OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE

MATHEMATICS (MEI) 4768

Statistics 3

Candidates answer on the Answer Booklet

OCR Supplied Materials:
- 8 page Answer Booklet
- Graph paper
- MEI Examination Formulae and Tables (MF2)

Other Materials Required:
None

Wednesday 20 January 2010
Afternoon

Duration: 1 hour 30 minutes

* 4 7 6 8 *

INSTRUCTIONS TO CANDIDATES

- Write your name clearly in capital letters, your Centre Number and Candidate Number in the spaces provided on the Answer Booklet.
- Use black ink. Pencil may be used for graphs and diagrams only.
- Read each question carefully and make sure that you know what you have to do before starting your answer.
- Answer **all** the questions.
- Do **not** write in the bar codes.
- 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 document consists of **4** pages. Any blank pages are indicated.

1 Coastal wildlife wardens are monitoring populations of herring gulls. Herring gulls usually lay 3 eggs per nest and the wardens wish to model the number of eggs per nest that hatch. They assume that the situation can be modelled by the binomial distribution $B(3, p)$ where p is the probability that an egg hatches. A random sample of 80 nests each containing 3 eggs has been observed with the following results.

Number of eggs hatched	0	1	2	3
Number of nests	7	23	29	21

(i) Initially it is assumed that the value of p is $\frac{1}{2}$. Test at the 5% level of significance whether it is reasonable to suppose that the model applies with $p = \frac{1}{2}$. **[10]**

(ii) The model is refined by estimating p from the data. Find the mean of the observed data and hence an estimate of p. **[2]**

(iii) Using the estimated value of p, the value of the test statistic X^2 turns out to be 2.3857. Is it reasonable to suppose, at the 5% level of significance, that this refined model applies? **[3]**

(iv) Discuss the reasons for the different outcomes of the tests in parts **(i)** and **(iii)**. **[2]**

2 **(a)** A continuous random variable, X, has probability density function

$$f(x) = \begin{cases} \frac{1}{72}(8x - x^2) & 2 \leqslant x \leqslant 8, \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find $F(x)$, the cumulative distribution function of X. **[3]**

(ii) Sketch $F(x)$. **[3]**

(iii) The median of X is m. Show that m satisfies the equation $m^3 - 12m^2 + 148 = 0$. Verify that $m \approx 4.42$. **[3]**

(b) The random variable in part **(a)** is thought to model the weights, in kilograms, of lambs at birth. The birth weights, in kilograms, of a random sample of 12 lambs, given in ascending order, are as follows.

3.16 3.62 3.80 3.90 4.02 4.72 5.14 6.36 6.50 6.58 6.68 6.78

Test at the 5% level of significance whether a median of 4.42 is consistent with these data. **[10]**

3 Cholesterol is a lipid (fat) which is manufactured by the liver from the fatty foods that we eat. It plays a vital part in allowing the body to function normally. However, when high levels of cholesterol are present in the blood there is a risk of arterial disease. Among the factors believed to assist with achieving and maintaining low cholesterol levels are weight loss and exercise.

A doctor wishes to test the effectiveness of exercise in lowering cholesterol levels. For a random sample of 12 of her patients, she measures their cholesterol levels before and after they have followed a programme of exercise. The measurements obtained are as follows.

Patient	A	B	C	D	E	F	G	H	I	J	K	L
Before	5.7	5.7	4.0	6.8	7.4	5.5	6.7	6.4	7.2	7.2	7.1	4.4
After	5.8	4.0	5.2	5.7	6.0	5.0	5.8	4.2	7.3	5.2	6.4	4.1

(i) A t test is to be used in order to see if, on average, the exercise programme seems to be effective in lowering cholesterol levels. State the distributional assumption necessary for the test, and carry out the test using a 1% significance level. **[11]**

(ii) A second random sample of 12 patients gives a 95% confidence interval of (−0.5380, 1.4046) for the true mean reduction (before – after) in cholesterol level. Find the mean and standard deviation for this sample. How might the doctor interpret this interval in relation to the exercise programme? **[7]**

4 The weights of a particular variety (A) of tomato are known to be Normally distributed with mean 80 grams and standard deviation 11 grams.

(i) Find the probability that a randomly chosen tomato of variety A weighs less than 90 grams. **[3]**

The weights of another variety (B) of tomato are known to be Normally distributed with mean 70 grams. These tomatoes are packed in sixes using packaging that weighs 15 grams.

(ii) The probability that a randomly chosen pack of 6 tomatoes of variety B, including packaging, weighs less than 450 grams is 0.8463. Show that the standard deviation of the weight of single tomatoes of variety B is 6 grams, to the nearest gram. **[5]**

(iii) Tomatoes of variety A are packed in fives using packaging that weighs 25 grams. Find the probability that the total weight of a randomly chosen pack of variety A is greater than the total weight of a randomly chosen pack of variety B. **[5]**

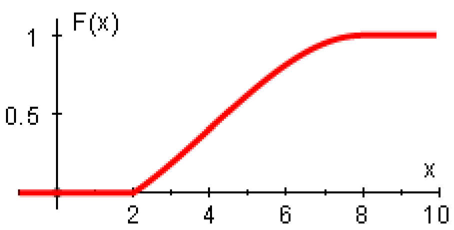
(iv) A new variety (C) of tomato is introduced. The weights, c grams, of a random sample of 60 of these tomatoes are measured giving the following results.

$$\Sigma c = 3126.0 \qquad \Sigma c^2 = 164\,223.96$$

Find a 95% confidence interval for the true mean weight of these tomatoes. **[5]**

4768 Statistics 3

1 (i)	H_0: The number of eggs hatched can be modelled by B(3, ½) H_1: The number of eggs hatched cannot be modelled by B(3, ½)	B1 B1		
	With p = ½ Probability: 0.125, 0.375, 0.375, 0.125 Exp'd frequency: 10, 30, 30, 10 Obs'd frequency: 7, 23, 29, 21	M1 A1	Probs × 80 for expected frequencies. All correct.	
	$X^2 = 0.9 + 1.6333 + 0.0333 + 12.1$ $= 14.666(7)$	M1 A1	Calculation of X^2. c.a.o.	
	Refer to χ_3^2.	M1	Allow correct df (= cells – 1) from wrongly grouped table and ft. Otherwise, no ft if wrong. $P(X^2 > 14.667) = 0.00212$.	
	Upper 5% point is 7.815. Significant. Suggests it is reasonable to suppose model with p = ½ does not apply.	A1 A1 A1	No ft from here if wrong. ft only c's test statistic. ft only c's test statistic.	**[10]**
(ii)	$\bar{x} = \frac{144}{80} = 1.8$ $\therefore \hat{p} = \frac{1.8}{3} = 0.6$	B1 B1	C.a.o. Use of E(X) = np. ft c's mean, provided $0 < \hat{p} < 1$.	**[2]**
(iii)	Refer to χ_2^2. Upper 5% point is 5.991. Suggests it is reasonable to suppose model with estimated p does apply.	M1 A1 A1	Allow df 1 less than in part (i). No ft if wrong. No ft if wrong. ft provided previous A mark awarded.	**[3]**
(iv)	For example: Estimating p leads to an improved fit … … at the expense of the loss of 1 degree of freedom. The model in (i) fails due to a large underestimate for $X = 3$.	E2	Reward any two sensible points for E1 each.	**[2]**
			Total	**[17]**

2 (a)	$\text{f}(x) = \frac{1}{72}\left(8x - x^2\right)$, $2 \le x \le 8$			
(i)	$\text{F}(x) = \int_2^x \frac{1}{72}\left(8t - t^2\right)\text{d}t$	M1	Correct integral with limits (which may be implied subsequently).	
	$= \frac{1}{72}\left[4t^2 - \frac{t^3}{3}\right]_2^x$	A1	Correctly integrated	
	$= \frac{1}{72}\left(4x^2 - \frac{x^3}{3} - 16 + \frac{8}{3}\right) = \frac{12x^2 - x^3 - 40}{216}$	A1	Limits used. Accept unsimplified form.	**[3]**
(ii)		G1	Correct shape; nothing below $y = 0$; non-negative gradient.	
		G1	Labels at (2, 0) and (8, 1).	
		G1	Curve (horizontal lines) shown for $x < 2$ and $x > 8$.	**[3]**
(iii)	$\text{F}(m) = ½ \quad \therefore \frac{12m^2 - m^3 - 40}{216} = \frac{1}{2}$	M1	Use of definition of median. Allow use of c's F(x).	
	$\therefore 12m^2 - m^3 - 40 = 108$ $\therefore m^3 - 12m^2 + 148 = 0$	A1	Convincingly rearranged. Beware: answer given.	
	Either F(4.42) = 0.5003(977) ≈ 0.5 Or $4.42^3 - 12 \times 4.42^2 + 148 = -0.0859(12) \approx 0$ $\therefore m \approx 4.42$	E1	Convincingly shown, e.g. 4.418 or better seen.	**[3]**

<table>
<tr><td>2 (b)</td><td>H_0: $m = 4.42$ H_1: $m \neq 4.42$
where m is the population median</td><td>B1
B1</td><td>Both. Accept hypotheses in words.
Adequate definition of m to include "population".</td><td></td></tr>
<tr><td></td><td>

Weights	− 4.42	Rank of |diff|
3.16	−1.26	7
3.62	−0.80	6
3.80	−0.62	4
3.90	−0.52	3
4.02	−0.40	2
4.72	0.30	1
5.14	0.72	5
6.36	1.94	8
6.50	2.08	9
6.58	2.16	10
6.68	2.26	11
6.78	2.36	12

</td><td>M1
M1
A1</td><td>for subtracting 4.42.
for ranks.
ft if ranks wrong.</td><td></td></tr>
<tr><td></td><td>$W_- = 2 + 3 + 4 + 6 + 7 = 22$</td><td>B1</td><td>($W_+ = 1 + 5 + 8 + 9 + 10 + 11 + 12 = 56$)</td><td></td></tr>
<tr><td></td><td>Refer to Wilcoxon single sample tables for $n = 12$.</td><td>M1</td><td>No ft from here if wrong.</td><td></td></tr>
<tr><td></td><td>Lower 2½% point is 13 (or upper is 65 if 56 used).</td><td>A1</td><td>i.e. a 2-tail test. No ft from here if wrong.</td><td></td></tr>
<tr><td></td><td>Result is not significant.</td><td>A1</td><td>ft only c's test statistic.</td><td></td></tr>
<tr><td></td><td>Evidence suggests that a median of 4.42 is consistent with these data.</td><td>A1</td><td>ft only c's test statistic.</td><td>[10]</td></tr>
<tr><td></td><td></td><td></td><td>Total</td><td>[19]</td></tr>
</table>

3 (i)	Must assume • Normality of population … • … of differences. H_0: $\mu_D = 0$ H_1: $\mu_D > 0$	 B1 B1 B1	 Both. Accept alternatives e.g. $\mu_D < 0$ for H_1, or $\mu_B - \mu_A$ etc provided adequately defined. Hypotheses in words only must include "population". Do NOT allow "$\bar{X} = \ldots$" or similar unless $\bar{X}$ is clearly and explicitly stated to be a population mean.	
	Where μ_D is the (population) mean reduction/difference in cholesterol level.	B1	For adequate verbal definition. Allow absence of "population" if correct notation μ is used.	
	MUST be PAIRED COMPARISON t test. Differences (reductions) (before – after) are: –0.1 1.7 –1.2 1.1 1.4 0.5 0.9 2.2 –0.1 2.0 0.7 0.3		Allow "after – before" if consistent with alternatives above.	
	$\bar{x} = 0.7833$ $s_{n-1} = 0.9833(46)$ $(s_{n-1}^{\ 2} = 0.966969)$	B1	Do not allow $s_n = 0.9415$ ($s_n^{\ 2} = 0.8864$)	
	Test statistic is $\dfrac{0.7833 - 0}{\frac{0.9833}{\sqrt{12}}}$	M1	Allow c's $\bar{x}$ and/or s_{n-1}. Allow alternative: 0 + (c's 2.718) $\times \dfrac{0.9833}{\sqrt{12}}$ (= 0.7715) for subsequent comparison with $\bar{x}$. (Or $\bar{x}$ – (c's 2.718) $\times \dfrac{0.9833}{\sqrt{12}}$ (= 0.0118) for comparison with 0.)	
	= 2.7595.	A1	c.a.o. but ft from here in any case if wrong. Use of $0 - \bar{x}$ scores M1A0, but ft.	
	Refer to t_{11}.	M1	No ft from here if wrong. P(t > 2.7595) = 0.009286.	
	Single-tailed 1% point is 2.718. Significant. Seems mean cholesterol level has fallen.	A1 A1 A1	No ft from here if wrong. ft only c's test statistic. ft only c's test statistic.	**[11]**
(ii)	CI is $\bar{x}$ ± 2.201 $\times \dfrac{s}{\sqrt{12}}$ = (−0.5380, 1.4046)	M1 B1 A1	Overall structure, seen or implied. From t_{11}, seen or implied. Fully correct pair of equations using the given interval, seen or implied.	
	$\bar{x}$ = ½(1.4046 − 0.5380) = 0.4333 $s = (1.4046 - 0.4333) \times \dfrac{\sqrt{12}}{2.201}$ = 1.5287	B1 M1 A1	 Substitute $\bar{x}$ and rearrange to find s. c.a.o.	
	Using this interval the doctor might conclude that the mean cholesterol level did not seem to have been reduced.	E1	Accept any sensible comment or interpretation of this interval.	**[7]**
			Total	**[18]**

4	$A \sim N(80,\ \sigma = 11)$ $B \sim N(70,\ \sigma = v)$		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)	$P(A < 90) = P\left(Z < \frac{90-80}{11} = 0.9091\right)$ $= 0.8182$	M1 A1 A1	For standardising. Award once, here or elsewhere. c.a.o.	**[3]**
(ii)	$W_B = B_1 + B_2 + \ldots + B_6 + 15 \sim N(435,$ $\sigma^2 = v^2 + v^2 + \ldots + v^2 = 6v^2)$ $P(\text{this} < 450) = P\left(Z < \frac{450-435}{v\sqrt{6}}\right) = 0.8463$ $\therefore \frac{450-435}{v\sqrt{6}} = \Phi^{-1}(0.8463) = 1.021$ $\therefore v = \frac{15}{1.021 \times \sqrt{6}} = 5.9977 = 6$ grams (nearest gram)	B1 B1 M1 B1 A1	Mean. Expression for variance. Formulation of the problem. Inverse Normal. Convincingly shown, beware A.G.	**[5]**
(iii)	$W_A = A_1 + A_2 + \ldots + A_5 + 25 \sim N(425,$ $\sigma^2 = 11^2 + 11^2 + \ldots + 11^2 = 605\ )$ $D = W_A - W_B \sim N(-10,$ $605 + 216 = 821)$ Want $P(W_A > W_B) = P(W_A - W_B > 0)$ $= P\left(Z > \frac{0-(-10)}{\sqrt{821}} = 0.3490\right) = 1 - 0.6365 = 0.3635$	B1 M1 A1 M1 A1	Mean. Accept "$B - A$". Variance. Accept sd (= 28.65). c.a.o.	**[5]**
(iv)	$\bar{x} = \frac{3126.0}{60} = 52.1,$ $s = \sqrt{\frac{164223.96 - 60 \times 52.1^2}{59}} = 4.8$ CI is given by $52.1 \pm$ 1.96 $\times \frac{4.8}{\sqrt{60}}$ $= 52.1 \pm 1.2146 = (50.885(4), 53.314(6))$	B1 M1 B1 M1 A1	Both correct. c.a.o. Must be expressed as an interval.	**[5]**
			Total	**[18]**