

**ADVANCED SUBSIDIARY GCE UNIT
MEI STATISTICS**

G241/01

Statistics 1 (Z1)

TUESDAY 5 JUNE 2007

Afternoon

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.
- 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.

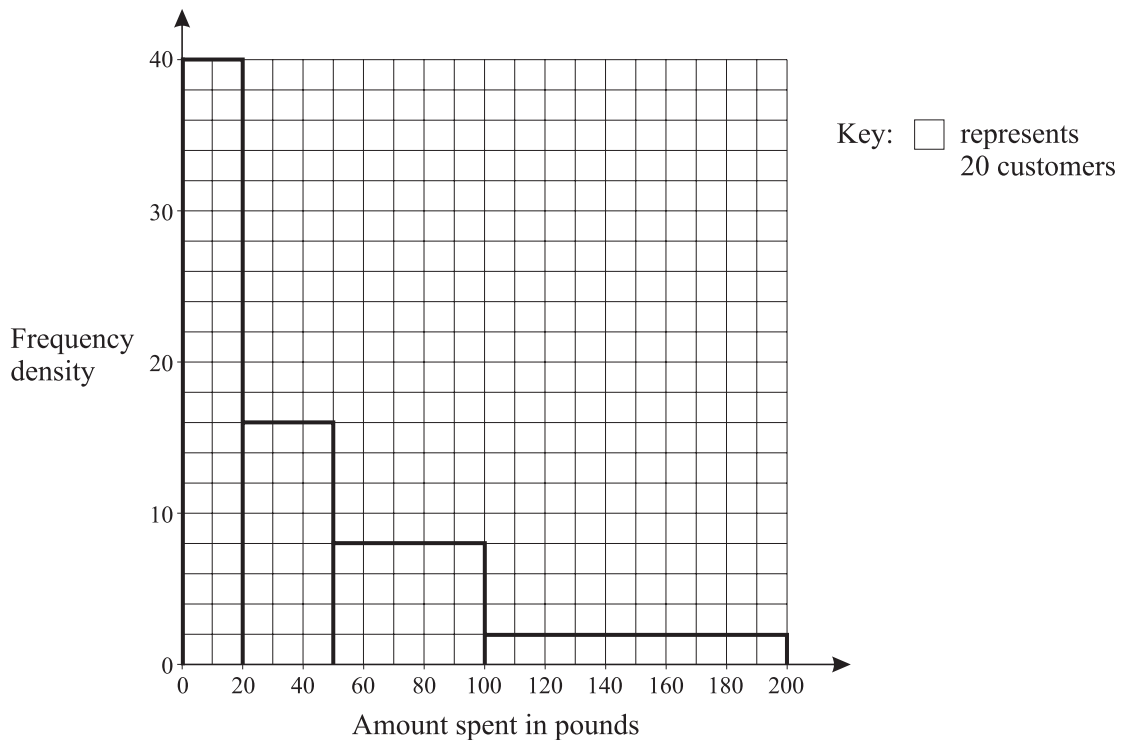
ADVICE TO CANDIDATES

- Read each question carefully and make sure you know what you have to do before starting your answer.
- 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 **7** printed pages and **1** blank page.

Section A (36 marks)

- 1 A girl is choosing tracks from an album to play at her birthday party. The album has 8 tracks and she selects 4 of them.
- (i) In how many ways can she select the 4 tracks? [2]
- (ii) In how many different orders can she arrange the 4 tracks once she has chosen them? [1]
- 2 The histogram shows the amount of money, in pounds, spent by the customers at a supermarket on a particular day.

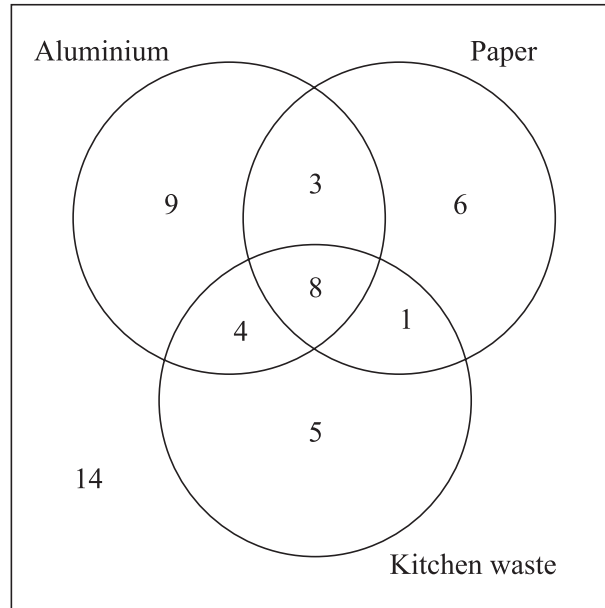


- (i) Express the data in the form of a grouped frequency table. [2]
- (ii) Use your table to estimate the total amount of money spent by customers on that day. [2]
- 3 The marks x scored by a sample of 56 students in an examination are summarised by

$$n = 56, \quad \Sigma x = 3026, \quad \Sigma x^2 = 178\,890.$$

- (i) Calculate the mean and standard deviation of the marks. [3]
- (ii) The highest mark scored by any of the 56 students in the examination was 93. Show that this result may be considered to be an outlier. [2]
- (iii) The formula $y = 1.2x - 10$ is used to scale the marks. Find the mean and standard deviation of the scaled marks. [3]

- 4 A local council has introduced a recycling scheme for aluminium, paper and kitchen waste. 50 residents are asked which of these materials they recycle. The numbers of people who recycle each type of material are shown in the Venn diagram.



One of the residents is selected at random.

- (i) Find the probability that this resident recycles

(A) at least one of the materials, [1]

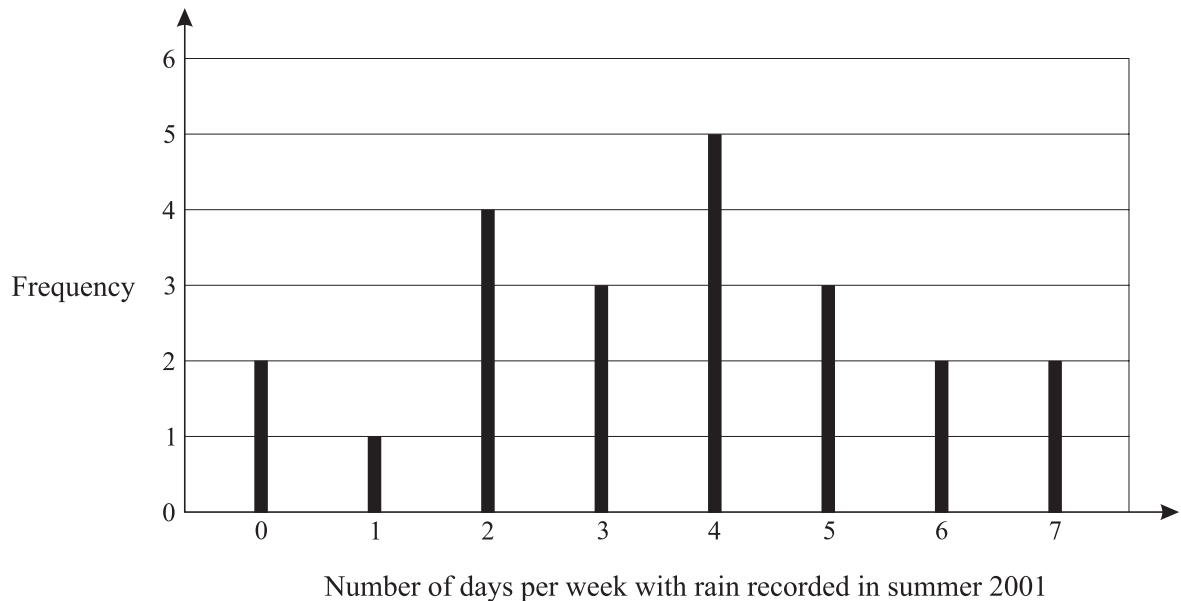
(B) exactly one of the materials. [2]

- (ii) Given that the resident recycles aluminium, find the probability that this resident does not recycle paper. [2]

Two residents are selected at random.

- (iii) Find the probability that exactly one of them recycles kitchen waste. [3]

- 5 A GCSE geography student is investigating a claim that global warming is causing summers in Britain to have more rainfall. He collects rainfall data from a local weather station for 2001 and 2006. The vertical line chart shows the number of days per week on which some rainfall was recorded during the 22 weeks of summer 2001.



- (i) Show that the median of the data is 4, and find the interquartile range. [3]
- (ii) For summer 2006 the median is 3 and the interquartile range is also 3. The student concludes that the data demonstrate that global warming is causing summer rainfall to decrease rather than increase. Is this a valid conclusion from the data? Give two brief reasons to justify your answer. [3]
- 6 In a phone-in competition run by a local radio station, listeners are given the names of 7 local personalities and are told that 4 of them are in the studio. Competitors phone in and guess which 4 are in the studio.

- (i) Show that the probability that a randomly selected competitor guesses all 4 correctly is $\frac{1}{35}$. [2]

Let X represent the number of correct guesses made by a randomly selected competitor. The probability distribution of X is shown in the table.

r	0	1	2	3	4
$P(X = r)$	0	$\frac{4}{35}$	$\frac{18}{35}$	$\frac{12}{35}$	$\frac{1}{35}$

- (ii) Find the expectation and variance of X . [5]

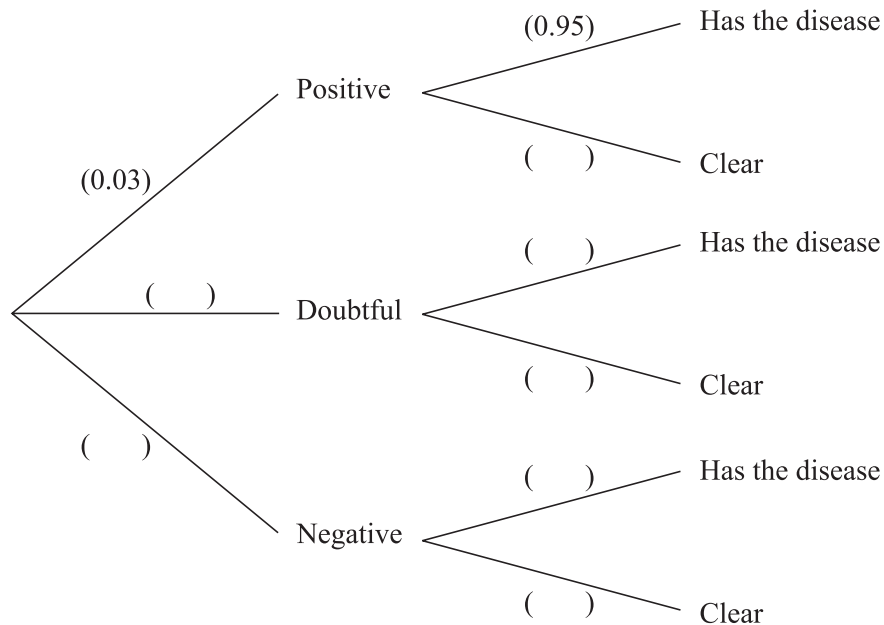
Section B (36 marks)

- 7 A screening test for a particular disease is applied to everyone in a large population. The test classifies people into three groups: 'positive', 'doubtful' and 'negative'. Of the population, 3% is classified as positive, 6% as doubtful and the rest negative.

In fact, of the people who test positive, only 95% have the disease. Of the people who test doubtful, 10% have the disease. Of the people who test negative, 1% actually have the disease.

People who do not have the disease are described as 'clear'.

- (i) Copy and complete the tree diagram to show this information. [4]



- (ii) Find the probability that a randomly selected person tests negative and is clear. [2]
- (iii) Find the probability that a randomly selected person has the disease. [3]
- (iv) Find the probability that a randomly selected person tests negative **given** that the person has the disease. [3]
- (v) Comment briefly on what your answer to part (iv) indicates about the effectiveness of the screening test. [2]

Once the test has been carried out, those people who test doubtful are given a detailed medical examination. If a person has the disease the examination will correctly identify this in 98% of cases. If a person is clear, the examination will always correctly identify this.

- (vi) A person is selected at random. Find the probability that this person either tests negative originally or tests doubtful and is then cleared in the detailed medical examination. [4]

- 8** A multinational accountancy firm receives a large number of job applications from graduates each year. On average 20% of applicants are successful.

A researcher in the human resources department of the firm selects a random sample of 17 graduate applicants.

- (i) Find the probability that at least 4 of the 17 applicants are successful. [3]
- (ii) Find the expected number of successful applicants in the sample. [2]
- (iii) Find the most likely number of successful applicants in the sample, justifying your answer. [3]

It is suggested that mathematics graduates are more likely to be successful than those from other fields. In order to test this suggestion, the researcher decides to select a new random sample of 17 mathematics graduate applicants. The researcher then carries out a hypothesis test at the 5% significance level.

- (iv) (A) Write down suitable null and alternative hypotheses for the test.
(B) Give a reason for your choice of the alternative hypothesis. [4]
- (v) Find the critical region for the test at the 5% level, showing all of your calculations. [4]
- (vi) Explain why the critical region found in part (v) would be unaltered if a 10% significance level were used. [2]

BLANK PAGE

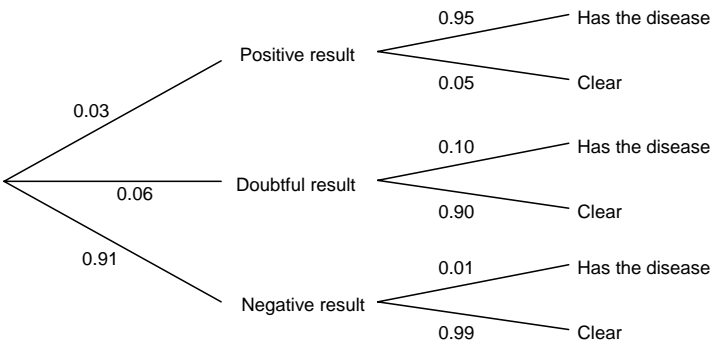
Permission to reproduce items where third-party owned material protected by copyright is included has been sought and cleared where possible. Every reasonable effort has been made by the publisher (OCR) to trace copyright holders, but if any items requiring clearance have unwittingly been included, the publisher will be pleased to make amends at the earliest possible opportunity.

OCR is part of the Cambridge Assessment Group. Cambridge Assessment is the brand name of University of Cambridge Local Examinations Syndicate (UCLES), which is itself a department of the University of Cambridge.

Mark Scheme 4766
June 2007

Q1 (i)	$\binom{8}{4}$ ways to select = 70	M1 for $\binom{8}{4}$ A1 CAO	2										
(ii)	$4! = 24$	B1 CAO	1										
		TOTAL	3										
Q2 (i)	<table border="1"> <thead> <tr> <th>Amount</th> <th>0- <20</th> <th>20- <50</th> <th>50- <100</th> <th>100- <200</th> </tr> </thead> <tbody> <tr> <td>Frequency</td> <td>800</td> <td>480</td> <td>400</td> <td>200</td> </tr> </tbody> </table>	Amount	0- <20	20- <50	50- <100	100- <200	Frequency	800	480	400	200	B1 for amounts B1 for frequencies	2
Amount	0- <20	20- <50	50- <100	100- <200									
Frequency	800	480	400	200									
(ii)	Total \approx $10 \times 800 + 35 \times 480 + 75 \times 400 + 150 \times 200 = \text{£}84800$	M1 for their midpoints \times their frequencies A1 CAO	2										
		TOTAL	4										
Q3 (i)	Mean = $\frac{3026}{56} = 54.0$ $S_{xx} = 178890 - \frac{3026^2}{56} = 15378$ $s = \sqrt{\frac{15378}{55}} = 16.7$	B1 for mean M1 for attempt at S_{xx} A1 CAO	3										
(ii)	$\bar{x} + 2s = 54.0 + 2 \times 16.7 = 87.4$ So 93 is an outlier	M1 for their $\bar{x} + 2 \times$ their s A1 FT for 87.4 and comment	2										
(iii)	New mean = $1.2 \times 54.0 - 10 = 54.8$ New $s = 1.2 \times 16.7 = 20.1$	B1 FT M1A1 FT	3										
		TOTAL	8										
Q4 (i)	(A) $P(\text{at least one}) = \frac{36}{50} = \frac{18}{25} = 0.72$ (B) $P(\text{exactly one}) = \frac{9+6+5}{50} = \frac{20}{50} = \frac{2}{5} = 0.4$	B1 aef M1 for $(9+6+5)/50$ A1 aef	3										
(ii)	$P(\text{not paper} \mid \text{aluminium}) = \frac{13}{24}$	M1 for denominator 24 or $24/50$ or 0.48 A1 CAO	2										
(iii)	$P(\text{one kitchen waste}) = 2 \times \frac{18}{50} \times \frac{32}{49} = \frac{576}{1225} = 0.470$	M1 for both fractions M1 for $2 \times$ product of both, or sum of 2 pairs A1	3										
		TOTAL	8										

Q5 (i)	11 th value is 4, 12 th value is 4 so median is 4 Interquartile range = 5 – 2 = 3	B1 M1 for either quartile A1 CAO	3
(ii)	No, not valid any two valid reasons such as : <ul style="list-style-type: none"> the sample is only for two years, which may not be representative the data only refer to the local area, not the whole of Britain even if decreasing it may have nothing to do with global warming more days with rain does not imply more total rainfall a five year timescale may not be enough to show a long term trend 	B1 E1 E1	3
		TOTAL	6
Q6 (i)	Either $P(\text{all 4 correct}) = \frac{4}{7} \times \frac{3}{6} \times \frac{2}{5} \times \frac{1}{4} = \frac{1}{35}$ or $P(\text{all 4 correct}) = \frac{1}{{}^7C_4} = \frac{1}{35}$	M1 for fractions, or 7C_4 seen A1 NB answer given	2
(ii)	$E(X) = 1 \times \frac{4}{35} + 2 \times \frac{18}{35} + 3 \times \frac{12}{35} + 4 \times \frac{1}{35} = \frac{80}{35} = 2\frac{2}{7} = 2.29$ $E(X^2) = 1 \times \frac{4}{35} + 4 \times \frac{18}{35} + 9 \times \frac{12}{35} + 16 \times \frac{1}{35} = \frac{200}{35} = 5.714$ $\text{Var}(X) = \frac{200}{35} - \left(\frac{80}{35}\right)^2 = \frac{24}{49} = 0.490 \text{ (to 3 s.f.)}$	M1 for $\sum rp$ (at least 3 terms correct) A1 CAO M1 for $\sum x^2p$ (at least 3 terms correct) M1dep for – their $E(X)^2$ A1 FT their $E(X)$ provided $\text{Var}(X) > 0$	5
		TOTAL	7

Section B			
Q7 (i)		<p>G1 probabilities of result</p> <p>G1 probabilities of disease</p> <p>G1 probabilities of clear</p> <p>G1 labels</p>	4
(ii)	$P(\text{negative and clear}) = 0.91 \times 0.99$ $= 0.9009$	<p>M1 for their 0.91×0.99</p> <p>A1 CAO</p>	2
(iii)	$P(\text{has disease}) = 0.03 \times 0.95 + 0.06 \times 0.10 + 0.91 \times 0.01$ $= 0.0285 + 0.006 + 0.0091$ $= 0.0436$	<p>M1 three products</p> <p>M1 <i>dep</i> sum of three products</p> <p>A1 FT their tree</p>	3
(iv)	$P(\text{negative} \mid \text{has disease})$ $= \frac{P(\text{negative and has disease})}{P(\text{has disease})} = \frac{0.0091}{0.0436} = 0.2087$	<p>M1 for their 0.01×0.91 or 0.0091 on its own or as numerator M1 <i>indep</i> for their 0.0436 as denominator</p> <p>A1 FT their tree</p>	3
(v)	<p>Thus the test result is not very reliable.</p> <p>A relatively large proportion of people who have the disease will test negative.</p>	<p>E1 FT for idea of 'not reliable' or 'could be improved', etc</p> <p>E1 FT</p>	2
(vi)	$P(\text{negative or doubtful and declared clear})$ $= 0.91 + 0.06 \times 0.10 \times 0.02 + 0.06 \times 0.90 \times 1$ $= 0.91 + 0.00012 + 0.054 = 0.96412$	<p>M1 for their $0.91 +$</p> <p>M1 for either triplet</p> <p>M1 for second triplet</p> <p>A1 CAO</p>	4
TOTAL			18

Q8	$X \sim B(17, 0.2)$		
(i)	$P(X \geq 4) = 1 - P(X \leq 3)$ $= 1 - 0.5489 = 0.4511$	B1 for 0.5489 M1 for 1 – their 0.5489 A1 CAO	3
(ii)	$E(X) = np = 17 \times 0.2 = 3.4$	M1 for product A1 CAO	2
(iii)	$P(X = 2) = 0.3096 - 0.1182 = 0.1914$ $P(X = 3) = 0.5489 - 0.3096 = 0.2393$ $P(X = 4) = 0.7582 - 0.5489 = 0.2093$ So 3 applicants is most likely	B1 for 0.2393 B1 for 0.2093 A1 CAO <i>dep</i> on both B1s	3
(iv)	(A) Let p = probability of a randomly selected maths graduate applicant being successful (for population) $H_0: p = 0.2$ $H_1: p > 0.2$ (B) H_1 has this form as the suggestion is that mathematics graduates are <u>more</u> likely to be successful.	B1 for definition of p in context B1 for H_0 B1 for H_1 E1	4
(v)	Let $X \sim B(17, 0.2)$ $P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.8943 = 0.1057 > 5\%$ $P(X \geq 7) = 1 - P(X \leq 6) = 1 - 0.9623 = 0.0377 < 5\%$ So critical region is $\{7,8,9,10,11,12,13,14,15,16,17\}$	B1 for 0.1057 B1 for 0.0377 M1 for at least one comparison with 5% A1 CAO for critical region <i>dep</i> on M1 and at least one B1	4
(vi)	Because $P(X \geq 6) = 0.1057 > 10\%$ Either: comment that 6 is still outside the critical region Or comparison $P(X \geq 7) = 0.0377 < 10\%$	E1 E1	2
		TOTAL	18

4766: Statistics 1

General Comments

The paper attracted a fairly wide range of responses, although there were relatively few scripts with very low scores. There was no evidence to suggest that candidates had insufficient time to attempt all questions. As in recent sessions, answers were often well presented but once again many candidates did not appear to appreciate the implications of using rounded answers in subsequent calculations.

Good answers were seen from many candidates in questions 1, 2, 3(i),(ii), 4(i),(i)i, 5(i), 6, 7(i)-(iii) and 8(i),(i)i. Candidates' work on Venn diagrams was much better than in recent papers, although in this paper candidates had to use a given diagram, rather than complete their own and perhaps this assisted them to perform well.

Candidates' responses to Q3(iii) suggest that more attention should be given to finding mean and standard deviation of transformed data. Calculation and interpretation of conditional probability as in Q7 continues to cause difficulties. In hypothesis testing, the work generally continues to improve; the use of point probabilities rather than tail probabilities seems to be declining, although many candidates are still not meeting the requirement to define p in words. There were a number of centres where candidates who scored well on the rest of the paper appeared to have minimal knowledge of hypothesis testing, possibly suggesting that this topic has only been covered superficially.

Comments on Individual Questions

Section A

1 Album tracks; combinations and arrangements

- (i) Many totally correct answers were seen although candidates occasionally evaluated 8P_4 .
- (ii) Again very many correct answers were seen with the most frequent error being an answer of 16, often from 4^2 .

2 Customer spending; frequency table and total from histogram.

- (i) Most candidates correctly stated the group limits, although occasionally boundaries such as 19 or 21 instead of 20 were seen. Answers to the frequencies were less successful with a significant number of candidates giving the frequency density in place of frequency or doubling or halving each frequency.
- (ii) Most candidates realised the necessity for finding the sum of the frequencies multiplied by the interval mid-point, although a few simply gave the sum of the frequencies as their answer. Others multiplied the mid-points by the frequency density. A few decided that the question required an estimation of the mean amount of money spent.

3 Exam marks; mean, standard deviation, outliers, linear transformation.

- (i) Virtually all candidates obtained the mean correctly although some were less successful with the standard deviation. Errors here included use of an incorrect formula for S_{xx} but only occasionally division by n rather than $(n-1)$.
- (ii) There were many fully correct answers although there was occasionally use of 1.5s rather than 2s.
- (iii) Many candidates were totally successful with the mean and standard deviation of the scaled data. The most frequent error was to calculate $s_y = 1.2s_x - 10$ instead of $s_y = 1.2s_x$. Some candidates decided to calculate the transformed summary statistics and then use these to find the new mean and standard deviation. Quite often this did lead to a correct new mean but almost without exception they were unable to adapt this approach to find the new standard deviation. The fact that only 2 marks were available should have alerted candidates that this did not warrant a further 2 pages of calculations.

4 Recycling; Venn diagram, conditional probability.

- (i) Most candidates answered both parts entirely correctly, demonstrating their abilities to correctly read and interpret a Venn diagram.
- (ii) A pleasing number of correct answers were seen to a question on a topic which candidates often struggle with. The idea was to use the Venn diagram to write down the probability without any calculation, but some chose to use the conditional probability formula which was of course equally acceptable. There was nonetheless a variety of errors leading to answers such as 13/50, 11/50 and 24/50, effectively missing the conditional nature of the question.
- (iii) Correct answers to this part were conspicuous by their absence. Invariably answers such as $2 \times 18/50 \times 32/50$ or $18/50 \times 32/50$ were given, with candidates not realizing that the second selection was from 49. Indeed sight of a second fraction with a denominator of 49 was a rarity, even from very high scoring candidates. This type of decreasing probability question has been set many times in the past and candidates should ask themselves a simple question – are the events independent or dependent?

5 Rainfall and global warming, median and interquartile range, discussion.

- (i) A considerable proportion of candidates stated that the 11th value was the median rather than the average of the 11th and 12th. They were more successful with the interquartile range although the use of $(7+1)/2$ for the lower quartile was not unusual. A very few candidates treated the data as continuous and constructed a cumulative frequency curve, gaining no credit.
- (ii) Full marks in this part were very rare. Many candidates, even those who overall scored highly, answered this as a question about summer rainfall, ignoring all reference to global warming being the cause. Such candidates thought that the conclusion was valid based on the median falling by 1 day and the IQR staying the same. This gained no credit.

6) **Telephone competition; probability, calculation of $E(X)$ and $\text{Var}(X)$.**

- (i) Most candidates answered correctly, either by using a probability argument or by considering combinations. A few tried to justify the given value by using the other probabilities given in the table.
- (ii) Most candidates calculated both expectation and variance correctly, although some inaccuracy was seen when candidates used decimal probabilities. Some candidates correctly found $E(X^2)$ thus scoring some credit, but then omitted the subtraction of $[E(X)]^2$ or used $[E(X)]$ only in calculating $\text{Var}(X)$. There are still some candidates who insist in dividing either $E(X)$ or $\text{Var}(X)$ or both by divisors n or $(n-1)$. Such actions are penalised. Overall this question was a rich source of marks for many candidates.

Section B

7 Screening test; tree diagram, probability, conditional probability, interpretation.

- (i) Almost all candidates gained all 4 marks here.
- (ii) Again the vast majority of candidates were successful here.
- (iii) Most candidates were again successful although a few multiplied instead of added the relevant products.
- (iv) Many candidates were successful here although some candidates were unable to find this conditional probability. Common errors included answers of 0.0091, 0.0436/0.91, 0.0436/0.0091 and $(0.0436 \times 0.0091)/0.0436$.
- (v) The attempts at commenting on the answer to part iv) were very mixed with some candidates thinking that the larger the value of their answer, the more effective the test. A significant number of answers referred to a proportion of negative results rather than a proportion of those with the disease.
- (vi) There were a few excellent answers but, without a complete tree diagram to assist them, most candidates failed to identify all the required possibilities. Common errors included partially correct answers such as $0.91 + 0.06 \times 0.9 = 0.964$, as well as entirely incorrect answers such as $0.91 \times 0.99 + 0.06 \times 0.9 = 0.9549$.

8 Job applications; binomial distribution, expected frequency, highest probability, hypothesis test, critical region.

- (i) Relatively few candidates were able to find this relatively straightforward upper tail probability correctly. Most failed to realise what was required by "at least". Answers of $P(X = 4) = 0.2093$, $P(X \geq 4) = 0.5489$ or 0.7582 , $P(X \geq 4) = 1 - 0.2093$ or $= 1 - 0.7582$ appeared with regularity.
- (ii) Most answers to part ii) were correct although few candidates resisted the urge to round their answer of 3.4 to an integer. Others insisted erroneously that $E(X) = 3$ or that $E(X) = 17 \times 0.4511$ (or their probability in part (i))
- (iii) Answers to this part were disappointing, with many candidates stating that 3 was the most likely number of applicants as that value was closest to the expectation. Although the value with highest probability in the binomial distribution is close to the expectation, it is necessary to calculate probabilities both sides of the expectation to confirm the maximum. With 3 marks available, candidates should realise that more than this is required. Full credit could only be given when candidates had found both $P(X = 3)$ and $P(X = 4)$, (and also preferably $P(X=2)$) but some were content to make their judgement based on $P(X=3)$ alone. Those who did not calculate any probabilities earned no marks at all. Again this type of question has been set in the past and the required methodology has been commented on in previous reports.
- (iv) Many candidates correctly stated their hypotheses in symbolic form. However, many incorrect notations were 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. As in previous papers, still very few candidates realise the need to define the parameter 'p' and thus most lose a third mark, even if they have stated their hypotheses correctly. Previous reports have referred to the importance of this. However the reason for the form of the alternative hypothesis was explained well by many candidates

- (v) There was also an improvement here on earlier papers, with fewer candidates using point probabilities. However, a common error was to evaluate lower tail probabilities, despite having the correct upper tail hypothesis. Amongst candidates who did find an upper tail probability, a very common error was to state correctly that $P(X \geq 6) = 0.1057 > 5\%$ and $P(X \geq 7) = 0.0377 < 5\%$ before giving a wrong critical region of $X \geq 6$. Other answers obviously along the right lines failed to include any probabilities as justification, for example $P(X \geq k) < 0.05$, $P(X \leq k-1) > 0.95$, $k - 1 = 6$, $k = 7$, critical region is 7 and above. Candidates are expected to give numerical probabilistic justification for their answers. A further frequent omission was the failure to provide an explicit numerical comparison of the tail probabilities with the significance level of 5%, which again is always a requirement in hypothesis tests.
- (vi) This was usually answered correctly by those candidates who had already shown an understanding of hypothesis testing in part (v).