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

Monday **19 JANUARY 2004** Morning 1 hour 20 minutes

Additional materials:
Answer booklet
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 Over a period of time, a teacher recorded the number of times, x, each of the 20 students in the mathematics class was absent. The distribution was as follows.

Number of times absent, x	0	1	2	3	4	5	6	7	8	9	10	11 or more
Number of students, f	4	6	3	2	0	2	0	1	1	0	1	0

$$\Sigma f = 20 \qquad \Sigma fx = 53 \qquad \Sigma fx^2 = 299$$

(i) Illustrate the data using a suitable diagram. [2]

(ii) State the mode and find the median for the data set. [2]

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

During this period of time, there were 30 mathematics lessons. The teacher needs to analyse the distribution of the number of times each student was *present* during the 30-lesson session.

(iv) Without creating a new frequency distribution, deduce values for the mean and standard deviation of the numbers of times students were present. Describe the shape of the new distribution. [3]

There are 12 boys and 8 girls in the class. The mean of the numbers of times boys were *absent* was 3, and the standard deviation was also 3.

(v) Show that the mean of the numbers of times girls were absent is 2.125. [2]

(vi) Find the standard deviation of the numbers of times girls were absent. [3]

2 In a certain city, there are just two political parties, the Red Party and the Blue Party. Each year one person from each party seeks to be elected as mayor of the city. If the city has a Red Party mayor one year, the probability that it has a Red Party mayor the next year is 0.5. If the city has a Blue Party mayor one year, the probability that it has a Blue Party mayor the next year is 0.7.

In the year 2004 the mayor belongs to the Blue Party.

(i) Illustrate the possible outcomes for mayor in the years 2005, 2006 and 2007 on a probability tree diagram. [3]

(ii) Find the probability that

(*A*) the mayors in 2005 and 2006 belong to the same party, [3]

(*B*) the mayor in 2007 belongs to the Red Party, [3]

(*C*) the mayor for just one of the three years after 2004 belongs to the Blue Party. [3]

(iii) Find the conditional probability that the mayor in 2005 belongs to the Red Party, given that the mayor in 2007 belongs to the Red Party. [3]

3 A school has 1600 pupils on the roll. A student wishes to take a sample of size 80 to survey which method pupils use most often in travelling to school.

(i) Describe in detail how the student could select a systematic sample of the pupils. [4]

(ii) State one advantage and one disadvantage of sampling as opposed to conducting a survey of the whole population. [2]

The results of the sample were as follows.

Method of travel	Bus	Car	Bicycle	Walk
Number of pupils	25	16	9	30

(iii) Use the sample information to estimate the number of pupils in the school who do not walk to school. [2]

Of the 30 pupils in the sample who walk to school, 20 are girls and 10 are boys. From these 30 pupils, 5 are chosen at random for an interview.

(iv) In how many ways may they be chosen? [2]

Let X represent the number of girls in the group of 5 pupils.

(v) Explain why the binomial distribution is *not* a suitable model for X. [1]

(vi) Find $P(X = 3)$. [4]

Turn over for Question 4

4 A cross-channel ferry company runs a daily service from England to France. Records show that on average 85% of the ferry crossings leave on time, and the rest leave late.

During the next three weeks there will be 21 departures.

(i) State a suitable distribution to model the number of times the ferry leaves on time and one assumption for the model to be valid. [2]

(ii) For these three weeks, find the probability that

(A) all departures leave on time, [2]

(B) exactly 3 departures leave late, [2]

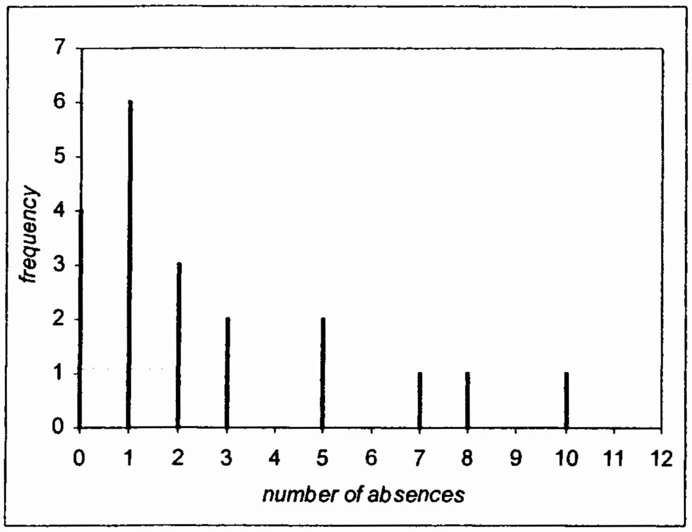
(C) during each week, no more than 1 of the 7 departures leaves late. [4]

During the summer season it is suspected that fewer than 85% of ferry crossings will leave on time. In a random sample of 15 summer sailings, just 10 leave on time.

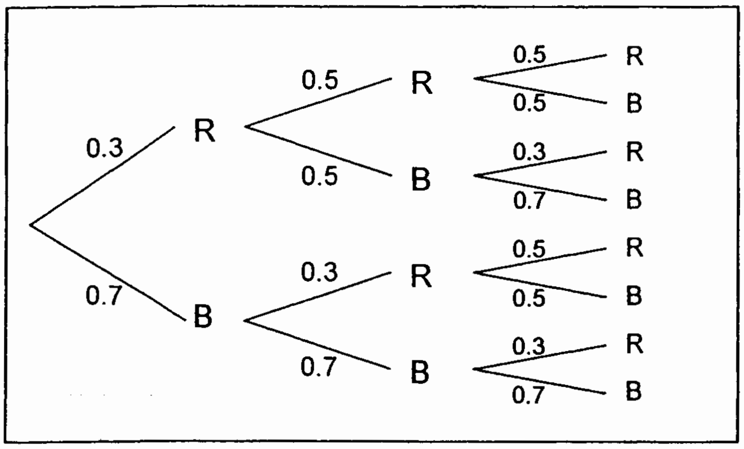
(iii) Carry out a suitable hypothesis test to examine the company's claim that 85% of summer sailings leave on time. Use a 5% significance level and state your hypotheses and conclusions carefully. [5]

Mark Scheme

Question 1

(i)		G1 vertical line chart with horizontal scale G1 for correct heights *Condone bar chart with gaps between bars and labels central underneath*	**2**
(ii)	Mode = 1 Median = 1.5	B1 B1	**2**
(iii)	Mean $= \frac{53}{20}$ *or* 2.65 s.d. $= \sqrt{\frac{299}{20} - 2.65^2} = \sqrt{7.9275} = 2.82$ (to 3 s.f.) $= 2.8$ (to 2 s.f.)	B1 M1 for variance A1 **cao**	**3**
(iv)	Mean number of students present = 30 – 2.65 = 27.35 Standard deviation = 2.82 (i.e. remains unchanged) Distribution of students present has a negative skew	B1 B1 E1	**3**
(v)	Mean no. of girls absent $= \frac{53-(12\times3)}{8} = \frac{17}{8} = 2.125$	B1 for "53 – (12 × 3)" B1 for "17 / 8" [SOI]	**2**
(vi)	s.d.(boys) = 3 $\Rightarrow 9 = \frac{\Sigma x^2}{12} - 9 \Rightarrow \Sigma x^2 = 216$ Hence s.d.(girls) $= \sqrt{\frac{299-216}{8} - 2.125^2}$ $= \sqrt{5.859375} = 2.42$ (to 3 s.f.) $= 2.4$ (to 2 s.f.)	B1 for Σx^2 M1 for new variance using "their 216" A1 **cao**	**3**
			15

Question 2

(i)		B1 for 1st pair branches B1 for 2nd set branches B1 for 3rd set branches	3
(ii)	(*A*) P(mayors in 2005 and 2006 belong to same party) $= 0.3 \times 0.5 + 0.7 \times 0.7 = 0.15 + 0.49$ $= 0.64$ (*B*) P(Red Party has the mayor in 2007) $= 0.3 \times 0.5 \times 0.5 + 0.3 \times 0.5 \times 0.3$ $+ 0.7 \times 0.3 \times 0.5 + 0.7 \times 0.7 \times 0.3$ $= 0.075 + 0.045 + 0.105 + 0.147$ $= 0.372 = 0.37$ (to 2 s.f.) (*C*) P(Blue Party has the mayor for just 1 of next 3 years) $= 0.3 \times 0.5 \times 0.5 + 0.3 \times 0.5 \times 0.3 + 0.7 \times 0.3 \times 0.5$ $= 0.075 + 0.045 + 0.105$ $= 0.225 = 0.23$ (to 2 s.f.)	M1 for 1 correct pair M1 for sum of 2 logically correct pairs A1 M1 for 1 correct triple M1 for sum of 4 logically correct triples A1 M1 for 1 correct triple M1 for sum of 3 logically correct triples A1	9
(iii)	P(Red Party mayor in 2005 | Red Party mayor in 2007) $= \dfrac{\text{P(Red Party mayor in 2005 and in 2007)}}{\text{P(Red Party mayor in 2007)}}$ $= \dfrac{0.3 \times 0.5 \times 0.5 + 0.3 \times 0.5 \times 0.3}{0.372} = \dfrac{0.12}{0.372}$ $= 0.32$ (2 s.f.) *or* 0.323 (3 s.f.)	M1 for 0.12 seen M1 for quotient with "their 0.372" on denominator A1 **cao**	3
			15

Question 3

(i)	Number the pupils 1 to 1600 (e.g. alphabetical order) [*or equivalent*] Choose a random number in the range 1 to 20, say r, as a starting value [*or equivalent*] Choose the r^{th}, $(r+20)^{\text{th}}$, $(r+40)^{\text{th}}$, ..., $(r+1580)^{\text{th}}$ pupil	E1 for method E1 for random start E1 for completion E1 for remaining choices	**4**
(ii)	Advantage of sampling: quicker to collect data Disadvantage of sampling: sample may not be representative of the population *or* may be biased *Accept equivalent distinct advantages/disadvantages*	E1 E1	**2**
(iii)	Number of pupils in sample who do not walk to school = 50 Hence estimate of number of pupils in school who do not walk to school $= 20 \times 50$ *or* $\frac{5}{8}$ of 1600 = 1000.	 M1 A1	**2**
(iv)	No. ways in which pupils can be chosen = $\binom{30}{5} = 142506$	M1 A1	**2**
(v)	Binomial distribution is not suitable since choices of pupil not independent *or* probability of choosing a boy (girl) is non-constant *or* sampling is *without* replacement.	E1	**1**
(vi)	$P(X=3) =$ $$\frac{\binom{20}{3}\times\binom{10}{2}}{\binom{30}{5}} = \frac{1140\times 45}{142506} = \frac{950}{2639} = 0.36 \text{ (2 s.f.)}$$ *or* $$\frac{20}{30}\times\frac{19}{29}\times\frac{18}{28}\times\frac{10}{27}\times\frac{9}{26}\times 10 = \frac{950}{2639} = 0.36 \text{ (2 s.f.)}$$	B1 for $\binom{20}{3}$ or $\binom{10}{2}$ M1 for quotient with their part (iv) on denominator M1 for numerator A1 **cao** *or* M1 for products on numerator M1 for products on denominator M1 for "$\times$ 10 or ${}^5C_3$" A1 **cao**	**4**
			15

Question 4

(i)	$X \sim B(21, 0.85)$ *or* 'binomial' if parameters stated or used in part (ii) Departure times are independent of each other *or* Probability of "success" ($p = 0.85$) is the same for each trial	B1 for distribution E1 for assumption	2
(ii)	(*A*) P(all departures leave on time) $= 0.85^{21} = 0.0329$ (to 3 s.f.) $= 0.033$ (to 2 s.f.)	M1 for probability A1	2
	(*B*) P(exactly 3 departures are late leaving) $= \binom{21}{18} \times 0.85^{18}\, 0.15^3 = 0.241$ (to 3 s.f.) $= 0.24$ (to 2 s.f.)	M1 for "$0.85^{18} \times 0.15^3$" M1 for "$\binom{21}{18} \times \ldots$" A1	3
	(*C*) P(during a week $\le$ 1 departure late leaving) $= 0.7166$ [from tables using B(7, 0.15)] $\Rightarrow$ P(during each week $\le$ 1 departure late leaving) $= (0.7166)^3 = 0.368$ (to 3 s.f.) $= 0.37$ (to 2 s.f.)	B1 for 0.7166 M1 for probability A1 **cao**	3
(iii)	H_0: $p = 0.85$ H_1: $p < 0.85$ [Assuming $X \sim B(15, 0.85)$] $P(X \le 10) = 0.0617$ (from tables) Since $0.0617 > 0.05$, [accept H_0] There is not enough evidence to accept the hypothesis that the proportion of sailings leaving on time has decreased. *Allow approach via critical regions:* Diagram with division between 9 and 10 "10 is not in the critical region" *or equivalent*	B1 for H_0 B1 for H_1 B1 for probability M1 (dep) for comparison A1 for conclusion in words *Third and fourth marks:* B1 M1 (dependent)	5
			15

Examiner’s Report

2613 Statistics 1

General Comments

Overall the performance of candidates was better than in the Summer 2003 paper. There were fewer very weak scripts and a number of outstanding submissions. The stronger candidates scored highly on all the questions with only the final probability in question 3 causing regular problems. Weaker candidates tended to gain the majority of their marks in questions 1 and 2 although a fair number did not recall work from GCSE. The probability calculations in question 2 were arguably more straightforward, and therefore accessible to weaker candidates, than in corresponding questions in recent papers.

Most answers were well presented and generally supported by sensible working and explanations. The vogue, in recent years, of quoting probabilities as percentages seems to have now, thankfully, disappeared. Virtually all candidates appeared to have some knowledge of all the topics although, as usual, hypothesis testing techniques were sketchy in some centres.

Comments on Individual Questions

Question 1 (Data analysis; vertical line diagram, measures of central tendency and standard deviation; absences in a class with gender considerations)

A good starting question for most candidates. The first three parts were usually well answered. The final three parts discriminated well between stronger and weaker candidates. It was pleasing to see that a substantial number of solutions showed some insight into the relationship between the distribution of absences and presences, and between the distribution of boys' and girls' absences.

(i) There were many totally correct answers. Most constructed a vertical line chart or used separate bars. A significant number used touching bars, treating the data as if it were continuous rather than discrete. Weaker candidates drew a frequency polygon or a cumulative frequency diagram with a small minority simply plotting the data as though it was a scatter diagram.

(ii) The mode was nearly always correct. The median was mostly correct but sometimes '1' or '2' or 'between 1 and 2' were seen. Occasionally, 'median =10.5' was seen, confusing the point of location of the median with its actual true x-value.

(iii) The mean and standard deviation were mostly correct. Quite a lot of candidates ignored the *given* summations and worked out the mean and standard deviation from the frequency distribution, thus wasting valuable time.

(iv) **Students present:** Many did not realise that '$p = 30 - a$' and recalculated from a new frequency distribution – often correctly but, once again, wasting time. They then realised that they could have got the answer by a much better way. Some of these candidates then went on to give the correct standard deviation but the examiners felt that quoting 'the standard deviation is the same' was by just pure luck rather than any real understanding of the question. There was a lot of explanation of the shape of the distribution but without the words 'negative skew'.

(v) **Girls absent:** This was mostly well done with $53 - 36 = 17$ often seen. There were some circular arguments used, for example: $8 \times 2.125 = 17$ followed by $\frac{17}{8} = 2.125$.

(vi) This part proved to be difficult for the majority of candidates. A few candidates who had performed well in part (v) failed here because they attempted to use the 'additive property' of standard deviation. They used it correctly. It was a pity for them that the sum of the square roots is not the same as the square root of the sums. Many simply thought that the overall standard deviation was the average of the boys' and girls' standard deviations. Those who knew the method frequently failed to rearrange $\frac{\Sigma x^2}{12} - 9 = 9$ correctly, often getting $\Sigma x^2 = (9 \times 12) + 9 = 117$.

(ii) mode = 1, median = 1.5;
(iii) mean = 2.65, s.d. = 2.82 (3 s.f.);
(iv) mean = 27.35, s.d. = 2.82 (3 s.f.), negative skew;
(vi) s.d. = 2.42 (3 s.f.)

Question 2 (Probability; tree diagram, combining probabilities by multiplication and addition, conditional probability; election of a mayor in different years)

This probability question turned out to be accessible for even the weakest of candidates. The construction and use of a standard probability tree diagram meant that the majority of candidates scored all of the first 12 marks available.

(i) The tree diagrams were usually correct. There were some 'double sized' trees seen and some labelled their branches with 0.7 and 0.5, or 0.3 and 0.5 – i.e. probabilities not adding up to 1.

(ii) Most scored the full 9 marks here or just made the odd slip. In (*A*) $0.15 \times 0.49 = 0.0735$ was seen. [They multiplied because it said 'and'.]

(iii) The conditional probability was not well answered. Most knew it was 'something' over 'something'. Common errors were $\frac{0.3}{0.372} = 0.806$ and $\frac{0.12}{0.3} = 0.4$. A few got 0.12, thinking it was P(red in 2005 and 2007), and went no further.

(ii) (*A*) 0.64, (*B*) 0.372, (*C*) 0.225; (iii) 0.323 (3 s.f.)

Question 3 (Systematic sampling; combinations, probability of events without replacement on methods of transport to school)

This question was a low scoring one for the majority of candidates. The main error was the lack of precision in the answers given for the descriptive parts. The probability parts met mixed responses, with most candidates confusing sampling *with* and sampling *without* replacement.

(i) For systematic sampling most said 'select every 20th from the school register' which was enough to gain 2 of the 4 marks available. However, some said every 10th or 80th or nth without specifying the value of n. Listing followed by the need for a random start and completion of the sampling cycle usually had at least one omission – sometimes all three. A few thought that systematic and stratified sampling were identical.

(ii) Many scored 2 marks for 'quicker' & 'not representative of the population / biased' although some thought this part referred to systematic sampling, which it clearly did not.

(iii) Most got 1000 with 50, 600, or even 62.5%, being the most common wrong answers.

(iv) ${}^{30}C_5 = 142506$ was often correct although ${}^{30}P_5$, 30! and 5! were seen.

(v) Not often answered correctly although most of the answers which were correct were concise and acknowledged the changing probability resulting from the previous selection. Incorrect answers included that the values were not given in the binomial tables; numbers of males and females were not equal; and there were more than two possibilities.

(vi) Having tried to explain why the Binomial distribution was *not* suitable nearly all then used it here assuming $B(5, \frac{2}{3})$ to give the incorrect 0.329. Ironically the few who did not offer an explanation in part (v) but realised the non-constant probabilities usually got the correct 0.360 but some 'forgot' to multiply by $^5C_3$, giving 0.036.

(iii) 1000; (iv) 142506; (vi) 0.360 (3 s.f.)

Question 4 (Binomial distribution and hypothesis testing; modelling departures for a cross-channel ferry)

A full spectrum of marks was seen. This remains a challenging topic for some candidates, where many misconceptions abound.

(i) Many did not understand what was required. A calculation of expectation as $21 \times 0.85 = 17.85$ was often seen as an answer. Very few candidates could state that a Binomial distribution was suitable. Even fewer could identify the associated parameters. Further, very few candidates appreciated the assumption of 'independent trials' for a binomial distribution.

(ii) There were many good answers to parts (*A*) and (*B*). The main error in part (*A*) was to quote 0.03 instead of 0.033. In part (*B*) some candidates forgot the combination factor of 10. Part (*C*) caused many more problems, with many candidates assuming that 'less than or equal to 1 each week for 3 weeks' would be the same as 'less than or equal to 3 in total' and thus used B (21, 0.85) again. Those who saw the need to use B (7, 0.15) often overlooked the possibility of tables to help them and used the more tedious calculation from the binomial formula. Some did get 0.7166 and stopped or then failed to cube it. Some did $[P(\text{exactly } 1)]^3$. Otherwise correct solutions often contained rounding errors. A fair number of candidates omitted this part of the question completely.

(iii) The examiners still saw a lot of '$H_0 = 0.85$' or '$H_0 = p = 0.85$'. Although most candidates scored both marks for stating the hypotheses, a significant minority still lose both marks for sloppy notation. Despite commenting on this in previous reports, it is very disappointing to see such unnecessary loss of credit. Many candidates correctly achieved $0.0617 > 0.05$. Often there was no conclusion beyond 'Accept H_0/Reject H_1'. even by the better candidates. It is important that candidates realise that a final mark is awarded for putting the conclusion into sensible context, such as 'There is not enough evidence to accept the hypothesis that the proportion of sailings leaving on time has decreased.'

(i) $X \sim B(21, 0.85)$, independence of departure times;
(ii) (*A*) 0.0329 (3 s.f.), (*B*) 0.241 (3 s.f.), (*C*) 0.368 (3 s.f.);
(iii) H_0: $p = 0.85$, H_1: $p < 0.85$, $0.0617 > 0.05$, conclusion