OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE UNIT **4767/01**

MATHEMATICS (MEI)

Statistics 2

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

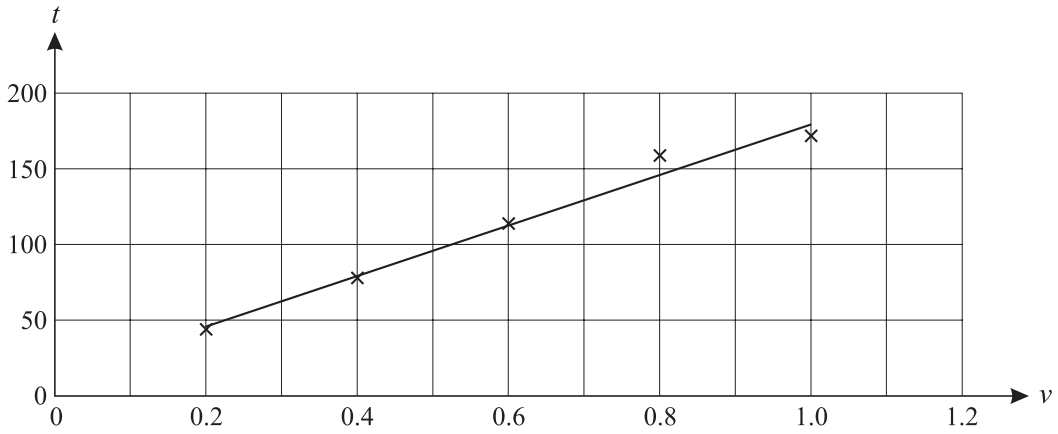
- Read each question carefully and make sure you know what you have to do before starting your answer.
- You are advised that an answer may receive **no marks** unless you show sufficient detail of the working to indicate that a correct method is being used.

This document consists of **6** printed pages and **2** blank pages.

1 In a science investigation into energy conservation in the home, a student is collecting data on the time taken for an electric kettle to boil as the volume of water in the kettle is varied. The student's data are shown in the table below, where v litres is the volume of water in the kettle and t seconds is the time taken for the kettle to boil (starting with the water at room temperature in each case). Also shown are summary statistics and a scatter diagram on which the regression line of t on v is drawn.

v	0.2	0.4	0.6	0.8	1.0
t	44	78	114	156	172

$n = 5,\ \Sigma v = 3.0,\ \Sigma t = 564,\ \Sigma v^2 = 2.20,\ \Sigma vt = 405.2.$



(i) Calculate the equation of the regression line of t on v, giving your answer in the form $t = a + bv$. [5]

(ii) Use this equation to predict the time taken for the kettle to boil when the amount of water which it contains is

(*A*) 0.5 litres,

(*B*) 1.5 litres.

Comment on the reliability of each of these predictions. [4]

(iii) In the equation of the regression line found in part **(i)**, explain the role of the coefficient of v in the relationship between time taken and volume of water. [2]

(iv) Calculate the values of the residuals for $v = 0.8$ and $v = 1.0$. [4]

(v) Explain how, on a scatter diagram with the regression line drawn accurately on it, a residual could be measured and its sign determined. [3]

2 **(a)** A farmer grows Brussels sprouts. The diameter of sprouts in a particular batch, measured in mm, is Normally distributed with mean 28 and variance 16. Sprouts that are between 24 mm and 33 mm in diameter are sold to a supermarket.

(i) Find the probability that the diameter of a randomly selected sprout will be within this range. [4]

(ii) The farmer sells the sprouts in this range to the supermarket for 10 pence per kilogram. The farmer sells sprouts under 24 mm in diameter to a frozen food factory for 5 pence per kilogram. Sprouts over 33 mm in diameter are thrown away. Estimate the total income received by the farmer for the batch, which weighs 25 000 kg. [3]

(iii) By harvesting sprouts earlier, the mean diameter for another batch can be reduced to k mm. Find the value of k for which only 5% of the sprouts will be above 33 mm in diameter. You may assume that the variance is still 16. [3]

(b) The farmer also grows onions. The weight in kilograms of the onions is Normally distributed with mean 0.155 and variance 0.005. He is trying out a new variety, which he hopes will yield a higher mean weight. In order to test this, he takes a random sample of 25 onions of the new variety and finds that their total weight is 4.77 kg. You should assume that the weight in kilograms of the new variety is Normally distributed with variance 0.005.

(i) Write down suitable null and alternative hypotheses for the test in terms of μ. State the meaning of μ in this case. [2]

(ii) Carry out the test at the 1% level. [6]

[Turn over

3 An electrical retailer gives customers extended guarantees on washing machines. Under this guarantee all repairs in the first 3 years are free. The retailer records the numbers of free repairs made to 80 machines.

Number of repairs	0	1	2	3	>3
Frequency	53	20	6	1	0

(i) Show that the sample mean is 0.4375. [1]

(ii) The sample standard deviation s is 0.6907. Explain why this supports a suggestion that a Poisson distribution may be a suitable model for the distribution of the number of free repairs required by a randomly chosen washing machine. [2]

The random variable X denotes the number of free repairs required by a randomly chosen washing machine. For the remainder of this question you should assume that X may be modelled by a Poisson distribution with mean 0.4375.

(iii) Find $P(X = 1)$. Comment on your answer in relation to the data in the table. [4]

(iv) The manager decides to monitor 8 washing machines sold on one day. Find the probability that there are at least 12 free repairs in total on these 8 machines. You may assume that the 8 machines form an independent random sample. [3]

(v) A launderette with 8 washing machines has needed 12 free repairs. Why does your answer to part **(iv)** suggest that the Poisson model with mean 0.4375 is unlikely to be a suitable model for free repairs on the machines in the launderette? Give a reason why the model may not be appropriate for the launderette. [3]

The retailer also sells tumble driers with the same guarantee. The number of free repairs on a tumble drier in three years can be modelled by a Poisson distribution with mean 0.15. A customer buys a tumble drier and a washing machine.

(vi) Assuming that free repairs are required independently, find the probability that

(*A*) the two appliances need a total of 3 free repairs between them,

(*B*) each appliance needs exactly one free repair. [5]

4 Two educational researchers are investigating the relationship between personal ambitions and home location of students. The researchers classify students into those whose main personal ambition is good academic results and those who have some other ambition. A random sample of 480 students is selected.

(i) One researcher summarises the data as follows.

Observed		Home location	
		City	Non-city
Ambition	Good results	102	147
	Other	75	156

Carry out a test at the 5% significance level to examine whether there is any association between home location and ambition. State carefully your null and alternative hypotheses. Your working should include a table showing the contributions of each cell to the test statistic. [9]

(ii) The other researcher summarises the same data in a different way as follows.

Observed		Home location		
		City	Town	Country
Ambition	Good results	102	83	64
	Other	75	64	92

(*A*) Calculate the expected frequencies for both 'Country' cells. [2]

(*B*) The test statistic for these data is 10.94. Carry out a test at the 5% level based on this table, using the same hypotheses as in part **(i)**. [3]

(*C*) The table below gives the contribution of each cell to the test statistic. Discuss briefly how personal ambitions are related to home location. [2]

Contribution to the test statistic		Home location		
		City	Town	Country
Ambition	Good results	1.129	0.596	3.540
	Other	1.217	0.643	3.816

(iii) Comment briefly on whether the analysis in part **(ii)** means that the conclusion in part **(i)** is invalid. [2]

BLANK PAGE

BLANK PAGE

Mark Scheme 4767
January 2007

Question 1

(i)	$\bar{t}$ = 112.8, $\bar{v}$ = 0.6 $b = \frac{Svt}{Svv} = \frac{405.2 - 3\times 564/5}{2.20 - 3^2/5} = \frac{66.8}{0.4}$ = 167 OR $b = \frac{405.2/5 - 0.6\times 112.8}{2.20/5 - 0.6^2} = \frac{13.36}{0.08}$ = 167 hence least squares regression line is: $t - \bar{t} = b(v - \bar{v})$ $\Rightarrow$ $t - 112.8 = 167(v - 0.6)$ $\Rightarrow$ $t = 167v + 12.6$	B1 for $\bar{t}$ and $\bar{v}$ used (SOI) M1 for attempt at gradient (b) A1 for 167 CAO M1 for equation of line A1 FT	**5**
(ii)	(A) For 0.5 litres, predicted time = = 167 × 0.5 + 12.6 = 96.1 seconds (B) For 1.5 litres, predicted time = = 167 ×1.5 + 12.6 = 263.1 seconds Any valid relevant comment relating to each prediction such as eg: 'First prediction is fairly reliable as it is interpolation and the data is a good fit' 'Second prediction is less certain as it is an extrapolation'	M1 for at least one prediction attempted A1 for both answers (FT their equation if b>0) NB for reading predictions off the graph only award A1 if accurate to nearest whole number E1 (first prediction) E1 (second prediction)	**4**
(iii)	The v-coefficient is the number of additional seconds required for each extra litre of water	E1 for indication of rate wrt v E1 *dep* for specifying ito units	**2**
(iv)	v = 0.8 $\Rightarrow$ predicted t = 167 × 0.8 + 12.6 = 146.2 Residual = 156 – 146.2 = 9.8 v = 1.0 $\Rightarrow$ predicted t = 167 × 1.0 + 12.6 = 179.6 Residual = 172 – 179.6 = –7.6	M1 for either prediction M1 for either subtraction A1 *CAO* for absolute value of both residuals B1 for both signs correct.	**4**
(v)	The residuals can be measured by finding the vertical distance between the plotted point and the regression line. The sign will be negative if the point is below the regression line (and positive if above).	E1 for distance E1 for vertical E1 for sign	**3**
			18

Question 2

(a) (i)	$X \sim N(28,16)$ $P(24 < X < 33) = P\left(\frac{24-28}{4} < Z < \frac{33-28}{4}\right)$ $= P(-1 < Z < 1.25)$ $= \Phi(1.25) - (1 - \Phi(1))$ $= 0.8944 - (1 - 0.8413)$ $= 0.8944 - 0.1587$ $= 0.7357$ (4 s.f.) *or* 0.736 (to 3 s.f.)	M1 for standardizing A1 for 1. 25 and -1 M1 for prob. with tables and correct structure A1 CAO (min 3 s.f., to include use of difference column)	**4**
(ii)	$25000 \times 0.7357 \times 0.1 = £1839$ $25000 \times 0.1587 \times 0.05 = £198$ Total = £1839 + £198 = £2037	M1 for either product, (with or without price) M1 for sum of both products with price A1 CAO awrt £2040	**3**
(iii)	$X \sim N(k, 16)$ From tables $\Phi^{-1}(0.95) = 1.645$ $\frac{33-k}{4} = 1.645$ $33 - k = 1.645 \times 4$ $k = 33 - 6.58$ $k = 26.42$ (4 s.f.) *or* 26.4 (to 3 s.f.)	B1 for ±1.645 seen M1 for correct equation in k with positive z-value A1 *CAO*	**3**
(b) (i)	H_0: $\mu = 0.155$; H_1: $\mu > 0.155$ Where μ denotes the mean weight in kilograms of the population of onions of the new variety	B1 for both correct & ito μ B1 for definition of μ	**2**
(ii)	Mean weight = 4.77/25 = 0.1908 Test statistic = $\frac{0.1908-0.155}{\sqrt{0.005}/\sqrt{25}} = \frac{0.0358}{0.01414}$ = 2.531 1% level 1-tailed critical value of z = 2.326 2.531 > 2.236 so significant. There is sufficient evidence to reject H_0 It is reasonable to conclude that the new variety has a higher mean weight.	B1 M1 must include $\sqrt{25}$ A1FT B1 for 2.326 M1 For sensible comparison leading to a conclusion A1 for correct, consistent conclusion in words and in context	**6**
			18

Question 3

(i)	Mean = $\frac{\Sigma xf}{n} = \frac{0+20+12+3}{80} = \frac{35}{80}$ (= 0.4375)	B1 for mean ***NB answer given***	**1**
(ii)	Variance = $0.6907^2 = 0.4771$ So Poisson distribution may be appropriate, since mean is close to variance	B1 for variance E1*dep on squaring s*	**2**
(iii)	P(X = 1) = $e^{-0.4375}\frac{0.4375^1}{1!}$ = 0.282 (3 s.f.) *Either:* Thus the expected number of 1's is 22.6 which is reasonably close to the observed value of 20. *Or*: This probability compares reasonably well with the relative frequency 0.25	M1 for probability calc. M0 for tables unless interpolated (0.2813) A1 B1 for expectation of 22.6 or r.f. of 0.25 E1 for comparison	**4**
(iv)	$\lambda = 8\times0.4375 = 3.5$ Using tables: P($X \geq 12$) = 1 – P($X \leq 11$) = 1 – 0.9997 = 0.0003	B1 for mean (SOI) M1 for using tables to find 1 – P($X \leq 11$) A1 FT	**3**
(v)	The probability of at least 12 free repairs is very low, so the model is not appropriate. This is probably because the mean number of free repairs in the launderette will be much higher since the machines will get much more use than usual.	E1 for 'at least 12' E1 for very low E1	**3**
(vi)	*(A)* $\lambda = 0.4375 + 0.15 = 0.5875$ P(X = 3) = $e^{-0.5875}\frac{0.5875^3}{3!}$ = 0.0188 (3 s.f.) *(B)* P(Drier needs 1) = $e^{-0.15}\frac{0.15^1}{1!}$ = 0.129 P(Each needs just 1) = 0.282 $\times$0.129 = 0.036	B1 for mean (SOI) M1 A1 B1 for 0.129 (SOI) B1FT for 0.036	**3** **2**
			18

Question 4

(i)

H_0: no association between ambition and home location;
H_1: some association between ambition and home location;

Observed		Home location	
		City	Non-city
Ambition	Good results	102	147
	Other	75	156

M1 A1 for attempt at expected values

Expected		Home location	
		City	Non-city
Ambition	Good results	91.82	157.18
	Other	85.18	145.82

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

Contribution to the test statistic		Home location	
		City	Non-city
Ambition	Good results	1.129	0.659
	Other	1.217	0.711

A1CAO for X^2

$X^2 = 3.716$
Refer to χ_1^2
Critical value at 5% level = 3.841
Result is not significant
There is insufficient evidence to conclude that there is any association between home location and ambition.
NB if H_0 H_1 reversed, or 'correlation' mentioned, do not award first B1or final B1 or final E1

B1 for 1 dof SOI
B1 CAO for cv
B1 *dep on attempt at cv*
E1 conclusion in context

4

4

(ii) (*A*)

Expected Country, Results = 249 * 156 / 480 = 80.93
Expected Country, Other = 231 * 156 / 480 = 75.08

B1
B1

2

(*B*)

Refer to χ_2^2
Critical value at 5% level = 5.991
Result is significant
There is evidence to conclude that there is association between home location and ambition.

B1 for 2 dof SOI
B1 CAO for cv
E1 for conclusion in context

3

(*C*)

'Country' students are much less likely than city or town to have 'Results' as their main ambition. Low contributions show that city and town students do not appear to differ markedly in their ambitions.

E1 for correct obs^n for 'Country'
E1 for additional correct observation (must refer to contributions)

2

(iii)

Conclusion in (i) is valid if only categorizing home location into city and non-city. However if non-city is subdivided into town and country this additional subdivision gives the data more precision and allows the relationship in part (ii) (*C*) to be revealed.

E1

E1

2

18

4767 - Statistics 2

General Comments

Most candidates were well prepared for this examination, demonstrating a good command of the necessary calculation techniques, and were able to complete all questions within the allowed time. None of the questions stood out as being either noticeably difficult or easy. Few candidates scored all of the available marks for explanation.

Comments on Individual Questions

Section A

1 (i) Very well answered, mostly producing full marks. Most lost marks occurred in the calculation of the gradient of the regression line, usually through the use of an incorrect method. Some candidates obtained the correct gradient but did not use the centroid of the data to find the t-intercept. Some candidates relying on calculators gave the incorrect equation t = 12.6v + 167.

(ii)A & B Most candidates scored both marks for the predictions. Many candidates gave suitable comments regarding the reliability of the predictions. Comments which failed to provide reasons for reliability/unreliability of the predictions scored no marks.

(iii) Few candidates scored marks on this part of the question. Many simply pointed out that the coefficient was the gradient of the line. Some managed to explain that it gave an indication of the rate of change of time taken for the kettle to boil with respect to the volume of water in the kettle. Very few mentioned units of time &/or volume.

(iv) Many scored full marks. Most knew to find the difference between the predicted and observed values but were not always sure of the signs of the residuals.

(v) Many candidates scored full marks. Marks were lost for failing to explain that the *distance* that needed to be measured was *vertical*. Some candidates did not realise that the question was asking how to measure residuals from a diagram and simply explained how to find residuals from an equation. Most provided an acceptable explanation of how to obtain the sign of the residual.

2 (a)(i) This question was well answered with many scoring full marks. Common errors included use of variance instead of standard deviation, and unnecessary continuity corrections.

(ii) Well answered with many candidates working to a suitable degree of accuracy and gaining full marks.

(iii) Well answered. A few candidates lost marks through using -1.645 instead of +1.645 in their equation although candidates who used $-1.645 \times 4 = (k - 33)$ were given the benefit of the doubt. A small number of candidates failed to use 33, with 28 and 24 seen in its place on several occasions.

(b)(i) Most candidates provided correct hypotheses. Few candidates identified μ as the population mean.

(ii) Well answered with full marks awarded to a reasonable proportion of candidates. Many lost marks through failing to use an appropriate test statistic despite help being available in the formula booklet. Omissions of square root signs were common. A small number failed to recognise that the value, 4.77 kg, was a *total weight* when calculating their test statistic – those preferring to work with total weight throughout could still obtain full marks. Most candidates now appreciate the requirement to provide conclusions in context.

3 (i) Nearly all candidates provided an acceptable justification of the given answer.

(ii) Many candidates lost marks on this question through failing to calculate the variance. Many gave the incorrect reason "the mean is approximately equal to the standard deviation" to support the Poisson model. A small number gained no marks for stating "events occur randomly and independently with a uniform mean rate", and/or "n is large and p is small"

(iii) Nearly all candidates obtained the correct value for P(X = 1) and most then went on to make a suitable comparison to receive full marks. A few lost marks for not providing enough detail – e.g. finding the expected number of 1s as 22.6 but not specifying the value in the table with which it was being compared.

(iv) Most candidates scored full marks. A small number mistakenly thought that P($X \geq 12$) was the same as 1 – P($X \leq 12$). A similar number used the Poisson p.d.f. to find P(X = 12) and used 1 – P(X = 12), which gained no credit.

(v) Most candidates picked up two of the three available marks – usually for noticing that the previous answer was small, and for explaining that in the laundrette the machines will be used more often than in the home. The mark for appreciating that a "tail" probability was used tended to be the mark dropped.

(vi)*A* Well answered, with most scoring full marks. Some candidates lost marks for failing to use the correct mean. Those who failed to combine the means and use a single Poisson distribution, preferring to work with separate distributions, often lost marks – usually for failing to identify all four combinations – though some scored full marks with this method.

(vi)*B* Not so well answered, with many adding rather than multiplying their probabilities. Most managed to obtain P(Drier needs 1 repair).

4 (i) Well answered. Most managed to provide correct hypotheses. In the calculation of X^2, some lost marks through excessive rounding of their expected frequencies. Candidates should be encouraged to work to at least 2dp when finding expected frequencies. Nearly all candidates used 1 degree of freedom as required, and found the correct critical value. Some lost marks for making the wrong conclusion. As ever, those failing to provide context in their conclusion were penalised. Simply stating "there is no evidence of association (between the two/variables)" did not earn the mark.

(ii)*A* Well answered

(ii)*B* Well answered, with those who scored the last four marks in part (i) usually gaining full marks.

(ii)*C* Poorly answered. One requirement in such questions is for candidates to identify the large contributions, indicating strong association (which most candidates can do), and to distinguish between positive and negative association (which tends to be neglected). Another requirement is to identify the small contributions, which show little association between the categories. Candidates commonly fail to refer to the contributions at all. Many referred only to “strong ambition” for those living in the country, without distinguishing between the two categories of ambition.

(iii) Poorly answered.