EXAMINATIONS

2 April 2001 (pm)

Subject 101 — Statistical Modelling

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Write your surname in full, the initials of your other names and your Candidate's Number on the front of the answer booklet.*

2. *Mark allocations are shown in brackets.*

3. *Attempt all 16 questions, beginning your answer to each question on a separate sheet.*

Graph paper is required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet and this question paper.

In addition to this paper you should have available Actuarial Tables and an electronic calculator.

1 The following amounts are the sizes of claims (£) on house insurance policies for a certain type of repair.

198	221	215	209	224	210	223	215	203	210
220	200	208	212	216					

Determine the lower quartile, median, upper quartile and interquartile range of these claim amounts. [2]

2 A certain medical test either gives a positive or negative result. The positive test result is intended to indicate that a person has a particular (rare) disease, while a negative test result is intended to indicate that they do not have the disease. Suppose, however, that the test sometimes gives an incorrect result: 1 in 100 of those who do not have the disease have positive test results, and 2 in 100 of those having the disease have negative test results.

If 1 person in 1000 has the disease, calculate the probability that a person with a positive test result has the disease. [2]

3 Suppose that the occurrence of events which give rise to claims in a portfolio of motor policies can be modelled as follows: the events occur through time at random, at rate μ per hour. Then the number of events which occur in a given period of time has a Poisson distribution *(you are given this).*

Show that the time between two consecutive events occurring has an exponential distribution with mean $1/\mu$ hours. [3]

4 For a certain type of policy the probability that a policyholder will make a claim in a year is 0.001. If a random sample of 10,000 policyholders is selected, calculate an approximate value for the probability that not more than 5 will make a claim next year. [2]

5 Show that the probability generating function for a binomial (n, p) distribution is

$$G_X(t) = (1 - p + pt)^n .$$

Deduce the moment generating function. [3]

6 Let X have a normal distribution with mean μ and standard deviation σ, and let the i^{th} cumulant of the distribution of X be denoted κ_i.

Assuming the moment generating function of X, determine the values of κ_2, κ_3, and κ_4. [3]

7 The number of policies (N) in a portfolio at any one time is modelled as a Poisson random variable with mean 10.

The number of claims (X_i) arising on a policy is also modelled as a Poisson random variable with mean 2, independently for each policy and independent of N.

Determine the moment generating function for the total number of claims, $\sum_{i=1}^{N} X_i$, arising for the portfolio of policies. [2]

8 Consider two independent lives A and B. The probabilities that A and B die within a specified period are 0.1 and 0.2 respectively. If A dies you lose £50,000, whether or not B dies. If B dies you lose £30,000, whether or not A dies.

(i) Calculate the mean and standard deviation of your total losses in the period. [4]

(ii) Calculate your expected loss within the period, given that one, and only one, of A and B dies. [3]

[Total 7]

9 The movement of a stock price is modelled as follows:

In each time period, the stock either goes up 1 with probability 0.35 , stays the same with probability 0.35, or goes down 1 with probability 0.30.

The change in the stock price after 500 time periods is being considered.

(i) Assuming that changes in successive time periods are independent, explain why the normal distribution can be used as an approximate model. [1]

(ii) Calculate an approximate value for the probability that, after 500 time periods, the stock will be up by more than 20 from where it started. [4]

[Total 5]

10 Claim amounts of a certain type are modelled using a normal distribution with an unknown mean and a known standard deviation σ = £20.

For a random sample of 20 claim amounts all that is known is that 5 of them are greater than £200.

(i) Let θ be the probability that a claim amount is greater than £200. Write down the maximum likelihood estimate of θ. [1]

(ii) Determine θ in terms of μ and hence calculate the maximum likelihood estimate of μ. [3]

[Total 4]

11 Previous years' records give that the standard deviation of claim size for a certain class of policy is £75. The distribution of claim size is to be modelled as a normal random variable.

(i) Determine the minimum sample size required to estimate the mean claim size such that a 95% confidence interval is of width ±£10. [2]

(ii) Calculate the corresponding sample size such that a 99% confidence interval is of width ±£10. [2]

[Total 4]

12 In 1998 a market research organisation conducted a survey in a large city. A random sample of 400 households showed that 68 were such that at least one person in the household was a member of a health/fitness club.

(i) An estimate of the true proportion of households in the city with at least one person being a member of a health/fitness club is therefore $\frac{68}{400} = 0.17$. Calculate 95% confidence limits for the true proportion.

[2]

(ii) A similar survey was conducted in 1999 and a random sample of 400 households showed that 80 were such that at least one person in the household was a member of a health/fitness club.

Perform a one-sided test to investigate whether the true proportion increased from 1998 to 1999. In particular determine the p-value of the test and state your conclusion. [3]

[Total 5]

13 A random sample of insurance policies on the contents of private houses written by each of three companies was examined and the sum insured (y) under each policy noted. The results are summarised below (in units of £100).

Company i	*Sample size* n_i	*Sample mean* $\bar{y}_i$	*Sample variance* s_i^2
1	15	115.13	196.41
2	10	95.00	341.33
3	12	135.42	609.36

For all 37 observations: $\Sigma y = 4302$, $\Sigma y^2 = 521662$.

Consider the model

$$Y_{ij} = \mu + \tau_i + e_{ij}, \qquad i = 1,2,3\,;\, j = 1,2,\ldots,n_i$$

where Y_{ij} is the jth sum insured for company i, n_i is the number of responses available for company i, the e_{ij} are independent normal variables, each with zero mean and variance σ^2, and $\sum_i n_i \tau_i = 0$.

(i) Calculate the values of the least squares/maximum likelihood estimates of μ and τ_i, $i = 1, 2, 3$. [3]

(ii) Perform an analysis of variance to investigate whether real differences exist among the means of the sums insured under such policies issued by the three companies. [5]

[Total 8]

14 Consider a group of 1000 policyholders, all of the same age, and each of whose lives is insured under one or more policies. The following frequency distribution gives the number of claims per policyholder in 1999 for this group.

Number of claims per policyholder (i)	0	1	2	3	≥4
Number of policyholders (f_i)	826	128	39	7	0

A statistician argues that an appropriate model for the distribution of X, the number of claims per policyholder, is $X \sim$ Poisson. Under this proposal, the frequencies expected are as follows *(you are not required to verify these)*:

Number of claims per policyholder	0	1	2	3	≥4
Expected number of policyholders	796.9	180.9	20.5	1.6	0.1

A second statistician argues that a more appropriate model for the distribution of X is given by:

$$P(X = x) = p(1 - p)^x, \quad x = 0, 1, 2, \ldots$$

(i) Without doing any further calculations, comment on the first statistician's proposed model for the data. [2]

Consider the second statistician's proposed model.

(ii) Verify that the mean of the distribution of X is $(1 - p)/p$ and hence calculate the method of moments estimate of p. [4]

(Note: this estimate is also the maximum likelihood estimate.)

(iii) Verify that the frequencies expected under the second statistician's proposed model are as follows:

Number of claims per policyholder	0	1	2	3	≥4
Expected number of policyholders	815.0	150.8	27.9	5.2	1.2

[3]

(iv) (a) Test the goodness-of-fit of the second statistician's proposed model to the data, quoting the p-value of your test statistic and your conclusion.

(b) Assuming that you had been asked to test the goodness-of-fit "at the 1% level", state your conclusion.

[7]

[Total 16]

15 It has been decided to model a claim amount distribution using a gamma distribution with parameters $\alpha = 4$ and λ (unknown), that is, with density

$$f(x; \lambda) = \frac{1}{6}\ \lambda^4\, x^3\, e^{-\lambda x}: 0 < x < \infty.$$

(i) A random sample of n claim amounts, $X_1, X_2, \ldots, X_n$, is selected and it is required to estimate the parameter λ.

(a) (1) Determine the method of moments estimator of λ.

(2) Show that the maximum likelihood estimator of λ is the same as the method of moments estimator.

(b) (1) Write down the moment generating function of X_i and derive the moment generating function of $Y = 2n\lambda\bar{X}$, where $\bar{X}$ is the sample mean. Explain why the distribution of Y is χ^2_{8n}.

(2) Using $Y = 2n\lambda\bar{X}$ as a pivotal quantity, derive a 95% confidence interval for λ. [12]

(ii) A random sample of 10 claim amounts yields a sample mean of $\bar{x} = £242$ and a standard deviation of $s = £112$.

(a) Using part (i), calculate a 95% confidence interval for λ.

(b) (1) Use the Central Limit theorem to obtain an approximate 95% confidence interval for the population mean.

(2) Calculate this confidence interval and convert it into an approximate confidence interval for λ.

(3) Comment briefly on the comparison of this interval with that obtained in part (ii)(a). [6]

[Total 18]

16 The table below gives data on the lean body mass (the weight without fat) and resting metabolic rate for twelve women who were the subjects in a study of obesity. The researchers suspected that metabolic rate is related to lean body mass.

Lean body mass (kg) x	*Resting metabolic rate* y
36.1	995
54.6	1425
48.5	1396
42.0	1418
50.6	1502
42.0	1256
40.3	1189
33.1	913
42.4	1124
34.5	1052
51.1	1347
41.2	1204

$$\sum x = 516.4, \qquad \sum x^2 = 22741.34$$
$$\sum y = 14821, \qquad \sum y^2 = 18695125$$
$$\sum xy = 650264.8$$

(i) Draw a scatter plot of the resting metabolic rate against lean body mass and comment briefly on any relationship. [2]

(ii) Calculate the least squares fit regression line in which resting metabolic rate is modelled as the response and the lean body mass as the explanatory variable. [3]

(iii) Determine a 95% confidence interval for the slope coefficient of the model. State any assumptions made. [5]

(iv) Use the fitted model to construct 95% confidence intervals for the mean resting metabolic rate when:

(a) the lean body mass is 50kg
(b) the lean body mass is 75kg [4]

(v) Comment on the appropriateness of each of the confidence intervals given in (iv). [2]

[Total 16]