OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE **4767/01**

MATHEMATICS (MEI)

Statistics 2

TUESDAY 15 JANUARY 2008 Morning

Time: 1 hour 30 minutes

Additional materials: Answer Booklet (8 pages)
Graph paper
MEI Examination Formulae and Tables (MF2)

INSTRUCTIONS TO CANDIDATES

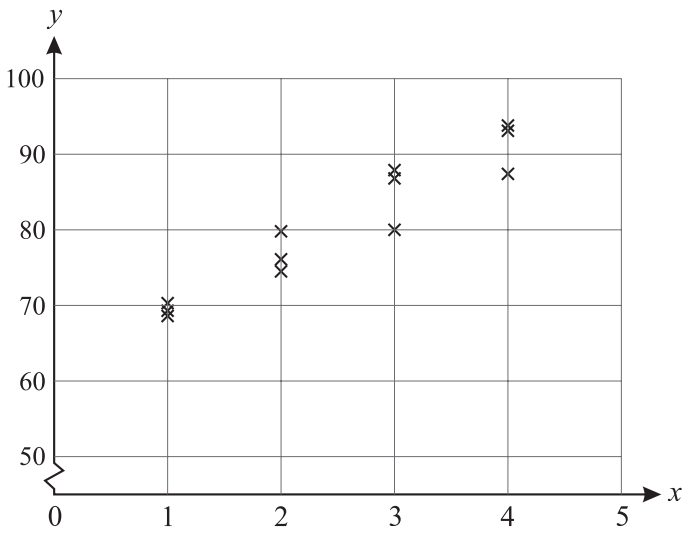
- Write your name, centre number and candidate number in the spaces provided on the answer booklet.
- Read each question carefully and make sure you know what you have to do before starting your answer.
- Answer **all** the questions.
- You are permitted to use a graphical calculator in this paper.
- 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 biology student is carrying out an experiment to study the effect of a hormone on the growth of plant shoots. The student applies the hormone at various concentrations to a random sample of twelve shoots and measures the growth of each shoot. The data are illustrated on the scatter diagram below, together with the summary statistics for these data. The variables x and y, measured in suitable units, represent concentration and growth respectively.



$$n = 12,\ \Sigma x = 30,\ \Sigma y = 967.6,\ \Sigma x^2 = 90,\ \Sigma y^2 = 78\,926,\ \Sigma xy = 2530.3.$$

(i) State which of the two variables x and y is the independent variable and which is the dependent variable. Briefly explain your answers. [3]

(ii) Calculate the equation of the regression line of y on x. [5]

(iii) Use the equation of the regression line to calculate estimates of shoot growth for concentrations of

(A) 1.2,

(B) 4.3.

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

(iv) Calculate the value of the residual for the data point where $x = 3$ and $y = 80$. [3]

(v) In further experiments, the student finds that using concentration $x = 6$ results in shoot growths of around $y = 20$. In the light of all the available information, what can be said about the relationship between x and y? [3]

2 A large hotel has 90 bedrooms. Sometimes a guest makes a booking for a room, but then does not arrive. This is called a 'no-show'. On average 10% of bookings are no-shows. The hotel manager accepts up to 94 bookings before saying that the hotel is full. If at least 4 of these bookings are no-shows then there will be enough rooms for all of the guests. 94 bookings have been made for each night in August. You should assume that all bookings are independent.

(i) State the distribution of the number of no-shows on one night in August. [2]

(ii) State the conditions under which the use of a Poisson distribution is appropriate as an approximation to a binomial distribution. [2]

(iii) Use a Poisson approximating distribution to find the probability that, on one night in August,

(*A*) there are exactly 4 no-shows,

(*B*) there are enough rooms for all of the guests who do arrive. [5]

(iv) Find the probability that, on all of the 31 nights in August, there are enough rooms for all of the guests who arrive. [2]

(v) (*A*) In August there are $31 \times 94 = 2914$ bookings altogether. State the exact distribution of the total number of no-shows during August. [2]

(*B*) Use a suitable approximating distribution to find the probability that there are at most 300 no-shows altogether during August. [5]

3 In a large population, the diastolic blood pressure (DBP) of 5-year-old children is Normally distributed with mean 56 and standard deviation 6.5.

(i) Find the probability that the DBP of a randomly selected 5-year-old child is between 52.5 and 57.5. [4]

The DBP of young adults is Normally distributed with mean 68 and standard deviation 10.

(ii) A 5-year-old child and a young adult are selected at random. Find the probability that the DBP of one of them is over 62 and the other is under 62. [5]

(iii) Sketch both distributions on a single diagram. [4]

(iv) For another age group, the DBP is Normally distributed with mean 82. The DBP of 12% of people in this age group is below 62. Find the standard deviation for this age group. [4]

[Question 4 is printed overleaf.]

4 **(a)** A researcher believes that there may be some association between a student's sex and choice of certain subjects at A-level. A random sample of 250 A-level students is selected. The table below shows, for each sex, how many study either or both of the two subjects, Mathematics and English.

	Mathematics only	English only	Both	Neither	**Row totals**
Male	38	19	6	32	**95**
Female	42	55	9	49	**155**
Column totals	**80**	**74**	**15**	**81**	**250**

Carry out a test at the 5% significance level to examine whether there is any association between a student's sex and choice of subjects. State carefully your null and alternative hypotheses. Your working should include a table showing the contributions of each cell to the test statistic. [12]

(b) Over a long period it has been determined that the mean score of students in a particular English module is 67.4 and the standard deviation is 8.9. A new teaching method is introduced with the aim of improving the results. A random sample of 12 students taught by the new method is selected. Their mean score is found to be 68.3. Carry out a test at the 10% level to investigate whether the new method appears to have been successful. State carefully your null and alternative hypotheses. You should assume that the scores are Normally distributed and that the standard deviation is unchanged. [7]

4767 Statistics 2

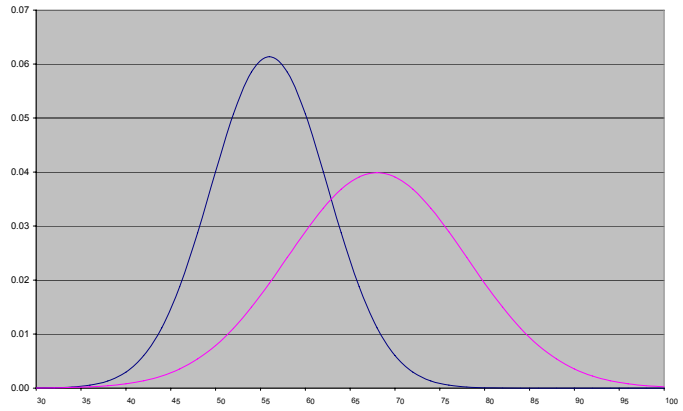
Question 1

(i)	*x* is independent, *y* is dependent since the values of *x* are chosen by the student but the values of *y* are dependent on *x*	B1 E1 dep E1 dep	**3**
(ii)	$\bar{x}$ = 2.5, $\bar{y}$ = 80.63 $b = \frac{Sxy}{Sxx} = \frac{2530.3 - 30 \times 967.6/12}{90 - 30^2/12} = \frac{111.3}{15} = 7.42$ OR $b = \frac{2530.3/12 - 2.50 \times 80.63}{90/12 - 2.50^2} = \frac{9.275}{1.25} = 7.42$ Hence least squares regression line is: $y - \bar{y} = b(x - \bar{x})$ $\Rightarrow y - 80.63 = 7.42(x - 2.5)$ $\Rightarrow y = 7.42x + 62.08$	B1 for $\bar{x}$ and $\bar{y}$ used (SOI) M1 for attempt at gradient (*b*) A1 for 7.42 **cao** M1 for equation of line A1 FT (*b*>0) for complete equation	**5**
(iii)	*(A)* For *x* = 1.2, predicted growth = 7.42 × 1.2 + 62.08 = 71.0 *(B)* For *x* = 4.3, predicted growth = 7.42 × 4.3 + 62.08 = 94.0 Valid relevant comments relating to the predictions such as : Comment re interpolation/extrapolation Comment relating to the fact that *x* = 4.3 is only just beyond the existing data. Comment relating to size of residuals near each predicted value (need not use word 'residual')	M1 for at least one prediction attempted. A1 for both answers (FT their equation if *b*>0) E1 (first comment) E1 (second comment)	**4**
(iv)	*x* = 3 $\Rightarrow$ predicted *y* = 7.42 × 3 + 62.08 = 84.3 Residual = 80 − 84.3 = −4.3	M1 for prediction M1 for subtraction A1 FT (*b*>0)	**3**
(v)	This point is a long way from the regression line. The line may be valid for the range used in the experiment but then the relationship may break down for higher concentrations, or the relationship may be non linear.	E1 E1 for valid in range E1 for *either* 'may break down' *or* 'could be non linear' *or* other relevant comment	**3**
			18

Question 2

(i)	Binomial (94,0.1)	B1 for binomial B1 dep for parameters	**2**
(ii)	*n* is large and *p* is small	B1, B1 Allow appropriate numerical ranges	**2**
(iii)	$\lambda = 94 \times 0.1 = 9.4$ *(A)* $P(X = 4) = e^{-9.4}\ \frac{9.4^4}{4!} = 0.0269$ (3 s.f.) or from tables $= 0.0429 - 0.0160 = 0.0269$ ***cao*** *(B)* Using tables: $P(X \geq 4) = 1 - P(X \leq 3)$ $= 1 - 0.0160 = 0.9840$ ***cao***	B1 for mean M1 for calculation or use of tables A1 M1 for attempt to find $P(X \geq 4)$ A1 **cao**	**5**
(iv)	P(sufficient rooms throughout August) $= 0.9840^{31} = 0.6065$	M1 A1 FT	**2**
(v)	(*A*) 31 × 94 = 2914 Binomial (2914,0.1) (*B*)Use Normal approx with $\mu = np = 2914 \times 0.1 = 291.4$ $\sigma^2 = npq = 2914 \times 0.1 \times 0.9 = 262.26$ $P(X \leq 300.5) = P\left(Z \leq \frac{300.5 - 291.4}{\sqrt{262.26}}\right)$ $= P(Z \leq 0.5619) = \Phi(0.5619) = 0.7130$	B1 for binomial B1 dep, for parameters B1 B1 B1 for continuity corr. M1 for probability using correct tail A1 **cao,** (but FT wrong or omitted CC)	**2** **5**
			18

Question 3

(i)	$X \sim N(56, 6.5^2)$ $P(52.5 < X < 57.5) = P\left(\frac{52.5-56}{6.5} < Z < \frac{57.5-56}{6.5}\right)$ $= P(-0.538 < Z < 0.231)$ $= \Phi(0.231) - (1 - \Phi(0.538))$ $= 0.5914 - (1 - 0.7046)$ $= 0.5914 - 0.2954$ $= 0.2960$ (4 s.f.) *or* 0.296 (to 3 s.f.)	M1 for standardizing A1 for -0.538 and 0.231 M1 for prob. with tables and correct structure A1 CAO (min 3 s.f., to include use of difference column)	**4**
(ii)	$P(\text{5-year-old} < 62) = P\left(Z < \frac{62-56}{6.5}\right)$ $= \Phi(0.923) = 0.8220$ $P(\text{young adult} < 62) = P\left(Z < \frac{62-68}{10}\right)$ $= \Phi(-0.6) = 1 - 0.7257 = 0.2743$ P(One over, one under) $= 0.8220 \times 0.7257 + 0.1780 \times 0.2743$ $= 0.645$	B1 for 0.8220 or 0.1780 B1 for 0.2743 or 0.7257 M1 for either product M1 for sum of both products A1 CAO	**5**
(iii)	 	G1 for shape G1 for means, shown explicitly or by scale G1 for lower max height in young adults G1 for greater variance in young adults	**4**
(iv)	$Y \sim N(82, \sigma^2)$ From tables $\Phi^{-1}(0.88) = 1.175$ $\frac{62-82}{\sigma} = -1.175$ $-20 = -1.175\,\sigma$ $\sigma = 17.0$	B1 for 1.175 seen M1 for equation in σ with z-value M1 for correct handling of LH tail A1 **cao**	**4**
			17

Question 4

(i)

H_0: no association between sex and subject;
H_1: some association between sex and subject;

OBS	Maths	English	Both	Neither	**Row sum**
Male	38	19	6	32	**95**
Female	42	55	9	49	**155**
Col sum	**80**	**74**	**15**	**81**	**250**

EXP	Maths	English	Both	Neither	**Row sum**
Male	30.40	28.12	5.70	30.78	**95**
Female	49.60	45.88	9.30	50.22	**155**
Col sum	**80**	**74**	**15**	**81**	**250**

CONT	Maths	English	Both	Neither
Male	1.900	2.958	0.016	0.048
Female	1.165	1.813	0.010	0.030

$X^2 = 7.94$

Refer to χ^2_3
Critical value at 5% level = 7.815
Result is significant
There is evidence to suggest that there is some association between sex and subject choice.
NB if H_0 H_1 reversed, or 'correlation' mentioned, do not award first B1 or final E1

B1

M1 A2 for expected values
(allow A1 for at least one row or column correct)

M1 for valid attempt at $(O-E)^2/E$
A1
NB These M1 A1 marks cannot be implied by a correct final value of X^2

M1 for summation
A1 **cao** for X^2

B1 for 3 deg of f
B1 CAO for cv

B1

E1

1

7

4

(ii)

H_0: $\mu = 67.4$; H_1: $\mu > 67.4$
Where μ denotes the mean score of the population of students taught with the new method.

Test statistic = $\dfrac{68.3 - 67.4}{8.9/\sqrt{12}} = \dfrac{0.9}{2.57}$
= 0.35

10% level 1 tailed critical value of z = 1.282
0.35 < 1.282 so not significant.
There is insufficient evidence to reject H_0
There is insufficient evidence to conclude that the mean score is increased by the new teaching method.

B1 for both correct

B1 for definition of μ

M1

A1 **cao**

B1 for 1.282

M1 for comparison

A1 for conclusion in words and in context

7

19

4767: Statistics 2

General Comments

The majority of candidates were well-prepared for this examination, continuing the pattern of recent years. It was evident that no question stood out as being either more difficult or more straightforward than the others. In general, candidates' abilities to structure answers to questions involving hypothesis tests, using correct notation and terminology, have shown improvement. As in recent sessions, many candidates struggled to obtain marks for explanation/interpretation, but otherwise scored well. The overall standard was high.

Comments on Individual Questions

Section A

1) (i) The majority correctly identified x as the independent variable, realising that growth depended on the hormone concentration. Few candidates stated that x was controlled.

(ii) Well answered with most candidates gaining full marks. Those leaving the equation of the regression line in unsimplified form were penalised. No extra credit was given to those candidates who calculated the p.m.c.c.

(iii) Most candidates successfully used their equations to obtain estimates of shoot growth, and the comments on the reliability of their estimates were generally as required. A number of candidates commented that the estimates were similar to the values on the graph, gaining no credit. The most successful used the idea of interpolation/extrapolation.

(iv) Most managed to obtain two of the three available marks, losing out on the final mark by providing a positive rather than negative residual.

(v) This part was poorly answered with only a few candidates obtaining full marks. Most candidates commented entirely about the context, completely avoiding discussion about the mathematical model and its suitability in the range given.

2) (i) Well answered. Several candidates used $n = 90$, leading to problems in the later parts of the question. In such questions it is expected that candidates will provide parameters and not just quote "binomial".

(ii) Well answered. Some candidates missed the point of this question and simply churned out comments relating to the conditions for a Poisson model to be used, generally, and not as an approximation to the binomial distribution.

(iii) *A* Well answered, with many candidates scoring full marks.

(iii) *B* Most candidates realised what was required, but some failed to correctly obtain the value of $P(X \geq 4)$.

(iv) Well answered.

(v) *A* Some candidates missed out on the marks here by writing down the Normal approximating distribution at this stage, bypassing the binomial distribution.

(v) *B* The Normal distribution was handled well. Many candidates failed to use an appropriate continuity correction and were penalised. Many used a Normal approximation to the Poisson distribution, leading to loss of accuracy.

3) (i) Mostly well answered, but many candidates lost the final accuracy mark through using *z*-values rounded to 2 d.p..

(ii) This proved difficult for many. Inappropriate attempts at continuity corrections were seen. Nonetheless, many managed to complete the question using the correct probability calculation with their values.

(iii) Well answered on the whole. Most candidates managed to draw a diagram containing two Normal curves and correctly label their means on a horizontal axis. Many managed to draw sketches which highlighted the difference in variance, but did not realise that this meant the curve for adults would have a lower maximum value.

(iv) Well answered. Some candidates lost marks through failing to handle the negative *z*-value correctly. A small number gave a negative value for σ.

4) (i) Well answered. Most candidates provided correct hypotheses. In calculating the test statistic, most candidates managed to work to an appropriate level of accuracy, helped by the lack of a need to round expected frequencies, and gained full marks for this. Some candidates failed to provide a table (or list) showing individual contributions to the test statistic despite being requested in the question, and hence lost marks. Most candidates correctly identified the correct number of degrees of freedom and critical value, and went on to make an appropriate conclusion.

(ii) Reasonably well answered. Most candidates scored a mark for providing correct hypotheses, but candidates still find the mark for defining μ as the population mean elusive. Indeed, many defined μ as the sample mean. Most candidates managed to obtain the correct test statistic and critical value then make an appropriate conclusion. A small number of inappropriate comparisons were seen, usually involving comparing a *z*-value with a probability. Several candidates treated the value 68.3 as a single observation rather than a sample mean and were penalised.