OCR
RECOGNISING ACHIEVEMENT

ADVANCED SUBSIDIARY GCE

MEI STATISTICS

G241

Statistics 1 (Z1)

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 15 June 2009
Afternoon

Duration: 1 hour 30 minutes

* G 2 4 1 *

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 **8** pages. Any blank pages are indicated.

Section A (36 marks)

1 In a traffic survey, the number of people in each car passing the survey point is recorded. The results are given in the following frequency table.

Number of people	1	2	3	4
Frequency	50	31	16	5

(i) Write down the median and mode of these data. **[2]**

(ii) Draw a vertical line diagram for these data. **[2]**

(iii) State the type of skewness of the distribution. **[1]**

2 There are 14 girls and 11 boys in a class. A quiz team of 5 students is to be chosen from the class.

(i) How many different teams are possible? **[2]**

(ii) If the team must include 3 girls and 2 boys, find how many different teams are possible. **[3]**

3 Dwayne is a car salesman. The numbers of cars, x, sold by Dwayne each month during the year 2008 are summarised by

$$n = 12, \qquad \Sigma x = 126, \qquad \Sigma x^2 = 1582.$$

(i) Calculate the mean and standard deviation of the monthly numbers of cars sold. **[3]**

(ii) Dwayne earns £500 each month plus £100 commission for each car sold. Show that the mean of Dwayne's monthly earnings is £1550. Find the standard deviation of Dwayne's monthly earnings. **[3]**

(iii) Marlene is a car saleswoman and is paid in the same way as Dwayne. During 2008 her monthly earnings have mean £1625 and standard deviation £280. Briefly compare the monthly numbers of cars sold by Marlene and Dwayne during 2008. **[2]**

4 The table shows the probability distribution of the random variable X.

r	10	20	30	40
$\mathrm{P}(X = r)$	0.2	0.3	0.3	0.2

(i) Explain why $\mathrm{E}(X) = 25$. **[1]**

(ii) Calculate $\mathrm{Var}(X)$. **[3]**

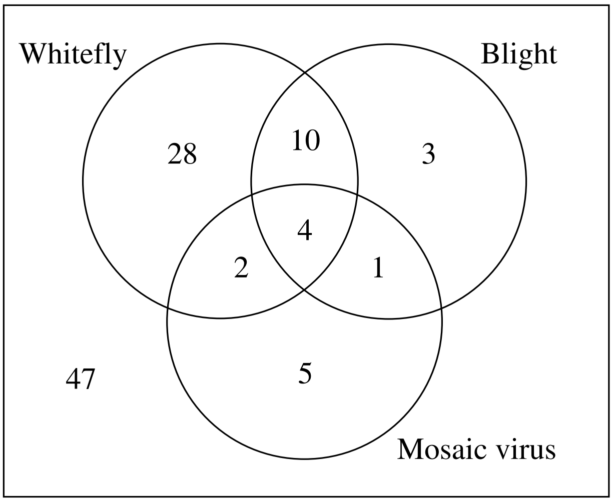
5 The frequency table below shows the distance travelled by 1200 visitors to a particular UK tourist destination in August 2008.

Distance (d miles)	$0 \leqslant d < 50$	$50 \leqslant d < 100$	$100 \leqslant d < 200$	$200 \leqslant d < 400$
Frequency	360	400	307	133

(i) Draw a histogram on graph paper to illustrate these data. **[5]**

(ii) Calculate an estimate of the median distance. **[3]**

6 Whitefly, blight and mosaic virus are three problems which can affect tomato plants. 100 tomato plants are examined for these problems. The numbers of plants with each type of problem are shown in the Venn diagram. 47 of the plants have none of the problems.



(i) One of the 100 plants is selected at random. Find the probability that this plant has

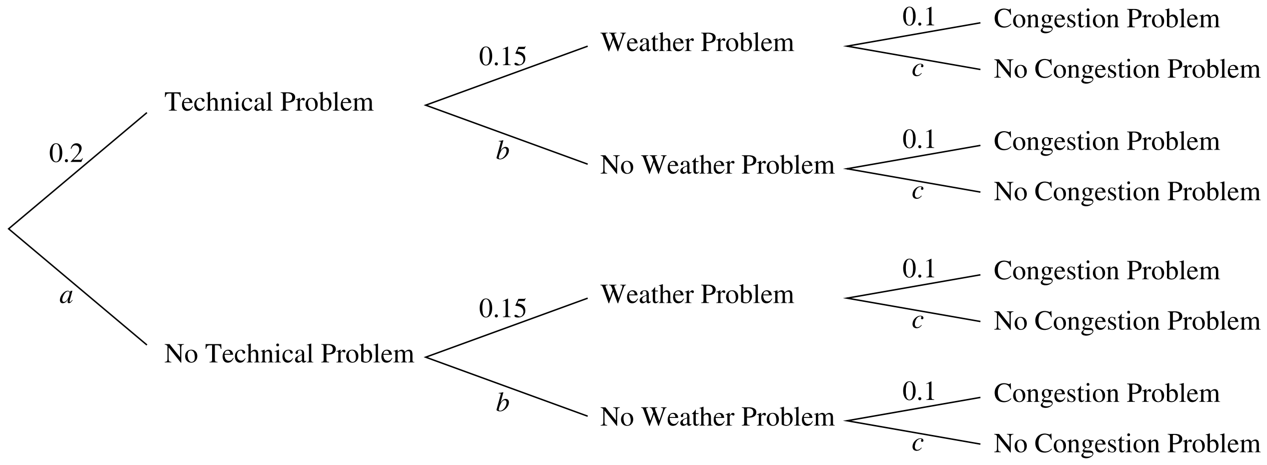
(A) at most one of the problems, **[1]**

(B) exactly two of the problems. **[2]**

(ii) Three of the 100 plants are selected at random. Find the probability that all of them have at least one of the problems. **[3]**

Section B (36 marks)

7 Laura frequently flies to business meetings and often finds that her flights are delayed. A flight may be delayed due to technical problems, weather problems or congestion problems, with probabilities 0.2, 0.15 and 0.1 respectively. The tree diagram shows this information.



(i) Write down the values of the probabilities *a*, *b* and *c* shown in the tree diagram. **[2]**

One of Laura's flights is selected at random.

(ii) Find the probability that Laura's flight is not delayed and hence write down the probability that it is delayed. **[4]**

(iii) Find the probability that Laura's flight is delayed due to just one of the three problems. **[4]**

(iv) Given that Laura's flight is delayed, find the probability that the delay is due to just one of the three problems. **[3]**

(v) Given that Laura's flight has no technical problems, find the probability that it is delayed. **[3]**

(vi) In a particular year, Laura has 110 flights. Find the expected number of flights that are delayed. **[2]**

8 The Department of Health 'eat five a day' advice recommends that people should eat at least five portions of fruit and vegetables per day. In a particular school, 20% of pupils eat at least five a day.

(i) 15 children are selected at random.

(*A*) Find the probability that exactly 3 of them eat at least five a day. **[3]**

(*B*) Find the probability that at least 3 of them eat at least five a day. **[3]**

(*C*) Find the expected number who eat at least five a day. **[2]**

A programme is introduced to encourage children to eat more portions of fruit and vegetables per day. At the end of this programme, the diets of a random sample of 15 children are analysed. A hypothesis test is carried out to examine whether the proportion of children in the school who eat at least five a day has increased.

(ii) (*A*) Write down suitable null and alternative hypotheses for the test.

(*B*) Give a reason for your choice of the alternative hypothesis. **[4]**

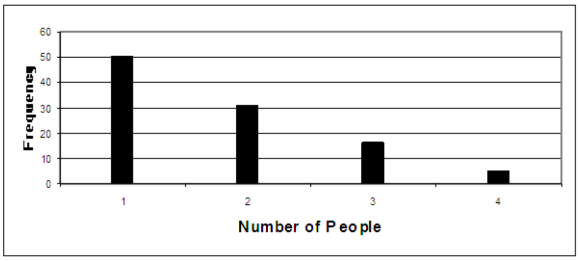
(iii) Find the critical region for the test at the 10% significance level, showing all of your calculations. Hence complete the test, given that 7 of the 15 children eat at least five a day. **[6]**

BLANK PAGE

BLANK PAGE

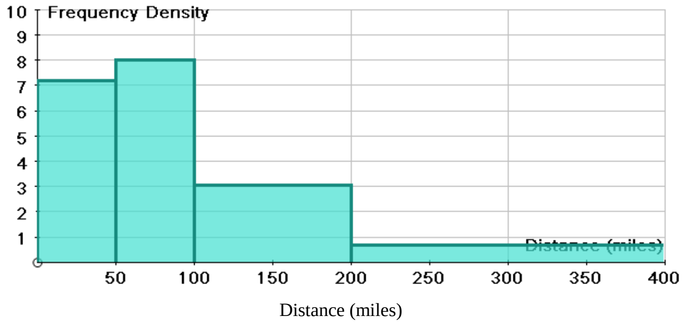
G241 Statistics 1

Q1 (i)	Median = 2
	Mode = 1
		B1 CAO
		B1 CAO
			2

(ii)		S1 labelled linear scales on both axes H1 heights	**2**
(iii)	Positive	B1	**1**
		TOTAL	**5**
Q2 (i)	$\binom{25}{5}$ different teams = 53130	M1 for $\binom{25}{5}$ A1 CAO	**2**
(ii)	$\binom{14}{3}\times\binom{11}{2} = 364\times 55 = 20020$	M1 for either combination M1 for product of both A1 CAO	**3**
		TOTAL	**5**

Q3 (i)	Mean = $\frac{126}{12}$ = 10.5
	Sxx = $1582 - \frac{126^2}{12}$ = 259
	s = $\sqrt{\frac{259}{11}}$ = 4.85
		B1 for mean
		M1 for attempt at *Sxx*
		A1 CAO
			3

(ii)	New mean = 500 + 100 × 10.5 = 1550 New s = 100 × 4.85 = 485	B1 ANSWER GIVEN M1A1FT	**3**
(iii)	On average Marlene sells more cars than Dwayne. Marlene has less variation in monthly sales than Dwayne.	E1 E1FT	**2**
		TOTAL	**8**

<table>
<tr><td><b>Q4 (i)</b></td><td>E(X) = 25 because the distribution is symmetrical.
Allow correct calculation of Σrp</td><td>E1 <u>ANSWER GIVEN</u></td><td><b>1</b></td></tr>
<tr><td><b>(ii)</b></td><td>E(X^2) = $10^2 \times 0.2 + 20^2 \times 0.3 + 30^2 \times 0.3 + 40^2 \times 0.2$=730
Var(X) = 730 – 25^2 = 105</td><td>M1 for $\Sigma r^2 p$ (at least 3 terms correct)
M1dep for – 25^2
A1 CAO</td><td><b>3</b></td></tr>
<tr><td></td><td></td><td><b>TOTAL</b></td><td><b>4</b></td></tr>
<tr><td><b>Q5 (i)</b></td><td>
<table>
<tr><th>Distance</th><th>freq</th><th>width</th><th>f dens</th></tr>
<tr><td>0-</td><td>360</td><td>50</td><td>7.200</td></tr>
<tr><td>50-</td><td>400</td><td>50</td><td>8.000</td></tr>
<tr><td>100-</td><td>307</td><td>100</td><td>3.070</td></tr>
<tr><td>200-400</td><td>133</td><td>200</td><td>0.665</td></tr>
</table>


</td><td>M1 for fds
A1 CAO
Accept any suitable unit for fd such as eg freq per 50 miles.
L1 linear scales on both axes and label
W1 width of bars
H1 height of bars</td><td><b>5</b></td></tr>
<tr><td><b>(ii)</b></td><td>Median = 600th distance
Estimate = 50 + $^{240}/_{400}$ × 50 = 50 + 30 = 80</td><td>B1 for 600th
M1 for attempt to interpolate
A1 CAO</td><td><b>3</b></td></tr>
<tr><td></td><td></td><td><b>TOTAL</b></td><td><b>8</b></td></tr>
<tr><td><b>Q6 (i)</b></td><td>(<i>A</i>) P(at most one) $= \frac{83}{100} = 0.83$
(<i>B</i>) P(exactly two) $= \frac{10+2+1}{100} = \frac{13}{100} = 0.13$</td><td>B1 aef
M1 for (10+2+1)/100
A1 aef</td><td><b>1</b>
<b>2</b></td></tr>
<tr><td><b>(ii)</b></td><td>P(all at least one) $= \frac{53}{100} \times \frac{52}{99} \times \frac{51}{98} = \frac{140556}{970200} = 0.145$</td><td>M1 for $\frac{53}{100} \times$
M1<i>dep</i> for product of next 2 correct fractions
A1 CAO</td><td><b>3</b></td></tr>
<tr><td></td><td></td><td><b>TOTAL</b></td><td><b>6</b></td></tr>
</table>

Q7 (i)	$a = 0.8$, $b = 0.85$, $c = 0.9$.	B1 for any one B1 for the other two	**2**
(ii)	P(Not delayed) = $0.8 \times 0.85 \times 0.9 = 0.612$ P(Delayed) = $1 - 0.8 \times 0.85 \times 0.9 = 1 - 0.612 = 0.388$	M1 for product A1 CAO M1 for 1 – P(delayed) A1FT	**4**
(iii)	P(just one problem) $= 0.2\times0.85\times0.9 + 0.8\times0.15\times0.9 + 0.8\times0.85\times0.1$ $= 0.153 + 0.108 + 0.068 = 0.329$	B1 one product correct M1 three products M1 sum of 3 products A1 CAO	**4**
(iv)	P(Just one problem | delay) $= \dfrac{\text{P(Just one problem and delay)}}{\text{P(Delay)}} = \dfrac{0.329}{0.388} = 0.848$	M1 for numerator M1 for denominator A1FT	**3**
(v)	P(Delayed | No technical problems) *Either* $= 0.15 + 0.85 \times 0.1 = 0.235$ *Or* $= 1 - 0.9 \times 0.85 = 1 - 0.765 = 0.235$ *Or* $= 0.15 \times 0.1 + 0.15 \times 0.9 + 0.85 \times 0.1 = 0.235$ *Or (using conditional probability formula)* $\dfrac{\text{P(Delayed and no technical problems)}}{\text{P(No technical problems)}}$ $= \dfrac{0.8\times0.15\times0.1+0.8\times0.15\times0.9+0.8\times0.85\times0.1}{0.8}$ $= \dfrac{0.188}{0.8} = 0.235$	M1 for 0.15 + M1 for second term A1CAO M1 for product M1 for 1 – product A1CAO M1 for all 3 products M1 for sum of all 3 products A1CAO M1 for numerator M1 for denominator A1CAO	**3**
(vi)	Expected number = $110 \times 0.388 = 42.7$	M1 for product A1FT	**2**
		TOTAL	**18**

Q8 (i)	X ~ B(15, 0.2) (***A***) $P(\boldsymbol{X} = 3) = \binom{15}{3} \times 0.2^3 \times 0.8^{12} = 0.2501$ OR from tables $0.6482 - 0.3980 = 0.2502$	M1 $0.2^3 \times 0.8^{12}$ M1 $\binom{15}{3} \times p^3 q^{12}$ A1 CAO OR: M2 for 0.6482 – 0.3980 A1 CAO	**3**
	(***B***) $P(\boldsymbol{X} \geq 3) = 1 - 0.3980 = 0.6020$	M1 $P(X \leq 2)$ M1 $1 - P(X \leq 2)$ A1 CAO	**3**
	(***C***) $E(X) = np = 15 \times 0.2 = 3.0$	M1 for product A1 CAO	**2**
(ii)	(*A*) Let *p* = probability of a randomly selected child eating at least 5 a day H_0: $p = 0.2$ H_1: $p > 0.2$ (*B*) H_1 has this form as the proportion who eat at least 5 a day is expected to increase.	B1 for definition of *p* in context B1 for H_0 B1 for H_1 E1	**4**
(iii)	Let X ~ B(15, 0.2) $P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.8358 = 0.1642 > 10\%$ $P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.9389 = 0.0611 < 10\%$ So critical region is {6,7,8,9,10,11,12,13,14,15}	B1 for 0.1642 B1 for 0.0611 M1 for at least one comparison with 10% A1 CAO for critical region *dep* on M1 and at least one B1	**6**
	7 lies in the critical region, so we reject null hypothesis and we conclude that there is evidence to suggest that the proportion who eat at least five a day has increased.	M1 *dep* for comparison A1 *dep* for decision and conclusion **in context**	
		TOTAL	**18**

G241 Statistics 1

General Comments

The level of difficulty of the paper appeared to be entirely appropriate for the candidates. The more able candidates scored heavily on all questions and the weaker candidates often picked up some marks on all questions with question 7 on probability contributing significantly to their total.

Most candidates supported their numerical answers with appropriate explanations and working although some rounding errors were noted. The possible exception was in question 8 where the procedure for distinguishing between hypotheses did not always include specific comparisons with 10% and where the construction of the critical region was often sketchy. There was a surprising inability to use the given numerical data in question 3 to find the standard deviation.

Weaker candidates often scored a significant proportion of their marks from question 1, the first three parts of the probability question (question 7) and from the initial parts of question 8. Amongst lower scoring candidates, there was evidence of the use of point probabilities in question 8. Also in question 8, many candidates are still not meeting the requirement to define p in words.

There seemed to be no trouble in completing the paper within the time allowed and no obvious misinterpretations of the rubric, although a very small number of candidates ignored the instruction to use graph paper for the histogram. It would also be very helpful if candidates could write down the question numbers on the front of the question paper.

Comments on Individual Questions

1) (i) The mode was usually correct, and most candidates also found the median correctly. However some candidates quoted locations rather than actual values and others thought that the median was 1 or 1.5. There were occasional errors such as thinking that there was a total of 180 (using $\sum fx$) rather than 102 cars in the survey. Some weaker candidates found the mean instead of the median.

   (ii) Most line diagrams were correct although a small number joined the lines in one manner or another. Some others forgot to label at least one of their axes.

   (iii) The majority identified the positive skewness of the distribution, but a significant number of candidates thought that the skewness was negative.

2) (i) Many totally correct answers were seen although candidates occasionally tried to use permutations.

   (ii) This part was rather less well answered, although a good number of fully correct answers were seen. The most common error was the use of addition instead of multiplication giving ${}^{14}C_3 + {}^{11}C_2$ and this occurred very frequently.

3) (i) Almost all candidates found the mean, but a large number of candidates did not know the formula for finding the standard deviation. Those who knew how to find S_{xx} usually went on to complete part (i) successfully. However there were many incorrect attempts at S_{xx} with common variations including $1582 – 10.5^2$ or $1582 – 12×10.5$. Others gave the standard deviation as 1582/11 or $\sqrt{(1582/11)}$ and some had no idea what to do with the

numbers they were given. Rather fewer candidates than in recent sessions divided by 12 rather than 11 and found the RMSD.

(ii) Almost all showed that Dwayne's monthly earnings were £1550. However the majority of candidates did not realize that all they needed to do was to multiply their standard deviation by 100, but instead tried to start again in finding the new standard deviation, almost always without success. Of those that did multiply by 100, a few then could not also resist the addition of 500.

(iii) The explanation regarding the means was usually correct but that for the standard deviations was either ignored or candidates failed to explain in context. A lack of context in explaining the means was condoned, but not in the case of the standard deviations.

4) (i) Almost all candidates correctly explained why E(X) = 25, although hardly any commented on the symmetry of the distribution, but instead calculated $\Sigma xp(x)$.

(ii) Very many correct answers were seen. Some candidates just found E(X^2), failing to subtract 25^2 to find the variance, and occasionally candidates found the correct answer but then went on to do further calculations. Several candidates tried to work out $10 \times 0.2^2 + 20 \times 0.3^2 +$…., ie squaring the probabilities rather than the x values.

5) (i) Although a number of fully correct histograms were seen, there were also many errors. Candidates should always draw a table to show the frequency densities even if such a table is not specifically asked for in the question. Common errors included a simple frequency diagram, frequency ÷ midpoint, frequency × class width, vertical axis not labelled correctly, 3.07 plotted as 3.7 and more rarely 0.665 plotted as 0.0665. The label on the vertical axis of the histogram was not always in agreement with the bars drawn; for example bars drawn at 360, 400, 153.5 and 33.25 were described as frequency density rather than frequency per 50 miles or sometimes as both. A horizontal scale consisting of inequalities was another common error.

(ii) In estimating the median, many candidates identified that the median was the 600th value or 600½ th value and then identified the correct interval from 50 to 100 but usually gave an answer of 75 rather than attempting any interpolation. Some got as far as 30 but then forgot to add on the 50. In many centres not a single candidate attempted interpolation, suggesting that this is a topic which centres should pay attention to. A small number decided to estimate a mean distance instead.

6) (i) (A) Marks scored on this question were surprisingly low. Errors of 0.36 (plants with one problem only) or 0.53 were very frequent in this part.

(i) (B) The correct answer of 0.13 was frequently seen. There was also a wide variety of incorrect answers, perhaps 0.17 being the most common of these.

(ii) Many candidates (including a significant number of very high scoring candidates) treated part ii) as if it were "with replacement" giving an answer of 0.53^3. Another less but fairly frequent wrong answer was $1 – 0.47^3$. A small number interpreted it as being a binomial distribution of 100 trials.

7) (i) Almost all candidates answered this correctly.

(ii) Most candidates answered this correctly but some candidates chose to find P(delayed) first, meaning that lengthy calculations were needed.

(iii) Once again this was usually answered correctly.

(iv) Many candidates struggled with the conditional probability, making a variety of errors, including (0.329 × 0.388)/0.388, 0.388/0.329 or just 0.329.

(v) Many candidates attempted to use a conditional probability approach to this part, but then the majority of these gave answers such as 0.388/0.8, 0.176/0.8 or 0.235/0.8, rather than the correct 0.188/0.8 = 0.235. A good proportion of candidates calculated just the numerator (0.188) or miscalculated it as 0.176 (missing the triplet 0.8 × 0.15 × 0.1). Very few realized the direct methods available such as 1 – 0.9 × 0.85 = 0.235.

(vi) This was very well answered although candidates usually rounded their answer to 43. On this occasion this error was not penalized. A few candidates miscalculated 110 × 0.388 as 38.8 rather than 42.68.

8) (i) (A) This was usually answered correctly either by calculation or tables, with direct calculation being the more popular method..
(B) Again this was usually answered correctly, but some candidates made things difficult for themselves by calculating point probabilities and then either forgetting P(0) or including P(3) and with varying degrees of accuracy. Some used tables incorrectly finding $1 - P(X \leq 3)$, rather than $1 - P(X \leq 2)$.
(C) Once again this was usually correct but occasionally the mode was found rather than the expected number.

(ii) Many candidates correctly stated their hypotheses in symbolic form. However, much use of incorrect notation was also seen. The required notation is clearly given in the mark scheme and candidates should be trained to use this, leading to a straightforward two marks. Many candidates still do not realise the need to define the parameter ‘p’ and thus they lose a third mark, even if they have stated their hypotheses correctly. The reason for the form of the alternative hypothesis was not always well explained in context.

(iii) Some Centres do not seem to have taught how to find a critical region and candidates from such Centres often ignored the request for the critical region and went straight to the hypothesis test. Of those who did try to find the critical region, many made errors, including omission of probabilities, failure to compare the probabilities with 10%, confusion between $P(X \geq n)$ and $P(X > n)$, and even in a surprisingly large number of cases an attempt to do a two–tailed test despite having stated the correct alternative hypothesis. There are still a considerable number of candidates who attempt to use point probabilities for a hypothesis test. Although it is given in the mark scheme, it is worth repeating here the recommended method for comparing the probabilities with the significance level. Candidates should find the two upper tail (in this case) cumulative probabilities which straddle the significance level.
$P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.8358$ or $0.1642 > 10\%$
$P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.9389$ or $0.0611 < 10\%$
Irrespective of whether their critical region was correct, many candidates declined to use that information, but instead started again with $P(X \geq 7) = 0.0181 < 10\%$ and tackled the hypothesis test by that method. Those who did use their critical region sometimes did not make it clear that ‘7 lies in the critical region’. Candidates should also be advised that it is necessary not only to make a decision but give a conclusion in context.