OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE UNIT **4768/01**

MATHEMATICS (MEI)

Statistics 3

FRIDAY 12 JANUARY 2007 Morning

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.
- 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.

ADVICE TO CANDIDATES

- Read each question carefully and make sure you know what you have to do before starting your answer.
- 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 The continuous random variable X has probability density function

$$f(x) = k(1 - x) \quad \text{for } 0 \leqslant x \leqslant 1$$

where k is a constant.

(i) Show that $k = 2$. Sketch the graph of the probability density function. [4]

(ii) Find $E(X)$ and show that $\text{Var}(X) = \frac{1}{18}$. [5]

(iii) Derive the cumulative distribution function of X. Hence find the probability that X is greater than the mean. [4]

(iv) Verify that the median of X is $1 - \dfrac{1}{\sqrt{2}}$. [2]

(v) $\overline{X}$ is the mean of a random sample of 100 observations of X. Write down the approximate distribution of $\overline{X}$. [3]

2 The manager of a large country estate is preparing to plant an area of woodland. He orders a large number of saplings (young trees) from a nursery. He selects a random sample of 12 of the saplings and measures their heights, which are as follows (in metres).

0.63 0.62 0.58 0.56 0.59 0.62 0.64 0.58 0.55 0.61 0.56 0.52

(i) The manager requires that the mean height of saplings at planting is at least 0.6 metres. Carry out the usual t test to examine this, using a 5% significance level. State your hypotheses and conclusion carefully. What assumption is needed for the test to be valid? [11]

(ii) Find a 95% confidence interval for the true mean height of saplings. Explain carefully what is meant by a 95% confidence interval. [5]

(iii) Suppose the assumption needed in part **(i)** cannot be justified. Identify an alternative test that the manager could carry out in order to check that the saplings meet his requirements, and state the null hypothesis for this test. [2]

3 Bill and Ben run their own gardening company. At regular intervals throughout the summer they come to work on my garden, mowing the lawns, hoeing the flower beds and pruning the bushes. From past experience it is known that the times, in minutes, spent on these tasks can be modelled by independent Normally distributed random variables as follows.

	Mean	Standard deviation
Mowing	44	4.8
Hoeing	32	2.6
Pruning	21	3.7

(i) Find the probability that, on a randomly chosen visit, it takes less than 50 minutes to mow the lawns. [3]

(ii) Find the probability that, on a randomly chosen visit, the total time for hoeing and pruning is less than 50 minutes. [3]

(iii) If Bill mows the lawns while Ben does the hoeing and pruning, find the probability that, on a randomly chosen visit, Ben finishes first. [4]

Bill and Ben do my gardening twice a month and send me an invoice at the end of the month.

(iv) Write down the mean and variance of the **total** time (in minutes) they spend on mowing, hoeing and pruning per month. [2]

(v) The company charges for the **total** time spent at 15 pence per minute. There is also a fixed charge of £10 per month. Find the probability that the total charge for a month does not exceed £40. [6]

[Turn over

4 **(a)** An amateur weather forecaster has been keeping records of air pressure, measured in atmospheres. She takes the measurement at the same time every day using a barometer situated in her garden. A random sample of 100 of her observations is summarised in the table below. The corresponding expected frequencies for a Normal distribution, with its two parameters estimated by sample statistics, are also shown in the table.

Pressure (a atmospheres)	Observed frequency	Frequency as given by Normal model
$a \leqslant 0.98$	4	1.45
$0.98 < a \leqslant 0.99$	6	5.23
$0.99 < a \leqslant 1.00$	9	13.98
$1.00 < a \leqslant 1.01$	15	23.91
$1.01 < a \leqslant 1.02$	37	26.15
$1.02 < a \leqslant 1.03$	21	18.29
$1.03 < a$	8	10.99

Carry out a test at the 5% level of significance of the goodness of fit of the Normal model. State your conclusion carefully and comment on your findings. [9]

(b) The forecaster buys a new digital barometer that can be linked to her computer for easier recording of observations. She decides that she wishes to compare the readings of the new barometer with those of the old one. For a random sample of 10 days, the readings (in atmospheres) of the two barometers are shown below.

Day	A	B	C	D	E	F	G	H	I	J
Old	0.992	1.005	1.001	1.011	1.026	0.980	1.020	1.025	1.042	1.009
New	0.985	1.003	1.002	1.014	1.022	0.988	1.030	1.016	1.047	1.025

Use an appropriate Wilcoxon test to examine at the 10% level of significance whether there is any reason to suppose that, on the whole, readings on the old and new barometers do not agree. [9]

Mark Scheme 4768
January 2007

Q1	$f(x) = k(1-x) \qquad 0 \le x \le 1$			
(i)	$\int_0^1 k(1-x)\mathrm{d}x = 1$ $\therefore k[x - \frac{1}{2}x^2]_0^1 = 1$ $\therefore k(1-\frac{1}{2}) - 0 = 1$ $\therefore k = 2$ Labelled sketch: straight line segment from (0,2) to (1,0).	M1 E1 G1 G1	Integral of f(x), including limits (possibly implied later), equated to 1. Convincingly shown. Beware printed answer. Correct shape. Intercepts labelled.	4
(ii)	$\mathrm{E}(X) = \int_0^1 2x(1-x)\mathrm{d}x$ $= [x^2 - \frac{2}{3}x^3]_0^1 = (1-\frac{2}{3}) - 0 = \frac{1}{3}$ $\mathrm{E}(X^2) = \int_0^1 2x^2(1-x)\mathrm{d}x$ $= [\frac{2}{3}x^3 - \frac{2}{4}x^4]_0^1 = (\frac{2}{3} - \frac{1}{2}) - 0 = \frac{1}{6}$ $\mathrm{Var}(X) = \frac{1}{6} - (\frac{1}{3})^2$ $= \frac{1}{18}$	M1 A1 M1 M1 A1	Integral for E(X) including limits (which may appear later). Integral for E(X^2) including limits (which may appear later). Convincingly shown. Beware printed answer.	5
(iii)	$\mathrm{F}(x) = \int_0^x 2(1-t)\mathrm{d}t$ $= [2t - t^2]_0^x = (2x - x^2) - 0 = 2x - x^2$ $\mathrm{P}(X > \mu) = \mathrm{P}(X > \frac{1}{3}) = 1 - \mathrm{F}(\frac{1}{3})$ $= 1 - (2\times\frac{1}{3} - (\frac{1}{3})^2) = 1 - \frac{5}{9} = \frac{4}{9}$	M1 A1 M1 A1	Definition of cdf, including limits, possibly implied later. Some valid method must be seen. [for $0 \le x \le 1$; do not insist on this.] For 1 – c's F(μ). ft c's E(X) and F(x). If answer only seen in decimal expect 3 d.p. or better.	4
(iv)	$\mathrm{F}\left(1-\frac{1}{\sqrt{2}}\right) = 2\left(1-\frac{1}{\sqrt{2}}\right) - \left(1-\frac{1}{\sqrt{2}}\right)^2$ $= 2 - \frac{2}{\sqrt{2}} - 1 + \frac{2}{\sqrt{2}} - \frac{1}{2} = \frac{1}{2}$ **Alternatively:** $2m - m^2 = \frac{1}{2}$ $\therefore m^2 - 2m + \frac{1}{2} = 0$ $\therefore m = 1 \pm \frac{1}{\sqrt{2}}$ so $m = 1 - \frac{1}{\sqrt{2}}$	M1 E1 M1 E1	Substitute $m = 1 - \frac{1}{\sqrt{2}}$ in c's cdf. Convincingly shown. Beware printed answer. Form a quadratic equation $\mathrm{F}(m) = \frac{1}{2}$ and attempt to solve it. ft c's cdf provided it leads to a quadratic. Convincingly shown. Beware printed answer.	2
(v)	$\overline{X} \sim \mathrm{N}(\frac{1}{3}, \frac{1}{1800})$	B1 B1 B1	Normal distribution. Mean. ft c's E(X). Correct variance.	3
				18

Q2				
(i)	$H_0 : \mu = 0{\cdot}6$ $H_1 : \mu < 0{\cdot}6$ Where μ is the (population) mean height of the saplings.	B1 B1 B1	Allow absence of "population" if correct notation μ is used, but do NOT allow "$\bar{X}=...$" or similar unless $\bar{X}$ is clearly and explicitly stated to be a <u>population</u> mean. Hypotheses in words only must include "population".	
	$\bar{x} = 0{\cdot}5883$, $s_{n-1} = 0{\cdot}03664$ ($s_{n-1}^2 = 0{\cdot}00134$)	B1	Do not allow $s_n = 0{\cdot}03507$ ($s_n^2 = 0{\cdot}00123$).	
	Test statistic is $\dfrac{0{\cdot}5883 - 0{\cdot}6}{\left(\dfrac{0{\cdot}03664}{\sqrt{12}}\right)}$	M1	Allow c's $\bar{x}$ and/or s_{n-1}. Allow alternative: 0·6 ± (c's –1·796) $\times \frac{0{\cdot}03664}{\sqrt{12}}$ (=0·5810, 0·6190) for subsequent comparison with $\bar{x}$. (Or $\bar{x}$ ± (c's –1·796) $\times \frac{0{\cdot}03664}{\sqrt{12}}$ (=0·5693, 0·6073) for comparison with 0·6.)	
	$= -1{\cdot}103$	A1	c.a.o. but ft from here in any case if wrong. Use of 0·6 – $\bar{x}$ scores M1A0, but ft.	
	Refer to t_{11}.	M1	No ft from here if wrong.	
	Lower 5% point is –1·796.	A1	No ft from here if wrong. Must be –1·796 unless it is clear that absolute values are being used.	
	–1·103 > –1·796, ∴ Result is not significant.	E1	ft only c's test statistic.	
	Seems mean height of saplings meets the manager's requirements.	E1	ft only c's test statistic.	11
	Underlying population is Normal.	B1		
(ii)	CI is given by $0{\cdot}5883 \pm 2{\cdot}201 \times \dfrac{0{\cdot}03664}{\sqrt{12}}$	M1 B1 M1	ft c's $\bar{x} \pm$. ft c's s_{n-1}.	
	$= 0{\cdot}5883 \pm 0{\cdot}0233 = (0{\cdot}565(0), 0{\cdot}611(6))$	A1	c.a.o. Must be expressed as an interval. ZERO if not same distribution as test. Same wrong distribution scores maximum M1B0M1A0. Recovery to t_{11} is OK.	

	In repeated sampling, 95% of intervals constructed in this way will contain the true population mean.	E1		5
(iii)	Could use the Wilcoxon test. Null hypothesis is “Median = 0.6”.	E1 E1		2
				18

Q3	$M \sim N(44,\ 4{\cdot}8^2)$ $H \sim N(32,\ 2{\cdot}6^2)$ $P \sim N(21,\ 3{\cdot}7^2)$		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)	$P(M < 50) = P(Z < \frac{50-44}{4{\cdot}8} = 1{\cdot}25)$ $= 0{\cdot}8944$	M1 A1 A1	For standardising. Award once, here or elsewhere.	3
(ii)	$H + P \sim N(32 + 21 = 53,$ $2{\cdot}6^2 + 3{\cdot}7^2 = 20{\cdot}45)$ $P(H + P < 50) = P(Z < \frac{50-53}{\sqrt{20{\cdot}45}} = -0{\cdot}6634)$ $= 1 - 0{\cdot}7465 = 0{\cdot}2535$	B1 B1 A1	Mean. Variance. Accept sd = $\sqrt{20{\cdot}45}$ = 4·522... c.a.o.	3
(iii)	Want $P(M > H + P)$ i.e. $P(M - (H + P) > 0)$ $M - (H + P) \sim N(44 - (32 + 21) = -9,$ $4{\cdot}8^2 + 2{\cdot}6^2 + 3{\cdot}7^2 =$ 43·49) $P(\text{this} > 0) = P(Z > \frac{0-(-9)}{\sqrt{43{\cdot}49}} = 1{\cdot}365)$ $= 1 - 0{\cdot}9139 = 0{\cdot}0861$	M1 B1 B1 A1	Allow $H + P - M$ provided subsequent work is consistent. Mean. Variance. Accept sd = $\sqrt{43{\cdot}49}$ = 6·594... c.a.o.	4
(iv)	Mean = 44 + 44 + 32 + 32 + 21 + 21 = 194 Variance = $4{\cdot}8^2 + 4{\cdot}8^2 + 2{\cdot}6^2 + 2{\cdot}6^2 + 3{\cdot}7^2 +$ $3{\cdot}7^2$ = 86·98	 B1 B1	 (sd = 9·3263…)	2
(v)	$C \sim N(194 \times 0{\cdot}15 + 10 = 39{\cdot}10,$ $86{\cdot}98 \times 0{\cdot}15^2 = 1{\cdot}957)$ $P(C \le 40) = P(Z \le \frac{40-39{\cdot}10}{\sqrt{1{\cdot}957}} = 0{\cdot}6433)$ $= 0{\cdot}7400$ **Alternatively:** $P(C \le 40) = P(\text{total time} \le \frac{40-10}{0.15} = 200$ minutes) $= P(Z \le \frac{200-194}{\sqrt{86{\cdot}98}} = 0{\cdot}6433)$	M1 M1 A1 M1 A1 A1 M1 M1 A1 M1 A1	c's mean in (iv) × 0·15 + 10 (or subtract 10 from 40 below) ft c's mean in (iv). c's variance in (iv) × $0{\cdot}15^2$ ft c's variance in (iv). c.a.o. – 10 ÷ 0.15 c.a.o. Correct use of c's variance in (iv). ft c's mean and variance in (iv).	6

	= 0·7400	A1	c.a.o.	
				18

Q4				
(a)	<table><tr><td>Obs</td><td>Exp</td></tr><tr><td>10</td><td>6·68</td></tr></table>	M1	Combine first two rows.	
	$\therefore X^2 = \frac{(10-6\cdot68)^2}{6\cdot68} + \text{etc}$ = 1·6501 + 1·7740 + 3·3203 + 4·5018 + 0·4015 + 0·8135	M1		
	= 12·46(12)	A1		
	d.o.f. = 6 – 3 = 3		Require d.o.f. = No. cells used – 3.	
	Refer to χ^2_3.	M1	No ft from here if wrong.	
	Upper 5% point is 7·815	A1	No ft from here if wrong.	
	12·46 > 7·815 ∴ Result is significant.	E1	ft only c's test statistic.	
	Seems the Normal model does not fit the data at the 5% level.	E1	ft only c's test statistic.	
	E.g. • The biggest discrepancy is in the class 1·01 < *a* ≤ 1·02	E1		
	• The model overestimates in classes …, but underestimates in classes …	E1	Any two suitable comments.	9
(b)	Old – New: 0·007 0·002 –0·001 –0·003 0·004 –0·008 –0·010 0·009 –0·005 –0·016 Rank of |diff| 6 2 1 3 4 7 9 8 5 10			
		M1	For differences. ZERO in this section if differences not used.	
		M1	For ranks of |difference|.	
		A1	All correct. ft from here if ranks wrong.	
	W_+ = 6 + 2 + 4 + 8 = 20	B1	Or W_- = 1 + 3 + 7 + 9 + 5 + 10 = 35	
	Refer to Wilcoxon single sample (/paired) tables for *n* = 10.	M1	No ft from here if wrong.	
	Lower two-tail 10% point is … … 10.	M1 A1	Or, if 35 used, upper point is 45. No ft from here if wrong.	
	20 > 10 ∴ Result is not significant.	E1	Or 35 < 45. ft only c's test statistic.	
	Seems there is no reason to suppose the barometers differ.	E1	ft only c's test statistic.	9
				18

4768 - Statistics 3

General Comments

Once again the overall standard of the scripts seen was pleasing: many candidates appeared well prepared for the paper. However, as in the past, the quality of their comments, interpretations and explanations was consistently below that of the rest of the work.

It was noticeable that candidates' use of correct mathematical notation was often poor. For example: integrals written without the terminator "dx" and interchanging the symbols "=" and "$\Rightarrow$". Also many candidates showed a lack of appreciation of the level of detail of arithmetic required to convince the examiner that an answer printed in the question has been obtained genuinely.

Invariably all four questions were attempted, and attempted well on the whole. Questions 1 and 3 were found to be particularly high scoring. There was no evidence to suggest that candidates found themselves short of time at the end.

Comments on Individual Questions

1 **Continuous random variables; no context.**

(i) Although the accuracy of notation left much to be desired (as noted above) virtually all candidates were able to establish the value of k satisfactorily.
Also, most candidates sketched the graph of f(x) with little difficulty. The only note of disappointment was the number of candidates who neglected to draw a sketch at all.

(ii) The value of E(X) was almost always obtained with no difficulty. Similarly Var(X) was found correctly. Candidates need to be aware, however, that a little more effort is appropriate when establishing the exact value printed in the question.

(iii) As in the past, candidates did not acquit themselves at all well when attempting to find the cumulative distribution function (c.d.f.). They must be encouraged to realise that a definite integral (with suitable limits) is expected.
It was then interesting to see that an appreciable proportion of candidates seemed not to know how to use and/or interpret their c.d.f. Instead, in both this part and the next, they set up and evaluated integrals that were completely unnecessary. Having said that there were very many who did eventually find the correct probability of X greater than the mean.

(iv) Perhaps fewer than half of the candidates used their c.d.f. and substitution to verify that the given value was the median. The majority (including those who first integrated) obtained and solved a quadratic equation for m, and this left them needing to remember to distinguish between the two roots.

(v) Most candidates were able to write down the correct distribution here, based on the Central Limit Theorem.

2 **The *t* distribution: hypothesis test for the population mean; confidence interval for a population mean; heights of saplings.**

(i) The null hypothesis was usually correct, although some used "$\geq$" instead of "=", but there were fewer correct alternative hypotheses. Furthermore it remains the case that too many candidates neglect to define in words the symbol "μ".
At this level it is expected that candidates are going to use the built-in statistical functions of their calculators for the mean and sample variance. There were an appreciable number of scripts where this did not seem to be the case, and so the accuracy of their results suffered a little from premature approximation. Nonetheless the test statistic was usually worked out correctly. Similarly the test was carried out and concluded correctly, the most common problem being the use of the wrong critical value (-2.201 instead of -1.796). When the test is one-tailed, requiring the lower tail critical value and involving a negative test statistic, candidates are often less than clear and careful about the negative signs. Centres are advised that the "special case", shown in past mark schemes to allow for a particular form of misreading the tables, will not be applied from June 2007 onwards.

(ii) Most candidates showed that they were familiar with how to construct a confidence interval, and did so successfully. Unsurprisingly, there were a number who seemed to forget that they should still be using the *t* distribution. Clear and accurate descriptions of the meaning of a confidence interval were disappointingly rare.

(iii) There were many correct responses to this part. However it was also quite common to see answers that were only partially correct, for example by identifying the Wilcoxon single sample rank sum test but then suggesting a null hypothesis that was inconsistent with it.

3 **Combinations of Normal distributions; times of gardening tasks.**

This question was very well answered with very many scoring full marks. Candidates seemed well prepared for it and understood what was expected. In many cases their answers were concise and to the point. 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) This part, too, was almost always correct.

(iii) Again, correct answers were often seen here too, but this time weaker candidates experienced difficulty with the formulation of the requirement.

(iv) Usually the mean total time was correct, but often the variance was not. Typically the error came about through a lack of proper understanding of the difference between $\text{Var}(2X)$ $(= 2^2\text{Var}(X))$ and $\text{Var}(X_1 + X_2)$ $(=\text{Var}(X_1) + \text{Var}(X_2))$. Here the former was used when it should have been the latter.

(v) There were many good answers to this part, and they were evenly split between those who adapted their answers to part (iv) to obtain the probability distribution of the monthly charges and hence the probability of the charge not exceeding £40, and those who found the time (in minutes) corresponding to a total charge of £40 and then the probability of not exceeding that time.

4 **Chi-squared test of goodness of fit; Wilcoxon paired sample test for a difference in population medians; air pressures.**

(a) Although there were plenty of good attempts at this part of the question many broke down in one or more of the following ways. Some, but not very many, candidates neglected to merge the first two classes. Quite a few used the wrong number of degrees of freedom, usually because they forgot to allow for the two estimated parameters, and hence their critical value was inappropriate. Following the conclusion of the test, many simply neglected to comment on their findings. For this last point it is expected that candidates will undertake a brief discussion of what can be deduced by looking at the data in order to explain the outcome of the test.

(b) There were very many good answers to this part and most of these scored full or nearly full marks for it. It was a rare script indeed where the candidate did not know to take differences and then rank the absolute values. An occasional slip with the arithmetic was seen here. The vast majority of candidates found and used the correct test statistic and compared it with the correct critical value, which led to a correct conclusion.