OCR
RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2613

Statistics 1

Thursday **5 JUNE 2003** Morning 1 hour 20 minutes

Additional materials:
Answer paper
Graph paper
MEI Examination Formulae and Tables (MF12)

TIME 1 hour 20 minutes

INSTRUCTIONS TO CANDIDATES

- Write your Name, Centre Number and Candidate Number in the spaces provided on the answer booklet.
- Answer **all** questions.
- You are permitted to use a graphical calculator in this paper.

INFORMATION FOR CANDIDATES

- The allocation 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 sufficient detail of the working is shown 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 60.

This question paper consists of 4 printed pages.

1 A train company runs a non-stop service from Oxbridge to Camford. The numbers of passengers on the 0730 service on 20 weekdays were as follows.

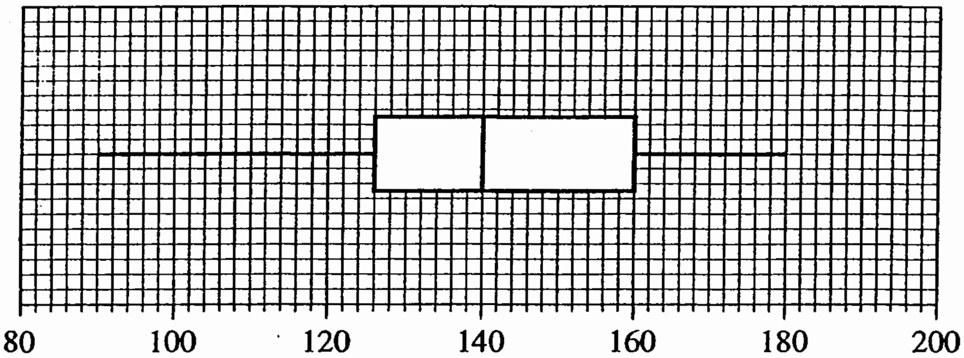
184	193	195	189	173
175	171	178	174	163
184	162	171	154	199
217	187	169	183	186

(i) Construct a sorted stem and leaf diagram to represent the data, taking as the stem values 150, 160, … . Hence describe the shape of the distribution. [4]

(ii) Calculate the median and the inter-quartile range. [3]

(iii) Using the inter-quartile range, show that there is just one outlier. Find the effect of its removal on the median and the inter-quartile range. [4]

The numbers of passengers on the 1030 service on another 20 days are summarised in the following box and whisker diagram.



(iv) Find the median and inter-quartile range. Hence compare the numbers travelling on the 0730 service and the 1030 service from Oxbridge to Camford. [4]

2 Next September I intend to buy a new car. Its registration plate will be of the form

HW 53 MSD

where HW is the local area code for the Isle of Wight, 53 represents the second half of the year 2003, and the last three letters are chosen at random.

The five parts of the question refer to the last three letters of the registration plate. You may assume that all 26 letters in the alphabet, of which 5 are vowels and 21 are consonants, are used for each of the random choices.

(i) Find the probability that the random letters on the plate are MSD, appearing in that order. [2]

(ii) Find the probability that the letters are M, S, D in any order. [2]

(iii) Find the probability that just two of the letters are the same. [3]

(iv) Find the probability that just one of the letters is a vowel. [4]

(v) Given that just one of the three letters is a vowel, find the probability that the first and last letters are the same. [4]

3 (i) Give one similarity and one difference between stratified and quota sampling. [2]

A road safety organisation wishes to survey cyclists according to which of three age groups (under 18, 18 to 35, over 35) they are in.

(ii) Briefly describe how a quota sample of size 300 might be obtained. [3]

For the sample from the under 18 age group, the numbers of days they had used their bicycles in the last week were recorded as follows.

Number of days (x)	0	1	2	3	4	5	6	7
Frequency (f)	15	10	9	5	7	24	8	2

(iii) State the mode and give a possible reason why it takes this value. [2]

(iv) Calculate the mean and standard deviation of the data. [3]

Three cyclists from the under 18 age group sample are to be chosen at random.

(v) In how many ways can this be done? [2]

(vi) Find the probability that all three used their bicycles on exactly 5 days. [3]

4 A motoring organisation reports that the proportion of drivers who would fail a basic sight test is 1 in 6.

For parts **(i)** to **(iii)** you may assume that this report is correct.

(i) Write down the value of p, the probability that a driver chosen at random would **pass** the sight test. [1]

(ii) A random sample of 30 drivers is taken.

(A) How many would be expected to pass the test?

(B) Find the probability that exactly this number pass the test. [4]

(iii) A random sample of n drivers is taken.

(A) Find the probability that all pass the sight test when $n = 13$.

(B) Find the smallest value of n such that the probability of all drivers passing is less than 5%. [4]

A journalist wishes to test the accuracy of the motoring organisation's report by checking the sight of a random sample of 15 drivers.

(iv) Write down suitable hypotheses for this test. Find the critical region for the test at the 10% significance level and illustrate it on a number line. [6]

Mark Scheme

Question 1

(i)	15 | 4 16 | 2 3 9 17 | 1 1 3 4 5 8 18 | 3 4 4 6 7 9 19 | 3 5 9 20 | 21 | 7 Key: 15 | 4 means 154 Shape of distribution is (roughly) symmetrical. *Allow "Normal" or "bell-shaped" or unimodal*	G1 for stem G1 for leaves in vertical alignment G1 for order [*condone at most 1 error*] *Note:* key not necessary E1 description	**4**
(ii)	Median = 180.5 Inter-quartile range = 188 – 171 = 17 *or* = 188.5 – 171 = 17.5	B1 **cao** for median M1 for sensible attempt at IQR A1	**3**
(iii)	$Q_1 - 1.5 \times IQR = 171 - 1.5 \times 17 = 145.5$ $Q_3 + 1.5 \times IQR = 188 + 1.5 \times 17 = 213.5$ Hence only data item outside the interval [145.5, 213.5] is 217. If 217 is removed, median drops to 178 IQR becomes 187 – 171 = 16 *or* 187.5 – 171 = 16.5 *or* 186.25 – 170.5 = 15.75	E1 for showing 217 is > 1.5 IQR above Q_3 E1 for showing there are no values < 1.5 IQR below Q_1 B1 **cao** for effect on median B1 **cao** for effect on IQR	**4**
(iv)	For the 20 journeys on the 10:30 train: Median number of passengers per journey = 140 Inter-quartile range = 160 – 126 = 34 On average the number of passengers travelling on 10:30 trains is lower, but there is a greater variability in the numbers of passengers on these trains.	B1 **cao** for median B1 **cao** for IQR E1 for comparison of no. of passengers E1 for comparison of variability in context	**4**
			15

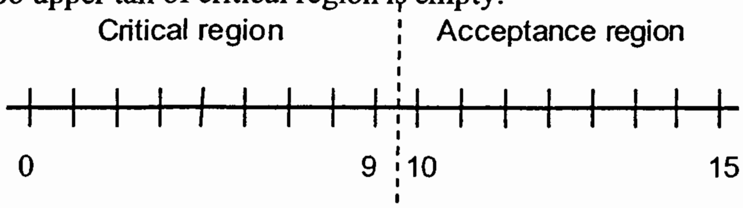
Question 2

(i)	P(letters on plate are MSD, in that order) $= \left(\frac{1}{26}\right)^3 = \frac{1}{17576} = 0.000057$ (to 2 s.f.)	M1 for probability A1 **cao**	**2**
(ii)	P(letters on plate are M, S, D in any order) $= 3! \times \left(\frac{1}{26}\right)^3 = \frac{6}{17576} = \frac{3}{8788} = 0.0034$ (to 2 s.f.)	M1 for probability [3! × their part (i)] A1	**2**
(iii)	P(exactly two of the letters are the same) $= 3\times\frac{1}{26}\times\frac{25}{26}$ *or* $\frac{26\times25\times3}{26^3}$ $= \frac{75}{676} = 0.111$ (to 3 s.f.) $= 0.11$ (to 2 s.f.)	M1 for $\frac{1}{26}$ seen M1 for $\frac{1}{26}\times\frac{25}{26}$ *or* M1 for 26×25 M1 for "divide by 26^3" A1 **cao**	**3**
(iv)	P(just one of the letters is a vowel) $= 3\times\frac{5}{26}\times\frac{21}{26}\times\frac{21}{26}$ $= \frac{6615}{17576} = 0.376$ (to 3 s.f.) $= 0.38$ (to 2 s.f.)	B1 for $\frac{5}{26}$ *or* $\frac{21}{26}$ seen M1 for $\frac{5}{26}\times\frac{21}{26}\times\frac{21}{26}$ M1 for $3\times p\times q^2$ A1 **cao**	**4**
(v)	P(first and last letters are the same, given that just one of the three letters is a vowel) $= \dfrac{\frac{21}{26}\times\frac{5}{26}\times\frac{1}{26}}{3\times\frac{5}{26}\times\frac{21}{26}\times\frac{21}{26}}$ *or* $\frac{21\times5}{3\times5\times21\times21} = \frac{1}{21}\times\frac{1}{3}$ $= \frac{1}{63} = 0.0159$ (to 3 s.f.) *or* 0.016 (to 2 s.f.)	M1 for $\frac{21}{26}\times$ M1 for $\frac{5}{26}\times\frac{1}{26}$ M1 for division by their part (iv) *or* M1 for "21×5", M1 for "$3\times5\times21\times21$" M1 for quotient A1 **cao**	**4**
			15

Question 3

(i)	Similarity: **both** divide population into strata, sub –groups or categories, etc. Difference: quota sampling is non-random, whereas the sampling within each stratum for stratified sampling is, *Or* quota sampling could be biased, stratified is not, *Or* with quota sampling the questioner chooses the samples, but in stratified sampling samples are chosen at random. *Or* stratified uses a sampling frame, quota does not.	B1 for similarity B1 for difference	2
(ii)	Divide sample of 300 into (age) groups, According to proportion in population or some pre-defined rule. Interview using opportunity sample, e.g. in a town centre at a cycle park.	E1 for age groups E1 for proportions E1 for method	3
(iii)	Mode = 5 days Could take this value since many cyclists use their bicycle regularly for school, college, work, etc.	B1 for mode E1	2
(iv)	Mean $= \frac{253}{80} = 3.16$ days (to 3 s.f.) $= 3.2$ days (to 2 s.f.) Standard deviation $= \sqrt{\frac{1189}{80} - 3.1625^2}$ $= \sqrt{4.86109} = 2.20$ (to 3 s.f.) $= 2.2$ (to 2 s.f.)	B1 **cao** for mean M1 for sensible attempt at variance A1	3
(v)	Number of ways of choosing the 3 cyclists $= {}^{80}C_3 = 82160$	M1 for ${}^{n}C_3$ A1 **cao**	2
(vi)	P(all 3 used their bicycle on at exactly 5 days) $= \frac{{}^{24}C_3}{{}^{80}C_3}$ *or* $\frac{24}{80} \times \frac{23}{79} \times \frac{22}{78}$ $= \frac{253}{10270} = 0.0246$ (to 3 s.f.) $= 0.025$ (to 2 s.f.)	M1 for ${}^{24}C_3$ M1 for division by ${}^{80}C_3$ *or* B1 for $\frac{24}{80}$ [= 0.3] M1 for product of probs. A1 **cao**	3
			15

Question 4

(i)	$p = \frac{5}{6}$	B1 for probability	**1**
(ii)	(*A*) Expected number to pass the test $= 30 \times \frac{5}{6} = 25$ (*B*) P(exactly 25 pass the test) $= {}^{30}C_{25} \times \left(\frac{5}{6}\right)^{25} \times \left(\frac{1}{6}\right)^{5}$ $= 0.192$ (to 3 s.f.) $= 0.19$ (to 2 s.f.)	B1 for expected number M1 for $(5/6)^r \times (1/6)^{30-r}$ M1 for ${}^{30}C_r \times \ldots$ A1 **cao**	**4**
(iii)	(*A*) P(all pass sight test) $= 1 - 0.9065 = 0.0935$ [*using tables*] *or* $= \left(\frac{5}{6}\right)^{13} = 0.0935$ (to 3 s.f.) $= 0.093$ (to 2 s.f.) (*B*) Searching for appropriate *n*: for $n = 16$: P(all pass) $= 1 - 0.9459$ [*tables*] *or* $\left(\frac{5}{16}\right)^{16} = 0.0541$ for $n = 17$: P(all pass) $= 1 - 0.9549$ [*tables*] *or* $\left(\frac{5}{16}\right)^{17} = 0.0451$ *or* Using logarithms: $n \log\left(\frac{5}{6}\right) < \log 0.05 \Rightarrow n > 16.43$ hence smallest sample size is where $n = \mathbf{17}$	M1 for use of tables or working out A1 M1 for attempt at search or use of logarithms A1 for answer	**4**
(iv)	H_0: $p = \frac{5}{6}$; H_1: $p \neq \frac{5}{6}$ Using binomial tables for $n = 15$: $P(X \leq 9) = \mathbf{0.0274} < 0.05$, but $P(X \leq 10) = \mathbf{0.0898} > 0.05$ So lower tail of critical region is {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} $P(X \geq 15) = 1 - P(X < 14) = 1 - 0.9351 = \mathbf{0.0649} > 0.05$ So upper tail of critical region is empty.  *Allow corresponding argument based on* H_0: $p = \frac{1}{6}$; H_1: $p \neq \frac{1}{6}$	B1 for H_0, B1 for H_1 M1 for at least one comparison A1 B1 for statement or evidence G1 **cao** for illustration	**6**
			15

Examiner’s Report

2613 Statistics 1

General Comments

Generally the performance of candidates on this paper was disappointing, especially at the upper end of the mark range. This was undoubtedly mainly due to the difficulty of the probability situation in question 2, which proved less accessible than intended. Candidates also did less well than expected in the data analysis of question 1. Here there seemed much confusion over the calculation of the inter-quartile range and its use in determining possible outliers.

It was encouraging to see a generally better understanding of the sampling procedures in question 3, but their comparing stratified with quota sampling still proved a stiff challenge for many candidates. The binomial distribution of question 4 was also well understood, in terms of probability calculations, but the task of finding a critical region for this two-tail test proved to be a step beyond the capabilities of most candidates. This was sometimes due to shortage of time, but mostly due to lack of a clear method of tackling the problem.

Comments on Individual Questions

Q.1 Data analysis; distribution of numbers of passengers on train journeys: median and inter-quartile range, outliers, box and whisker diagram

There was a very variable performance on this question, with marks being lost in all parts. The only discernible trend was that only a minority could calculate the inter-quartile range correctly and very few candidates indeed knew the test for outliers using the inter-quartile range.

(i) The stem and leaf diagram was usually constructed accurately, but there were several cases of the leaves being misaligned vertically. In addition to the required answer of "roughly symmetrical", answers of 'Normal', 'bell-shaped' or 'unimodal' were accepted. Answers *just* referring to skewness were marked as incorrect.

(ii) Whilst the majority of candidates identified the median correctly, many struggled to find the inter-quartile range, using either of the acceptable methods for a small data set. Sadly, some did not even subtract the lower quartile from the upper quartile, preferring just to quote a range from Q_1 to Q_3. It was disappointing to see a topic which was first encountered at GCSE so poorly attempted

(iii) Furthermore, the rule for finding outliers using the boundary values $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$ seemed to be virtually unknown by nearly all candidates. Many used their own devices, such as $Q_2 \pm 2 \times \text{IQR}$, mirroring the definition which involves mean and standard deviation. Occasionally the latter method was used instead. Thankfully, nearly all candidates identified the outlier, 217, correctly, and were therefore often able to score the final two marks. Problems encountered in finding the original IQR were often repeated when the outlier was excluded.

(iv) The median and inter quartile range were often found more consistently using the given box and whisker diagram, but a minority of candidates demonstrated poor reading of the scales and so quoted an incorrect IQR. Pleasingly, comments given were usually sensible and mostly in context, as expected.

(i) stem-and-leaf diagram, roughly symmetrical; (ii) median = 180.5, IQR = 17 *or* 17.5;
(iii) $171 - 1.5\times 17 = 145.5$, $188 + 1.5\times 17 = 213.5$, 217 only value outside [145.5, 213.5],
median = 178, IQR = 16
(iv) median = 140, IQR = 34, comment comparing medians and IQRs in context.

Q.2 Probability; letters on a number plate: calculating and combining probabilities, conditional probability

The style of this question proved to be a stumbling block for the majority of candidates. It was never expected that attempts using permutations and combinations should be used, but this proved to be the case on a large number of occasions. The phrase in the stem "You may assume that all 26 letters ... are used for each of the random choices" was largely ignored by most candidates, assuming that letters could not be repeated, a possibility made quite explicit in part (iii). The conditional probability in part (v) was really a closed book, with virtually no one getting this part right. The complexity of using the formula for conditional probability in this context was beyond the vast majority of candidates.

(i) Incorrectly thinking that all the letters had to be different often led to the answer $\frac{1}{26}\times\frac{1}{25}\times\frac{1}{24}$ rather than the required $\left(\frac{1}{26}\right)^3$. The incorrect answer was often derived from $\frac{1}{{}^{26}P_3}$.

(ii) The consequence of assuming letter dependence often led to the answer $6\times\frac{1}{26}\times\frac{1}{25}\times\frac{1}{24}$, rather than the required $6\times\left(\frac{1}{26}\right)^3$. Candidates were not penalised twice, however, and often scored both marks here on follow through, provided their answer was 6 times their answer to part (i).

(iii) The correct answer was rarely seen, with an extra factor of $\frac{1}{26}$ being present. Provided this was the only error, candidates could score both of the method marks. The final factor of multiplying by 3 was included in the accuracy mark.

(iv) There was limited success in this part. Candidates who recognised the binomial nature of the required probability often scored all 4 marks. However, a common error was to omit the factor of multiplying by 3. Here this was a method mark, so such candidates were limited to just the first two marks.

(v) Correct answers were hardly ever seen. Many candidates simply calculated the numerator of the fraction $\frac{P(A\cap B)}{P(B)}$ and so gained partial credit. Many others were confused by the numerator, but did gain some follow through credit by dividing by P(B), i.e. their answer to part (iv). Many solutions showed the candidates' frustration with this question with much crossed-out work and re-attempts at various parts.

(i) $\frac{1}{17576}$; (ii) $\frac{3}{8788}$; (iii) $\frac{75}{676}$; (iv) $\frac{6615}{17576}$ = 0.376 (to 3 s.f.); (v) $\frac{1}{63}$.

Q.3 Sampling and probability; dividing cyclists into age-groups

There were some pleasing solutions to this question, but the majority of candidates demonstrated some confusion with sampling methods. In the final part, most candidates assumed independence of probabilities, but should have sampled *without* replacement. This was often coupled with similar misinterpretation of question 2, part (i).

(i) Many candidates managed to gain at least 1 of the two marks available for contrasting stratified and quota sampling. From the answers given, however, it was evident that knowledge of these techniques is only sketchy. Most surprising was the lack of all correct answers for the similarity, where the answer "both divide the population into groups" was not seen as often as expected. Equivalent answers for the differences, such as "stratified chooses samples at random, but in quota sampling the questioner chooses the samples", "stratified uses a sampling frame, but quota does not", were quite acceptable.

(ii) Nearly all candidates divided the sample up into age groups – even when they failed to gain credit for the similarity in part (i). Most divided the 300 into three groups of 100, which was allowed, but few candidates discussed the idea of the number in each group being either "according to proportion in the population" or "according to some pre-defined rule". The third mark was available for describing the method of taking the sample, but was rarely gained.

(iii) Nearly all candidates found the mode correctly, but a small minority quoted the reason as "the most frequently occurring, rather than giving a reason in context why the mode should be 5.

(iv) Most candidates found the mean and standard deviation correctly, with a small number of errors in finding the standard deviation. These included taking $n = 8$ or the use of $(\Sigma fx)^2$ in the standard deviation formula.

(v) The correct answer of ${}^{80}C_3$ was seen less often than expected. An unexpected wrong answer was the use of ${}^{18}C_3$, presumably taking the '18' from the previous line. Generously, a method mark could be gained in this case.

(vi) In contrast to question 2, part (i), $\left(\frac{24}{80}\right)^3$ was invariably seen instead of the correct $\frac{24}{80} \times \frac{23}{79} \times \frac{22}{78}$. This only highlights the confusion which students have between dependent and independent probabilities.

(i) similarity and difference between stratified and quota sampling;
(ii) divide 300 into 3 age groups using pre-defined rule, interview in suitable context;
(iii) 5 days and comment; (iv) mean = 3.16, s.d. = 2.20 (to 3 s.f.);
(v) 82160; (vi) 0.0246 (to 3 s.f.).

Q.4 Binomial distribution and hypothesis testing; defective tyres on lorries and vans

Responses to question 4 were variable; most candidates handled the probability calculations reasonably well, but once again the hypothesis testing section provided responses ranging from being largely correct to utter nonsense.

(i) The vast majority of candidates successfully quoted $\frac{5}{6}$ (i.e. $1 - \frac{1}{6}$).

(ii) Again, nearly all candidates knew how to find the expected number of students passing the test (25 out of 30) and its associated probability.

(iii) This part was again usually tackled well by the majority of candidates. Most correctly found the probability of all students passing the test when $n = 13$, and a pleasingly large number extended this idea to that of finding the minimum value of n such that this probability was less than 0.05. Most solutions were carried out by calculation rather than using tables.

(iv) Many previous reports have flagged up the need to use correct notation in stating null and alternative hypotheses, as given in the solution below. It is depressing to see the large number of candidates who still do not use the correct notation involving probability p. The omission of p was prevalent and often lost both marks for stating the hypotheses. It was very disappointing to see only a very few correct solutions for the critical region, even when the alternative hypothesis for this two-tail test was stated correctly. Systematic correct use of cumulative binomial tables was rarely seen, especially to show that the upper tail was empty. Prevalent errors included the using of 10% rather than splitting the two tails into 5% for the significance levels. An improvement, however, was the much less frequent use of point probabilities for hypothesis testing.

(i) $\frac{5}{6}$; (ii) (*A*) 25, (*B*) 0.192 (to 3 s.f.); (iii) (*A*) 0.0935 (to 3 s.f.), (*B*) $n = 17$;
(iv) $H_0: p = \frac{5}{6}$, $H_1: p \neq \frac{5}{6}$, critical region {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, number line