EXAMINATIONS

17 April 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 14 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 Let ${}_tp_x$ denote the probability that a life aged x survives for at least a further t years, and consider three independent lives aged 40, 50 and 60 years such that ${}_{10}p_{40} = 0.95$, ${}_{10}p_{50} = 0.85$, ${}_{10}p_{60} = 0.70$.

(i) Determine the probability that exactly one of these three lives survives ten years. [2]

(ii) Determine the probability that it is the youngest life that survives, given that exactly one life survives ten years. [1]

[Total 3]

2 Claim amounts are modelled as an exponential random variable with mean £1,000.

(i) Calculate the probability that one such claim amount is greater than £5,000. [1]

(ii) Calculate the probability that a claim amount is greater than £5,000 given that it is greater than £1,000. [2]

[Total 3]

3 A magazine claims that 25% of its readers are students. A random sample of 200 readers is taken and is found to contain 42 students.

Calculate the probability of obtaining 42 or fewer student readers, assuming that the magazine's claim is correct. [3]

4 The percentage return on an investment of a particular type over a period of one year is to be modelled as a normal random variable X with mean μ and variance 1. A potential investor is interested in the chance that the return on such an investment will exceed 9%.

A random sample of ten such returns have values

7.3, 8.9, 8.3, 6.2, 9.8, 7.7, 9.4 , 7.9, 9.1, and 7.4 .

Calculate the maximum likelihood estimate of $\theta = P(X > 9)$. [3]

5 In a quality control test, a random sample of 100 items is selected, of which 90 are found to be satisfactory. The value of p, the probability of an item being satisfactory, is unknown.

Write down the probability of observing 90 satisfactory items out of 100 as a function of p, and thus derive the maximum likelihood estimate of p. [3]

6 A national survey research company has past data which indicate that the interview time for a consumer opinion study has a standard deviation of 6 minutes.

Calculate the size of the sample that should be taken if the company requires a 99% confidence interval for the mean interview time to be within ±2 minutes. [2]

7 The following information on white blood cell count (WBCC) was collected from subjects one week after the start of chemotherapy treatment. One group of subjects (A) received steroids in addition to the chemotherapy treatment and the other group (B) received a placebo in addition to the chemotherapy. The subjects were assigned to the groups at random.

Group A — Steroid
WBCC (millions of cells per ml.)

12.4	15.2	12.7	15.9	12.2	14.2	12.9	14.2	12.4	14.6
12.7	13.6	12.5	13.3	12.1	13.9	17.1	13.6	17.2	13.1

Group B — Placebo
WBCC (millions of cells per ml.)

17.0	13.5	15.4	14.1	15.4	14.8	12.9	14.4	13.2	13.1
12.9	13.9	13.0	13.6	13.0	13.4	12.9	13.1	14.4	13.8

(i) Construct stem and leaf diagrams for Group A and Group B separately. [2]

(ii) Comment on the results in the context of investigating an association between WBCC and the treatment with or without steroids. [2]

[Total 4]

8 Consider a negative binomial variable X with probability function given by

$$P(X = x) = \binom{k+x-1}{x} p^k q^x \quad : x = 0, 1, 2, \ldots \text{ where } 0 < p < 1 \text{ and } q = 1 - p.$$

(i) Show that the moment generating function is given by

$$M(t) = \left(\frac{p}{1 - qe^t} \right)^k \text{ for } qe^t < 1. \qquad [2]$$

(ii) Determine $E(X)$ and $E(X^2)$ by expanding $M(t)$ as a power series as far as the term in t^2, and hence verify that the mean and variance of X are given by

$$\frac{kq}{p} \text{ and } \frac{kq}{p^2}, \text{ respectively.} \qquad [3]$$

[Total 5]

9 Let X and Y be independent random variables. Let V and W be the random variables defined by

$$V = \max \{X, Y\}, \quad W = \min \{X, Y\}$$

i.e. V is the larger, and W is the smaller, of the observations of X and Y.

Let F_X, F_Y, F_V, F_W denote the distribution functions of X, Y, V, W respectively.

(i) Show that $F_V(t) = F_X(t)F_Y(t)$. [2]

(ii) Show that $F_W(t) = F_X(t) + F_Y(t) - F_X(t)F_Y(t)$. [3]

(iii) The random variable X has an exponential distribution with parameter 4 and, independently, Y has an exponential distribution with parameter 4. Obtain the distribution function of the minimum of X and Y and state its mean. [3]

[Total 8]

10 A company wants to estimate the percentage of its customers who are willing to shop on the internet. It decides to do so by calculating a symmetrical 95% two-sided confidence interval for the unknown percentage.

(i) Show that, based on a random sample of 200 of the company's customers, the required confidence interval will have a width which is no greater than 13.9%. [4]

(ii) Calculate the sample size required which will ensure that, whatever the true percentage, the width of the confidence interval will be no greater than 10%. [2]

[Total 6]

11 Consider the following model for aggregate claim amounts S:

$$S = X_1 + X_2 + \ldots + X_N$$

where the X_i are independent, identically distributed random variables representing individual claim amounts and N is a random variable, independent of the X_i, and representing the number of claims. Let X have mean μ_X and let N have mean μ_N and variance σ_N^2.

(i) Show that

$$E(SN) = \mu_X(\mu_N^2 + \sigma_N^2)$$

by considering expected values conditional on the value of N. [3]

(ii) Hence derive an expression for the covariance between S and N. [2]

[Total 5]

12 A simple model for the movement of a stock price is such that, independently in each time period, the stock either:

goes up with probability $(\frac{1}{4} - \theta)$;

stays the same with probability $(\frac{5}{8} + 2\theta)$;

goes down with probability $(\frac{1}{8} - \theta)$.

(i) Determine the range of admissible values of the parameter θ. [2]

(ii) (a) Calculate the probability that the stock goes down in one time period, in the case $\theta = 0.1$.

(b) Calculate the probability that the stock stays the same for two consecutive time periods, in the case $\theta = 0$.

(c) Calculate the probability that, in four time periods, the stock goes up twice and stays the same twice, in the case $\theta = -0.2$. [4]

(iii) Data are collected for 80 consecutive time periods and yield the following observed frequencies:

change in stock	up	same	down
no. of time periods	24	35	21

(a) (1) Write down an expression for $L(\theta)$, the likelihood of these data, and show that

$\frac{\partial}{\partial\theta}\log L(\theta) = 0$ reduces to the quadratic equation

$5120\theta^2 - 468\theta - 95 = 0$

(2) Explain why one of the roots of this quadratic yields the maximum likelihood estimate of θ and hence determine this estimate. [5]

(b) (1) Calculate the expected frequencies using the model with the maximum likelihood estimate of θ.

(2) Hence perform a χ^2 goodness of fit test of the model and state your conclusion clearly. [6]

(c) Comment briefly on what additional information would be needed for these data in order to investigate the validity of the assumption of independence used in this model, and comment briefly on how the validity might be checked. [2]

[Total 19]

13 Six insurance companies were being compared with regard to premiums being charged for house contents insurance for houses in a particular postcode region. Independent random samples of five policies from each company are examined and the premiums (in £) were recorded.

Company	*A*	*B*	*C*	*D*	*E*	*F*
	151	152	175	149	123	145
	168	141	155	148	132	131
	128	129	162	137	142	155
	167	120	186	138	161	172
	134	115	148	169	152	141
Totals	748	657	826	741	710	744

$$\sum\sum y_{ij} = 4,426 \quad ; \quad \sum\sum y_{ij}^2 = 661,796$$

(i) Compute an ANOVA table for these data, and show that there are no significant differences, at the 5% level, between mean premiums being charged by each company. [5]

(ii) State the assumptions required for the above analysis of variance, and, by drawing a suitable diagram of these data, comment briefly on the validity of these assumptions. [4]

(iii) Calculate a 95% confidence interval for the underlying common standard deviation σ, using $\frac{SS_R}{\sigma^2}$ as a pivotal quantity with a χ^2 distribution. [4]

(iv) A colleague points out that company C has the largest mean premium of £165.20 and that Company B has the smallest mean premium of £131.40 and suggests performing a *t*-test to compare these two companies.

(a) Perform this *t*-test, using the estimate of variance from the ANOVA table, and in particular show that there is a significant difference at the 1% level.

(b) Your colleague states that there is therefore a significant difference between the six companies.

Discuss the apparent contradiction with your conclusion in part (i) and explain the flaw in your colleague's argument. [5]

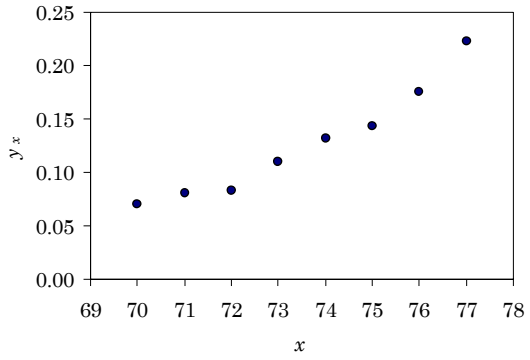
[Total 18]

14 The table below gives the numbers of deaths n_x in a year in groups of women aged x years. The exposures of the groups, denoted E_x, are also given (the exposure is essentially the number of women alive for the year in question). The values of the death rates y_x, where $y_x = n_x/E_x$, and the log(death rates), denoted w_x, are also given.

age x	*number of deaths* n_x	*exposure* E_x	$y_x = n_x/E_x$	$w_x = \log y_x$
70	30	426	0.07042	–2.6532
71	38	471	0.08068	–2.5173
72	38	454	0.08370	–2.4805
73	53	482	0.10996	–2.2077
74	59	445	0.13258	–2.0205
75	61	423	0.14421	–1.9365
76	82	468	0.17521	–1.7417
77	96	430	0.22326	–1.4994

$\Sigma x = 588 \quad \Sigma x^2 = 43260 \quad \Sigma w = -17.0568 \quad \Sigma w^2 = 37.5173 \quad \Sigma xw = -1246.7879$

(i) A scatter plot of y_x against x is shown below.



Draw a scatter plot of w_x against x and comment briefly on the two scatter plots and the relationships displayed. [3]

(ii) (a) Calculate the least squares fit regression line in which w_x is modelled as the response and x as the explanatory variable.

(b) Draw the fitted line on your scatter plot of w_x against x.

(c) Calculate a 95% confidence interval for the slope coefficient of the regression model of w_x on x, adopting the assumptions of the usual "normal regression model".

(d) Calculate the fitted values for the number of deaths for the group aged 71 years and the group aged 76 years. [12]

(iii) Explain briefly the relationship between the fitting procedure used in part (ii) and a model which states that the number of deaths N_x is a random variable with mean $E_x bc^x$ for some constants b and c. [3]

[Total 18]