RECOGNISING ACHIEVEMENT

ADVANCED GCE

MATHEMATICS (MEI) 4767

Statistics 2

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

Monday 25 January 2010
Morning

Duration: 1 hour 30 minutes

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 A pilot records the take-off distance for his light aircraft on runways at various altitudes. The data are shown in the table below, where a metres is the altitude and t metres is the take-off distance. Also shown are summary statistics for these data.

a	0	300	600	900	1200	1500	1800
t	635	704	776	836	923	1008	1105

$n = 7 \quad \Sigma a = 6300 \quad \Sigma t = 5987 \quad \Sigma a^2 = 8\,190\,000 \quad \Sigma t^2 = 5\,288\,931 \quad \Sigma at = 6\,037\,800$

(i) Draw a scatter diagram to illustrate these data. **[3]**

(ii) State which of the two variables a and t is the independent variable and which is the dependent variable. Briefly explain your answer. **[3]**

(iii) Calculate the equation of the regression line of t on a. **[5]**

(iv) Use the equation of the regression line to calculate estimates of the take-off distance for altitudes

(*A*) 800 metres,

(*B*) 2500 metres.

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

(v) Calculate the value of the residual for the data point where $a = 1200$ and $t = 923$, and comment on its sign. **[4]**

2 On average 2% of a particular model of laptop computer are faulty. Faults occur independently and randomly.

(i) Find the probability that exactly 1 of a batch of 10 laptops is faulty. **[3]**

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

(iii) A school buys a batch of 150 of these laptops. Use a Poisson approximating distribution to find the probability that

(*A*) there are no faulty laptops in the batch, **[3]**

(*B*) there are more than the expected number of faulty laptops in the batch. **[3]**

(iv) A large company buys a batch of 2000 of these laptops for its staff.

(*A*) State the exact distribution of the number of faulty laptops in this batch. **[2]**

(*B*) Use a suitable approximating distribution to find the probability that there are at most 50 faulty laptops in this batch. **[5]**

3 In an English language test for 12-year-old children, the raw scores, X, are Normally distributed with mean 45.3 and standard deviation 11.5.

(i) Find

(A) $\mathrm{P}(X < 50)$, **[3]**

(B) $\mathrm{P}(45.3 < X < 50)$. **[2]**

(ii) Find the least raw score which would be obtained by the highest scoring 10% of children. **[3]**

(iii) The raw score is then scaled so that the scaled score is Normally distributed with mean 100 and standard deviation 15. This scaled score is then rounded to the nearest integer. Find the probability that a randomly selected child gets a rounded score of exactly 111. **[4]**

(iv) In a Mathematics test for 12-year-old children, the raw scores, Y, are Normally distributed with mean μ and standard deviation σ. Given that $\mathrm{P}(Y < 15) = 0.3$ and $\mathrm{P}(Y < 22) = 0.8$, find the values of μ and σ. **[5]**

[Question 4 is printed overleaf.]

Turn over

4 A council provides waste paper recycling services for local businesses. Some businesses use the standard service for recycling paper, others use a special service for dealing with confidential documents, and others use both. Businesses are classified as small or large. A survey of a random sample of 285 businesses gives the following data for size of business and recycling service.

		Recycling Service		
		Standard	Special	Both
Size of business	Small	35	26	44
	Large	55	52	73

(i) Write down null and alternative hypotheses for a test to examine whether there is any association between size of business and recycling service used. **[1]**

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

		Recycling Service		
		Standard	Special	Both
Size of business	Small	0.1023	0.2607	0.0186
	Large	0.0597	0.1520	0.0108

The sum of these contributions is 0.6041.

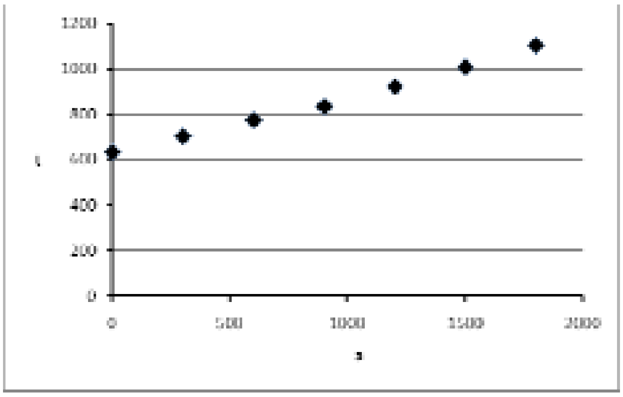
(ii) Calculate the expected frequency for large businesses using the special service. Verify the corresponding contribution 0.1520 to the test statistic. **[4]**

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

The council is also investigating the weight of rubbish in domestic dustbins. In 2008 the average weight of rubbish in bins was 32.8 kg. The council has now started a recycling initiative and wishes to determine whether there has been a reduction in the weight of rubbish in bins. A random sample of 50 domestic dustbins is selected and it is found that the mean weight of rubbish per bin is now 30.9 kg, and the standard deviation is 3.4 kg.

(iv) Carry out a test at the 5% level to investigate whether the mean weight of rubbish has been reduced in comparison with 2008. State carefully your null and alternative hypotheses. **[8]**

4767 Statistics 2

1	**(i)**		G1 For values of *a* G1 for values of *t* G1 for axes	**[3]**
	(ii)	*a* is independent, *t* is dependent since the values of *a* are not subject to random variation, but are determined by the runways which the pilot chooses, whereas the values of *t* are subject to random variation.	B1 E1dep E1dep	**[3]**
	(iii)	$\bar{a} = 900,\ \bar{t} = 855.2$ $b = \frac{S_{at}}{S_{aa}} = \frac{6037800 - 5987 \times 6300/7}{8190000 - 6300^2/7} = \frac{649500}{2520000} = 0.258$ OR $b = \frac{6037800/7 - 855.29 \times 900}{8190000/7 - 900^2} = \frac{92785}{360000} = 0.258$ hence least squares regression line is: $t - \bar{t} = b(a - \bar{a})$ $\Rightarrow\ t - 855.29 = 0.258\,(a - 900)$ $\Rightarrow\ t = 0.258a + 623$	B1 for $\bar{a}$ and $\bar{t}$ used (SOI) M1 for attempt at gradient (*b*) A1 for 0.258 **cao** M1 for equation of line A1 FT for complete equation	**[5]**
	(iv)	(*A*) For *a* = 800, predicted take–off distance $= 0.258 \times 800 + 623 = 829$ (*B*) For *a* = 2500, predicted take–off distance $= 0.258 \times 2500 + 623 = 1268$ Valid relevant comments relating to the predictions such as: First prediction is interpolation so should be reasonable Second prediction is extrapolation and may not be reliable	M1 for at least one prediction attempted A1 for both answers (FT their equation if *b*>0) E1 (first comment) E1 (second comment)	**[4]**
	(v)	$a = 1200 \Rightarrow$ predicted $t = 0.258 \times 1200 + 623 = 933$ Residual = 923 – 933 = –10 The residual is negative because the observed value is less than the predicted value.	M1 for prediction M1 for subtraction A1 FT E1	**[4]**
			Total	**[19]**

2	**(i)**	P(1 of 10 is faulty) $= \begin{pmatrix}10\\1\end{pmatrix} \times 0.02^1 \times 0.98^9 \ = 0.1667$	M1 for coefficient M1 for probabilities A1	**[3]**
	(ii)	n is large and p is small	B1, B1 Allow appropriate numerical ranges	**[2]**
	(iii)	$\lambda = 150 \times 0.02 = 3$ (A) P(X = 0) = $\tilde{e}^{-3}\dfrac{3^0}{0!}$ = 0.0498 (3 s.f.) or from tables = 0.0498	B1 for mean (soi) M1 for calculation or use of tables A1	**[3]**
		(B) Expected number = 3 Using tables: P(X > 3) = 1 – P($X \le$ 3) = 1 – 0.6472 = 0.3528	B1 expected no = 3 (soi) M1 A1	**[3]**
	(iv)	(A) Binomial(2000,0.02)	B1 for binomial B1 for parameters	**[2]**
		(B) Use Normal approx with $\mu = np = 2000 \times 0.02 = 40$ $\sigma^2 = npq = 2000 \times 0.02 \times 0.98 = 39.2$ $P(X \le 50) = P\left(Z \le \dfrac{50.5-40}{\sqrt{39.2}}\right)$ $= P(Z \le 1.677) = \Phi(1.677) = 0.9532$ NB Poisson approximation also acceptable for full marks	B1 B1 B1 for continuity corr. M1 for probability using correct tail A1 CAO	**[5]**
			Total	**[18]**

3	**(i)**	(*A*) $P(X < 50)$ $= P\left(Z < \frac{50-45.3}{11.5}\right)$ $= P(Z < 0.4087)$ $= \Phi(0.4087)$ $= 0.6585$	M1 for standardising M1 for correct structure of probability calc' A1 CAO inc use of diff tables NB When a candidate's answers suggest that (s)he appears to have neglected to use the difference column of the Normal distribution tables penalise the first occurrence only	**[3]**
		(*B*) $P(45.3 < X < 50)$ $= 0.6585 - 0.5$ $= 0.1585$	M1 A1	**[2]**
	(ii)	From tables $\Phi^{-1}(0.9) = 1.282$ $\frac{k-45.3}{11.5} = 1.282$ $k = 45.3 + 1.282 \times 11.5 = 60.0$	B1 for 1.282 seen M1 for equation in *k* A1 CAO	**[3]**
	(iii)	P(score = 111) $= P(110.5 < Y < 111.5)$ $= P\left(\frac{110.5-100}{15} < Z < \frac{111.5-100}{15}\right)$ $= P(0.7 < Z < 0.7667)$ $= \Phi(0.7667) - \Phi(0.7)$ $= 0.7784 - 0.7580$ $= 0.0204$	B1 for both continuity corrections M1 for standardising M1 for correct structure of probability calc' A1 CAO	**[4]**
	(iv)	From tables, $\Phi^{-1}(0.3) = -0.5244$, $\Phi^{-1}(0.8) = 0.8416$ $22 = \mu + 0.8416\,\sigma$ $15 = \mu - 0.5244\,\sigma$ $7 = 1.3660\,\sigma$ $\sigma = 5.124$, $\mu = 17.69$	B1 for 0.5244 or 0.8416 seen M1 for at least one equation in z, μ & σ A1 for both correct M1 for attempt to solve two appropriate equations A1 CAO for both	**[5]**
			TOTAL	**[17]**

4	(i)	H_0: no association between size of business and recycling service used. H_1: some association between size of business and recycling service used.	B1 for both	[1]
	(ii)	Expected frequency = 78/285 × 180 = 49.2632 Contribution = $(52 – 49.2632)^2$ / 49.2632 = 0.1520	M1 A1 M1 for valid attempt at $(O\text{-}E)^2/E$ A1 ***NB Answer given*** Allow 0.152	[4]
	(iii)	Test statistic X^2 = 0.6041 Refer to $\mathcal{X}_2^{\,2}$ Critical value at 5% level = 5.991 Result is not significant There is no evidence to suggest any association between size of business and recycling service used. NB if H_0 H_1 reversed, or 'correlation' mentioned in part (i), do not award B1in part (i) or E1 in part (iii).	B1 B1 for 2 deg of f(seen) B1 CAO for cv B1 for not significant E1	[5]
	(iv)	H_0: $\mu = 32.8$; H_1: $\mu < 32.8$ Where μ denotes the population mean weight of rubbish in the bins. Test statistic = $\dfrac{30.9-32.8}{3.4/\sqrt{50}} = -\dfrac{1.9}{0.4808} = -3.951$ 5% level 1 tailed critical value of z = –1.645 –3.951 < –1.645 so significant. There is sufficient evidence to reject H_0 There is evidence to suggest that the weight of rubbish in dustbins has been reduced.	B1 for use of 32.8 B1 for both correct B1 for definition of μ M1 must include √50 A1 B1 for ±1.645 M1 for sensible comparison leading to a conclusion A1 for conclusion in words in context	[8]
			TOTAL	[18]

4767 Statistics 2

General Comments

Once again a very good overall standard was seen. No single question stood out as being more difficult or more straightforward than the others, and there was no evidence that candidates were short of time in which to complete the examination. A variety of techniques for handling hypothesis tests was seen; in most cases, candidates demonstrated a decent understanding of the technique they were employing. The vast majority of candidates handled probability calculations efficiently and accurately, using appropriate probability distributions.

Comments on Individual Questions

1 Candidates were first required to draw a scatter diagram for some given data; full marks were awarded in most cases, with occasional marks lost for erroneous points or failing to label axes. Most correctly identified the independent variable and the dependent variable, but few obtained both of the marks for justifying their choice. Many recognised that *a* was in some way controlled by the pilot but few satisfactorily explained that *t* was subject to random variation - however, many gained credit for implying this without stating it explicitly. Most gained full marks for calculating the equation of the regression line *t* on *a*, using it to obtain estimates for *t* and commenting on their reliability. Odd marks were lost for writing equations in terms of *y* and *x* instead of *t* and *a*, working with too little accuracy or providing unsatisfactory comments such as 'the prediction for 800m is reliable as it lies on the line' or ' .... as it lies near the points in the scatter diagram'. Some candidates calculated the regression line *a* on *t* and were penalised quite severely although most of the remaining marks were available with appropriate working and comments. Most candidates realised that to calculate a residual they must first find an estimate for *t* then subtract it from 923, producing a negative value which indicated that the observed value was less than the predicted value; however, many showed a poor understanding of this area of the course.

2 This question was based around the binomial distribution and involved the use of suitable approximating distributions to calculate probabilities. Part (i) was well answered by those candidates realising that the binomial model was needed - frequently seen mistakes included using 0.2 for 2%, and in some cases 0.1 was used (presumably from the '1 of a batch of 10' mentioned in the question). In part (ii), most candidates scored both available marks for explaining when a Poisson distribution is appropriate as an approximation to the binomial distribution; those referring to 'the number' and 'the probability' instead of n and p were penalised as this was deemed imprecise. Part (iii) (*A*) was well answered; some candidates used 0.02 for their Poisson mean instead of 3. Part (iii) (*B*) was less well answered; P(X > 3) = 1 - P(X < 2) was seen regularly. In part (iv), most candidates managed to correctly identify the 'exact distribution' as Bin(2000, 0.02) in (*A*), and go on to use a Normal approximation to calculate the required probability in (*B*). Some candidates did not seem to understand what was meant by exact distribution and left this part blank, or wrote down the mean and variance. In (*B*), many candidates used a Normal approximation to the Poisson distribution rather than the required Normal approximation to the binomial distribution. Other common mistakes involved the lack of a continuity correction or use of the wrong continuity correction. Generally, the resulting Normal calculation was handled well. With more candidates using graphical calculators, it was not surprising to see the use of a Poisson approximation to the binomial distribution; this could lead to full marks. Candidates should note that in questions of this type they should provide evidence of their method; stating the distribution being used and giving some indication that such a calculator has been used is recommended as a minimum.

3 This question involving the use of the Normal distribution was well answered; most lost marks occurred in parts (iii) and (iv). Most candidates scored full marks in part (i) with the occasional mark lost through failure to work to a sufficient level of accuracy - either by premature rounding or by neglecting to use the difference column in the Normal probability tables. Part (ii) was well answered with few mistakes seen. In part (iii), many scored full marks, but a large number failed to realise that $P(X = 111)$ meant finding $P(110.5 < X < 111.5)$. Variations on this were seen (e.g. $P(110 < X < 112)$) and were given some credit. Part (iv) was generally well answered with most candidates able to obtain appropriate equations and solve them simultaneously. Common errors included using probabilities instead of z-values (e.g. $22 = \mu + 0.8\sigma$) or use of 1 - 0.5244 instead of -0.5244 (to give $15 = \mu + 0.4756\sigma$).

4 The first part of this question involved a Chi-squared test for association. It was pleasing to see the majority of candidates using appropriate terminology and providing sufficient explanation in their answers. Part (i) was well handled with only a few candidates mixing the null and alternative hypotheses. With the answer given in Part (ii), the onus was on the candidates to justify it by working at a sufficient level of accuracy; in many cases the candidates' working would not lead to the given answer. Candidates should be encouraged not to round expected frequencies too severely as this can have a large effect on the resulting chi-squared test statistic. Part (iii) was well answered by most candidates; however, some were unsure in their use of the phrase 'not significant' and many referred to 'two-tailed' tests or used the 97.5% value rather than the 2.5% value. In part (iv) most candidates provided the correct hypotheses but the definition of μ as the population mean was not commonly seen. Most candidates managed to successfully complete the test and received all the remaining marks. A common error in this part of the question involved treating the observed value of 30.9 as a single observation rather than the mean of a random sample of 50; however, this was seen less often than this type of mistake used to be. Some candidates provided a critical value from the product moment correlation coefficient table. Others found difficulty if they tried to calculate $P(Z < -3.951)$ as the provided tables do not cover this *z*-value; even so, those providing a sensible argument were given full credit. Other methods (e.g. confidence interval approach) were seen and could achieve full marks if handled correctly. Once again, some candidates made inappropriate comparisons such as $-3.951 < 1.645$ therefore we reject the null hypothesis, etc. but it is pleasing to note that this was seen less often than in previous years.