EXAMINATIONS

26 April 2004 (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 The probability that a claim is made on a certain type of policy in a particular year is 0.04. Five hundred policies are selected at random.

Use a suitable normal approximation to calculate the probability that no more than 30 of these will result in a claim during the year. [3]

2 The interquartile range of a continuous distribution with distribution function $F(x)$ is defined as $IQR = x_2 - x_1$ where $F(x_1) = 0.25$ and $F(x_2) = 0.75$.

Show that for a normal distribution with variance σ^2, the interquartile range is $IQR = 1.349\sigma$. [3]

3 The cumulant generating function of a random variable X is given by:

$$C_X(t) = \log M_X(t) = 2\left\{(1-t)^{-10} - 1\right\}$$

where $M_X(t)$ is the moment generating function.

Determine the mean and variance of the distribution of X. [3]

4 Claim sizes in a certain insurance situation are modelled by an exponential distribution with mean \$20,000. The insurer defines a claim to be a *large claim* if the claim size exceeds \$35,000.

State, with a reason, the expected size of a *large claim*. [2]

5 Suppose that $\hat{\theta}$ is an unbiased estimator of a parameter θ and has variance $\frac{\theta^2}{10}$.

Derive an expression for the mean square error of $k\hat{\theta}$, where k is a constant, and determine the value of k for which the mean square error is a minimum. [4]

6 Let X and Y be random variables which each takes values 1 and 2 only.

Calculate $E[X|Y = 2]$, given that $E[X] = 6/5$, $E[X|Y = 1] = 7/6$, and $P(Y = 1) = 3/5$. [2]

7 An insurance company has a portfolio of large industrial insurance risks. Three such independent risks, labelled A, B, C, are being investigated. The sums insured, in units of £100 million, are 2.5, 4.8, 7.2 for A, B, C, respectively. The company's expert risk assessors estimate that the probabilities of a claim arising in the next calendar year are 0.01, 0.02, 0.005 for A, B, C, respectively. If a claim does arise, then the full sum insured will be paid out and no further claims can then arise for that risk. The above probability estimates should be used throughout this question.

(i) Determine, for this group of three risks, using suitable indicator variables or otherwise, the mean and standard deviation of the total claim amount paid out in the next calendar year.

[3]

(ii) You are told that exactly one claim has arisen. Calculate the mean total claim amount and comment on your answer in relation to your answer in part (i).

[4]

[Total 7]

8 Let X_1 and X_2 be independent Poisson variables with respective parameters μ_1 and μ_2.

Let $S = X_1 + X_2$

(i) Show that S has a Poisson distribution with mean $\mu_1 + \mu_2$. [2]

(ii) Show that the conditional distribution of X_1 given $S = s$ is binomial, and state its parameters.

[3]

[Total 5]

9 In order to simulate an observation of a normal random variable it is suggested that

$$S = \sum_{i=1}^{n} X_i$$

is used, where $X_1, \ldots, X_n$ is a random sample from a continuous uniform distribution on the interval $(-\frac{1}{2}, \frac{1}{2})$.

(i) Determine the approximate distribution of S. [2]

(ii) Determine the value of n which should be used if S is required to represent a standard normal random variable. [1]

(iii) Explain why S has the same coefficient of skewness as a standard normal random variable. [1]

[Total 4]

10 The number of claims which arise in a year under a policy of a certain type follows a Poisson distribution with mean λ. It is required to test

$$H_0 : \lambda = 0.2 \text{ v } H_1 : \lambda > 0.2$$

and it is decided to reject H_0 in favour of H_1 if 15 or more claims arise in a year under a group of 50 independent such policies.

By using tables of Poisson distribution probabilities, calculate the power of this test in each of the cases:

(a) $\lambda = 0.3$, and
(b) $\lambda = 0.4$

and comment briefly on the results. [5]

11 The following table gives the sums insured (in units of £1,000) for a random sample of insurance policies on the contents of private houses for each of four insurance companies.

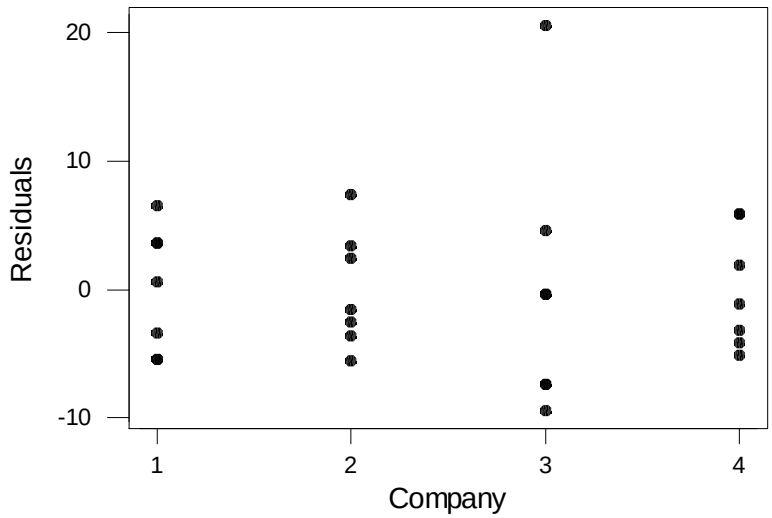
Company			
1	2	3	4
(y_1)	(y_2)	(y_3)	(y_4)
39	24	21	33
29	28	30	26
33	33	30	28
36	22	51	30
27	29	23	27
36	20	23	37
27	23	35	37

$\Sigma y_1 = 227$	$\Sigma y_2 = 179$	$\Sigma y_3 = 213$	$\Sigma y_4 = 218$
$\Sigma y_1^2 = 7{,}501$	$\Sigma y_2^2 = 4{,}703$	$\Sigma y_3^2 = 7{,}125$	$\Sigma y_4^2 = 6{,}916$

(i) Test whether there are any differences between the population means of the four companies, using one-way analysis of variance.

[5]

(ii) A plot of the residuals for the fitted one-way analysis of variance model is given below:



Comment on the adequacy of the model. [2]

[Total 7]

12 For the estimation of a binomial probability $p = P(\text{success})$, a series of n independent trials are performed and X represents the number of successes observed.

(i) Write down the likelihood function $L(p)$ and show that the maximum likelihood estimator (MLE) of p is $\hat{p} = \dfrac{X}{n}$.

[3]

(ii) (a) Determine the Cramer-Rao lower bound for the estimation of p.

(b) Show that the variance of the MLE is equal to the Cramer-Rao lower bound.

(c) Write down an approximate sampling distribution for $\hat{p}$ valid for large n.

[4]

(iii) In order to develop an approximate 95% confidence interval for p for large n, the following pivotal quantity is to be used

$$\frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \approx N(0,1) .$$

Assuming that this pivotal quantity is monotonic in p, show that re-arrangement of the inequality

$$-1.96 < \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} < 1.96$$

leads to a quadratic inequality in p, and hence determine an approximate 95% confidence interval for p of the form $p_L(\hat{p}) < p < p_U(\hat{p})$.

[5]

(iv) A simpler and more widely used approximate confidence interval is obtained by using the following pivotal quantity

$$\frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \approx N(0,1) .$$

Determine the resulting approximate 95% confidence interval using this. [2]

(v) In two separate applications the following data were observed:

(a) 4 successes out of 10 trials
(b) 80 successes out of 200 trials

In each case calculate the two approximate confidence intervals from parts (iii) and (iv) and comment briefly on your answers.

[6]
[Total 20]

13 A researcher studying the claims experience of a company (in a particular class of business) records the payments on 100 recent claims. The payments (in units of £1000, and sorted) are given below.

Payments

0.30	0.89	0.96	1.16	1.67	1.77	1.93	1.98	2.07	2.09	2.30	2.48
2.58	2.78	3.00	3.19	3.21	3.21	3.25	3.31	3.34	3.37	3.66	3.95
4.16	4.18	4.60	4.72	4.73	4.76	5.01	5.17	5.21	5.63	5.72	6.00
6.13	6.17	6.24	6.37	6.47	6.48	6.87	7.05	7.16	7.21	7.51	7.72
7.74	8.00	8.00	8.03	8.04	8.54	9.11	9.18	9.49	9.59	10.00	10.36
10.85	11.08	11.22	11.27	11.38	11.45	11.69	11.78	12.27	12.30	12.50	13.04
13.28	13.43	13.48	13.85	14.27	14.31	14.49	14.55	14.62	14.68	14.70	14.83
15.67	15.70	15.77	16.28	16.44	17.17	17.89	18.03	18.12	20.72	22.00	24.33
25.41	28.30	31.00	32.80								

For these 100 observations: $\Sigma x = 952.75$, $\Sigma x^2 = 13584.5217$.

The researcher wants to examine whether or not the exponential distribution provides a good description of the distribution of the payments.

(i) (a) Calculate the sample mean and standard deviation for the 100 payments, and comment briefly on whether or not you think the exponential distribution will provide an acceptable fit to the data.

(b) Specify the fitted exponential distribution, giving a brief justification of your approach.

Note: you are not required to give any mathematical derivation in your justification.

[5]

(ii) The researcher decides to conduct a formal chi-squared goodness-of-fit test of the exponential distribution to the data, using five equi-probable intervals (i.e. intervals each with associated probability 0.2).

(a) Show that the value x which is exceeded with probability p by an exponential variable with mean μ satisfies $x = -\mu \log p$.

(b) Calculate the values which divide the positive real numbers into five equi-probable intervals for your fitted exponential distribution.

[4]

(iii) Conduct the formal goodness-of-fit test outlined in part (ii) above and comment on the result. [5]

[Total 14]

14 Forensic scientists use various methods for determining the likely time of death from post-mortem examination of human bodies. A recently suggested objective method uses the concentration of a compound (3-methoxytyramine or 3-MT) in a particular part of the brain.

In a study of the relationship between post-mortem interval and the concentration of 3-MT, samples of the appropriate part of the brain were taken from coroners' cases for which the time of death had been determined from eye-witness accounts. The intervals (x; in hours) and concentrations (y; in parts per million) for 18 individuals who were found to have died from organic heart disease are given in the following table. For the last two individuals (numbered 17 and 18 in the table), there was no eye-witness testimony directly available, and the time of death was established on the basis of other evidence including knowledge of the individuals' activities.

Observation number	*Interval (x)*	*Concentration (y)*
1	5.5	3.26
2	6.0	2.67
3	6.5	2.82
4	7.0	2.80
5	8.0	3.29
6	12.0	2.28
7	12.0	2.34
8	14.0	2.18
9	15.0	1.97
10	15.5	2.56
11	17.5	2.09
12	17.5	2.69
13	20.0	2.56
14	21.0	3.17
15	25.5	2.18
16	26.0	1.94
17	48.0	1.57
18	60.0	0.61

$\Sigma x = 337 \quad \Sigma x^2 = 9854.5 \quad \Sigma y = 42.98 \quad \Sigma y^2 = 109.7936 \quad \Sigma xy = 672.8$

In this investigation you are required to explore the relationship between concentration (regarded as the response/dependent variable) and interval (regarded as the explanatory/independent variable).

(i) Construct a scatterplot of the data. Comment on any interesting features of the data and discuss briefly whether linear regression is appropriate to model the relationship between concentration of 3-MT and the interval from death.

[5]

(ii) Calculate the correlation coefficient for the data, and use it to test the null hypothesis that the population correlation coefficient is equal to zero.

[5]

(iii) Calculate the equation of the least-squares fitted regression line, and use it to estimate the concentrations of 3-MT:

(a) after 1 day and
(b) after 2 days

Comment briefly on the reliability of these estimates. [5]

(iv) Calculate a 99% confidence interval for the slope of the regression line. Using this confidence interval, test the hypothesis that the slope of the regression line is equal to zero. Comment on your answer in relation to the answer given in part (ii) above.

[6]
[Total 21]

END OF PAPER