RECOGNISING ACHIEVEMENT

Oxford Cambridge and RSA Examinations

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2614/1

Statistics 2

Wednesday **16 JANUARY 2002** 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 approximate 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 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 60.

This question paper consists of 3 printed pages and 1 blank page.

[Turn over

1 A football reporter claims that there is a positive association between the position of a team in its division and the attendance at its next home game. On a Saturday during a season he records the position and attendance for the eleven home teams in Division One, as given in the following table.

Team	Position	Attendance
Barnsley	9th	14 831
Birmingham	4th	17 191
Bolton	3rd	15 585
Burnley	7th	16 107
Grimsby	23rd	4 911
Portsmouth	17th	13 376
Preston	6th	14 511
Sheffield United	8th	12 921
Watford	2nd	17 488
West Bromwich	5th	16 511
Wimbledon	11th	9 030

(i) Rank the positions and the attendances. Carry out an appropriate hypothesis test at the 1% level to test the reporter's claim. State your hypotheses and conclusions clearly, justifying the form of the alternative hypothesis. [10]

(ii) State an assumption about the sample data for the test to be valid. Explain whether or not you think it is appropriate in this case. [2]

(iii) The reporter concludes that "to increase the attendance at matches, all a team has got to do is climb to a higher position in the division". Comment critically on the reporter's conclusion. [3]

2 A supermarket takes delivery of bags of potatoes with nominal weight 5 kg. A large number of such bags are weighed with the result that the mean weight is 5.5 kg and 10% of the bags are below nominal weight. You may assume that the weights, X kg, of bags of potatoes are modelled by the Normal distribution $N(\mu, \sigma^2)$.

(i) Illustrate the information on a diagram and show that an estimate of the standard deviation is about 0.4 kg. [5]

(ii) Taking $\mu = 5.5$ and $\sigma = 0.4$, find the probability that a bag chosen at random weighs between 5.3 kg and 5.8 kg. [3]

(iii) Assuming the mean remains the same, find the required standard deviation in order that at most 2% of bags are below nominal weight. [3]

(iv) A customer chooses two of the original bags at random. You may assume that the total weight is modelled by the Normal distribution $N(2\mu, 2\sigma^2)$, where $\mu = 5.5$ and $\sigma = 0.4$. Find the probability that she gets a total weight of at least 10 kg of potatoes. [4]

3 The manufacturers of Jupiter Jellybabies have launched a promotion to boost sales. One per cent of bags, chosen at random, contain a prize. A school tuck shop takes delivery of 500 bags of Jupiter Jellybabies. Let X represent the number of bags in the delivery which contain a prize.

(i) State clearly the distribution which X takes. [2]

(ii) Using a Poisson approximating distribution, find $\mathrm{P}(3 \leqslant X \leqslant 7)$. [3]

The values of the prizes are in the following proportions.

Value of prize	£10	£100	£1000
Proportion	90%	9%	1%

(iii) Suppose the tuck shop receives 5 bags which contain prizes. Find the probability that at least one of these prizes has value £1000. [2]

A supermarket orders a consignment of 7500 bags of Jupiter Jellybabies.

(iv) Using a suitable approximating distribution, find the probability that this consignment contains at least 70 but not more than 80 prizes. [4]

(v) What is the expected total value of the prizes in the consignment? [4]

4 In a statistical survey of cars coming into a city centre during the morning rush hour, the number of occupants, X, is modelled by the probability distribution

$$\mathrm{P}(X = r) = \frac{k}{r} \quad \text{for } r = 1, 2, 3, 4.$$

(i) Tabulate the probability distribution and determine the value of k. [2]

(ii) Illustrate the distribution using a suitable diagram. [2]

(iii) Calculate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. [4]

(iv) Calculate the probability that, for five consecutive cars entering the city centre, at least two have no occupants other than the driver. [4]

During a campaign by the city council to reduce the volume of traffic, pairs of single occupant drivers are put in touch with each other and encouraged to share their journeys.

(v) Without further calculations, state, with reasons, the effect on each of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. [3]

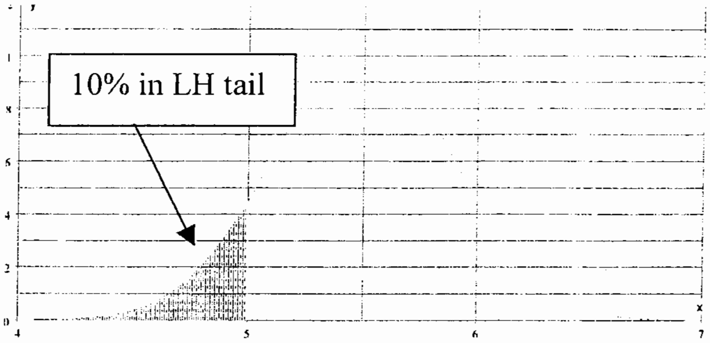
Mark Scheme

January 2002 2614 MEI Statistics 2

Question 1

<table>
<tr>
<td>(i)</td>
<td>
<table>
<tr><th><em>Team</em></th><th><em>Pos. rank</em></th><th><em>Att. Rank</em></th><th>d^2</th></tr>
<tr><td>Barnsley</td><td>8</td><td>6</td><td>4</td></tr>
<tr><td>Birmingham</td><td>3</td><td>2</td><td>1</td></tr>
<tr><td>Bolton</td><td>2</td><td>5</td><td>9</td></tr>
<tr><td>Burnley</td><td>6</td><td>4</td><td>4</td></tr>
<tr><td>Grimsby</td><td>11</td><td>11</td><td>0</td></tr>
<tr><td>Portsmouth</td><td>10</td><td>8</td><td>4</td></tr>
<tr><td>Preston</td><td>5</td><td>7</td><td>4</td></tr>
<tr><td>Sheffield United</td><td>7</td><td>9</td><td>4</td></tr>
<tr><td>Watford</td><td>1</td><td>1</td><td>0</td></tr>
<tr><td>West Bromwich</td><td>4</td><td>3</td><td>1</td></tr>
<tr><td>Wimbledon</td><td>9</td><td>10</td><td>1</td></tr>
</table>
$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 32}{11\times 120}$
$= 0.85$ (to 2 s.f.) [<em>or</em> 0.854 or 0.855 to 3 s.f.]
H$_0$: $\rho = 0$ and H$_1$: $\rho > 0$
Looking for positive association:
critical value at 1% level is 0.7091
Since 0.85 [<em>or</em> 0.854 or 0.855] &gt; 0.7091, there is sufficient evidence to reject H$_0$,
i.e. conclude that there is positive association between league position and attendance (at the next match).
</td>
<td>B1 for ranks
B1 for d^2 f.t. their ranks
M1 for r_s
A1 f.t. for $|r_s| < 1$
B1 for H$_0$, B1 for H$_1$
E1
B1 for ± 0.7091
M1 for comparison with c.v., provided $|r_s| < 1$
A1 for conclusion in words</td>
<td><strong>10</strong></td>
</tr>
<tr>
<td>(ii)</td>
<td>Modelling assumption is that the sample is take at random from the population (e.g. of all such pairs of positions and attendances for the season);
e.g. not really a random sample, just a random Saturday; could be representative of the season, however.</td>
<td>B1 for <em>random</em> or <em>representative</em> sample
E1 for explanation [dependant on B1]</td>
<td><strong>2</strong></td>
</tr>
<tr>
<td>(iii)</td>
<td>A higher position might not <em>cause</em> a higher attendance;
there will certainly be <em>other factors</em>,
such as population density in the surrounding area, counter-attractions, etc.</td>
<td>B1 for ref. to <em>cause</em>
B1 for <em>other factors</em>
E1 for <em>example</em> of other factors or other good reason</td>
<td><strong>3</strong></td>
</tr>
<tr>
<td></td><td></td><td></td><td><strong>15</strong></td>
</tr>
</table>

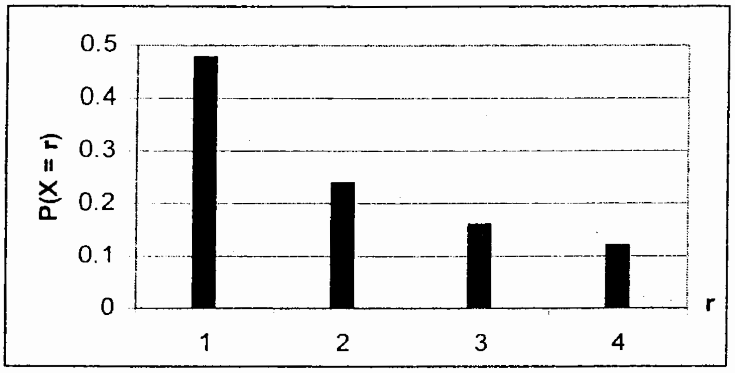
Question 2

(i)	 Taking $\mu = 5.5$, let standard deviation = σ, then: $-1.282\sigma + \mu = 5$ $\Rightarrow \quad \sigma = \dfrac{5.5-5.0}{1.282} \;(= 0.39) \;\approx\; 0.4$ (kg.)	G1 for shape and mode G1 for LH tail, suitably labelled B1 for ± 1.282 seen M1 for setting up equation with negative z-value A1 for $\dfrac{5.5-5.0}{1.282}$	**5**
(ii)	$P(5.3 < X < 5.8) = P\left(\dfrac{5.3-5.5}{0.4} < Z < \dfrac{5.8-5.5}{0.4}\right)$ $= P(-0.5 < Z < 0.75)$ $= 0.7734 - (1 - 0.6915)$ $= 0.4649$ *or* 0.465 (to 3 s.f.) *or* 0.46 (to 2 s.f.)	M1 for standardizing M1 for probability calc. A1 (to at least 2 s.f.)	**3**
(iii)	Let σ be the standard deviation such that at most 2% of bags are underweight: $P(X < 5) = 0.02 \;\Rightarrow\; P\left(Z < \dfrac{5.0-5.5}{\sigma}\right) = 0.02$ but from tables: $P(Z < -2.054) = 2$ hence $\dfrac{5.0-5.5}{\sigma} = -2.054$ $\Rightarrow \quad \sigma = 0.243$ (to 3.s.f.) *or* 0.24 (to 2 s.f.) (kg)	M1 for 1st statement M1 for equation A1 (to at least 2 s.f.)	**3**
(iv)	Taking $\mu = 2 \times 5.5 = 11$ and $\sigma^2 = 2 \times 0.4^2 = 0.32 \;\Rightarrow\; \sigma = 0.5657$ $P(Y \geq 10) = P(Z \geq -1.768) = 0.9615$ or 0.962 (to 3 s.f.) *or* $P(Z \geq -1.77) = 0.9616$ or 0.962 (to 3 s.f.)	B1 for new mean B1 for new variance M1 for probability calc. A1 for value **cao** (to at least 3 s.f.)	**4**
			15

Question 3

(i)	Distribution: $X \sim B(500, 0.01)$	B1 for *binomial* B1 for both parameters	**2**
(ii)	Using $\lambda = 500 \times 0.01 = 5$: $P(3 \le X \le 7) = P(X \le 7) - P(X \le 2)$ $= 0.8666 - 0.1247$ [using tables] $= 0.7419$ *or* 0.742 (to 3 s.f.) *or* 0.74 (to 2 s.f.)	B1 for λ; their np (SOI) M1 for statement (SOI) A1 for calculation	**3**
(iii)	P(at least one of these contains a £1000 certificate) $= 1 - 0.99^5 = 0.049$ (to 2 s.f.)	M1 for probability calc. A1 (to at least 2 s.f.)	**2**
(iv)	For $n = 7500$ and $p = 0.01$ use a Normal approximation: $Y \sim N(75, 74.25)$; $P(69.5 < Y < 80.5)$ $= P\left(\frac{69.5-75}{\sqrt{74.25}} < Z < \frac{80.5-75}{\sqrt{74.25}}\right)$ $= P(-0.638 < Z < 0.638)$ $= 2 \times (0.7384 - 0.5)$ *or* $0.7384 - 0.2616$ $= 0.4768$ *or* 0.477 (to 3 s.f.) *Allow equivalent solution using* $N(75, 75)$: $P(69.5 < Y < 80.5) = 0.475$ (to 3 s.f.)	B1 for Normal approx. B1 for both continuity corrections M1 for probability A1 for value **cao** (to at least 3 s.f.)	**4**
(v)	Expected prize value $= £(10 \times 0.9 + 100 \times 0.09 + 1000 \times 0.01)$ [= £28] [Expected number of prizes = 75] Hence expected value of prizes = £(75 × 28) = £2100 ***or*** Expected prize per packet $= £(0 \times 0.99 + 10 \times 0.009 + 100 \times 0.0009 + 1000 \times 0.000)$ [= £0.28] Hence expected value of prizes = £(7500 × 0.28) = £2100	M1 for expected p. v. A1 M1 for product of 75 and "their 28" A1 *or* M1 for expected p. p. p. A1 M1 A1	**4**
			15

Question 4

(i)	$\begin{array}{c\|cccc} r & 1 & 2 & 3 & 4 \\ \hline \mathrm{P}(X=r) & k & \tfrac{1}{2}k & \tfrac{1}{3}k & \tfrac{1}{4}k \end{array}$ Now $k + \frac{1}{2}k + \frac{1}{3}k + \frac{1}{4}k = 1$ $\Rightarrow \quad k = \frac{12}{25} = 0.48$	B1 for tabulation (SOI) B1 for value of k	**2**
(ii)		G1 for horizontal scale and attempt at representing data G1 for lines in proportion **cao**	**2**
(iii)	$\mathrm{E}(X) = 1 \times 0.48 + 2 \times 0.24 + 3 \times 0.16 + 4 \times 0.12$ $= 1.92$ $\mathrm{E}(X^2) = 1 \times 0.48 + 4 \times 0.24 + 9 \times 0.16 + 16 \times 0.12$ Hence $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$ $= 4.8 - 1.92^2$ $= 4.8 - 3.6864$ $= 1.1136$ *or* 1.11 (to 3 s.f.)	B1 for E(X) [provided $\Sigma p = 1$] M1 for E(X^2) [$\Sigma p = 1$] M1 for positive variance A1 **cao** (to at least 3 s.f.)	**4**
(iv)	P(at least two have driver only) = 1 – P(0 or 1 have driver only) $= 1 - (0.52^5 + 5 \times 0.52^4 \times 0.48)$ = 1 – 0.2135 = 0.787 (to 3 s.f.) *or* 0.79 (to 2 s.f.)	M1 for P(1) M1 for P(0 or 1) M1 for 1 – their P(0 or 1) A1 **cao**	**4**
(v)	If single occupant car drivers share, then there will be a smaller proportion of cars in the $X = 1$ category and a greater proportion in the $X = 2$ category, which will increase E(X) and reduce Var (X)	E1 for change in proportions B1 for effect on E(X) B1 for effect on Var(X)	**3**
			15

Examiner’s Report

Statistics 2 (2614)

General Comments

The overall standard was slightly better than in recent examinations, with a larger than usual number of candidates achieving marks above 45 and rather less weak candidates than normal. Most candidates were able to achieve a good degree of success in each question, with better attempts at questions on the Normal distribution and Poisson distribution (Questions 2 and 3) than in recent examinations. As ever it was the discussion and interpretation questions which proved the stumbling block for even the most successful candidates, who frequently were only able to score one or two of the five such marks in Question 1, despite losing only two or three marks in the whole of the rest of the paper. Premature approximation was occasionally seen, leading to unnecessary loss of marks. In particular, candidates should be advised against rounding of z-values to 2 decimal places, prior to looking them up in the tables.

Comments on Individual Questions

Question 1 (Bivariate data; rank correlation: league positions and match attendances for football teams)

(i) The majority of candidates were able to rank the data correctly and go on to calculate Spearman's coefficient of rank correlation, and full credit was available to candidates who ranked the position data correctly in either order. A variety of errors were seen in the application of the formula for r_s. A few candidates ignored the instruction to rank the data, instead calculating Pearson's product moment correlation coefficient.

Most candidates were able to state hypotheses correctly in terms of ρ or in some cases ρ_s, but few justified the form of the alternative hypothesis as being a result of the reporter's claim. However a significant number of candidates stated the hypotheses in words or used incorrect notation, thus losing at least one of the two easy marks available. Confusion arose for candidates who had ranked the positions from highest (numerically) to lowest and thus arrived a negative value of r_s. Such candidates sometimes gave the alternative hypothesis as $H_1: \rho < 0$, which gained credit only if accompanied by a clear explanation of how this could be in line with the reporter's claim.

Most candidates knew how to carry out the hypothesis test and, although candidates were usually able to read the tables correctly, many lost the final mark because they did not interpret their conclusion in context. Candidates who simply conclude 'accept H_1, there is positive correlation' lose credit. They must mention the context; in this case stating for example that 'there is positive correlation*between the league position and the attendance*'.

(ii) The essential requirement was for the sample to be 'random' or 'representative'. Independence was an often quoted property, which did not gain credit. Relatively few candidates gained credit and in many cases answers were clearly discussing the nature of the variables, not of the sample. A surprising number thought that bivariate Normality was a requirement, despite their earlier use of rank correlation.

(iii) Relatively few candidates were aware that this was testing their understanding of the difference between correlation and causation. A number of factors which could cause variation in *one* of the variables were identified, but rarely were these cited as a possible *cause* of the association between the two variables.

[(i) r_s = 0.855; Ho: $\rho = 0$, H_1: $\rho > 0$, 1% c.v. = 0.7091, hence reject H_0;
(ii) comments on suitability of test; (iii) comments on reporter's conclusion]

Question 2 (Normal distribution; weights of bags of potatoes)

(i) The graph was usually correctly drawn, and candidates who found it difficult to draw a correctly shaped Normal curve were not penalised if they made a reasonable attempt. It was pleasing to see that most candidates attempted to calculate the standard deviation from the given data, rather than using $\sigma = 0.4$ to establish that 10% are below nominal weight. Most were also able to handle the use of inverse tables with a ϕ value below 0.5, with some good use of correct notation. Candidates who set up an equation in σ which led to a negative value of σ, but who subsequently conveniently 'lost' the negative sign, were penalised.

(ii) Most candidates achieved the full three marks on this question, although errors such as the introduction of a spurious continuity correction or errors in handling the cumulative probabilities read from the tables were occasionally seen.

(iii) Candidates who were successful in establishing the value of σ in part (i) were usually also successful here, although again a number of them set up an equation leading to a negative standard deviation, but then solved this and ended up with a positive value of σ. Conversion of 2% into a decimal caused difficulty to some, who used 0.2 rather than 0.02.

(iv) A good number of correct solutions were seen, and candidates who actually calculated the numerical value of $2\sigma^2 = 0.32$ and then used this as their variance usually gained full marks. Unfortunately many took the incorrect shortcut of simply doubling the standard deviation, thus denying themselves two of the four marks available. Others tried to standardise without square rooting their variance. A small proportion of candidates found the wrong tail.

[(i) diagram, 0.39 kg (to 2 d.p.); (ii) 0.465; (iii) 0.243 kg; (iv) 0.9615]

Question 3 (Poisson and Normal approximations to the binomial; prizes in bags of sweets)

(i) Generally well done, with the commonest error being to think that the variable had a Poisson distribution.

(ii) Almost all candidates were able to find the value of the mean $\lambda = 5$, but a very large number could not apply the tables correctly to the problem and tried to find $P(3 \leqslant X \leqslant 7)$ by using $P(X \leqslant 7) - P(X \leqslant 3)$, thus losing two of the three marks available. A small number summed the five Poisson point probabilities, usually reaching the correct answer.

(iii) Many correct solutions were seen, but equally many candidates made errors, either by calculating $P(X = 1)$ or using $P(X \geqslant 1) = 1 - P(X \leqslant 1)$ or using a Poisson approximation with $n = 5$, none of which gained credit.

(iv) Correct Normal approximations were often seen here, but the continuity correction was then either omitted, or more often applied correctly at one tail but incorrectly at the other. Most candidates who identified the correct Normal approximation were able to handle the cumulative probabilities read from the tables without errors, scoring at least three of the four marks available. No penalty was applied to candidates who used a variance of 75.

(v) Many candidates scored well in this part despite difficulties earlier in the question. The most common error was the use of an incorrect multiplication factor, often 7500, applied to the correct terms $10 \times 0.9 + 100 \times 0.09 + 1000 \times 0.01$, either before or after the summation of these had been completed. Some candidates found the expected number of prizes of each value, but unfortunately then rounded them to whole numbers, before correctly multiplying them by the respective prize values.

[(i) $X \sim B(500, 0.001)$; (ii) 0.7419; (iii) 0.049; (iv) 0.477; (v) £2100]

Question 4 (Discrete random variable; number of occupants of cars in a traffic survey)

(i) Although many candidates handled this very well, a disappointingly large number were unable to solve their equation correctly, resulting in a variety of incorrect k values, often 0.1, but occasionally greater than 1. A brief check of their probabilities would have indicated that they did not sum to one.

(ii) Most candidates drew a vertical line diagram as was required to illustrate a discrete probability distribution, but some were careless with the heights of their lines. A sketch of a discrete probability distribution in Statistics should be roughly to scale, unlike that of a graph in Pure Mathematics, where scale may be less relevant. There is also still a misconception amongst some candidates that such a distribution should be illustrated by means of a bar chart, or worse still a curve.

(iii) The vast majority of candidates knew how to find the mean and variance, although some lost marks due to their probabilities having a sum not equal to one. It was pleasing to see that very few candidates forgot to subtract $[E(X)]^2$ from $E(X^2)$.

(iv) Candidates found this part to be very demanding. Many had little idea where to start, and even where candidates recognised (explicitly or implicitly) that a binomial distribution was appropriate, there were many errors. Successful candidates usually calculated $P(X \geqslant 2) = 1 - P(X \leqslant 1)$, rather than adding the probabilities of 2, 3, 4 and 5 cars having only the driver. The many errors seen included attempts to calculate $P(X \geqslant 2)$ using $1 - P(X \leqslant 2)$ or $1 - P(X = 1)$; calculating $P(X = 2)$ rather than $P(X \geqslant 2)$; interchanging 0.48 and 0.52; omission of the binomial coefficients.

(v) The majority of candidates correctly stated that the mean would increase, although their reasoning was often rather vague, if not wrong. The effect on the variance was less clear to many candidates, who tried to assess the effect on $E(X^2)$ and $[E(X)]^2$, often concluding that both would rise and so the variance would not change. Consideration instead of the amount of variation of the new distribution as compared to the old was the approach which usually led to success. There were also some spurious attempts to apply formulae such as $E(aX + bY) = aE(X) + bE(Y)$.

[(i) $25k = 12 \Rightarrow k = 0.48$; (ii) vertical line chart; (iii) $E(X) = 1.92$, $\text{Var}(X) = 1.11$; (iv) 0.787; (v) increase $E(X)$ and reduce $\text{Var}(X)$]