EXAMINATIONS

17 April 2000 (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 graph paper, Actuarial Tables and an electronic calculator.

1 Fourteen economists were asked to provide forecasts for the percentage rate of inflation for the third quarter of 2002. They produced the forecasts given below.

1.2	1.4	1.5	1.5	1.7	1.8	1.8
1.9	1.9	2.1	2.7	3.2	3.9	5.0

Calculate the median and the upper and lower quartiles of these forecasts. [2]

2 Insurance policies providing car insurance are such that the sizes of claims are normally distributed with mean £1,870 and standard deviation £610. In one month 50 claims are made. Assuming that claims are independent, calculate the probability that the total of the claim sizes is more than £100,000. [3]

3 In an investigation into the proportion (θ) of lapses in the first year of a certain type of policy, the uncertainty about θ is modelled by taking θ to have a beta distribution with parameters $\alpha = 1$ and $\beta = 9$, that is, with density

$$f(\theta) = 9(1 - \theta)^8 : 0 < \theta < 1.$$

Using this distribution, calculate the probability that θ exceeds 0.2. [2]

4 Consider the following three probability statements concerning an F variable with 6 and 12 degrees of freedom.

(a) $P(F_{6,12} > 0.250) = 0.95$

(b) $P(F_{6,12} < 4.821) = 0.99$

(c) $P(F_{6,12} < 0.130) = 0.01$

State, with reasons, whether each of these statements is true. [3]

5 An insurance company's records suggest that experienced drivers (those aged over 21) submit claims at a rate of 0.1 per year, and inexperienced drivers (those 21 years old or younger) submit claims at a rate of 0.15 per year. A driver can submit more than one claim a year. The company has 40 experienced and 20
inexperienced drivers insured with it.

The number of claims for each driver can be modelled by a Poisson distribution, and claims are independent of each other. Calculate the probability the company will receive three or fewer claims in a year. [3]

6 The number of claims which arise under a policy of a particular type in a year is to be modelled as a Poisson(λ) random variable. A random sample of 600 such policies gave rise to a total of 72 claims in 1999.

Determine the approximate probability value of this result in a test of

$$H_0 : \lambda = 0.14 \quad \text{v} \quad H_1 : \lambda < 0.14 \, .$$ [3]

7 Suppose that X and Y are continuous random variables.

Prove that $E(X) = \int_{-\infty}^{\infty} E(X \mid Y = y) f_Y(y) dy$. [3]

8 A device contains an electronic component which has a lifetime modelled by a distribution with mean 3.6 hours and standard deviation 2.6 hours. On failure a new component is automatically and instantaneously inserted as a replacement.

Consider the operation of the device with 100 such components used one after the other. Determine the approximate probability that the resulting total lifetime of the device will be greater than 400 hours. [4]

9 The discrete random variable X has the following probability function:

$$P(X = i) = 0.2 + \alpha i \quad : i = -2, -1, 0, 1, 2.$$

(i) State the possible values that α can take. [1]

(ii) Given a random sample $x_1, x_2, ..., x_n$ from this distribution, determine the method of moments estimate of α and show that this can result in inadmissible estimates (i.e. estimates outside the range of possible values of α). [4]

[Total 5]

10 Under a particular model for the evolution of the size of a population over time, the probability generating function of X_t, the size at time t, $G_{X_t}(s)$, is given by

$$G_{X_t}(s) = \left\{ \frac{s + \lambda t(1-s)}{1 + \lambda t(1-s)} \right\} \quad \text{where } G_{X_t}(s) = E(s^{X_t}) \text{ and } \lambda > 0 \, .$$

If the population dies out, it remains in this extinct state for ever.

(i) Show that the expected size of the population at any time t is 1. [3]

(ii) Show that the probability that the population has become extinct by time t is given by $\lambda t / (1 + \lambda t)$. [2]

(iii) Comment briefly on the future prospects for the population. [1]

[Total 6]

11 A charity sent eight hundred of its supporters information packs about its activities and asked for donations to help it continue its work. Two hundred were sent a pack about its work in education, two hundred a pack about its work in health care, and four hundred a pack about its anti-poverty programmes. Ninety-seven of the eight hundred people responded with donations; the breakdown is shown in the table below.

	Education	*Health*	*Poverty*	*Total*
Donate	26	31	40	97
Don't donate	174	169	360	703
Total	200	200	400	800

Perform a χ^2 test on this table to investigate whether the proportions who send donations are affected by the type of pack received. [5]

12 A random sample of 200 pairs of observations (x, y) from a discrete bivariate distribution (X, Y) is as follows:

the observation (–2, 2) occurs 50 times
the observation (0, 0) occurs 90 times
the observation (2, –1) occurs 60 times.

Calculate the sample correlation coefficient for these data. [4]

13 Suppose that the distribution of a physical coefficient, X, can be modelled using a uniform distribution on (0, 1). A researcher is interested in the distribution of Y, an adjusted form of the reciprocal of the coefficient, where $Y = (1 / X) – 1$.

(i) Show that the probability density function of Y is given by

$$f_Y(y) = 1/(1 + y)^2 \ , \ y > 0$$ [3]

(ii) Show that the mean of Y does not exist. [2]

[Total 5]

14 An insurance company issues house buildings policies for houses of similar size in four different post-code regions A, B, C and D.

(i) An insurance agent takes independent random samples of 10 house buildings policies for houses of similar size in regions A and B. The annual premiums (£) were as follows:

Region A: 229 241 270 256 241 247 261 243 272 219
($\Sigma x = 2{,}479$; $\Sigma x^2 = 617{,}163$)

Region B: 261 269 284 268 249 255 237 270 269 257
($\Sigma x = 2{,}619$; $\Sigma x^2 = 687{,}467$)

(a) Perform a two-sample t-test at the 5% level to compare the premiums for these two regions.

(b) Present the data in a simple diagram and hence comment briefly on the validity of the assumptions required for the above t-test.

(c) Calculate a 95% confidence interval for the underlying common standard deviation σ of such premiums. [10]

(ii) The agent takes further independent random samples of 10 such policies from the other two regions C and D. The annual premiums were as follows:

Region C: 253 247 244 245 221 229 245 256 232 269
($\Sigma x = 2{,}441$; $\Sigma x^2 = 597{,}607$)

Region D: 279 268 290 245 281 262 287 257 262 246
($\Sigma x = 2{,}677$; $\Sigma x^2 = 718{,}973$)

(a) Perform a one-way analysis of variance at the 5% level to compare the premiums for all four regions.

(b) Present the new data in a simple diagram and hence comment briefly on the validity of the assumptions required for the analysis of variance.

(c) Calculate a 95% confidence interval for the underlying common standard deviation σ of such premiums in the four regions. [10]

(iii) Comment briefly on your two confidence intervals in (i)(c) and (ii)(c) above. [1]

[Total 21]

15 An engineer is interested in estimating the probability that a particular electrical component will last at least 12 hours before failing. In order to do this, a random sample of n components is tested to destruction and their failure times, $x_1, x_2, ..., x_n$, are recorded. The engineer models failure times by assuming that they come from a distribution with distribution function, *F,* and probability density function, *f,* given below.

$$F(x) = 1 - \frac{1}{(1+x)^{\alpha-1}}, \; f(x) = \frac{\alpha - 1}{(1+x)^{\alpha}}, \qquad \alpha > 1, \; x > 0.$$

(i) Determine $\hat{\alpha}$, the maximum likelihood estimator of α, and, assuming n is large, use asymptotic theory to show that an approximate 95% confidence interval for α is given by $\hat{\alpha} \pm 1.96 \frac{\hat{\alpha} - 1}{\sqrt{n}}$. [8]

(ii) A sample of size $n = 80$ leads to a maximum likelihood estimate of α of 1.56. Use this figure to

(a) estimate the probability a component will fail before 12 hours,

(b) determine an approximate upper 95% one-sided confidence interval for α, and

(c) hence determine an approximate 95% one-sided confidence interval which provides an upper bound for the probability in part (ii)(a) above. [6]

(iii) Sixty-one of the eighty components tested in part (ii) failed before 12 hours, so a second engineer estimates the failure probability by 61/80 = 0.7625, and constructs an upper 95% confidence interval based on the binomial distribution.

(a) Construct this interval, and

(b) comment on the advantages and disadvantages of this method when compared to the method of part (ii). [4]

[Total 18]

16 The table below contains measurements on the strengths of beams. The width and height of each beam was fixed but the lengths varied. Data are available on the length (cm) and strength (Newtons) of each beam.

Length l	$x = \log l$	*Strength p*	$y = \log p$	*Fitted value*	*Residual*
7	1.946	11775	9.374	9.379	–0.005
7	1.946	11275	9.330	9.379	–0.049
9	2.197	8400	9.036	9.055	–0.019
9	2.197	8200	9.012	9.055	–0.043
12	2.485	6100	8.716	8.684	0.032
12	2.485	6050	8.708	8.684	0.024
14	2.639	5200	8.556	8.486	0.070
18	2.890	3750	8.230	8.162	0.068
18	2.890	3650	8.202	8.162	0.040
20	2.996	3275	8.094	8.026	0.068
20	2.996	3175	8.063	8.026	0.037
24	3.178	2200	7.696	7.791	–0.095
24	3.178	2125	7.662	7.791	–0.129

$\Sigma x = 34.023, \quad \Sigma x^2 = 91.3978, \quad \Sigma y = 110.679, \quad \Sigma xy = 286.6299$

It is thought that P and L satisfy the law $P = k/L$ where k is a constant, so $\log P = \log k - \log L$, i.e. $Y = \log k - X$.

A graph of $\log P$ against $\log L$ is included.

The simple linear regression model $y = \alpha + \beta x$ has been fitted to the data, and the fitted values and residuals are recorded in the table above.

(i) Use the data summaries above to calculate the least squares estimates $\hat{\alpha}$ of α and $\hat{\beta}$ of β. [3]

(ii) Assuming the usual normal linear regression model

(a) estimate the error variance σ^2,

(b) calculate a 95% confidence interval for β, and

(c) discuss briefly whether the data are consistent with the relationship $P = k/L$. [7]

(iii) Plot the residuals of the model against X and comment on the information contained in the plot. [3]

[Total 13]