OCR
RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 4767

Statistics 2

Monday **22 MAY 2006** Morning 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 3 printed pages and 1 blank page.

Registered Charity 1066969
[Turn over

1 A low-cost airline charges for breakfasts on its early morning flights. On average, 10% of passengers order breakfast.

(i) Find the probability that, out of 8 randomly selected passengers, exactly 1 orders breakfast. [2]

(ii) Use a suitable Poisson approximating distribution to find the probability that the number of breakfasts ordered by 30 randomly selected passengers is

(*A*) exactly 6,

(*B*) at least 8. [5]

(iii) State the conditions under which the use of a Poisson distribution is appropriate as an approximation to a binomial distribution. [2]

(iv) The aircraft carries 120 passengers and the flight is always full. Find the mean μ and variance σ^2 of a Normal approximating distribution suitable for modelling the total number of passengers on the flight who order breakfast. [2]

(v) Use your Normal approximating distribution to calculate the probability that more than 15 breakfasts are ordered on a particular flight. [3]

(vi) The airline wishes to be at least 99% certain that the plane will have sufficient breakfasts for all passengers who order them. Find the minimum number of breakfasts which should be carried on each flight. [4]

2 The head circumference of 3-year-old boys is known to be Normally distributed with mean 49.7 cm and standard deviation 1.6 cm.

(i) Find the probability that the head circumference of a randomly selected 3-year-old boy will be

(*A*) over 51.5 cm,

(*B*) between 48.0 and 51.5 cm. [5]

(ii) Four 3-year-old boys are selected at random. Find the probability that exactly one of them has head circumference between 48.0 and 51.5 cm. [3]

(iii) The head circumference of 3-year-old girls is known to be Normally distributed with mean μ and standard deviation σ. Given that 60% of 3-year-old girls have head circumference below 49.0 cm and 30% have head circumference below 47.5 cm, find the values of μ and σ. [4]

A nutritionist claims that boys who have been fed on a special organic diet will have a larger mean head circumference than other boys. A random sample of ten 3-year-old boys who have been fed on this organic diet is selected. It is found that their mean head circumference is 50.45 cm.

(iv) Using the null and alternative hypotheses H_0: $\mu = 49.7$ cm, H_1: $\mu > 49.7$ cm, carry out a test at the 10% significance level to examine the nutritionist's claim. Explain the meaning of μ in these hypotheses. You may assume that the standard deviation of the head circumference of organically fed 3-year-old boys is 1.6 cm. [6]

3 A student is investigating the relationship between the length x mm and circumference y mm of plums from a large crop. The student measures the dimensions of a random sample of 10 plums from this crop. Summary statistics for these dimensions are as follows.

$$\sum x = 4715 \quad \sum y = 13\,175 \quad \sum x^2 = 2\,237\,725$$

$$\sum y^2 = 17\,455\,825 \quad \sum xy = 6\,235\,575 \quad n = 10$$

(i) Calculate the sample product moment correlation coefficient. [5]

(ii) Carry out a hypothesis test at the 5% significance level to determine whether there is any correlation between length and circumference of plums from this crop. State your hypotheses clearly, defining any symbols which you use. [6]

(iii) (*A*) Explain the meaning of a 5% significance level. [2]

(*B*) State one advantage and one disadvantage of using a 1% significance level rather than a 5% significance level in a hypothesis test. [2]

The student decides to take another random sample of 10 plums. Using the same hypotheses as in part **(ii)**, the correlation coefficient for this second sample is significant at the 5% level. The student decides to ignore the first result and concludes that there is correlation between the length and circumference of plums in the crop.

(iv) Comment on the student's decision to ignore the first result. Suggest a better way in which the student could proceed. [3]

4 A survey of a random sample of 250 people is carried out. Their musical preferences are categorized as pop, classical or jazz. Their ages are categorized as under 25, 25 to 50, or over 50. The results are as follows.

		Musical preference			**Row totals**
		Pop	Classical	Jazz	
Age group	Under 25	57	15	12	**84**
	25 – 50	43	21	21	**85**
	Over 50	22	32	27	**81**
Column totals		**122**	**68**	**60**	**250**

(i) Carry out a test at the 5% significance level to examine whether there is any association between musical preference and age group. State carefully your null and alternative hypotheses. Your working should include a table showing the contributions of each cell to the test statistic. [12]

(ii) Discuss briefly how musical preferences vary between the age groups, as shown by the contributions to the test statistic. [6]

Mark Scheme 4767
June 2006

(i)	$P(X = 1) = 8 \times 0.1^1 \times 0.9^7$ $= 0.383$	M1 for binomial probability $P(X=1)$ A1 (at least 2sf) CAO	**2**
(ii)	$\lambda = 30 \times 0.1 = 3$ (*A*) $P(X = 6) = e^{-3}\frac{3^6}{6!} = 0.0504$(3 s.f.) or from tables $= 0.9665 - 0.9161 = 0.0504$ (*B*) Using tables: $P(X \geq 8) = 1 - P(X \leq 7)$ $= 1 - 0.9881 = 0.0119$	B1 for mean SOI M1 for calculation or use of tables to obtain $P(X=6)$ A1 (at least 2sf) CAO M1 for correct probability calc' A1 (at least 2sf) CAO	**1** **2** **2**
(iii)	*n* is large and *p* is small	B1, B1 Allow appropriate numerical ranges	**2**
(iv)	$\mu = np = 120 \times 0.1 = 12$ $\sigma^2 = npq = 120 \times 0.1 \times 0.9 = 10.8$	B1 B1	**2**
(v)	$P(X > 15.5) = P\left(Z > \frac{15.5-12}{\sqrt{10.8}}\right)$ $= P(Z > 1.065) = 1 - \Phi(1.065) = 1 - 0.8566$ $= 0.1434$ NB Allow full marks for use of N(12,12) as an approximation to Poisson(12) leading to $1 - \Phi(1.010) = 1 - 0.8438 = 0.1562$	B1 for correct continuity correction. M1 for probability using correct tail A1 **cao, (**but FT wrong or omitted CC)	**3**
(vi)	From tables $\Phi^{-1}(0.99) = 2.326$ $\frac{x + 0.5 - 12}{\sqrt{10.8}} \geq 2.326$ $x = 11.5 + 2.326 \times \sqrt{10.8} \geq 19.14$ So 20 breakfasts should be carried NB Allow full marks for use of N(12,12) leading to $x \geq 11.5 + 2.326 \times \sqrt{12} = 19.56$	B1 for 2.326 seen M1 for equation in *x* and positive *z*-value A1 CAO (condone 19.64) A1FT for rounding appropriately (i.e. round up if c.c. used o/w rounding should be to nearest integer)	**4**
			18

Question 2

(i)	$X \sim N(49.7, 1.6^2)$ (*A*) $P(X > 51.5) = P\left(Z > \frac{51.5 - 49.7}{1.6}\right)$ $= P(Z > 1.125)$ $= 1 - \Phi(1.125) = 1 - 0.8696 = 0.1304$ (*B*) $P(X < 48.0) = P\left(Z < \frac{48.0 - 49.7}{1.6}\right)$ $= P(Z < -1.0625) = 1 - \Phi(1.0625)$ $= 1 - 0.8560 = 0.1440$ $P(48.0 < X < 51.5) = 1 - 0.1304 - 0.1440 = 0.7256$	M1 for standardizing M1 for prob. calc. A1 (at least 2 s.f.) M1 for appropriate prob' calc. A1 (0.725 – 0.726)	**5**
(ii)	P(one over 51.5, three between 48.0 and 51.5) $= \binom{4}{1} \times 0.7256 \times 0.2744^3 = 0.0600$	M1 for coefficient M1 for 0.7256×0.2744^3 A1 FT (at least 2 sf)	**3**
(iii)	From tables, $\Phi^{-1}(0.60) = 0.2533$, $\Phi^{-1}(0.30) = -0.5244$ $49.0 = \mu + 0.2533\sigma$ $47.5 = \mu - 0.5244\sigma$ $1.5 = 0.7777\sigma$ $\sigma = 1.929$, $\mu = 48.51$	B1 for 0.2533 or 0.5244 seen M1 for at least one correct equation μ & σ M1 for attempt to solve two correct equations A1 CAO for both	**4**
(iv)	Where μ denotes the mean circumference of the entire population of organically fed 3-year-old boys. $n = 10$, Test statistic $Z = \frac{50.45 - 49.7}{1.6/\sqrt{10}} = \frac{0.75}{0.5060} = 1.482$ 10% level 1 tailed critical value of z is 1.282 1.482 > 1.282 so significant. There is sufficient evidence to reject H_0 and conclude that organically fed 3-year-old boys have a higher mean head circumference.	E1 M1 A1(at least 3sf) B1 for 1.282 M1 for comparison leading to a conclusion A1 for conclusion in context	**6**
			18

Question 3

(i)	**EITHER:** $S_{xy} = \Sigma xy - \frac{1}{n}\Sigma x \Sigma y = 6235575 - \frac{1}{10} \times 4715 \times 13175$ $= 23562.5$ $S_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 2237725 - \frac{1}{10} \times 4715^2 =$ 14602.5 $S_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = 17455825 - \frac{1}{10} \times 13175^2 =$ 97762.5 $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{23562.5}{\sqrt{14602.5 \times 97762.5}} = 0.624$ **OR:** cov $(x,y) = \frac{\sum xy}{n} - \bar{x}\bar{y} = 6235575/10 - 471.5 \times 1317.5$ $= 2356.25$ rmsd(x) $= \sqrt{\frac{S_{xx}}{n}} = \sqrt{(14602.5/10)} = \sqrt{1460.25} = 38.21$ rmsd(y) $= \sqrt{\frac{S_{yy}}{n}} = \sqrt{(97762.5/10)} = \sqrt{9776.25} = 98.87$ $r = \frac{\text{cov}(x,y)}{rmsd(x)rmsd(y)} = \frac{2356.25}{38.21 \times 98.87} = 0.624$	M1 for method for S_{xy} M1 for method for at least one of S_{xx} or S_{yy} A1 for at least one of S_{xy}, S_{xx} or S_{yy} correct M1 for structure of r A1 (0.62 to 0.63) M1 for method for cov (x,y) M1 for method for at least one msd A1 for at least one of S_{xy}, S_{xx} or S_{yy} correct M1 for structure of r A1 (0.62 to 0.63)	**5**
(ii)	H_0: $\rho = 0$ H_1: $\rho \neq 0$ (two-tailed test) where ρ is the population correlation coefficient For n = 10, 5% critical value = 0.6319 Since 0.624 < 0.6319 we cannot reject H_0: There is not sufficient evidence at the 5% level to suggest that there is any correlation between length and circumference.	B1 for H_0, H_1 in symbols B1 for defining ρ B1FT for critical value M1 for sensible comparison leading to a conclusion A1 FT for result B1 FT for conclusion in context	**6**
(iii)	*(A)* This is the probability of rejecting H_0 when it is in fact true. *(B)* Advantage of 1% level – less likely to reject H_0 when it is true. Disadvantage of 1% level – less likely to accept H_1 when H_0 is false.	B1 for 'P(reject H_0)' B1 for 'when true' B1, B1 Accept answers in context	**2** **2**

(iv)	The student's approach is not valid. If a statistical procedure is repeated with a new sample, we should not simply ignore one of the two outcomes. The student could combine the two sets of data into a single set of twenty measurements.	E1 E1 – allow suitable alternatives. E1 for combining samples.	**3**
			18

Question 4

(i)

H_0: no association between musical preference and age;
H_1: some association between musical preference and age;

Observed		Musical preference			**Row totals**
		Pop	Classical	Jazz	
Age group	Under 25	57	15	12	**84**
	25 – 50	43	21	21	**85**
	Over 50	22	32	27	**81**
Column totals		**122**	**68**	**60**	**250**

Expected		Musical preference			**Row totals**
		Pop	Classical	Jazz	
Age group	Under 25	40.992	22.848	20.160	**84**
	25 – 50	41.480	23.120	20.400	**85**
	Over 50	39.528	22.032	19.440	**81**
Column totals		**122**	**68**	**60**	**250**

Contributions		Musical preference			
		Pop	Classical	Jazz	
Age group	Under 25	6.25	2.70	3.30	
	25 – 50	0.06	0.19	0.02	
	Over 50	7.77	4.51	2.94	

$X^2 = 27.74$

Refer to χ_4^2
Critical value at 5% level = 9.488
Result is significant
There is some association between age group and musical preference.
NB if H_0 H_1 reversed, or 'correlation' mentioned, do not award first B1or final E1

B1

M1 A2 for expected values (at least 1 dp) (allow A1 for at least one row or column correct)

M1 for valid attempt at $(O-E)^2/E$
A1 for all correct

M1dep for summation
A1 for X^2 (27.7 – 27.8)

B1 for 4 deg of f

B1 CAO for cv
B1FT
E1 (conclusion in context)

1

7

4

(ii)	The values of 6.25 and 7.77 show that under 25's have a strong positive association with pop whereas over 50's have a strong negative association with pop. The values of 4.51 and 2.94 show that over 50's have a reasonably strong positive association with both classical and jazz. The values of 2.70 and 3.30 show that under 25's have a reasonably strong negative associations with both classical and jazz. The 25-50 group's preferences differ very little from the overall preferences.	B1, B1 for specific reference to a value from the table of contributions followed by an appropriate comment B1, B1 (as above for second value) B1, B1 (as above for third value)	**6**
			18

4767 - Statistics 2

General Comments

The majority of candidates seemed to be well prepared and took the opportunity to display their knowledge of the specification with much success. Candidates were consistently very successful handling the variety of calculations needed and able to structure clearly their answers to hypothesis tests, but found interpretive parts of questions more of a challenge; in particular Q3 (iii) and Q4 (ii). The majority of candidates showed sufficient accuracy in their working, resisting the temptation to round-off values to make working easier. Candidates appeared to have ample time to complete the paper within the time allowed.

Comments on Individual Questions

Section A

1) (i) Well answered, mostly producing full marks. The most frequent error was to use the Poisson distribution.

(ii)A Well answered with most candidates scoring full marks. Very few failed to obtain the correct Poisson mean.

(ii)B Well answered. Some candidates used 1-P($X \leq 8$) which scored no marks.

(iii) Most candidates scored full marks. Frequently seen errors include comments relating to general conditions for a Poisson model, and "p is close to 0 or 1".

(iv) Well answered. Many candidates were penalised for stating $\sigma^2=12$.

(v) Many candidates failed to recognise the need for a continuity correction, and were penalised, but went on to secure the remaining marks.

(vi) Well answered with many candidates scoring full marks. Frequently seen errors include using -2.326, and inappropriate rounding for the final (integer) answer.

2) (i)*A* Well answered with most candidates working to a suitable level of accuracy and finding a probability using the correct tail of the Normal distribution. Some candidates lost the accuracy mark through premature rounding or spurious attempts at inappropriate continuity corrections.

(i)*B* Again, well answered. It was pleasing to see most candidates working with the correct form of calculation and to a suitable level of accuracy.

(ii) Well answered. A common error was to use B(3,p) – this was treated as a mis-read. Candidates could obtain full marks by using their answer to part (i)B in place of 0.7256.

(iii) Reasonably well answered with good attempts at solving simultaneous equations regularly seen. Common mistakes included using 0.5244 in place of -0.5244 despite this producing a negative value for σ. Other candidates used (1-0.5244) in place of -0.5244. Some failed to obtain z-values in their equations and scored no marks.

(iv) Very few candidates picked up the mark for stating that μ represents the population mean. With the remaining marks, many candidates scored highly. Some lost the final mark for failing to provide a conclusion in context. Many candidates were penalised for calculating their test statistic using a divisor of 1.6 rather than $1.6/\sqrt{10}$.

3) (i) Well answered. Most scored full marks. Some candidates lost the final accuracy mark through miscalculation.

(ii) Very few candidates obtained the mark for identifying ρ as the population correlation coefficient. Some candidates were penalised for using a one-tailed test. Most identified the correct critical value but a few candidates failed to make an explicit, sensible comparison before coming to their conclusion. Those candidates who failed to relate their conclusions to the context of the question lost the final mark.

(iii)*A* Poorly answered. Not many convincing answers were seen. Many candidates made vague, general comments about the area of the critical region.

(iii)*B* Poorly answered. Incorrect comments such as “it is more accurate” were common.

(iv) Most candidates did better on this part of the question than the previous part. Many lost the final mark through stating that the procedure could be improved by repeating the test. Many candidates realised that ignoring the first result was not a good idea but failed to provide a suitable reason.

4) (i) Well answered. Apart from slips in calculations, marks were commonly lost through failing to give hypotheses in context, rounding expected frequencies to the nearest integer, using the wrong value for ν and/or the wrong critical value. Candidates mentioning correlation were penalised. Some candidates ignored the instruction to show a table of contributions of each cell to the test statistic – these candidates were automatically penalised in part (ii)

(ii) Poorly answered. Very few candidates referred to the contributions to the test statistic despite the request in the question. Commonly, candidates made references only to expected and observed frequencies which could gain at most two marks.