RECOGNISING ACHIEVEMENT

ADVANCED GCE **4768/01**

MATHEMATICS (MEI)

Statistics 3

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, 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.

1 **(a)** Sarah travels home from work each evening by bus; there is a bus every 20 minutes. The time at which Sarah arrives at the bus stop varies randomly in such a way that the probability density function of X, the length of time in minutes she has to wait for the next bus, is given by

$$f(x) = k(20 - x) \text{ for } 0 \leqslant x \leqslant 20, \text{ where } k \text{ is a constant.}$$

(i) Find k. Sketch the graph of $f(x)$ and use its shape to explain what can be deduced about how long Sarah has to wait. [5]

(ii) Find the cumulative distribution function of X and hence, or otherwise, find the probability that Sarah has to wait more than 10 minutes for the bus. [4]

(iii) Find the median length of time that Sarah has to wait. [3]

(b) **(i)** Define the term 'simple random sample'. [2]

(ii) Explain briefly how to carry out cluster sampling. [3]

(iii) A researcher wishes to investigate the attitudes of secondary school pupils to pollution. Explain why he might prefer to collect his data using a cluster sample rather than a simple random sample. [2]

2 An electronics company purchases two types of resistor from a manufacturer. The resistances of the resistors (in ohms) are known to be Normally distributed. Type A have a mean of 100 ohms and standard deviation of 1.9 ohms. Type B have a mean of 50 ohms and standard deviation of 1.3 ohms.

(i) Find the probability that the resistance of a randomly chosen resistor of type A is less than 103 ohms. [3]

(ii) Three resistors of type A are chosen at random. Find the probability that their total resistance is more than 306 ohms. [3]

(iii) One resistor of type A and one resistor of type B are chosen at random. Find the probability that their total resistance is more than 147 ohms. [3]

(iv) Find the probability that the total resistance of two randomly chosen type B resistors is within 3 ohms of one randomly chosen type A resistor. [5]

(v) The manufacturer now offers type C resistors which are specified as having a mean resistance of 300 ohms. The resistances of a random sample of 100 resistors from the first batch supplied have sample mean 302.3 ohms and sample standard deviation 3.7 ohms. Find a 95% confidence interval for the true mean resistance of the resistors in the batch. Hence explain whether the batch appears to be as specified. [4]

3 **(a)** A tea grower is testing two types of plant for the weight of tea they produce. A trial is set up in which each type of plant is grown at each of 8 sites. The total weight, in grams, of tea leaves harvested from each plant is measured and shown below.

Site	A	B	C	D	E	F	G	H
Type I	225.2	268.9	303.6	244.1	230.6	202.7	242.1	247.5
Type II	215.2	242.1	260.9	241.7	245.5	204.7	225.8	236.0

(i) The grower intends to perform a *t* test to examine whether there is any difference in the mean yield of the two types of plant. State the hypotheses he should use and also any necessary assumption. [3]

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

(b) The tea grower deals with many types of tea and employs tasters to rate them. The tasters do this by giving each tea a score out of 100. The tea grower wishes to compare the scores given by two of the tasters. Their scores for a random selection of 10 teas are as follows.

Tea	Q	R	S	T	U	V	W	X	Y	Z
Taster 1	69	79	85	63	81	65	85	86	89	77
Taster 2	74	75	99	66	75	64	96	94	96	86

Use a Wilcoxon test to examine, at the 5% level of significance, whether it appears that, on the whole, the scores given to teas by these two tasters differ. [8]

4 **(a)** A researcher is investigating the feeding habits of bees. She sets up a feeding station some distance from a beehive and, over a long period of time, records the numbers of bees arriving each minute. For a random sample of 100 one-minute intervals she obtains the following results.

Number of bees	0	1	2	3	4	5	6	7	$\geqslant 8$
Number of intervals	6	16	19	18	17	14	6	4	0

(i) Show that the sample mean is 3.1 and find the sample variance. Do these values support the possibility of a Poisson model for the number of bees arriving each minute? Explain your answer. [3]

(ii) Use the mean in part **(i)** to carry out a test of the goodness of fit of a Poisson model to the data. [10]

(b) The researcher notes the length of time, in minutes, that each bee spends at the feeding station. The times spent are assumed to be Normally distributed. For a random sample of 10 bees, the mean is found to be 1.465 minutes and the standard deviation is 0.3288 minutes. Find a 95% confidence interval for the overall mean time. [4]

4768 Statistics 3

Q1	f(x) = k(20 – x) 0 ≤ x ≤ 20			
(a) (i)	$\int_0^{20} k(20-x)\mathrm{d}x = \left[k\left(20x - \frac{x^2}{2}\right)\right]_0^{20} = k \times 200 = 1$ $\therefore k = \frac{1}{200}$ Straight line graph with negative gradient, in the first quadrant. Intercept correctly labelled (20, 0), with nothing extending beyond these points. Sarah is more likely to have only a short time to wait for the bus.	M1 A1 G1 G1 E1	Integral of f(x), including limits (which may appear later), set equal to 1. Accept a geometrical approach using the area of a triangle. C.a.o.	5
(ii)	Cdf $\mathrm{F}(x) = \int_0^x \mathrm{f}(t)\mathrm{d}t$ $= \frac{1}{200}\left(20x - \frac{x^2}{2}\right)$ $= \frac{x}{10} - \frac{x^2}{400}$ P(X > 10) = 1 – F(10) = 1 – (1 – ¼) = ¼	M1 A1 M1 A1	Definition of cdf, including limits (or use of "+c" and attempt to evaluate it), possibly implied later. Some valid method must be seen. Or equivalent expression; condone absence of domain [0, 20]. Correct use of c's cdf. f.t. c's cdf. Accept geometrical method, e.g area = ½(20 – 10)f(10), or similarity.	4
(iii)	Median time, m, is given by F(m) = ½. $\therefore \frac{m}{10} - \frac{m^2}{400} = \frac{1}{2}$ $\therefore m^2 - 40m + 200 = 0$ $\therefore m = 5.86$	M1 M1 A1	Definition of median used, leading to the formation of a quadratic equation. Rearrange and attempt to solve the quadratic equation. Other solution is 34.14; no explicit reference to/rejection of it is required.	3

(b) (i)	A simple random sample is one where every sample of the required size has an equal chance of being chosen.	E2	S.C. Allow E1 for "Every member of the population has an equal chance of being chosen independently of every other member".	2
(ii)	Identify clusters which are capable of representing the population as a whole. Choose a random sample of clusters. Randomly sample or enumerate within the chosen clusters.	E1 E1 E1		3
(iii)	A random sample of the school population might involve having to interview single or small numbers of pupils from a large number of schools across the entire country. Therefore it would be more practical to use a cluster sample.	E1 E1	For "practical" accept e.g. convenient / efficient / economical.	2
				19

Q2	$A \sim \mathrm{N}(100,\ \sigma = 1.9)$ $B \sim \mathrm{N}(50,\ \sigma = 1.3)$		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(A < 103) = $\mathrm{P}\left(Z < \frac{103-100}{1.9} = 1.5789\right)$ = 0·9429	M1 A1 A1	For standardising. Award once, here or elsewhere. c.a.o.	 3
(ii)	$A_1 + A_2 + A_3 \sim \mathrm{N}(300,$ $\sigma^2 = 1.9^2 + 1.9^2 + 1.9^2 = 10.83\ )$ P(this > 306) = $\mathrm{P}\left(Z > \frac{306-300}{3\cdot291} = 1\cdot823\right) = 1 - 0\cdot9658 = 0.0342$	B1 B1 A1	Mean. Variance. Accept sd (= 3.291). c.a.o.	 3
(iii)	$A + B \sim \mathrm{N}(150,$ $\sigma^2 = 1.9^2 + 1.3^2 = 5.3\ )$ P(this > 147) = $\mathrm{P}\left(Z > \frac{147-150}{2\cdot302} = -1.303\right)$ = 0·9037	B1 B1 A1	Mean. Variance. Accept sd (= 2.302). c.a.o.	 3
(iv)	$B_1 + B_2 - A \sim N(0,$ $1\cdot3^2 + 1\cdot3^2 + 1\cdot9^2 = 6\cdot99)$ P(−3< this < 3) $= \mathrm{P}\left(\frac{-3-0}{2.644} < Z < \frac{3-0}{2.644}\right) = \mathrm{P}(-1\cdot135 < Z < 1\cdot135)$ = 2 × 0.8718 – 1 = 0.7436	B1 B1 M1 A1 A1	Mean. Or A – (B_1 + B_2). Variance. Accept sd (= 2.644). Formulation of requirement ... … two sided. c.a.o.	 5
(v)	Given $\bar{x} = 302.3 \quad s_{n-1} = 3.7$ CI is given by $302.3 \pm 1\cdot96 \times \frac{3.7}{\sqrt{100}}$ = 302·3 ± 0·7252 = (301·57(48), 303·02(52)) The batch appears not to be as specified since 300 is outside the confidence interval.	M1 B1 A1 E1	Correct use of 302.3 and $3.7/\sqrt{100}$. For 1·96 c.a.o. Must be expressed as an interval.	 4
				18

Q3				
(a) (i)	H_0: $\mu_D = 0$ (or $\mu_I = \mu_{II}$) H_1: $\mu_D \neq 0$ (or $\mu_{II} \neq \mu_I$) where μ_D is "mean for II – mean for I"	B1 B1	Both. Hypotheses in words only must include "population". For adequate verbal definition. Allow absence of "population" if correct notation μ is used, but do NOT allow "$\overline{X}_I = \overline{X}_{II}$" or similar unless $\overline{X}$ is clearly and explicitly stated to be a population mean.	
	Normality of differences is required.	B1		3
(ii)	MUST be PAIRED COMPARISON *t* test. Differences are: 10.0, 26.8, 42.7, 2.4, –14.9, –2.0, 16.3, 11.5			
	$\bar{d} = 11.6$ $\quad s_{n-1} = 17.707$	B1	s_n = 16.563 but do NOT allow this here or in construction of test statistic, but FT from there.	
	Test statistic is $\dfrac{11.6 - 0}{\frac{17.707}{\sqrt{8}}}$	M1	Allow c's $\bar{d}$ and/or s_{n-1}. Allow alternative: 0 + (c's 2.365) $\times \frac{17.707}{\sqrt{8}}$ (= 14.806) for subsequent comparison with $\bar{d}$. (Or $\bar{d}$ – (c's 2.365) $\times \frac{17.707}{\sqrt{8}}$ (=–3.206) for comparison with 0.)	
	= 1.852(92).	A1	c.a.o. but ft from here in any case if wrong. Use of $0 - \bar{d}$ scores M1A0, but ft.	
	Refer to t_7.	M1	No ft from here if wrong.	
	Double-tailed 5% point is 2.365.	A1	No ft from here if wrong.	
	Not significant.	A1	ft only c's test statistic.	
	Seems there is no difference between the mean yields of the two types of plant.	A1	ft only c's test statistic. Special case: (t_8 and 2.306) can score 1 of these last 2 marks if either form of conclusion is given.	7

(b)				

Diff	−5	4	−14	−3	6	1	−11	−8	−7	−9
Rank of |diff|	4	3	10	2	5	1	9	7	6	8

		M1 M1 A1	For differences. ZERO in this section if differences not used. For ranks. FT from here if ranks wrong	
	$W_+ = 1 + 3 + 5 = 9$ (or $W_- = 2+4+6+7+8+9+10 = 46$)	B1		
	Refer to tables of Wilcoxon paired (/single sample) statistic for $n = 10$.	M1	No ft from here if wrong.	
	Lower (or upper if 46 used) double-tailed 5% point is 8 (or 47 if 46 used).	A1	i.e. a 2-tail test. No ft from here if wrong.	
	Result is not significant.	A1	ft only c's test statistic.	
	No evidence to suggest the tasters differ on the whole.	A1	ft only c's test statistic.	8
				18

Q4				
(a) (i)	$\bar{x} = \frac{310}{100} = 3.1$ $s^2 = \frac{1288 - 100 \times 3.1^2}{99} = \frac{327}{99} = 3.303$ Evidence could support Poisson since the variance is fairly close to the mean.	B1 B1 E1		3
(ii)	see table below			

f_o	6	16	19	18	17	14	6	4	0
f_e	4.50	13.97	21.65	22.37	17.33	10.75	5.55	2.46	1.42
Merged	22 18.47						10 9.43		

	M1 A1 A1	Calculation of expected frequencies. Last cell correct. All others correct, but ft if wrong.	
	M1	Combining cells. (Condone if not combined as fully as shown above, but require top two cells combined as a minimum.)	
X^2 = 0.6747 + 0.3244 + 0.8537 + 0.0063 + 0.9826 + 0.0345	M1	Calculation of X^2.	
= 2.876(2)	A1	(Condone wrong last cell.) Depends on both of the preceding M marks.	
Refer to χ_4^2. e.g. Upper 10% point is 7.779.	M1	Allow correct df (= cells – 2) from wrongly grouped or ungrouped table, and FT. Otherwise, no FT if wrong.	
Not significant. Suggests Poisson model does fit … … at any reasonable level of significance.	A1 A1 A1	ft only c's test statistic. ft only c's test statistic. Or other sensible comment.	10

(b)	CI is given by 1.465 ±	M1	If <u>both</u> 1.465 and $0.3288/\sqrt{10}$ are correct.	
	2·262	B1 B1	**If t_9 used.** 95% 2-tail point for c's *t* distribution (Independent of previous mark).	
	$\times \frac{0.3288}{\sqrt{10}}$			
	= 1.465 ± 0.2352= (1.2298, 1.7002)	A1	c.a.o. Must be expressed as an interval.	4
				17