ADVANCED SUBSIDIARY GCE

G242/01

MEI STATISTICS

Statistics 2 (Z2)

WEDNESDAY 21 MAY 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 in capital letters, your 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 **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.
- 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.

OCR is an exempt Charity

[Turn over

1   An environmental group is monitoring the number of Scots pine seedlings on a large region of Scottish moorland. The moorland is divided into plots of equal area. The numbers of Scots pine seedlings in the plots are modelled using a Poisson distribution.

   (i) Write down the conditions for a Poisson model to apply.   [2]

Records for similar regions suggest the mean number of seedlings per plot is 8.

   (ii) Find the probability that in a randomly chosen plot there will be

      (*A*) exactly 7 seedlings,   [2]

      (*B*) more than 7 seedlings.   [2]

   (iii) Find the probability that, out of a random sample of 5 plots, at least one plot contains more than 7 seedlings.   [3]

The heights, in cm, of seedlings are modelled using a Normal distribution with mean 56 and standard deviation 20.

   (iv) Show that the probability that the height of a seedling exceeds 70 cm is 0.242.   [2]

   (v) Find the height exceeded by 95% of seedlings.   [3]

2   A coffee company is investigating the cholesterol-lowering effects of removing a particular oil in the production of a new coffee product. Over a period of several weeks, a group of 12 volunteers drink this new coffee product. For each volunteer, the decrease in cholesterol level is measured. The results, in suitable units, are as follows.

      182   194   202   209   187   165   168   171   178   201   190   173

   (i) Use these data to estimate the mean and variance of the underlying population.   [2]

The coffee company will promote the new coffee product if it believes that the population mean decrease in cholesterol level is more than 175.

   (ii) Use a *t* test at the 5% level of significance to examine whether this sample provides evidence that the mean decrease in cholesterol level is more than 175. State your null and alternative hypotheses clearly.   [9]

   (iii) What assumptions about the underlying population and the sample are necessary for the above test to be valid?   [2]

3 A medical research team is investigating whether there is an association between coronary artery disease and gum disease. They collect data from a random sample of 250 patients who have been tested for possible heart conditions. The results are as follows.

	Patients with gum disease	Patients without gum disease
Patients with coronary artery disease	40	45
Patients without coronary artery disease	103	62

   (i) Carry out a test, at the 5% level of significance, to examine whether these data provide evidence of an association between these classification factors. State clearly the null and alternative hypotheses. [10]

In a further investigation, patients with gum disease are classified by age and by occurrence of coronary artery disease. For a random sample of 160 such patients, the results are as follows.

	Age $\leqslant 45$	$45 <$ Age $\leqslant 60$	Age > 60
Patients with coronary artery disease	4	12	17
Patients without coronary artery disease	49	32	46

  (ii) These results are used to carry out a test for association between age and occurrence of coronary artery disease among people with gum disease. The calculated statistic for the χ^2 test is 8.2808. Given that no cells needed to be merged, what is the conclusion of the test when a 5% significance level is used? Justify your answer with an appropriate critical value. [4]

  (iii) In the light of the conclusions obtained in parts **(i)** and **(ii)**, discuss briefly what the results suggest about links between gum disease and coronary artery disease. [2]

4 It is suspected that, due to global warming, the population of a certain species of ant is decreasing. As part of an ongoing field survey regarding the distribution of this species of ant, a number of traps are located at different sites over a large area. One hour after each trap is set, the number of trapped ants is counted. The results for a sample of 12 traps this year are as follows.

        23   43   16   15   22   33   27   11   18   24   10   20

   (i) Records for this species of ant taken over a number of years show the median number of trapped ants to be 26. Use a Wilcoxon test to examine, at the 5% significance level, whether there has been a decrease in the median number of ants this year. State your null and alternative hypotheses clearly. [12]

  (ii) What assumptions about the variable being tested and the sample used are necessary when carrying out this Wilcoxon test? [2]

[Question 5 is printed overleaf.]

5 A pet products researcher is investigating the bedding material preferences of hamsters. He wishes to find out if hamsters prefer to sleep on a new man-made bedding material rather than the more traditional bedding materials. He has eight hamsters on which to carry out his research. Each hamster is given a choice of three bedding compartments, only one of which contains the new man-made material. The researcher records the numbers of hamsters, x, that choose the new material over a period of 150 days. The results are as follows.

Number of hamsters choosing new material, x	0	1	2	3	4	5	$\geqslant 6$
Frequency, f	10	31	42	34	19	8	6

   (i) Explain why, assuming the hamsters choose independently and randomly, the researcher should model the data using a binomial distribution with $p = \frac{1}{3}$. [1]

 (ii) Use the cumulative binomial probability table with $n = 8$ and $p = \frac{1}{3}$ to find the probabilities corresponding to the values of x in the table. Hence obtain the expected frequencies corresponding to the observed frequencies. [5]

(iii) Carry out a test of the goodness of fit of the above binomial model. Use a 5% significance level. State your null and alternative hypotheses clearly. [9]

G242 Statistics 2

1		(i)	Scots pine seedlings occur randomly and independently with uniform mean rate.	B1 B1	**2**
1	A	(ii)	$e^{-8} \times 8^7 \div 7!$ (0.4530-03134 from tables)	M1	
			0.1396	A1	**2**
	B		$1 - P(X \le 7)$	M1	
			$1 - 0.4530$		
			0.547	A1	**2**
1		(iii)	$(1 - (ii)B)^5$ [$= 0.453^5$]	M1	
			$1 - 0.453^5$	M1	
			0.9809	A1	**3**
1		(iv)	$P(\text{height} > 70) = P(Z > \dfrac{70 - 56}{20}) = P(Z > 0.7)$ $= 1 - \Phi(0.7) = 1 - 0.7580 \ (= 0.242 \ (\text{answer given}))$	M1 A1	**2**
1		(v)	z = 1.645	B1 (±)	
			-1.645×20 + 56	M1 (-ive z)	
			23.1	A1	**3**
2		(i)	Mean = 185	B1	
			Variance = 210.727	B1	**2**
2		(ii)	$H_0 : \mu = 175$ & $H_1 : \mu > 175$	B1	
			Where μ represents the mean decrease for the underlying population.	B1	
			$t = \dfrac{185 - 175}{SD / \sqrt{12}} = 2.39 \ (3\text{s.f.})$	M1 A1 FT	
			11 degrees of freedom	B1	
			At 5% level, critical value of t is 1.796	B1	
			2.39 > 1.796 so the result is significant.	M1A1	
			Evidence suggests the mean decrease in cholesterol level is more than 175.	A1	**9**
2		(iii)	The decrease in cholesterol level in the underlying population follows a Normal distribution. The sample is assumed to be random.	E1(Normal) E1(Random)	**2**

3		(i)	H_0: there is no association between gum disease and coronary artery disease H_1: there is an association between gum disease and coronary artery disease	B1	
			Expected frequencies	M1 A1	

	With g.d.	Without g.d.
With c.a.d	48.62	36.38
Without c.a.d.	94.38	70.62

Contributions to X^2 (without Yates' correction)

	With g.d.	Without g.d.
With c.a.d	1.52827	2.04245
Without c.a.d.	0.78729	1.05217

M1
M1 (summation)

Contributions to X^2 (with Yates' correction)

	With g.d.	Without g.d.
With c.a.d	1.35612	1.81238
Without c.a.d.	0.69861	0.93365

$X^2 = 5.4102$ (or 4.8008 with Yates' correction) — A1 CAO
1 degree of freedom — B1
Critical value at 5% level is 3.841 — B1
As 5.4102 (or 4.8008) > 3.841 the result is significant — M1
There is evidence of an association between gum disease and coronary artery disease — A1(in context)

10

3		(ii)	2 degrees of freedom	B1	
			Critical value at 5% level is 5.991	B1	
			8.2808 > 5.991 the result is significant	M1	
			There is evidence of an association between age and coronary artery disease	A1(in context)	**4**
3		(iii)	Suitable comments (in context)	E1 E1	**2**

4	(i)	H_0: population median = 26	B1	
		H_1: population median < 26	B1	
		Actual differences		
		-3 17 -10 -11 -4 7 1 -15 -8 -2 -16 -6	B1	
		Associated ranks		
		3 12 8 9 4 6 1 10 7 2 11 5	M1 A1	
		$T = 3 + 8 + 9 + 4 + 10 + 7 + 2 + 11 + 5 = 59$	B1	
		$T^+ = 12 + 6 + 1 = 19$		
			B1	
		$\therefore T = 19$	B1	
		From tables – at the 5% level of significance in a		
		one-tailed Wilcoxon signed rank test, the critical		
			B1	
		value of T is 17		
		$19 > 17 \therefore$ the result is not significant	M1 A1	
		The evidence does not suggest a decrease in the numbers of ants this year.	E1	
				12
4	(ii)	Variable - symmetry	E1	
		Sample - random	E1	**2**

5		(i)	If hamsters choose their bedding randomly then the three options will be equally likely to occur. Hence the probability that a hamster chooses the new material will be ⅓.	E1	1
5		(ii)	0.0390, 0.1561, 0.2731, 0.2732, 0.1707, 0.0682, 0.0197 5.85, 23.415, 40.965, 40.98, 25.605, 10.23, 2.955	M1 A2 M1 A1	5
5		(iii)	H_0: B(8, ⅓) is a good model H_1: B(8, ⅓) is not a good model Merge final two cells. $X^2 = \dfrac{(10-5.85)^2}{5.85} + \dfrac{(31-23.415)^2}{23.415} + \dfrac{(42-40.965)^2}{40.965} +$ $\dfrac{(34-40.98)^2}{40.98} + \dfrac{(19-25.605)^2}{25.605} + \dfrac{(14-13.185)^2}{13.185}$ = 8.370 5 degrees of freedom (6-1) Critical value at 5% level is 11.07 (FT their dof) 8.370 < 11.07 the result is not significant. No evidence to suggest the binomial distribution is not a good model.	B1 M1 M1 [$(o$-$e)^2/e$] M1 [sum] A1 B1 B1 M1 A1	9