI if differences not used. Accept negatives throughout.</p> <p data-bbox="858 1294 1374 1395">B1 Accept $s_{n-1}^2 = 731.62 \dots$ [$s_n = 25.0420$, but do <u>NOT</u> allow this here or in construction of CI.]</p> <p data-bbox="858 1429 1166 1529">M1 Allow c's $\bar{d} \pm \dots$ B1 B1 If t_6 used. 99% 2-tail point for c's t distribution. (Independent of previous mark.)</p> <p data-bbox="858 1664 1110 1697">M1 Allow c's s_{n-1}.</p> <p data-bbox="858 1720 1353 1821">A1 c.a.o. Must be expressed as an interval. [Upper boundary is 84.4694]</p> <p data-bbox="858 1888 1353 1989">E1 E1 Insist on "population", but allow "σ".</p>	9																				
			18																				

4768 - Statistics 3

General Comments

The overall standard of the scripts seen was pleasing: many candidates appeared well prepared for this paper. However, as in January, the quality of their comments, interpretations and explanations was consistently below that of the rest of the work. In particular, in Question 3 the standard of writing was felt to be disappointingly poor. Invariably all four questions were attempted. Question 2 was found to be very accessible and most candidates scored full or nearly full marks. The other Questions, 1, 3 and 4, were well answered, with many candidates scoring relatively high marks, though Question 3 was not quite as well answered as the others. There was no evidence to suggest that candidates found themselves short of time at the end.

Comments on Individual Questions

- 1) **Continuous random variables; Chi-squared hypothesis test for the goodness of fit; simulating loads on structures.**
In this question many candidates lost marks needlessly through carelessness and a lack of good exam technique.
- (i) The mean of the distribution was usually found with little difficulty. On the other hand, when they came to consider the mode, many candidates found that $f(x)$ possessed two stationary points but failed to provide any explanation at all for choosing one rather than the other. Some form of discrimination was required for full marks.
- (ii) It was disappointing just how often candidates' work for the cumulative distribution function was deficient. A valid method should involve a definite integral (complete with limits) or include "+ c" and an attempt to evaluate it. Also the verification of the values of $F(x)$ left very much to be desired. In far too many cases the work presented showed no evidence whatsoever that the candidate had bothered to perform the calculation. For instance candidates who write " $F(\frac{1}{4}) = 3(\frac{1}{4})^4 - 8(\frac{1}{4})^3 + 6(\frac{1}{4})^2 = \frac{67}{256}$ as required" (sic) should not be surprised to receive no credit.
- (iii) In this part the expected frequencies and the calculations up to and including the test statistic were usually correct. Most also went on to compare it to a critical value taken from the correct Chi-squared distribution. However very few appreciated that the test statistic was so large that it would be considered significant at every level of test available to them and that therefore this meant that the evidence should be described as "very highly significant" or words to that effect.
As in the past with questions of this type many candidates neglected to discuss their conclusions, as requested in the question. Perhaps they have not realised that the conclusion of the test informs but is not part of this discussion. On this occasion a comparison of the observed and expected frequencies should have led to an explanation of the exceptional size of the test statistic.
- 2) **Combinations of Normal distributions; journey times on bus routes.**
This question was very well answered with very many scoring full marks. Candidates seemed well prepared for it and understood what was expected. Those who take the trouble to provide simple sketch graphs of the standard Normal distribution do much to enhance the quality of their responses. There was evidence from some quarters of effective use of the built in functions on graphics calculators.

- (i) This part was almost always correct.
 - (ii) And so was this part.
 - (iii) If this part went wrong, as it did just occasionally, it was as a result of an incorrect variance (use of 0.85 instead of 0.85^2).
 - (iv) Again this part was done well. There were more errors with the variance than in part (iii). The question indicated that the scheduled time should be “given as a whole number of minutes” which meant that the final answer should have been rounded up. Quite a few candidates neglected to do this and so lost the final mark.
 - (v) Answers to this part were mostly correct, but a large minority used the wrong percentage point in their calculation. This was usually because they thought mistakenly that a small sample necessitated the use of a t distribution, and failed to realise that the standard deviation given was for the population.
- 3) **Sampling; the t distribution: hypothesis test for the population mean; confidence interval for a population mean; survey of staff opinions and wealth.**
- The answers to the early parts of this question left much to be desired. It was often not clear that candidates understood the different types of sampling. For example, their comments about one type were likely to be so vague that they could equally well apply to any other type and contained nothing that was characteristic of the type under consideration. Concepts of “fair”, “representative” and “bias” seemed poorly understood.
- (i) A common criticism offered for random sampling was that there would be more chosen from the largest group (administrative) than from the other two groups. Few seemed to realise that random sampling carried the risk of picking an exceptional sample. Judged by what they wrote candidates seemed to want equal numbers to be chosen from the three groups in order for the sample to be representative.
 - (ii) As mentioned above it was sometimes not possible to tell that the candidate was talking about a systematic sample, nor that he/she knew what a systematic sample involved. Responses were often confined to repeating the criticism of unrepresentativeness made about random sampling in part (i).
 - (iii) Led by the question, perhaps, candidates commented favourably about stratified sampling as the desirable method, to the effect that it would be representative because the proportions were the same as in the population (which was not always consistent with what they had said in part (i)).
 - (iv) This part was almost always correct.
 - (v) The calculations for the test and of the confidence interval were usually performed competently. Sometimes the final conclusion of the test was not as carefully expressed as is required.
- 4) **Wilcoxon paired sample test for a difference in population medians; paired sample confidence interval for a difference in population means using a t distribution; installation of new fences at factories.**
- (i) This part was very well answered by very many candidates who worked through the Wilcoxon test with obvious ease. Occasionally there were errors in the ranking process or with the critical value or with the wording of the final conclusion.

- (ii) The calculation of the confidence interval was usually correct, though quite often the wrong number of degrees of freedom and/or the wrong percentage point were used. The “wordy” bits of this question were regularly missing or incomplete. Candidates should note that Normality of the *differences* is required; Normality of the two separate populations is neither necessary nor sufficient for this. The word “population” is a necessary part of both the assumption for using the *t* distribution and the explanation for not using the Normal distribution.