OCR
RECOGNISING ACHIEVEMENT

General Certificate of Education
Advanced Supplementary (AS) and Advanced Level
former Oxford and Cambridge modular syllabus

MEI STRUCTURED MATHEMATICS — 5514/1

Statistics 2

Tuesday **16 JANUARY 2001** Afternoon 1 hour 20 minutes

Additional materials:
- Answer paper
- Graph paper
- Students' Handbook

TIME 1 hour 20 minutes

INSTRUCTIONS TO CANDIDATES

Write your name, Centre number and candidate number in the spaces provided on the answer paper/ answer booklet.

Answer **all** questions.

Write your answers on the separate answer paper provided.

If you use more than one sheet of paper, fasten the sheets together.

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 sufficient detail of the working is shown on the answer paper to indicate that a correct method is being used.

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

1 A car manufacturer is introducing a new model. The car is tested for fuel economy three times at each of four different speeds. The values of the fuel economy, y miles per gallon, at each of the speeds, x miles per hour, are displayed in the following table.

x	40	40	40	50	50	50	60	60	60	70	70	70
y	52.7	53.8	54.5	48.1	49.7	51.3	43.3	41.1	48.0	37.5	42.0	44.7

$$n = 12, \quad \sum x = 660, \quad \sum y = 566.7, \quad \sum x^2 = 37\,800, \quad \sum xy = 30\,533.$$

(i) Represent the data by a scatter diagram, drawn on graph paper. [2]

(ii) Calculate the equation of the regression line of y on x and plot it on your scatter diagram. [6]

(iii) Hence predict the fuel economy of the car at speeds of

(*A*) 45 mph,

(*B*) 65 mph. [3]

(iv) Use your scatter diagram to compare the reliability of your predictions in part **(iii)**.

What do your comments suggest about the validity of a least squares regression line for this data set? [4]

2 Scores on an IQ test are modelled by the Normal distribution with mean 100 and standard deviation 15. The scores are reported to the nearest integer.

(i) Find the probability that a person chosen at random scores

(*A*) exactly 105,

(*B*) more than 110. [5]

(ii) Only people with IQs in the top 2.5% are admitted to the organisation ***BRAIN***. What is the minimum score for admission? [3]

(iii) Find the probability that, in a random sample of 20 people, exactly 6 score more than 110. [3]

(iv) Find the probability that, in a random sample of 200 people, at least 60 score more than 110. [4]

3 An airline regularly sells more seats for its early morning flight from London to Paris than are available. On average, 5% of customers who have purchased tickets do not turn up. For this flight, the airline always sells 108 tickets. Let X represent the number of customers who do not turn up for this flight.

(i) State the distribution of X, giving one assumption you must make for it to be appropriate. [3]

There is room for 104 passengers on the flight. For the rest of the question use a suitable Poisson approximation.

(ii) Find the probability that

(*A*) there are exactly three empty seats on Monday's flight, [3]

(*B*) Tuesday's flight is full, [3]

(*C*) from Monday to Friday inclusive the flight is full on just one day. [2]

For every customer who turns up and finds no seat is available, the airline pays compensation of £250.

(iii) Calculate the expected amount of compensation per flight. [4]

4 A discrete random variable X has probability distribution defined by

$$P(X = r) = k(3r^2 - 3r + 1) \quad \text{for } r = 1, 2, 3, 4, 5, 6.$$

(i) Copy and complete the following table.

r	1	2	3	4	5	6
$P(X = r)$	k					$91k$

Hence show that $k = \frac{1}{216}$ and illustrate the distribution on a sketch. [4]

In one turn of the game of *Polypoly* a player throws three ordinary dice, the score being the largest of the numbers appearing face up. The score, X, is given by the probability distribution defined above.

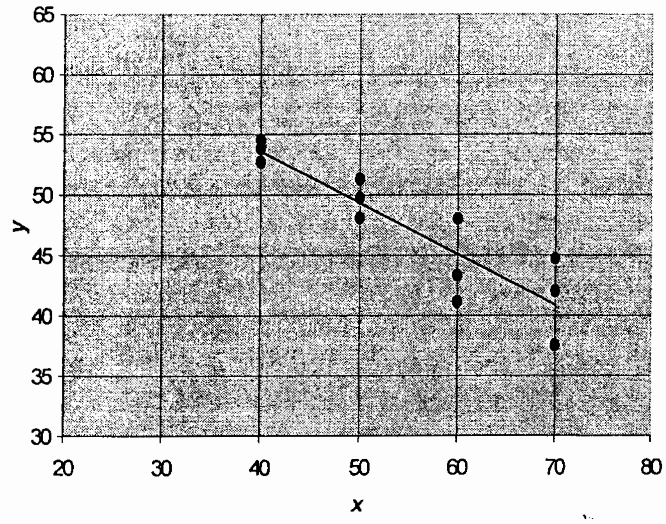
(ii) Find the mean and standard deviation of the score for one turn. [4]

(iii) Find the probability that the player will score a total of exactly 10 in two turns. [4]

(iv) Use a probability argument to verify that $P(X = 2) = \frac{7}{216}$. [3]

Mark Scheme

Question 1

(i)		G1 for linear scales G1 for visually correct points *Allow 1 or 2 errors* [> 2 *errors deduct* 1] *Accept axes interchanged if clearly labelled*	2
(ii)	$\bar{x} = 55,\ \bar{y} = 47.2(25)$ $m = \dfrac{30533 - 12 \times 55 \times 47.225}{37800 - 12 \times 55^2} = \dfrac{-635.5}{1500}$ *or* $m = \dfrac{30533/12 - 55 \times 47.225}{37800/12 - 55^2} = \dfrac{-52.958\dot{3}}{125}$ $y - \bar{y} = m(x - \bar{x})$ $\Rightarrow\ y - 47.225 = -0.424\,(x - 55)$ $\Rightarrow\ y = 70.527 - 0.424x$ *Allow ranges*: [69.896, 70.66] and [−0.426, −0.4126]	B1 for use of $(\bar{x}, \bar{y})$ M1 for attempt at *m* M1 A1 **cao** for equation G1 for line through $(\bar{x}, \bar{y})$ G1 for line with *negative* gradient *m*	6
(iii)	(*A*) When $x = 45$, $y = 70.527 - 0.424 \times 45$ $= 51.447$, i.e. 51.4(5) mpg (*B*) When $x = 65$, $y = 70.527 - 0.424 \times 65$ $= 42.967$, i.e. 43.0 mpg	M1 for estimates *by calculation or read from the graph* A1, A1 for *positive* mpg [dep. on *negative m*]	3
(iv)	Prediction of mpg at 45 mph liable to be more reliable that the prediction at 65 mph, since regression line is a better fit at lower speed (or worse fit at higher speed). As speeds increase the recorded mpg values have a wider spread, thus the use of a regression line for predictions, especially at the upper end, is dubious.	B1 E1 E1 for reference to shape of scatter E1 for conclusion on general reliability	4
			15

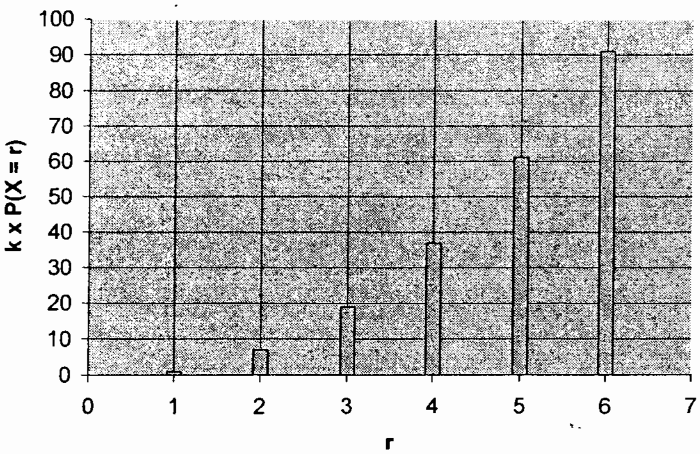
Question 2

(i)	$X \sim N(100, 15^2)$ (*A*) $P(104.5 < X < 105.5)$ $= P(0.3 < Z < 0.367)$ $= 0.6431 - 0.6179 = 0.025$ (2 s.f.) (B) $P(X > 110.5) = P(Z > 0.7)$ $= 1 - 0.7580 = 0.242$ (3 s.f.) *or* $P(X > 110) = P(Z > 2/3) = [0.252, 0.253]$	M1 for cont. correction [interval width = 1 inc 105] M1 for prob. calculation A1 **cao** M1 for upper tail calculation, A1	**5**
(ii)	$100 + 1.96 \times 15 = 129.4$ Hence minimum score = 130	B1 for ± 1.96 M1 for calc. with +ve z A1 inc. rounding up to nearest integer	**3**
(iii)	$^{20}C_6 \times 0.242^6 \times 0.758^{14} = 0.161$(3 s.f.) *Note:* $p = 0.252$ gives 0.170 (3 s.f.); $p = 0.253$ gives 0.171 (3 s.f.)	M1 for their "0.242^6 " M1 for binomial expression A1	**3**
(iv)	Using the approx. $X \sim N(np, npq) = N(48.4, 36.6872)$: P(at least 60 people score > 110) $= P(X > 59.5)$ $= P\left(Z > \frac{59.5 - 48.4}{\sqrt{36.6872}}\right) = 1 - P(Z < 1.833)$ $= 1 - 0.9666 = 0.0334$ *or* $P(X > 60) = 0.0277$ *Note:* Using $p = 0.252$ gives $N(50.4, 37.6992) \to 0.0692$ Using $p = 0.253$ gives $N(50.6, 37.7982) \to 0.0739$ ***Alternative solution:*** Exact answer using Binomial distribution from calculator: Using $p = 0.242$ gives $B(200, 0.242) \to 0.0358$ Using $p = 0.252$ gives $B(200, 0.252) \to 0.0711$ Using $p = 0.253$ gives $B(200, 0.253) \to 0.0757$	B1 for Normal approximation **soi** M1 for standardisation M1 for probability calculation A1 **cao** as given or in interval [0.0692, 0.0739] f.t. *must include correct continuity correction* M3 (by implication) A1 **cao** as given or in interval [0.0711, 0.0757]	**4**
			15

Question 3

(i)	Distribution: $X \sim B(108, 0.05)$ Assume independence of customers not turning up	B1 for "binomial" B1 for parameters E1 for explanation	3
(ii)	$\lambda = np = 108 \times 0.05 = 5.4$ *Using tables*: (*A*) $P(X = 7) = 0.8217 - 0.7017 = 0.12$ (2 s.f.) (*B*) $P(X \leq 4) = 0.37$ (2 s.f.) (*C*) $5 \times 0.3733 \times 0.6267^4 = 0.29$ (2 s.f.)	B1 for λ value **soi** B1 for choosing 7, 4 M1 for calculation, A1 allow $P(X = 3)$ f.t. M1 for tail probability A1 **cao** M1 for binomial expression A1	8
(iii)	Cost of reimbursing disappointed customers $= £250 \times P(X = 3) + £500 \times P(X = 2) + £750 \times P(X = 1) + £1000 \times P(X = 0)$ $= £250 \times 0.1185 + £500 \times 0.0659 + £750 \times 0.0244 + £1000 \times 0.0045$ = [£85, £86]	M1 for one "£ x prob." M1 for at least 2 products with correct pairings M1 for sum of 4 products inc. numerical point prob. A1	4
			15

Question 4

(i)	<table><tr><td>r</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td></tr><tr><td>$P(X=r)$</td><td>k</td><td>$7k$</td><td>$19k$</td><td>$37k$</td><td>$61k$</td><td>$91k$</td></tr></table> $\Rightarrow k(1+7+19+37+61+91)=1 \;\Rightarrow\; k=\dfrac{1}{216}$ 	B1 for complete table **soi** B1 for probability equation or equivalent G1 for lines in proportion *allow 1 slightly out* G1 for axes and horizontal scale, *dependent on a plot*	**4**
(ii)	Mean = $\dfrac{1}{216}(1 \times 1 + 2 \times 7 + \ldots + 6 \times 91) = \dfrac{1071}{216} = 4.96$ (3 s.f.) $\Sigma r^2 P(X=r) = \dfrac{1}{216}(1^2 \times 1 + 2^2 \times 7 + \ldots + 6^2 \times 91) = \dfrac{5593}{216}$ Variance $= \dfrac{5593}{216} - \left(\dfrac{1071}{216}\right)^2 = 1.308$ $\Rightarrow$ standard deviation $= \sqrt{1.308} = 1.14$ (3 s.f.)	B1 for mean *allow fraction as shown* M1 for $\Sigma r^2 P(X=r)$ M1 for +ve variance A1 for standard dev. *f.t. from mean only*	**4**
(iii)	P(score exactly 10 in 2 turns) = P(4, 6) + P(5, 5) + P(6, 4) $\dfrac{37}{216}\times\dfrac{91}{216}+\dfrac{61}{216}\times\dfrac{61}{216}+\dfrac{91}{216}\times\dfrac{37}{216} = 0.224$ (3 s.f.)	M1 for ≥ 2 pairs **soi** M1 for a product of 2 correct probabilities M1 for sum of 3 correct products A1 **cao**	**4**
(iv)	$P(X=2) = P(X \leq 2) - P(X \leq 1) = \left(\dfrac{2}{6}\right)^3 - \left(\dfrac{1}{6}\right)^3 = \dfrac{7}{216}$ ***Or*** $P(X=2) = P(2,2,2) + 3 \times P(2,2,1) + 3 \times P(2,1,1)$ $= \dfrac{1}{216}+\dfrac{3}{216}+\dfrac{3}{216}=\dfrac{7}{216}$	M1 for cumulative probabilities M1 for calculation A1 ***Or*** B1 for $\frac{1}{6}^3$ or 6^3 M1 for list of at least 3 possibilities A1 for correct addition of terms to give answer	**3**
			15

Examiner’s Report

Statistics 2 (5514)

General Comments

Candidates produced a range of responses to the questions set, the majority demonstrating a fair to good understanding of the topics covered. In general candidates' work was clear and showed sufficient detail in calculations. Some written comments were of a poor standard and lacked appropriate or precise statistical terminology. On the whole, questions 1 and 4 were fairly well answered, whereas questions 2 and 3 were found to be more difficult. Overall, the standard of candidates' work was roughly the same as in previous sessions.

Comments on Individual Questions

Question 1 (Bivariate data; regression; car fuel economy)

There were many good responses to the first three parts of this question, showing sound understanding of basic techniques of linear regression. However, the rather unfamiliar comment section in part (iv) was misinterpreted by all but the most astute candidates.

(i) Scatter diagrams were usually drawn neatly on graph paper. Occasional errors arose through inappropriate choice of scale or misreading of chosen scales.

(ii) A small number of candidates used the normal equations, but a large majority used the formula method for the calculation of a and b in the equation $y = a + bx$. There were a large number of fully correct equations, though some premature rounding resulted in inaccuracies in the calculation of a and/or b, and a few misinterpreted the sigma notation given in the handbook and expanded the brackets incorrectly e.g. $\Sigma(x-\bar{x})(y-\bar{y}) = \Sigma xy - \overline{xy}$. It is important that candidates become familiar with the particular format used in the handbook. Some mis-plotting of the regression line, either not passing through ($\bar{x}$, $\bar{y}$) or ignoring any axis breaks, was evident.

(iii) Predictions were generally well found to a suitable accuracy, either from the regression equation or from the graph.

(iv) There were disappointingly few correct responses to this "reliability" comment. Four marks were a guide to the number of comments required. Much rote learning was regurgitated and was not applied to this situation in particular. Some candidates wrote at considerable length, resulting in contradictions. There was evidence of confusion between the terms *accurate* and *valid*; many candidates merely compared the predictions in part (iii) from the graph and from the equation.

The candidates were expected to contrast the "goodness of fit" by noting that the prediction of the fuel economy (mpg) at 45 mph was liable to be more reliable than the prediction at 65 mph, since the regression line is a better model at the lower speed. As speeds increase, the recorded mpg values have a wider spread, thus the use of a regression line for predictions, especially at the higher speeds, is dubious.

(i) scatter diagram; (ii) $y = 70.53 - 0.42x$; (iii) 51.4 mpg, 43.0 mpg;
(iv) comments on reliability of the regression line as a predictor of fuel economy.

Question 2 (Normal distribution; IQ scores)

A large number of candidates confused *score* and *reported score* and hence made repeated errors throughout this question. Allowance was made for this in the mark scheme, with alternative solutions being allowed in part (i) (*A*) and part (iii). The answers given at the end of this question reflect the expected solutions using appropriate continuity corrections.

(i) (*A*) The probability of scoring exactly 105 was often calculated as $P(X < 105)$, with candidates presumably forgetting the reference to the *reported* score. Incorrect continuity corrections were given by many candidates.
(*B*) This was tackled much better than part (*A*) but accuracy was lost by those who did not use a z-value to 3 decimal places.

(ii) An appropriate z-value was used by the majority of candidates to find a suitable score, but it was rare to see this rounded up to the next integer.

(iii) The majority of candidates made good use of the binomial distribution to find a point probability. Full credit was given for following through their probability value from part (i) (*B*).

(iv) The idea of using a Normal approximation to a binomial situation, in a question which started with the Normal distribution, seemed to be beyond most candidates. The erroneous assumption that $P(Y > 60 \mid n = 200) = P(X > 6 \mid n = 20)$ was common and wholly unacceptable. Those who used a Normal approximation usually scored most of the marks. Here the omission of the continuity correction was penalised.

(i) (*A*) 0.025, (*B*) 0.0242; (ii) 130; (iii) 0.161; (iv) 0.0334.

Question 3 (Poisson and Normal approximations to the Binomial; airline tickets)

There were many good responses, which came from careful reading of the question. However, 108 and 104 were often confused. Very few candidates presented correct solutions to part (iii).

(i) A disappointingly large number of candidates were unable to write down the precise distribution of X using conventional notation. The mention of a Poisson approximation later suggested not using a Poisson distribution here, but many candidates insisted upon it. The assumption of "independence of customers not turning up" was often truncated to "independence". No mark was given unless the appropriate context was stated.

(ii) Consistent use of 104 instead of 108 gained some credit although it was recognised that this made (*B*) easier. Most candidates spotted that a binomial distribution was needed in (*C*) and used their answer to (*B*) correctly.

(iii) A fair attempt was made by most at multiplying a probability by a cost. The main error was to confuse the random variables "number of customers who do not turn up" and "number of customers who turn up and find no seat available". The minority who made use of a probability distribution table and who explained how probabilities were found usually did very well.

(i) $X \sim B(108, 0.05)$; (ii) (*A*) 0.12, (*B*) 0.37, (*C*) 0.29; (iii) £85.

Question 4 (Discrete random variable; largest score when throwing 3 fair dice)

Candidates appeared to be familiar with the techniques requires in this question. Parts (i), (ii) and (iii) were usually answered well. Part (iv) proved to be more challenging.

(i) Virtually all candidates completed the table correctly and used the property that $\Sigma P(X = r) = 1$. A surprising number omitted a sketch of the distribution. It is difficult to say whether or not this was due to lack of time. Of those who did make a sketch, some did not appreciate that this should have been to scale, preferably drawn with a ruler. **Whilst bars, with suitable labelling, were condoned, only a vertical line chart will be accepted for such a diagram in future papers.**

(ii) It was pleasing to see so many good attempts at $E(X)$ and $SD(X)$. A very small number of errors were generated by rounding decimals prematurely and by confusing $SD(X)$ with $Var(X)$.

(iii) Again, there were a large number of good attempts at this part. Common mistakes were with the number of times to count $P(X = 5).P(X = 5)$ and $P(X = 4).P(X = 6)$.

(iv) Candidates seem to have a better understanding of what is meant by a "probability argument" than in the past. However, there were still those who fudged the issue by merely using the formula for $X = 2$, or found $1 - P(X \neq 2)$. An argument based on either subtraction of cumulative probabilities *or* enumeration of favourable outcomes together with associated probabilities was expected.

(i) $216k = 1 \Rightarrow k = \frac{1}{216}$; (ii) mean = 4.96, s.d. = 1.14; (iii) 0.224.