OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE **4767/01**

MATHEMATICS (MEI)

Statistics 2

FRIDAY 23 MAY 2008 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.
- 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 researcher believes that there is a negative correlation between money spent by the government on education and population growth in various countries. A random sample of 48 countries is selected to investigate this belief. The level of government spending on education x, measured in suitable units, and the annual percentage population growth rate y, are recorded for these countries. Summary statistics for these data are as follows.

$$\Sigma x = 781.3 \quad \Sigma y = 57.8 \quad \Sigma x^2 = 14\,055 \quad \Sigma y^2 = 106.3 \quad \Sigma xy = 880.1 \quad n = 48$$

(i) Calculate the sample product moment correlation coefficient. [5]

(ii) Carry out a hypothesis test at the 5% significance level to investigate the researcher's belief. State your hypotheses clearly, defining any symbols which you use. [6]

(iii) State the distributional assumption which is necessary for this test to be valid. Explain briefly how a scatter diagram may be used to check whether this assumption is likely to be valid. [2]

(iv) A student suggests that if the variables are negatively correlated then population growth rates can be reduced by increasing spending on education. Explain why the student may be wrong. Discuss an alternative explanation for the correlation. [3]

(v) State briefly one advantage and one disadvantage of using a smaller sample size in this investigation. [2]

2 A public water supply contains bacteria. Each day an analyst checks the water quality by counting the number of bacteria in a random sample of 5 ml of water.

Throughout this question, you should assume that the bacteria occur randomly at a mean rate of 0.37 bacteria per 5 ml of water.

(i) Use a Poisson distribution to

(*A*) find the probability that a 5 ml sample contains exactly 2 bacteria, [2]

(*B*) show that the probability that a 5 ml sample contains more than 2 bacteria is 0.0064. [3]

(ii) The month of September has 30 days. Find the probability that during September there is at most one day when a 5 ml sample contains more than 2 bacteria. [4]

The daily 5 ml sample is the first stage of the quality control process. The remainder of the process is as follows.

- If the 5 ml sample contains more than 2 bacteria, then a 50 ml sample is taken.
- If this 50 ml sample contains more than 8 bacteria, then a sample of 1000 ml is taken.
- If this 1000 ml sample contains more than 90 bacteria, then the supply is declared to be 'questionable'.

(iii) Find the probability that a random sample of 50 ml contains more than 8 bacteria. [3]

(iv) Use a suitable approximating distribution to find the probability that a random sample of 1000 ml contains more than 90 bacteria. [4]

(v) Find the probability that the supply is declared to be questionable. [2]

3 A company has a fleet of identical vans. Company policy is to replace all of the tyres on a van as soon as any one of them is worn out. The random variable X represents the number of miles driven before the tyres on a van are replaced. X is Normally distributed with mean 27 500 and standard deviation 4000.

(i) Find $P(X > 25\,000)$. [4]

(ii) 10 vans in the fleet are selected at random. Find the probability that the tyres on exactly 7 of them last for more than 25 000 miles. [3]

(iii) The tyres of 99% of vans last for more than k miles. Find the value of k. [3]

A tyre supplier claims that a different type of tyre will have a greater mean lifetime. A random sample of 15 vans is fitted with these tyres. For each van, the number of miles driven before the tyres are replaced is recorded. A hypothesis test is carried out to investigate the claim. You may assume that these lifetimes are also Normally distributed with standard deviation 4000.

(iv) Write down suitable null and alternative hypotheses for the test. [3]

(v) For the 15 vans, it is found that the mean lifetime of the tyres is 28 630 miles. Carry out the test at the 5% level. [5]

[Question 4 is printed overleaf.]

4 A student is investigating whether there is any association between the species of shellfish that occur on a rocky shore and where they are located. A random sample of 160 shellfish is selected and the numbers of shellfish in each category are summarised in the table below.

		Location		
		Exposed	Sheltered	Pool
Species	Limpet	24	32	16
	Mussel	24	11	3
	Other	5	22	23

(i) Write down null and alternative hypotheses for a test to examine whether there is any association between species and location. [1]

The contributions to the test statistic for the usual χ^2 test are shown in the table below.

Contribution		Location		
		Exposed	Sheltered	Pool
Species	Limpet	0.0009	0.2585	0.4450
	Mussel	10.3472	1.2756	4.8773
	Other	8.0719	0.1402	7.4298

The sum of these contributions is 32.85.

(ii) Calculate the expected frequency for mussels in pools. Verify the corresponding contribution 4.8773 to the test statistic. [4]

(iii) Carry out the test at the 5% level of significance, stating your conclusion clearly. [5]

(iv) For each species, comment briefly on how its distribution compares with what would be expected if there were no association. [5]

(v) If 3 of the 160 shellfish are selected at random, one from each of the 3 types of location, find the probability that all 3 of them are limpets. [3]

4767 Statistics 2

Question 1

(i)	**EITHER:** $S_{xy} = \Sigma xy - \frac{1}{n}\Sigma x \Sigma y = 880.1 - \frac{1}{48} \times 781.3 \times 57.8$ $= -60.72$ $S_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 14055 - \frac{1}{48} \times 781.3^2 = 1337.7$ $S_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = 106.3 - \frac{1}{48} \times 57.8^2 = 36.70$ $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-60.72}{\sqrt{1337.7 \times 36.70}} = -0.274$ **OR:** cov $(x,y) = \frac{\sum xy}{n} - \bar{x}\bar{y} = 880.1/48 - 16.28 \times 1.204$ $= -1.265$ rmsd$(x) = \sqrt{\frac{S_{xx}}{n}} = \sqrt{}(1337.7/48) = \sqrt{}27.87 = 5.279$ rmsd$(y) = \sqrt{\frac{S_{yy}}{n}} = \sqrt{}(36.70/48) = \sqrt{}0.7646 = 0.8744$ $r = \frac{\text{cov}(x,y)}{rmsd(x)rmsd(y)} = \frac{-1.265}{5.279 \times 0.8744} = -0.274$	M1 for method for S_{xy} M1 for method for at least one of S_{xx} or S_{yy} A1 for at least one of S_{xy}, S_{xx}, S_{yy} correct M1 for structure of r A1 CAO (–0.27 to –0.28) M1 for method for cov (x,y) M1 for method for at least one msd A1 for at least one of cov/msd correct M1 for structure of r A1 CAO (–0.27 to –0.28)	**5**
(ii)	H_0: $\rho = 0$ H_1: $\rho < 0$ (one-tailed test) where ρ is the population correlation coefficient For $n = 48$, 5% critical value = 0.2403 Since **|** – 0.274 **|** > 0.2403 we can reject H_0: There is sufficient evidence at the 5% level to suggest that there is negative correlation between education spending and population growth.	B1 for H_0, H_1 in symbols B1 for defining ρ B1FT for critical value M1 for sensible comparison leading to a conclusion A1 for result (FT r<0) E1 FT for conclusion in words	**6**
(iii)	Underlying distribution must be bivariate Normal. If the distribution is bivariate Normal then the scatter diagram will have an elliptical shape.	B1 CAO for bivariate Normal B1 indep for elliptical shape	**2**
(iv)	• Correlation does not imply causation • There could be a third factor • increased growth could cause lower spending. Allow any sensible alternatives, including example of a possible third factor.	E1 E1 E1	**3**
(v)	Advantage – less effort or cost Disadvantage – the test is less sensitive (ie is less likely to detect any correlation which may exist)	E1 E1	**2**
			18

Question 2

(i)	(*A*) P(X = 2) = $e^{-0.37}\frac{0.37^2}{2!}$ = 0.0473 (*B*) P(X > 2) = 1 – ($e^{-0.37}\frac{0.37^2}{2!}$ + $e^{-0.37}\frac{0.37^1}{1!}$ + $e^{-0.37}\frac{0.37^0}{0!}$) = 1 – (0.0473 + 0.2556 + 0.6907) = 0.0064	M1 A1 (2 s.f.) M1 for P(X = 1) and P(X = 0) M1 for complete method A1 ***NB Answer given***	**5**
(ii)	P(At most one day more than 2) = $\binom{30}{1} \times 0.9936^{29} \times 0.0064 + 0.9936^{30}$ = = 0.1594 + 0.8248 = 0.9842	M1 for coefficient M1 for $0.9936^{29} \times 0.0064$ M1 for 0.993630 A1 CAO (min 2sf)	**4**
(iii)	λ = 0.37 ×10 = 3.7 P(X > 8) = 1 – 0.9863 = 0.0137	B1 for mean (SOI) M1 for probability A1 CAO	**3**
(iv)	Mean no. per 1000ml = 200 × 0.37 = 74 Using Normal approx. to the Poisson, X ~ N(74, 74) P(X > 90) = $P\left(Z > \frac{90.5-74}{\sqrt{74}}\right)$ = P(Z >1.918) = 1 – Φ(1.918) = 1 – 0.9724 = 0.0276	B1 for Normal approx. with correct parameters (SOI) B1 for continuity corr. M1 for probability using correct tail A1 CAO (min 2 s.f.)**, (**but FT wrong or omitted CC)	**4**
(v)	P(questionable) = 0.0064 × 0.0137 ×0.0276 = 2.42 × 10^{-6}	M1 A1 CAO	**2**
			18

Question 3

(i)	$X \sim N(27500, 4000^2)$ $P(X > 25000) = P\left(Z > \frac{25000 - 27500}{4000}\right)$ $= P(Z > -0.625)$ $= \Phi(0.625) = 0.7340$ (3 s.f.)	M1 for standardising A1 for -0.625 M1 *dep* for correct tail A1CAO (must include use of differences)	**4**
(ii)	P(7 of 10 last more than 25000) $= \binom{10}{7} \times 0.7340^7 \times 0.2660^3 = 0.2592$	M1 for coefficient M1 for $0.7340^7 \times 0.2660^3$ A1 FT (min 2sf)	**3**
(iii)	From tables $\Phi^{-1}(0.99) = 2.326$ $\frac{k - 27500}{4000} = -2.326$ $x = 27500 - 2.326 \times 4000 = 18200$	B1 for 2.326 seen M1 for equation in k and negative z-value A1 CAO for awrt 18200	**3**
(iv)	H_0: $\mu = 27500$; H_1: $\mu > 27500$ Where μ denotes the mean lifetime of the new tyres.	B1 for use of 27500 B1 for both correct B1 for definition of μ	**3**
(v)	Test statistic = $\frac{28630 - 27500}{4000/\sqrt{15}} = \frac{1130}{1032.8}$ = 1.094 5% level 1 tailed critical value of z = 1.645 1.094 < 1.645 so not significant. There is not sufficient evidence to reject H_0 There is insufficient evidence to conclude that the new tyres last longer.	M1 must include $\sqrt{15}$ A1 FT B1 for 1.645 M1 *dep* for a sensible comparison leading to a conclusion A1 for conclusion in words in context	**5**
			18

Question 4

(i)	H_0: no association between location and species. H_1: some association between location and species.	B1 for both	**1**
(ii)	Expected frequency = 38/160 × 42 = 9.975 Contribution = $(3 – 9.975)^2 / 9.975$ = 4.8773	M1 A1 M1 for valid attempt at $(O-E)^2/E$ A1 ***NB Answer given***	**4**
(iii)	Refer to χ^2_4 Critical value at 5% level = 9.488 Test statistic X^2 = 32.85 Result is significant There appears to be some association between location and species NB if H_0 H_1 reversed, or 'correlation' mentioned, do not award first B1or final E1	B1 for 4 deg of f(seen) B1 CAO for cv M1 Sensible comparison, using 32.85, leading to a conclusion A1 for correct conclusion (FT their c.v.) E1 conclusion in context	**5**
(iv)	• Limpets appear to be distributed as expected throughout all locations. • Mussels are much more frequent in exposed locations and much less in pools than expected. • Other shellfish are less frequent in exposed locations and more frequent in pools than expected.	E1 E1, E1 E1, E1	**5**
(v)	$\frac{24}{53} \times \frac{32}{65} \times \frac{16}{42} = 0.0849$	M1 for one fraction M1 for product of all 3 A1 CAO	**3**
			18

4767 Statistics 2

General Comments

In keeping with recent sessions, the majority of candidates demonstrated good understanding and high marks were plentiful. No one question, or part of a question, stood out as being particularly difficult or particularly easy. This year saw a noticeable improvement in answers where some form of explanation or interpretation was required.

Comments on Individual Questions

Section A

1) (i) Well answered. Common mistakes involved premature rounding – rounding of $\bar{y}$ to 1.20, seen frequently. A few candidates omitted the square root in the denominator.

(ii) Well answered. Many obtained at least 5 marks out of the available 6; typically, the lost mark was for failure to define ρ as the population correlation coefficient, despite being asked to 'define any symbols' used. Some candidates lost a mark for failing to provide a conclusion in context. Those candidates providing nonsensical comparisons, for example comparing a negative p.m.c.c. with a positive critical value, were heavily penalised.

(iii) The requirement for an underlying 'bivariate Normal distribution' for the test to be valid was not widely known. Candidates were more familiar with the idea that the points in the scatter diagram should fall within an elliptical shape. Many candidates commented that the test would be valid if the scatter diagram showed negative correlation.

(iv) A good variety of comments were seen. Very few gained full marks. Many scored a mark for pointing out that there may be other factors involved. Comments relating to 'causation' were less common. Many candidates commented on dependence/independence without fully answering the question. A mark was available for those who pointed out that higher spending on education could be due to lower population growth rate, i.e. reversing the dependency.

(v) Reasonably well answered, but little credit was given to vague answers – e.g. simply stating that the a smaller sample size leads to a 'less accurate' result gained no credit, but pointing out that a smaller sample could be 'less representative of the population' gained a mark. Candidates are advised to use statistical words/reasons in their comments as much as possible.

2) (i) (*A*) Well answered.

(i) (*B*) Accuracy errors often prevented candidates achieving the provided answer. Few candidates showed awareness that, with the printed answer being correct to 4 decimal places, working should have included, ideally, values showing 5 decimal places. Candidates should be aware that when answers are given, sufficient working must be seen.

(ii) A significant minority failed to use the appropriate binomial model. For those who did, many simply found P(X=1) then stopped. A common failure was omitting $^{30}C_1$. Those using Po(0.192) did so successfully in most cases.

(iii) Most candidates managed to use Po(3.7) tables to get the correct answer. The most common failure was saying P(X > 8) = 1 – P (X ≤ 7).

(iv) Most obtained N(74, 74) but other parameters were seen (e.g. $\sigma^2 = 200 \times 0.37 \times 0.63 = 46.62$). Knowledge of the need for, and how to apply, a continuity correction was lacking. Other frequent mistakes were use of σ = 74, not √74, and failure to use tables to a sufficient level of accuracy [i.e. 1 – Φ(1.918) not 1 – Φ(1.92)].

(v) Many candidates recognised the correct method. A significant number simply thought that the answer was the same as that given in part (iv).

3) (i) Well answered with most obtaining z = ± 0.625. A sizeable minority gave 0.266 as their answer (from using the wrong tail).

(ii) Nearly all showed an intention to use B(10, 0.734). [or B(10, their (i))] As in Q2, a common failure was omitting the ^nC_r term, here $^{10}C_7$. A few had just 10 instead of $^{10}C_7$. Sometimes the powers of p & q were reversed even though part (i) was correct.

(iii) Most found the correct z = 2.326 from the Inverse Normal table. Many proceeded to use this value of z and produced an incorrect upper tail value. A few bypassed Inverse Normal tables, using $(x – \mu) / \sigma$ = ±0.99 or ±0.01.

(iv) Hypotheses were usually correct. A precise definition of μ was lacking in most cases.

(v) Generally, well done. Those omitting √15 when standardising were heavily penalised. A number of candidates obtaining a test statistic of 1.094 went on to make a nonsensical comparison (e.g. 1.094 > 0.05 or 0.863 < 1.645). Though other approaches were seen, the majority of candidates kept to the approach in the published mark scheme.

4) (i) Well answered. A small number omitted the context from their hypotheses. Very few mentioned correlation or tried to use parameters in their hypotheses.

(ii) Well answered, although accuracy was an issue for a number of candidates. Obtaining the expected frequency of 9.975 was not a problem, but some rounded this to 9.98 which does not lead to the provided answer.

(iii) Well answered on the whole. Frequently seen mistakes included: use of a 2 ½ % critical value, using the wrong number of degrees of freedom, using $X^2 =$ 4.8773 instead of the value provided. For some reason, many candidates felt it necessary to calculate the test statistic themselves, despite it being provided in the question – there was no evidence that this prevented those candidates from finishing the paper.

(iv) Fully correct answers were seen, but despite the question stating clearly, "…compares with what would be expected…", many comments written did not address this. Many compared one expected frequency with another expected frequency, or one observed frequency with another. Some only made comments about the various contributions to the test statistic; of these, some recognised that the high contributions indicated some 'difference' between observed and expected, but did not explain whether there were more observed than expected or fewer. The candidates' attempts at this question were better than attempts at similar types seen in recent sessions.

(v) Well answered. Many were able to multiply the correct three fractions and score full marks. Putting the grand total on the bottom instead of the column total was, however, a fairly frequent error.