EXAMINATIONS

19 September 2002 (pm)

Subject 101 — Statistical Modelling

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all 13 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, with any additional sheets firmly attached, and this question paper.

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

1 A very crude model for the distribution of claim size, X, in a particular situation represents X as a discrete random variable which takes the values £5,000, £10,000, and £20,000 with probabilities 0.4, 0.5, and 0.1 respectively.

Calculate the probability that of five randomly selected claims, three are for £5,000 each and the other two are for larger amounts. [2]

2 The probability that a component in a rocket motor will fail when the motor is fired is 0.02. To achieve a greater reliability several similar components are to be fitted in parallel; the motor will then fail only if all the individual components fail simultaneously.

Determine the minimum number of components required to ensure that the probability the motor fails is less than one in a billion (i.e. less than 10^{-9}), assuming that components fail independently. [2]

3 A random variable X which can be used in certain circumstances as a model for claim sizes has cumulative distribution function

$$F(x) = \begin{cases} 0 & , \ x < 0 \\ 1 - \left(\dfrac{2}{2+x} \right)^3 & , \ x > 0 \end{cases}$$

Calculate the value of the conditional probability $P(X > 3 \,|\, X > 1)$. [3]

4 Suppose that the sums assured under policies of a certain type are modelled by a distribution with mean £8,000 and standard deviation £3,000. Consider a group of 100 independent policies of this type.

Calculate the approximate probability that the total sum assured under this group of policies exceeds £845,000. [3]

5 Suppose that a line is fitted by least squares to a set of data $\{(x_i, y_i), i = 1, 2, \ldots, n\}$ which has sample correlation coefficient r. Let the fitted value at x_i be denoted $\hat{y}_i$.

Show that the sample correlation coefficient of the data $\{(y_i, \hat{y}_i), i = 1, 2, \ldots, n\}$, that is, of the observed and fitted y values, is also equal to r. [3]

6 As part of an investigation an insurance company collected data for the year 2000 on claims sizes for all claims on a certain type of motor insurance policy. The resulting data are given below in the form of a grouped frequency distribution.

Claim size (£)	*Frequency*
≤ 100	862
> 100 and ≤ 200	608
> 200 and ≤ 300	1253
> 300 and ≤ 400	1066
> 400 and ≤ 500	558
> 500	1290
Total	5637

(i) Calculate the cumulative frequencies and draw a graph of the claim size distribution function (i.e. the cumulative frequencies against claim size). [3]

(ii) You have been asked to determine the proportion of claim sizes which are less than £250. Use linear interpolation to approximate this proportion to two decimal places. [2]

(iii) Use linear interpolation to approximate (to the nearest £) the value of the median of the claim size distribution. [2]

[Total 7]

7 Suppose that X is a continuous random variable uniformly distributed on (0, 1).

(i) Show that the moment generating function of $Y = -\log X$ is

$$M_Y(t) = 1/(1 - t) \quad \text{for} \quad t < 1.$$ [3]

(ii) Hence state the distribution of Y. [1]

[Total 4]

8 Assume that, for a group of insurance policies, the number of claims on each policy can be modelled by a Poisson distribution with the same rate λ per year, independently for each policy. Such a group of 1500 policies gave rise to a total of 183 claims during the last year.

(i) State the value of the maximum likelihood estimate of λ. [1]

(ii) Hence obtain estimates of the following quantities:

(a) The probability that a subset of 10 of these policies results in no claims over the next six months.

(b) The probability that a subset of 250 of these policies results in more than 40 claims over the next year. [4]

[Total 5]

9 Thirty employees working in the call centre of a bank were chosen at random from the workforce and agreed to assist in the assessment of two different training programmes (A and B).

Ten of the selected employees were randomly assigned to each of the two training programmes, while the other ten were to receive no additional training and act as controls.

Shortly after the training began, one of the employees on training programme B and two of the control group left the employment of the bank.

At the end of the training period each of the twenty-seven employees still involved was observed at work over a period of time by performance assessors. Each employee was then given a performance score (measuring a range of aspects of their work) expressed as a mark out of a possible maximum of 100.

The results were as follows:

	Performance scores										*Totals*
Control	55	74	64	62	37	78	50	44	—	—	464
Training programme A	63	79	60	75	89	58	75	72	84	69	724
Training programme B	64	55	57	73	51	60	62	78	68	—	568

$\Sigma\Sigma x_{ij}^2 = 118{,}128$.

(i) Plot the data in a simple and informative way which displays both the "within treatment" and "between treatment" variation. [2]

(ii) Conduct an analysis of variance to investigate whether differences exist among the three "treatments" and comment briefly on your findings. [7]

[Total 9]

10 The number of claims N which arise from a portfolio of business is modelled as a Poisson variable with mean μ. The claim amounts $X_i : i = 1, 2, ..., N$ are modelled as independent gamma variables each with parameters α and λ and are independent of N. Let S be the total claim amount arising from this portfolio.

(i) Obtain expressions for the mean and standard deviation of S in terms of μ, α and λ, using general results for the mean and variance of S, [2]

(ii) Consider the case where $\mu = 100$ and the individual claim amounts have mean £100 and standard deviation £50.

(a) Calculate the mean and standard deviation of the total claim amount S.

(b) Calculate an approximate value for the probability that the total claim amount S exceeds £12,500, giving a brief justification of your approach. [4]

[Total 6]

11 Consider a group of n males each aged 30 years living in a particular society. The lives may be assumed to be independent.

Suppose that x of the n men die by the end of a subsequent period of duration t_0 years and that $n - x$ survive the period.

Suppose that the lifetime of these men from age 30 is to be modelled as a random variable with an exponential distribution with mean $1/\lambda$ hours.

(i) State the distribution of the random variable X whose value x, the number of men who die within t_0 years, has been observed. [2]

(ii) (a) Verify that the log-likelihood of the observation x is given by:

$$\ell(\lambda) = \log L(\lambda) = x \log(1 - e^{-\lambda t_0}) - (n - x)\lambda t_0 + \textit{constant}$$

(b) Determine $\hat{\lambda}$, the maximum likelihood estimator of λ. [6]

(iii) Show that $E\left(-\frac{\partial^2 \ell}{\partial \lambda^2}\right) = \frac{n t_0^{\,2} e^{-\lambda t_0}}{1 - e^{-\lambda t_0}}$ [4]

(iv) (a) Consider the case n = 1,000, t_0 = 20 years, and x = 320.

Evaluate $\hat{\lambda}$, and show that the value of its standard error is approximately 0.00108.

(b) Hence, calculate an approximate 95% confidence interval for the mean lifetime from age 30 of such men. [7]

[Total 19]

12 A psychologist conducted an investigation into the effect of alcohol on reaction times using 10 male and 10 female subjects. Each subject was given two tests on different days, during which his/her reaction times were recorded.

Before each of the tests, the subject drank a glass of liquid. Some glasses contained a fixed quantity of alcohol and others contained a liquid which had a similar colour and taste but no alcohol. Each subject drank one glass of each type. The order of presentation was randomized, independently for each subject.

The data below give the reaction times, in units of 0.01 seconds. Also given is the difference between the reaction time with alcohol and the reaction time without alcohol for each subject (reaction time with alcohol minus reaction time without).

Males

Subject No.	1	2	3	4	5	6	7	8	9	10
With alcohol	45	51	35	43	51	54	51	49	44	52
Without alcohol	40	54	21	31	44	47	39	33	32	56
Difference	5	–3	14	12	7	7	12	16	12	–4

Females

Subject No.	1	2	3	4	5	6	7	8	9	10
With alcohol	47	54	58	48	60	46	55	74	56	49
Without alcohol	39	40	42	30	51	41	55	68	47	40
Difference	8	14	16	18	9	5	0	6	9	9

(i) (a) Construct a 95% confidence interval for the mean difference between the reaction times with and without alcohol for the males, using the 10 difference values.

(b) Construct a similar 95% confidence interval based on the female difference values.

(c) Comment briefly on the two confidence intervals. [8]

(ii) (a) Perform a two-sample *t*-test to investigate whether the alcohol effect differs between males and females.

(b) Show that the variances in the male and female samples are not significantly different at the 5% level, and comment briefly with reference to the test conducted in (ii)(a). [8]

[Total 16]

13 The table below gives the frequency of coronary heart disease by age group. The table also gives the age group midpoint (x) and $y = \log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right)$, where $\hat{\theta}$ denotes the proportion in an age group with coronary heart disease.

		Coronary Heart Disease			
Age group	x	*Yes*	*No*	n	y
20–29	25	1	9	10	–2.19722
30–34	32.5	2	13	15	–1.87180
35–39	37.5	3	9	12	–1.09861
40–44	42.5	5	10	15	–0.69315
45–49	47.5	6	7	13	–0.15415
50–54	52.5	5	3	8	0.51083
55–59	57.5	13	4	17	1.17865
60–69	65	8	2	10	1.38629

$\sum x = 360;\ \sum x^2 = 17437.5;\ \sum y = -2.9392;\ \sum y^2 = 13.615;\ \sum xy = -9.0429$

(i) (a) Calculate an estimate of the probability of having coronary heart disease under the assumption that the probability does not differ over the age groups.

(b) Construct an 8×2 contingency table with marginal totals and conduct a χ^2 test for differences in the probability of having coronary heart disease for the different age groups. [8]

(ii) Consider the regression model $y = \alpha + \beta x$.

(a) Draw a scatterplot of y against x, and comment on the appropriateness of the suggested model.

(b) Calculate the least squares fitted regression line of y on x.

(c) Calculate a 99% confidence interval for the slope parameter.

(d) Interpret the result obtained in (i)(b) with reference to the confidence interval obtained in (ii)(c). [13]

[Total 21]