EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA, 1999

Statistical Theory and Methods I

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* ${}^nC_r$ *and that* ln *stands for* $\log_e$.

This examination paper consists of 8 printed pages. This front cover is page 1. The reverse of the front cover, which is intentionally left blank, is page 2. Question 1 starts on page 3.

There are 8 questions altogether in the paper.

1. State *Bayes' Theorem*. (4)

A disorder of the pancreas, known as acute pancreatitis, occurs in two distinct forms, Type A and Type G. 60% of all cases of acute pancreatitis are of Type A, the remainder are of Type G. The causes of the two forms of the disorder, and hence the required treatment, are quite different, but it is not easy to distinguish between them by clinical examination alone. Consequently, a biochemical test has been developed that is based on the continuous random variable X, which is the logarithm of the level of alkaline phosphate in the blood.

In patients with the Type A form of this disorder, X has a Normal distribution with mean value 5.2 and standard deviation 0.25. In patients with the Type G form, X has a Normal distribution with mean value 5.7 and standard deviation 0.20.

(i) It has been decided to diagnose a patient suffering from acute pancreatitis as Type A whenever X is less than 5.5 and as Type G otherwise. Given that a patient suffering from acute pancreatitis is diagnosed as Type A, find the conditional probability that the patient is really of Type G.
(8)

(ii) Suppose now that a patient suffering from acute pancreatitis will be diagnosed as Type A whenever X is less than some value c and as Type G otherwise. Write down an expression for the probability that a randomly-selected sufferer from acute pancreatitis is correctly diagnosed. By treating this probability as a function of c, show that it is maximised when $c = 5.5$ (approximately).
(8)

2. Suppose that X and Y are independent binomial random variables, such that

$$P(X = x) = \binom{n}{x} \theta^{x} (1-\theta)^{n-x}, \qquad x = 0,1,..., n$$

$$P(Y = y) = \binom{m}{y} \theta^{y} (1-\theta)^{m-y}, \qquad y = 0,1,..., m$$

(i) Show that the random variable $X + Y$ also has a binomial distribution. (10)

(ii) Suppose now that $X + Y$ is known to equal z, where z is a fixed integer between 0 and $n + m$. Evaluate $P(X = x \mid X + Y = z)$, for $x = 0, 1, \ldots, z$, and hence show that the conditional distribution of X is hypergeometric.
(7)

(iii) A network consists of two sub-networks, the first consisting of 10 components and the second consisting of 30 components. Each component has probability 0.1 of failing within one year, independently of all the other components. After one year, it is found that five components have failed. Find the conditional probability that exactly two components in the first sub-network have failed.
(3)

Turn over

3. (i) Suppose that X and Y are continuous random variables with joint probability density function $f(x, y)$. Let $h(X)$ be any real-valued function of X. Prove that

$$E\{h(X).Y\} = E_X\{h(X).E_Y(Y \mid X)\}.$$

(5)

(ii) Suppose now that the regression of Y on X is linear, that is

$$E_Y(Y \mid X = x) = \alpha + \beta x \qquad \text{(for all possible } x\text{).}$$

Using the result of part (i), find expressions for $E(Y)$ and $E(XY)$. Hence express α and β in terms of the unconditional expected values and variances of X and Y and the correlation between X and Y.

(10)

(iii) Suppose that, in addition to the regression of Y on X being linear, the conditional variance of Y given X is constant, that is

$$\text{var}(Y \mid X = x) = \sigma^2 \qquad \text{(for all possible } x\text{).}$$

Express σ^2 in terms of the unconditional variance of Y and the correlation between X and Y.

(5)

4. A random sample of size n is drawn from a continuous distribution with cumulative distribution function $F(x)$ and probability density function $f(x)$. The ordered values of this sample are $X_1 \le X_2 \le \ldots \le X_n$. Let $F_j(x)$ and $f_j(x)$ denote, respectively, the cumulative distribution function and probability density function of X_j ($j = 1, 2, \ldots, n$).

(i) Show that

$$F_1(x) = 1 - \{1 - F(x)\}^n.$$

(3)

(ii) For $j = 1, 2, \ldots, n$, let $b_j(x) = \binom{n}{j}\{F(x)\}^j\{1 - F(x)\}^{n-j}$. Show that

$$F_j(x) = b_j(x) + b_{j+1}(x) + \ldots + b_n(x).$$

(5)

(iii) Suppose now that a random sample of size n is drawn from a uniform distribution on the interval (0, 1). Show that, for $j = 1, 2, \ldots, n - 1$,

$$\frac{db_j(x)}{dx} = \frac{n!}{(j-1)!(n-j)!}x^{j-1}(1-x)^{n-j} - \frac{n!}{j!(n-j-1)!}x^j(1-x)^{n-j-1}.$$

(6)

(iv) Hence show that, in this case,

$$f_j(x) = \frac{n!}{(j-1)!(n-j)!}x^{j-1}(1-x)^{n-j}, \qquad j = 1, 2, \ldots, n.$$

(6)

5. Suppose that X and Y are continuous random variables, with the following probability density functions:

$$f_X(x) = \frac{\theta^\alpha x^{\alpha-1} e^{-\theta x}}{\Gamma(\alpha)}, \qquad x > 0$$

$$f_Y(y) = \frac{\theta^\beta y^{\beta-1} e^{-\theta y}}{\Gamma(\beta)}, \qquad y > 0$$

(where α, β and θ are all greater than 0 and Γ denotes the gamma function). In other words, X and Y have gamma distributions with a common scale parameter θ.

Define new random variables U and V as follows:

$$U = \frac{X}{X+Y}, \qquad V = X + Y .$$

Show that V has a gamma distribution, also with scale parameter θ, that U has a beta distribution, and that U and V are independent.

A particular job consists of two consecutive tasks, whose duration times are independent and identically distributed exponential random variables. Deduce from one of the above results the distribution of W, the proportion of time spent on the first task (i.e. the ratio of the time spent on the first task to the total time for the job).

(20)

6. (i) The continuous random variable, X, follows the exponential distribution with probability density function

$$f(x) = \theta e^{-\theta x}, \qquad x > 0,$$

where $\theta > 0$. Show that X has moment-generating function

$$M_X(t) = \frac{\theta}{\theta - t}, \qquad t < \theta .$$

Hence find the expected value and variance of X.

(9)

(ii) Suppose that X_1, X_2, ..., X_n are independently distributed, each with the exponential distribution described in part (i). Find the moment-generating function of

$$Z = \frac{\theta}{\sqrt{n}}\left(X_1 + ... + X_n\right) - \sqrt{n}$$

and find its limiting form as $n \to \infty$. [Hint: consider taking the limit of the logarithm of the moment-generating function.] By recognising the limiting moment-generating function, name the limiting distribution of Z.

(11)

7. (a) Suppose that the continuous random variable X has a Normal distribution with mean value μ and standard deviation σ. Show that the random variable

$$Z = \frac{X - \mu}{\sigma}$$

has a standard Normal distribution.

(5)

(b) The following numbers are a random sample of real numbers from a uniform distribution between 0 and 1:

0.143 0.338 0.672 0.919.

Use these values to generate four random variates from each of the following distributions, and explain carefully the method you use:

(i) the Poisson distribution with expected value 2;

(5)

(ii) the standard Normal distribution;

(5)

(iii) the Normal distribution with expected value -3 and variance 0.25;

(3)

(iv) the chi-squared distribution with one degree of freedom.

(2)

8. Two players are playing a tennis match. In a particular game, player A serves every point. The probability that player A wins any point when serving is θ, and so the probability that player B wins any point when A is serving is $1 - \theta$. You may assume that points are won and lost independently of one another.

The players each win three of the first six points, reaching a score of "Deuce". The game must now continue until one of the players has scored two points more than the other, at which stage the player with more points wins the game. Let X_j denote the difference between the total number of points scored by player A and the total number of points scored by player B on completion of the $(6 + j)$th point of the game ($j = 0, 1, 2, \ldots$). We define $X_{j+1} = X_j$ whenever $X_j = 2$ or $X_j = -2$. Write down the one-step and two-step transition matrices of a Markov Chain model for $X_1, X_2, \ldots$.

(8)

Considering the initial state of this Markov Chain (X_0), write down all the possible values of X_{2m} (where m is any positive integer). Show that, for $m = 1, 2, \ldots$

$$P(X_{2m} = 0) = 2\theta(1-\theta)P(X_{2(m-1)} = 0)$$

$$P(X_{2m} = 2) = \theta^2 P(X_{2(m-1)} = 0) + P(X_{2(m-1)} = 2).$$

(6)

Hence show that the probability that player A eventually wins the game is

$$\frac{\theta^2}{\theta^2 + (1-\theta)^2} .$$

(6)