OCR
RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MATHEMATICS 4732

Probability & Statistics 1

Tuesday **18 JANUARY 2005** Afternoon 1 hour 30 minutes

Additional materials:
Answer booklet
Graph paper
List of Formulae (MF1)

TIME 1 hour 30 minutes

INSTRUCTIONS TO CANDIDATES

- Write your name, centre number and candidate number in the spaces provided on the answer booklet.
- Answer **all** the questions.
- Give non-exact numerical answers correct to 3 significant figures unless a different degree of accuracy is specified in the question or is clearly appropriate.
- You are permitted to use a graphical calculator in this paper.

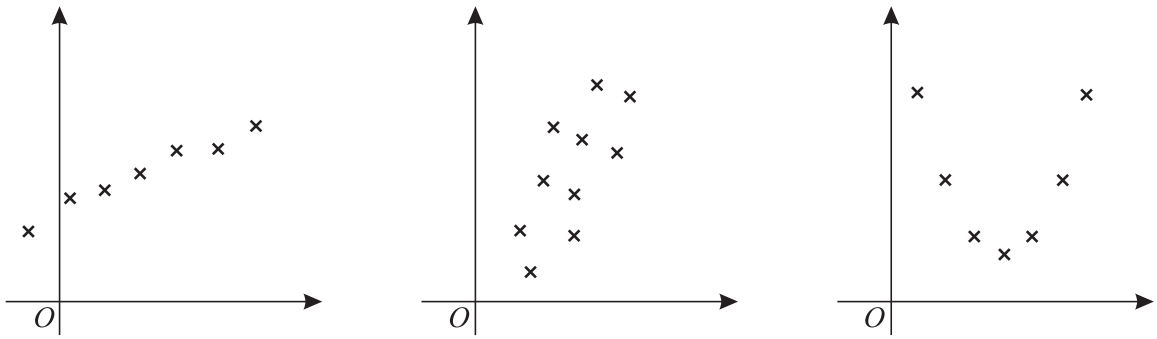
INFORMATION FOR CANDIDATES

- 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.
- Questions carrying smaller numbers of marks are printed earlier in the paper, and questions carrying larger numbers of marks later in the paper.
- **You are reminded of the need for clear presentation in your answers.**

This question paper consists of 4 printed pages.

Registered Charity Number: 1066969
[Turn over

1 The scatter diagrams below illustrate three sets of bivariate data, *A*, *B* and *C*.



Set *A* Set *B* Set *C*

State, with an explanation in each case, which of the three sets of data has

(i) the largest,

(ii) the smallest,

value of the product moment correlation coefficient. [4]

2 The back-to-back stem-and-leaf diagram below shows the number of hours of television watched per week by each of 15 boys and 15 girls.

Boys		Girls
8 7 7 6 6 4 4 3	0	0 5 5 6 6 7 7 8 8 9
2 2 0	1	0 0 4
6 5 4	2	2 7
5	3	

Key: 4 | 2 | 2 means a boy who watched 24 hours and a girl who watched 22 hours of television per week.

(i) Find the median and the quartiles of the results for the boys. [3]

(ii) Give a reason why the median might be preferred to the mean in using an average to compare the two data sets. [1]

(iii) State one advantage, and one disadvantage, of using stem-and-leaf diagrams rather than box-and-whisker plots to represent the data. [2]

3 Two commentators gave ratings out of 100 for seven sports personalities. The ratings are shown in the table below.

Personality	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Commentator I	73	76	78	65	86	82	91
Commentator II	77	78	79	80	86	89	95

(i) Calculate Spearman's rank correlation coefficient for these ratings. [5]

(ii) State what your answer tells you about the ratings given by the two commentators. [1]

4 The table below shows the probability distribution of the random variable X.

x	−2	−1	0	1	2
$\mathrm{P}(X = x)$	$\frac{1}{4}$	$\frac{1}{5}$	k	$\frac{2}{5}$	$\frac{1}{10}$

(i) Find the value of the constant k. [2]

(ii) Calculate the values of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. [5]

5 On average 1 in 20 members of the population of this country has a particular DNA feature. Members of the population are selected at random until one is found who has this feature.

(i) Find the probability that the first person to have this feature is

(a) the sixth person selected, [3]

(b) not among the first 10 people selected. [3]

(ii) Find the expected number of people selected. [2]

6 Louise and Marie play a series of tennis matches. It is given that, in any match, the probability that Louise wins the first two sets is $\frac{3}{8}$.

(i) Find the probability that, in 5 randomly chosen matches, Louise wins the first two sets in exactly 2 of the matches. [3]

It is also given that Louise and Marie are equally likely to win the first set.

(ii) Show that P(Louise wins the second set, given that she won the first set) = $\frac{3}{4}$. [2]

(iii) The probability that Marie wins the first two sets is $\frac{1}{3}$. Find

P(Marie wins the second set, given that she won the first set). [2]

7 It is known that, on average, one match box in 10 contains fewer than 42 matches. Eight boxes are selected, and the number of boxes that contain fewer than 42 matches is denoted by Y.

(i) State two conditions needed to model Y by a binomial distribution. [2]

Assume now that a binomial model is valid.

(ii) Find

(a) $P(Y = 0)$, [2]

(b) $P(Y \geqslant 2)$. [2]

(iii) On Wednesday 8 boxes are selected, and on Thursday another 8 boxes are selected. Find the probability that on one of these days the number of boxes containing fewer than 42 matches is 0, and that on the other day the number is 2 or more. [3]

8 An examination paper consists of 8 questions, of which one is on geometric distributions and one is on binomial distributions.

(i) If the 8 questions are arranged in a random order, find the probability that the question on geometric distributions is next to the question on binomial distributions. [3]

Four of the questions, including the one on geometric distributions, are worth 7 marks each, and the remaining four questions, including the one on binomial distributions, are worth 9 marks each. The 7-mark questions are the first four questions on the paper, but are arranged in random order. The 9-mark questions are the last four questions, but are arranged in random order. Find the probability that

(ii) the questions on geometric distributions and on binomial distributions are next to one another, [3]

(iii) the questions on geometric distributions and on binomial distributions are separated by at least 2 other questions. [4]

9 Five observations of bivariate data produce the following results, denoted as (x_i, y_i) for $i = 1, 2, 3, 4, 5$.

$$(13, 2.7) \qquad (13, 4.0) \qquad (18, 2.8) \qquad (23, 3.3) \qquad (23, 2.2)$$

$$[\Sigma x = 90,\ \Sigma y = 15.0,\ \Sigma x^2 = 1720,\ \Sigma y^2 = 46.86,\ \Sigma xy = 264.0.]$$

(i) Show that the regression line of y on x has gradient -0.06, and find its equation in the form $y = a + bx$. [4]

(ii) The regression line is used to estimate the value of y corresponding to $x = 20$, but the value $x = 20$ is accurate only to the nearest whole number. Calculate the difference between the largest and the smallest values that the estimated value of y could take. [3]

The numbers e_1, e_2, e_3, e_4, e_5 are defined by

$$e_i = a + bx_i - y_i \quad \text{for } i = 1, 2, 3, 4, 5.$$

(iii) The values of e_1, e_2 and e_3 are 0.6, -0.7 and 0.2 respectively. Calculate the values of e_4 and e_5. [2]

(iv) Calculate the value of $e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$ and explain the relevance of this quantity to the regression line found in part **(i)**. [2]

(v) Find the mean and the variance of e_1, e_2, e_3, e_4, e_5. [4]