OCR
RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 4767

Statistics 2

Thursday **9 JUNE 2005** Morning 1 hour 30 minutes

Additional materials:
Answer booklet
Graph paper
MEI Examination Formulae and Tables (MF2)

TIME 1 hour 30 minutes

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.

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.
- Final answers should be given to a degree of accuracy appropriate to the context.
- The total number of marks for this paper is 72.

This question paper consists of 5 printed pages and 3 blank pages.

[Turn over

1 A student is collecting data on traffic arriving at a motorway service station during weekday lunchtimes. The random variable X denotes the number of cars arriving in a randomly chosen period of ten seconds.

(i) State two assumptions necessary if a Poisson distribution is to provide a suitable model for the distribution of X. Comment briefly on whether these assumptions are likely to be valid. [4]

The student counts the number of arrivals, x, in each of 100 ten-second periods. The data are shown in the table below.

x	0	1	2	3	4	5	>5
Frequency, f	18	39	20	12	8	3	0

(ii) Show that the sample mean is 1.62 and calculate the sample variance. [3]

(iii) Do your calculations in part **(ii)** support the suggestion that a Poisson distribution is a suitable model for the distribution of X? Explain your answer. [1]

For the remainder of this question you should assume that X may be modelled by a Poisson distribution with mean 1.62.

(iv) Find $\mathrm{P}(X = 2)$. Comment on your answer in relation to the data in the table. [4]

(v) Find the probability that at least ten cars arrive in a period of 50 seconds during weekday lunchtimes. [3]

(vi) Use a suitable approximating distribution to find the probability that no more than 550 cars arrive in a randomly chosen period of one hour during weekday lunchtimes. [4]

2 The fuel economy of a car varies from day to day according to weather and driving conditions. Fuel economy is measured in miles per gallon (mpg).

The fuel economy of a particular petrol-fuelled type of car is known to be Normally distributed with mean 38.5 mpg and standard deviation 4.0 mpg.

(i) Find the probability that on a randomly selected day the fuel economy of a car of this type will be above 45.0 mpg. [4]

(ii) The manufacturer wishes to quote a fuel economy figure which will be exceeded on 90% of days. What figure should be quoted? [3]

The daily fuel economy of a similar type of car which is diesel-fuelled is known to be Normally distributed with mean 51.2 mpg and unknown standard deviation σ mpg.

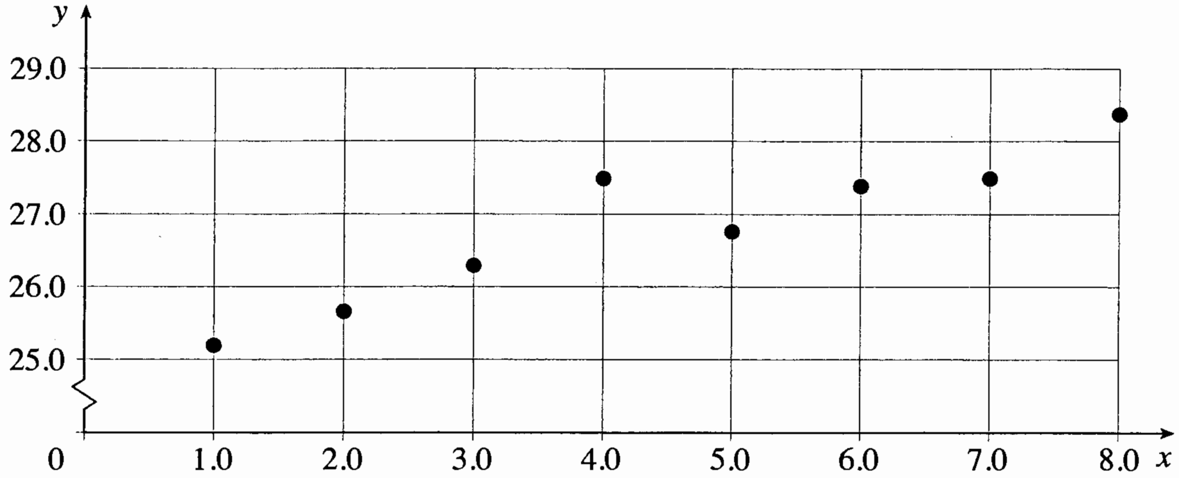
(iii) Given that on 75% of days the fuel economy of this type of car is below 55.0 mpg, show that $\sigma = 5.63$. [3]

(iv) Draw a sketch to illustrate both distributions on a single diagram. [4]

(v) Find the probability that the fuel economy of either the petrol or the diesel model (or both) will be above 45.0 mpg on a randomly selected day. You may assume that the fuel economies of the two models are independent. [4]

[Turn over

3 In a triathlon, competitors have to swim 600 metres, cycle 40 kilometres and run 10 kilometres. To improve her strength, a triathlete undertakes a training programme in which she carries weights in a rucksack whilst running. She runs a specific course and notes the total time taken for each run. Her coach is investigating the relationship between time taken and weight carried. The times taken with eight different weights are illustrated on the scatter diagram below, together with the summary statistics for these data. The variables x and y represent weight carried in kilograms and time taken in minutes respectively.



Summary statistics: $n = 8, \Sigma x = 36, \Sigma y = 214.8, \Sigma x^2 = 204, \Sigma y^2 = 5775.28, \Sigma xy = 983.6$.

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

On one of the eight runs, the triathlete was carrying 4 kilograms and took 27.5 minutes. On this run she was delayed when she tripped and fell over.

(ii) Calculate the value of the residual for this weight. [3]

(iii) The coach decides to recalculate the equation of the regression line without the data for this run. Would it be preferable to use this recalculated equation or the equation found in part **(i)** to estimate the delay when the triathlete tripped and fell over? Explain your answer. [2]

The triathlete's coach claims that there is positive correlation between cycling and swimming times in triathlons. The product moment correlation coefficient of the times of twenty randomly selected competitors in these two sections is 0.209.

(iv) Carry out a hypothesis test at the 5% level to examine the coach's claim, explaining your conclusions clearly. [5]

(v) What distributional assumption is necessary for this test to be valid? How can you use a scatter diagram to decide whether this assumption is likely to be true? [2]

4 **(a)** The selling prices of semi-detached houses in the suburbs of a particular city are known to be Normally distributed with mean £166 500 and standard deviation £14 200. A householder on one large estate claims that houses on her estate have a higher mean selling price. The selling prices of six randomly selected houses on her estate are

£180 000, £152 000, £156 500, £172 000, £189 000, £169 000.

(i) State suitable null and alternative hypotheses to test her claim. [2]

(ii) Carry out the test at the 5% level of significance, stating your conclusions clearly. You may assume that the standard deviation of the selling prices of houses on this estate is £14 200. [6]

(b) The manager of a restaurant undertakes a survey of the numbers and types of drinks ordered by a random sample of 400 customers. Customers are categorized as business, tourist or local. The drinks are categorized as alcoholic or soft drinks. A table of results of the survey is as follows.

		Type of drink		**Row totals**
		Alcoholic	Soft drinks	
Type of customer	Business	54	63	117
	Tourist	95	41	136
	Local	71	76	147
Column totals		220	180	400

Carry out a test at the 5% level of significance to examine whether there is any association between type of customer and type of drink. State carefully your null and alternative hypotheses.

[10]

Mark Scheme 4767
June 2005

GENERAL INSTRUCTIONS

Marks in the mark scheme are explicitly designated as **M**, **A**, **B**, **E** or **G**.

M marks ("method") are for an attempt to use a correct method (not merely for stating the method).

A marks ("accuracy") are for accurate answers and can only be earned if corresponding **M** mark(s) have been earned. Candidates are expected to give answers to a sensible level of accuracy in the context of the problem in hand. The level of accuracy quoted in the mark scheme will sometimes deliberately be greater than is required, when this facilitates marking.

B marks are independent of all others. They are usually awarded for a single correct answer. Typically they are available for correct quotation of points such as 1.96 from tables.

E marks ("explanation") are for explanation and/or interpretation. These will frequently be sub divisible depending on the thoroughness of the candidate's answer.

G marks ("graph") are for completing a graph or diagram correctly.

- Insert part marks in **right-hand** margin in line with the mark scheme. For fully correct parts tick the answer. For partially complete parts indicate clearly in the body of the script where the marks have been gained or lost, in line with the mark scheme.

- Please indicate incorrect working by ringing or underlining as appropriate.

- Insert total in **right-hand** margin, ringed, at end of question, in line with the mark scheme.

- Numerical answers which are not exact should be given to at least the accuracy shown. Approximate answers to a greater accuracy *may* be condoned.

- Probabilities should be given as fractions, decimals or percentages.

- FOLLOW-THROUGH MARKING SHOULD NORMALLY BE USED WHEREVER POSSIBLE. There will, however, be an occasional designation of '**c.a.o.**' for "correct answer only".

- Full credit MUST be given when correct alternative methods of solution are used. If errors occur in such methods, the marks awarded should correspond as nearly as possible to equivalent work using the method in the mark scheme.

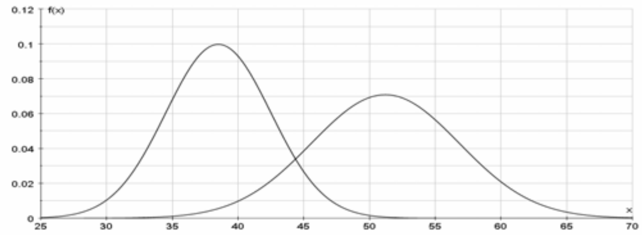
- The following notation should be used where applicable:

Question 1

(i)	Uniform average rate of occurrence; Successive arrivals are independent. Suitable arguments for/against each assumption: Eg Rate of occurrence could vary depending on the weather (any reasonable suggestion)	E1,E1 for suitable assumptions E1, E1 must be in context	**4**
(ii)	Mean $= \dfrac{\Sigma xf}{n} = \dfrac{39+40+36+32+15}{100} = \dfrac{162}{100} = 1.62$ Variance $= \dfrac{1}{n-1}\left(\Sigma fx^2 - n\bar{x}^2\right)$ $= \dfrac{1}{99}\left(430 - 100 \times 1.62^2\right) = 1.69$ (to 2 d.p.)	B1 for mean ***NB answer given*** M1 for calculation A1	**3**
(iii)	Yes, since mean is close to variance	B1FT	**1**
(iv)	$\mathrm{P}(X = 2) = \mathrm{e}^{-1.62}\dfrac{1.62^2}{2!}$ $= 0.260$ (3 s.f.) *Either:* Thus the expected number of 2's is 26 which is reasonably close to the observed value of 20. *Or*: This probability compares reasonably well with the relative frequency 0.2	M1 for probability calc. M0 for tables unless interpolated A1 B1 for expectation of 26 or r.f. of 0.2 E1	**4**
(v)	$\lambda = 5 \times 1.62 = 8.1$ Using tables: $\mathrm{P}(X \geq 10) = 1 - \mathrm{P}(X \leq 9)$ $= 1 - 0.7041 = 0.2959$	B1FT for mean (SOI) M1 for probability from using tables to find $1 - \mathrm{P}(X \leq 9)$ A1 FT	**3**
(vi)	Mean no. of items in 1 hour = 360 × 1.62 = 583.2 Using Normal approx. to the Poisson, $X \sim \mathrm{N}(583.2, 583.2)$: $\mathrm{P}(X \leq 550.5) = \mathrm{P}\left(Z \leq \dfrac{550.5 - 583.2}{\sqrt{583.2}}\right)$ $= \mathrm{P}(Z \leq -1.354) = 1 - \Phi(1.354) = 1 - 0.9121$	B1 for Normal approx. with correct parameters (SOI) B1 for continuity corr. M1 for probability	**4**

	$= 0.0879$ (3 s.f.)	using correct tail A1 CAO**, (**but FT wrong or omitted CC)	
			19

Question 2

(i)	$X \sim \mathrm{N}(38.5,16)$ $\mathrm{P}(X > 45) \;=\; \mathrm{P}\left(Z > \dfrac{45-38.5}{4}\right)$ $= \mathrm{P}(\,Z > 1.625)$ $= 1 - \Phi(1.625) \;=\; 1 - 0.9479$ $= 0.0521$ (3 s.f.) *or* 0.052 (to 2 s.f.)	M1 for standardizing A1 for 1.625 M1 for prob. with tables and correct tail A1 CAO (min 2 s.f.)	**4**
(ii)	From tables $\Phi^{-1}(\,0.90\,) = 1.282$ $\dfrac{x-38.5}{4} = -1.282$ $x = 38.5 - \;1.282 \times 4 = 33.37$ So 33.4 should be quoted	B1 for 1.282 seen M1 for equation in x and negative z-value A1 CAO	**3**
(iii)	$Y \sim \mathrm{N}(51.2,\, \sigma^2)$ From tables $\Phi^{-1}(\,0.75\,) = 0.6745$ $\dfrac{55-51.2}{\sigma} = 0.6745$ $3.8 = 0.6745\,\sigma$ $\sigma = 5.63$	B1 for 0.6745 seen M1 for equation in σ with z-value A1 ***NB answer given***	**3**
(iv)		G1 for shape G1 for means, shown explicitly or by scale G1 for lower max height in diesel G1 for higher variance in diesel	**4**
(v)	$\mathrm{P}(\text{Diesel} > 45) \;=\; \mathrm{P}\left(Z > \dfrac{45-51.2}{5.63}\right)$	M1 for prob. calc. for diesel	

	$= P(Z > -1.101) = \Phi(1.101) = 0.8646$ P(At least one over 45) = 1 – P(Both less than 45) = 1 - (1 - 0.0521) x (1 - 0.8646) = 1 - 0.9479 x 0.1354 = 0.8717 NB allow correct alternatives based on: P(D over, P under)+P(D under, P over)+ P(both over) *or* P(D over) + P(P over) – P(both over)	M1 for correct structure M1*dep* for correct probabilities A1 CAO (2 s.f. min)	**4**
			18

Question 3

(i)	$\bar{x} = 4.5,\ \ \bar{y} = 26.85$ $b = \frac{Sxy}{Sxx} = \frac{983.6 - 36 \times 214.8/8}{204 - 36^2/8} = \frac{17}{42} = 0.405$ OR $b = \frac{983.6/8 - 4.5 \times 26.85}{204/8 - 4.5^2} = \frac{2.125}{5.25} = 0.405$ hence least squares regression line is: $y - \bar{y} = b(x - \bar{x})$ $\Rightarrow\ y - 26.85 = 0.405(x - 4.5)$ $\Rightarrow\ y = 0.405x + 25.03$	B1 for $\bar{x}$ and $\bar{y}$ used (SOI) M1 for attempt at gradient (*b*) A1 for 0.405 **cao** M1 *indep* for equation of line A1FT for complete equation	**5**
(ii)	$x = 4 \Rightarrow$ predicted $y = 0.405 \times 4 + 25.03 = 26.65$ Residual = 27.5 – 26.65 = 0.85	M1 for prediction A1FT for ± 0.85 B1FT for sign (+)	**3**
(iii)	The new equation would be preferable, since the equation in part (i) is influenced by the unrepresentative point (4,27.5)	B1 E1	**2**
(iv)	H_0: $\rho = 0$; H_1: $\rho > 0$ where ρ represents the population correlation coefficient Critical value at 5% level is 0.3783 Since 0.209 < 0.3783, there is not sufficient evidence to reject H_0, i.e. there is not sufficient evidence to conclude that there is any correlation between cycling and swimming times.	B1 for H_0 and H_1 B1 for defining ρ B1 for 0.3783 M1 for comparison leading to conclusion A1*dep on cv* for conclusion in words	**5**

		in context	
(v)	Underlying distribution must be bivariate normal. The distribution of points on the scatter diagram should be approximately elliptical.	B1 E1	**2**
			17

Question 4

(a) (i)	H_0: $\mu = 166500$; H_1: $\mu > 166500$ Where μ denotes the mean selling price in pounds of the population of houses on the large estate	B1 for both correct B1 for definition of μ	**2**
(ii)	n = 6, $\Sigma x = 1018500$, $\bar{x} = £169750$ Test statistic = $\dfrac{169750 - 166500}{14200/\sqrt{6}} = \dfrac{3250}{5797}$ $= 0.5606$ 5% level 1 tailed critical value of z = 1.645 0.5606 < 1.645 so not significant. There is insufficient evidence to reject H_0 It is reasonable to conclude that houses on this estate are not more expensive than in the rest of the suburbs.	B1CAO M1 must include √6 A1FT B1 for 1.645 M1 for comparison leading to a conclusion A1 for conclusion in words in context	**6**

<table>
<tr>
<td><b>(b)</b></td>
<td>
H_0: no association between customer and drink types;
H_1: some association between customer and drink types;

<table>
<tr><th colspan="2" rowspan="2">Observed</th><th colspan="2">Type of drink</th><th rowspan="2">Row totals</th></tr>
<tr><th>Alcoholic</th><th>Soft drinks</th></tr>
<tr><td rowspan="3">Type of customer</td><td>Business</td><td>54</td><td>63</td><td>117</td></tr>
<tr><td>Tourist</td><td>95</td><td>41</td><td>136</td></tr>
<tr><td>Local</td><td>71</td><td>76</td><td>147</td></tr>
<tr><td colspan="2">Column totals</td><td>220</td><td>180</td><td>400</td></tr>
</table>

<table>
<tr><th colspan="2" rowspan="2">Expected</th><th colspan="2">Type of drink</th><th rowspan="2">Row totals</th></tr>
<tr><th>Alcoholic</th><th>Soft drinks</th></tr>
<tr><td rowspan="3">Type of customer</td><td>Business</td><td>64.35</td><td>52.65</td><td>117</td></tr>
<tr><td>Tourist</td><td>74.80</td><td>61.20</td><td>136</td></tr>
<tr><td>Local</td><td>80.85</td><td>66.15</td><td>147</td></tr>
<tr><td colspan="2">Column totals</td><td>220</td><td>180</td><td>400</td></tr>
</table>

<table>
<tr><th colspan="2" rowspan="2">Chi squared contribution</th><th colspan="2">Type of drink</th><th rowspan="2">Row totals</th></tr>
<tr><th>Alcoholic</th><th>Soft drinks</th></tr>
<tr><td rowspan="3">Type of customer</td><td>Business</td><td>1.665</td><td>2.035</td><td>3.699</td></tr>
<tr><td>Tourist</td><td>5.455</td><td>6.667</td><td>12.122</td></tr>
<tr><td>Local</td><td>1.200</td><td>1.467</td><td>2.667</td></tr>
</table>

$X^2 = 18.49$

Refer to $\mathcal{X}_2^2$
Critical value at 5% level = 5.991
Result is significant
There is some association between customer type and type of drink.
NB if H_0 H_1 reversed, or ‘correlation’ mentioned, do not award first B1or final B1 or final E1
</td>
<td>B1
M1 A1 for expected values (to 2dp)
M1 for valid attempt at $(O\text{-}E)^2/E$
M1dep for summation
A1CAO for X^2
B1 for 2 deg of f
B1 CAO for cv
B1*dep on cv*
E1</td>
<td><b>6</b>
<b>4</b></td>
</tr>
<tr><td></td><td></td><td></td><td><b>18</b></td></tr>
</table>

4767 - MEI Statistics 2

General Comments

On the whole candidates scored well on this paper, many probably being Further Maths students taking this A2 unit in Year 12. Most candidates demonstrated a good level of knowledge and understanding of all of the topics and there were many scripts in which candidates gave very good responses to all four questions. Very few candidates appeared to have been inappropriately entered for the paper. Question 4 which examined the new topics in the specification (contingency tables and the hypothesis test for the mean of a Normal distribution) was answered well, with many candidates gaining nearly full marks. Most parts of the first three questions also elicited good responses, although candidates struggled to give two valid assumptions in Question 1 part (i). Question 2 part (v), although not exceptionally demanding, did prove to be beyond the majority of candidates. Hypothesis testing was generally well done, except for a failure to define the parameter used in the hypotheses (very frequently seen) and a failure to give the final conclusion in context. It appeared that most candidates had adequate time to complete the paper, with the possible exception of a few who adopted extremely time consuming methods, such as the calculation of ten separate Poisson probabilities, rather than the use of tables in Question 1 part (v).

Comments on Individual Questions

1) (i) This standard request was the least well done part of Question 1, even by very high-scoring candidates. In this case independence (of events) and a uniform mean rate of occurrence were the correct assumptions. Many candidates quoted the former but fewer quoted the latter, sometimes instead mentioning a ‘known’ mean rate, but more often randomness or mention of large *n* and small *p* were suggested. Randomness rather than a deterministic situation is a requirement of every statistical distribution, not specifically of the Poisson. Many candidates who mentioned independence were able to make a suitable comment which indicated that they understood the meaning although this was true of less of those who mentioned the second assumption.

(ii) Most candidates scored either full marks or lost just one mark due to the use of divisor *n* rather than *n* – 1 in the sample variance. In the new specification a divisor of *n* is used in finding *msd*, not variance.

(iii) This was usually correct.

(iv) Once again this was well answered with only a few candidates rounding λ to 1.6 and then using tables, which is not acceptable. Many were able to go on to compare their result with the frequency of $x = 2$ in the table. Some candidates thought that ‘the table’ referred to the cumulative Poisson probability tables.

(v) Most candidates correctly multiplied 1.62 by 5 to find the new mean and then used tables, but at this stage a few made errors of the form P(X ≥ 10) = 1 – P(X ≤ 10). Some used entirely spurious methods, or occasionally did not use tables but instead calculated ten separate point probabilities and then subtracted their sum from one, usually making some calculation error on the way.

(vi) Most candidates realised that a Normal approximation was required and found the parameters correctly. Continuity corrections were often omitted and sometimes the wrong correction, 549.5 instead of 550.5, was used. Relatively few candidates miscalculated the parameters.

2) (i) This was well answered, with just a few candidates using variance instead of standard deviation or giving an inaccurate final answer due to premature approximation. A few of the candidates who used graphical calculator built-in probability functions did not appear to know how to do this correctly since their answer was wrong and thus they could be given no credit since there was no working shown.

(ii) Most candidates realised that an inverse Normal calculation was required, but many did not realise that a negative z-value was appropriate and so obtained a final answer which was on the wrong side of the mean. As has been stated in reports on the legacy specification 2614, candidates are advised to draw a sketch if there is any doubt in their mind as to which tail is involved. Alternatively a mental check of their final answer in relation to the value of the mean should indicate if they have made an error in the sign of z.

(iii) This was very well answered. Most candidates scored full marks, although a number lost one mark by rounding 0.6745 to 0.675, which then does not lead to the given answer. Candidates should realise that given answers are correct to the number of decimal places stated and if they get a different answer then they have made an error. Some candidates, having gained credit for a correct equation in σ, then failed to show any working whatsoever to simplify their equation and simply quoted the given value of σ.

(iv) Few fully correct sketches were seen. In some cases both curves were shown centred around the same mean, or just one curve was drawn. In other cases the means were clearly different but the standard deviations were not. However some candidates produced very clear sketches, including the more subtle point that the maximum height of the diesel curve should be lower than that of the petrol, since both areas are equal to 1.

(v) Only a small proportion of candidates answered correctly. Most started off by finding the probability that the diesel model is above 45.0, gaining one mark. However candidates then either stopped at that, multiplied by the probability for petrol, or in many cases found the sum P(diesel > 45) + P(petrol > 45) + P(both > 45), whereas P(both > 45) should of course be subtracted from the sum of the former two.

3) (i) Most candidates found the equation of the regression line correctly and many of those who made errors appeared to have made a slip rather than not knowing what to do.

(ii) In past sessions many candidates have had little knowledge of residuals. Happily this situation has improved, and the vast majority scored full marks here.

(iii) Most candidates realised and were able to explain that the recalculated equation is preferable as it excludes the result which is not representative of the triathlete's usual performance. A few felt that this was a genuine result and therefore should be included. This argument was not worthy of credit, since the result may have been genuine but is not representative.

(iv) The hypothesis test was generally done well with most candidates scoring 4 marks out of 5. However, despite regular reference to this in examiners reports for the legacy 2614, a correct definition of ρ as the 'population correlation coefficient' was very rarely seen. Pleasingly, most candidates gave the concluding statement in context, rather than simply stating that 'there is no correlation'. Few candidates thought that a two-tailed test was appropriate, although such candidates could follow through and lose just one mark.

(v) Many candidates failed to quote the required assumption of a bivariate Normal distribution. As in the legacy 2614 this failing was again often strongly linked to centres, with many centres in which no candidates gained this mark, and others where almost all did so. The fact that an elliptical scatter diagram can be used as an indication that the test is valid was better known, although by no means universally so, and again the knowledge thereof was strongly linked to centres. Following the removal of coursework from the Statistics 2 assessment, centres need to place more emphasis on ensuring that candidates learn these assumptions, given that they no longer meet them as part of their coursework.

4) (a)(i) Many hypotheses were given in words or occasionally in terms of $\bar{x}$ or ρ rather than in terms of μ as is required. Those candidates who did use μ rarely defined μ as the mean of the population (ie of all houses on the large estate) thus losing credit.

(ii) It is pleasing to report that many correct responses were seen on this new topic. The majority of candidates who were successful found the test statistic in the form of a *z*-value and then compared this to the critical *z*-value. A much smaller number compared two probabilities. However many candidates failed to divide the standard deviation by $\sqrt{6}$, thus in effect simply using the distribution of *X* and this error was heavily penalised. There was a variety of other errors, the most common of which was to calculate a probability and then compare it with a *z*-value or vice versa.

(b) Once again this new topic was generally dealt with very well. In a contingency table test of association, hypotheses should be given in words and most candidates did so, although some mentioned correlation rather than association. A few candidates had no idea how to proceed and some others knew they had to calculate expected frequencies, but not how to do so. However most knew what to do and did it correctly, with surprisingly little evidence of premature approximation. Having calculated the test statistic, most candidates went on to complete the comparison and conclusion correctly, but a few lost marks, either by making an error in the calculation of the number of degrees of freedom, or by using the wrong figure from the tables, or by making the comparison based on correct figures, but coming to the wrong conclusion. Some candidates failed to give their result in context.