EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

(formerly the Examinations of the Institute of Statisticians)

HIGHER CERTIFICATE IN STATISTICS, 2002

CERTIFICATE IN OFFICIAL STATISTICS, 2002

Paper I : Statistical Theory

Time Allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use silent, cordless, non-programmable electronic calculators.

Where a calculator is used the ***method*** *of calculation should be stated in full.*

Note that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$ *and that* ln *stands for* $\log_e$.

This examination paper consists of 9 pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. A standard pack of playing cards consists of 4 suits (Clubs, Diamonds, Hearts and Spades), each of 13 cards numbered in ascending order as 2, 3, 4, …, 10, Jack, Queen, King, Ace. A well-shuffled pack is dealt to 4 players so that each receives 13 cards.

(i) Find, correct to 4 significant figures, the probability that a given player receives no Spades.

(4)

In answering the following questions, you may leave your answers in terms of factorials. Find the probabilities that

(ii) a given player receives a Yarborough, i.e. a hand all of whose cards are in the range 2, 3, …, 9,

(4)

(iii) a given player receives exactly 6 Hearts,

(5)

(iv) each player receives at least 3 Clubs.

(7)

2. Among the inhabitants of Altamania, height, H say, is distributed Normally with mean 160 cm and standard deviation 4 cm, i.e. N(160, 16).

(i) Find the proportion of the population whose heights are within one standard deviation of the mean. Find also the proportion of the population who are more than 168 cm tall.

(5)

(ii) The Altamanian Police Force (APF) is restricted to persons who are more than 168 cm tall, and may be assumed to consist of a random sample of Altamanians satisfying this condition.

Find

(a) the median height of members of the APF,

(b) the proportion of members of the APF who are more than 170 cm tall.

(10)

(iii) Assume that the mean and standard deviation of height among members of the APF are 169.5 cm and 1.352 cm respectively. Find an approximate value for the probability that the mean height of a random sample of 25 members of the APF is more than 170 cm.

(5)

3. (i) In a tasting trial, a taster is told that she will be presented with 10 samples, each of which is equally likely to be butter or margarine independently of the rest (so that there will not necessarily be five of each), and asked to identify them. State the probability distribution of the total number of correctly identified samples on the basis of random guessing, and write down its mean and variance.

(5)

(ii) Suppose instead that the taster is told that she will be presented with 5 samples of butter and 5 of margarine, all in random order, and is asked to identify them. The taster is unable to distinguish between butter and margarine, so she chooses 5 at random from the 10 samples and states that she thinks these 5 are butter. Let X be the number of samples which she correctly identifies as butter. Show that the probability distribution $P(X = x)$ is given by:

$$p(x) = \frac{\binom{5}{x}\binom{5}{5-x}}{\binom{10}{5}}, \qquad x = 0,\ 1,\ \ldots,\ 5.$$

Evaluate $p(x)$ for $x = 0, 1, \ldots, 5$ and obtain the mean and variance of this distribution. Also write down the mean and variance of the total number of the 10 samples which are correctly identified.

(15)

4. (i) Fatal accidents occur at random at a known 'black spot', following a Poisson process with mean 4 per year. Draw a diagram of the probability mass function of X, the actual annual number of fatal accidents, and calculate the probabilities of the following events.

(a) In a given year there is at most one fatal accident.

(b) In a given 6-month period there are no fatal accidents.

(c) In a given 18-month period there are no fatal accidents.

(10)

(ii) The annual number, Y say, of non-fatal accidents at the same place may be assumed to be a Poisson random variable with mean 12, and X and Y are independent. State the distribution of $V = X + Y$ and write down the mean and variance of V. Given that there were in total 20 accidents one year, find the probability that 5 of these were fatal.

(5)

(iii) State the distribution of W, the total number of accidents in a given 4-year period. Use a suitable approximation to calculate the probability that there will be more than 70 accidents in the next 4 years.

(5)

Turn over

5. State *Bayes' Theorem* and explain its use in practice.

(5)

A signal corps has to use an information channel, which is faulty. All messages are sent as a sequence of six binary digits. The receiver knows that the message is one of the following: 'Advance' (A), or 'Retreat' (R), or 'Stay where you are' (S). From past experience he expects these messages in the respective ratios 1:4:5. The three messages are sent as

A: 0 1 0 1 0 0; R: 0 1 1 0 1 1; S: 1 0 1 0 0 1.

Independently for each character in the message, the fault causes '0' to be sent in place of '1' with probability p, and '1' to be sent in place of '0' with equal probability p, the probability that any given character is transmitted correctly being $1 - p$. The message is received as 0 1 1 1 0 0. Show that

$$P(011100 \text{ is received} \mid \text{A: } 010100 \text{ is sent}) = p(1-p)^5$$

and obtain similar expressions for

$$P(011100 \text{ is received} \mid \text{R: } 011011 \text{ is sent})$$

and

$$P(011100 \text{ is received} \mid \text{S: } 101001 \text{ is sent}).$$

(5)

Deduce that

$$P(\text{A: } 010100 \text{ is sent} \mid 011100 \text{ is received}) = \frac{(1-p)^3}{(1-p)^3 + 4p^2(1-p) + 5p^3},$$

and write down similar expressions for

$$P(\text{R: } 011011 \text{ is sent} \mid 011100 \text{ is received})$$

and

$$P(\text{S: } 101001 \text{ is sent} \mid 011100 \text{ is received}).$$

(5)

If it is assumed that p is at most 0.1, which interpretation of the message is most likely to be correct?

(5)

6. (i) The random variable X follows the exponential distribution with rate parameter λ, so that the probability density function (pdf) of X is given by

$$f(x) = \lambda e^{-\lambda x}, \;\; x > 0, \;\lambda > 0.$$

Show that the moment generating function of X, $M_X(t)$ say, is given by

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-1}, \; t < \lambda,$$

and hence show that the mean and variance of X are given by $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$ respectively.

(10)

(ii) A random sample $x_1, \ldots, x_n$ is taken from this distribution. Obtain the likelihood, $L(\lambda)$, as a function of λ, and show that the maximum likelihood estimate (MLE) of λ is given by

$$\hat{\lambda} = \frac{1}{\bar{x}},$$

where $\bar{x}$ denotes the mean of the sample $x_1, \ldots, x_n$.

(5)

(iii) By applying the central limit theorem to the distribution of $\bar{X}$, or otherwise, explain why

$$\frac{1}{\bar{x}} - \frac{1.96}{\bar{x}\sqrt{n}} < \lambda < \frac{1}{\bar{x}} + \frac{1.96}{\bar{x}\sqrt{n}}$$

is an approximate 95% confidence interval for λ if n is large.

(5)

7. The bivariate probability distribution of the random variables X and Y is summarised in the following table.

		Y			
		0	1	2	3
X	0	k	$6k$	$9k$	$4k$
	1	$8k$	$18k$	$12k$	$2k$
	2	k	$6k$	$9k$	$4k$

(i) Find k.

(3)

(ii) Obtain the marginal distributions of X and Y.

(4)

(iii) Find the conditional distribution of X given $Y = 2$.

(3)

(iv) Calculate the correlation coefficient between X and Y.

(8)

(v) State with a reason whether or not X and Y are independent.

(2)

Turn over

8. The ages in years, x, and scores, y, of nine trainees in an aptitude test are tabulated below. Plot these data on a graph, comment on their suitability for correlation and regression analysis, and calculate the product-moment correlation coefficient, r_{xy} say.

(7)

Trainee	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Age, x	16	16	17	19	19	19	20	20	25
Score, y	46	53	50	48	53	55	50	53	69

Note: $\sum x^2 = 3309$, $\sum y^2 = 25633$ and $\sum xy = 9183$.

Fit a simple linear regression of score on age to these data, and use the fitted line to estimate the average score for a trainee aged

(a) 18,

(b) 26.

(6)

You are now told that trainee I is the only trainee to have experienced higher education. Recalculate the correlation and regression without this person, and comment on your results.

(7)