EXAMINATIONS

8 September 2003 (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 15 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.

© Faculty of Actuaries
© Institute of Actuaries

1 In a large corporation 50 new employees joined the company's pension scheme during the last year. It is assumed that each new employee has a probability of 0.40 of remaining in the scheme for at least 10 years, independently for each new employee.

Calculate an approximate value for the probability that more than half of last year's 50 new employees remain in the pension scheme for at least 10 years. [3]

2 Let $(X_1, X_2, \ldots, X_9)$ be a random sample from a $N(0,\sigma^2)$ distribution. Let $\bar{X}$ and S^2 denote the sample mean and variance respectively.

Find the approximate value of $P\left(\bar{X} > S\right)$ by referring to an appropriate statistical table. [3]

3 A random sample of size 200 is taken from a large group of motor policies. The proportion of policies on which claims arose during the last year is $\hat{\theta} = 0.16$.

Calculate approximate 95% confidence limits for the true proportion θ for the whole group of policies. [2]

4 A random sample of 16 values, $x_1, x_2, \ldots, x_{16}$, was drawn from a normal population and gave the following summary statistics:

$$\sum_{i=1}^{16} x_i = 51.2, \quad \sum_{i=1}^{16} x_i^2 = 243.19.$$

Calculate a 95% confidence interval for the population mean. [3]

5 Ten independent hypothesis tests are to be conducted, each at the 5% significance level.

Calculate the probability that at least one of the tests produces a significant result, assuming that the null hypothesis for each of the 10 tests is true, and comment briefly on the value. [3]

6 Show that the slope of the regression line fitted by least squares to the three points

$$(0,0)\ ,\ (1,y)\ ,\ (2,2)$$

is 1 for all values of y. [2]

7 Claims arise through time on a portfolio of policies one after another, at random, and at a constant rate λ per week. Claim sizes are to be modelled as a $N(\mu, \sigma^2)$ random variable, independent of the times of occurrence and the accumulated numbers of claims.

The moment generating function of S, the total size of all claims which occur in a period of k weeks (where k is a positive integer), is to be used in a theoretical report being written by two students.

One student (A) suggests that the moment generating function of S is given by:

$$\textit{Suggestion } A: \quad M_S(t) = \exp\left[k\lambda\left\{\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) - 1\right\}\right]$$

while the other (B) disagrees and suggests that it is in fact given by:

$$\textit{Suggestion } B: \quad M_S(t) = \exp\left[k\lambda \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) - 1\right]$$

One of the students is correct. Determine which one this is. [3]

8 A student actuary is about to collect data from a batch of insurance proposal forms. It is known that on average one in every fifty such forms contains incomplete information, and the student wants to be 95% sure of having at least 500 forms which are complete.

Show that he must use a batch of at least 516 forms. (Use a suitable model and appropriate tables for the distribution of the number of forms containing incomplete information.) [4]

9 Suppose that the joint probability distribution of two random variables X and Y is given by the following table:

		Y		
		2	*4*	*6*
	1	0.2	0.0	0.2
X	*2*	0.0	0.2	0.0
	3	0.2	0.0	0.2

(i) Show that X and Y are uncorrelated, but are not independent. [3]

(ii) Leaving the probabilities in the first and third rows of the table the same, change the entries in the second row so that X and Y are independent. [2]

[Total 5]

10 The number of claims on a portfolio of policies was observed as follows:

Number of claims per day	*0*	*1*	*2*	*3*	*4*	*5*	*Total*
Frequency	48	32	17	2	0	1	100

Use a χ^2 goodness-of-fit test to test the hypothesis that the number of claims each day follows a Poisson distribution. [7]

11 The following table contains 10 claim amounts for repair costs arising from a particular type of storm damage to private houses, for each of four different postcode regions:

	Regions			
	A	*B*	*C*	*D*
Claim amount (£)	961	1,507	1,303	1,022
	1,263	1,349	959	997
	1,304	1,521	1,297	1,335
	1,532	1,134	1,051	1,216
	1,294	1,293	1,163	1,277
	1,605	993	993	1,135
	1,308	1,126	978	1,273
	1,393	1,140	891	1,244
	1,255	1,305	1,177	1,105
	1,131	1,224	1,153	1,524
Sums	13,046	12,592	10,965	12,128
Sums of squares	17,322,090	16,116,822	12,208,021	14,929,994

(i) Show that a one-way analysis of variance to compare the mean claim amounts for the regions produces a significant result at the 5% level, but not at the 1% level. [5]

(ii) Compare the mean claim amounts for the regions A, B, C, and D by using a least significant difference approach with a significance level of 5%. [4]

[Total 9]

12 Let X denote the number of accidents a manual worker in a particular factory has in a year. For a given worker the distribution of X is modelled as a Poisson distribution with unknown parameter u that varies across the workforce. U is regarded as a random variable which has a gamma distribution with parameters α and λ, i.e.

$X \mid (U = u) \sim \text{Poisson}(u)$,
$U \sim \text{gamma}(\alpha, \lambda)$.

(i) Show that the marginal distribution of X has mean $\frac{\alpha}{\lambda}$ and variance $\frac{\alpha}{\lambda} + \frac{\alpha}{\lambda^2}$. [3]

(ii) A dataset has a sample mean of $\bar{x}$ and a sample variance of s^2. Show that α and λ may be estimated by

$$\hat{\alpha} = \frac{\bar{x}^2}{s^2 - \bar{x}} \ , \ \hat{\lambda} = \frac{\bar{x}}{s^2 - \bar{x}}$$

using the method of moments. [2]

(iii) State the circumstances under which the method of moments produce inadmissible estimates of α and λ. [1]

[Total 6]

13 In an air pollution monitoring study undertaken in a residential area near an industrial plant a sample of 20 sites is chosen and an observation at each site is made of the content (in parts per million) of a particular contaminant. The data recorded are given in the table below together with some summaries:

76	78	76	78	84	79	79	81	85	76
78	79	75	83	87	80	78	77	81	77

$\Sigma x = 1{,}587 \quad \Sigma x^2 = 126{,}131$

(i) (a) Present these data graphically using a dotplot and comment briefly on the shape of the distribution.

(b) Calculate a 95% confidence interval for the mean contaminant content for the residential area from which the sites were selected.

(c) Comment briefly on the validity of this confidence interval in the light of your answer to part (a). [8]

(ii) After some modifications by the operators of the industrial plant designed to reduce the level of pollution due to this contaminant, another observation was made at each of the same sites. The data recorded are given in the table below with the sites in the same order as in the table above:

74	74	76	79	83	76	76	81	84	76
81	77	74	83	89	78	77	72	79	78

(a) Calculate the difference (before – after) in contaminant content for each site, present these differences graphically and comment briefly on the shape of the distribution.

(b) Perform an appropriate test to investigate whether the modification has led to a reduction in the contaminant content.

(c) Comment briefly on the validity of this test in the light of your answer to part (ii)(a). [9]

[Total 17]

14 Consider a linear regression model in which responses Y_i are uncorrelated and have expectations βx_i and common variance $\sigma^2 (i = 1, \ldots, n)$, i.e. Y_i is modelled as a linear regression through the origin:

$$E(Y_i \mid x_i) = \beta x_i \quad \text{and} \quad V(Y_i \mid x_i) = \sigma^2 \quad (i = 1, \ldots, n).$$

(i) (a) Show that the least squares estimator of β is $\hat{\beta}_1 = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2$.

(b) Derive the expectation and variance of $\hat{\beta}_1$ under the model. [5]

(ii) An alternative to the least squares estimator in this case is:

$$\hat{\beta}_2 = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} x_i = \overline{Y} / \overline{x}.$$

(a) Derive the expectation and variance of $\hat{\beta}_2$ under the model.

(b) Show that the variance of the estimator $\hat{\beta}_2$ is at least as large as that of the least squares estimator $\hat{\beta}_1$. [4]

(iii) Now consider an estimator $\hat{\beta}_3$ of β which is a linear function of the responses, i.e. an estimator which has the form $\hat{\beta}_3 = \sum_{i=1}^{n} a_i Y_i$, where $a_1, \ldots, a_n$ are constants.

(a) Show that $\hat{\beta}_3$ is unbiased for β if $\sum_{i=1}^{n} a_i x_i = 1$, and that the variance of $\hat{\beta}_3$ is $\sum_{i=1}^{n} a_i^2 \sigma^2$.

(b) Show that the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ above may be expressed in the form $\hat{\beta}_3 = \sum_{i=1}^{n} a_i Y_i$ and hence verify that $\hat{\beta}_1$ and $\hat{\beta}_2$ satisfy the condition for unbiasedness in (iii)(a).

(c) It can be shown that, subject to the condition $\sum_{i=1}^{n} a_i x_i = 1$, the variance of $\hat{\beta}_3$ is minimised by setting $a_i = \dfrac{x_i}{\sum_{i=1}^{n} x_i^2}$.

Comment on this result. [7]

[Total 16]

15 Let $S(0)$ denote the price of a certain security, and let $S(n)$ denote the price of the security at the end of n successive weeks for $n = 1, 2, 3, \ldots$. A model for the changes in these prices is such that the price ratios $\frac{S(n)}{S(n-1)}$ for $n \geq 1$ are independent identically distributed random variables with a lognormal distribution.

[*Note*: The random variable Y is lognormal with parameters μ and σ^2 if $\log(Y)$ is normal $N(\mu,\sigma^2)$, that is, Y is lognormal if it can be expressed as $Y = e^X$ where $X \sim N(\mu,\sigma^2)$. The mean and variance of Y are given by $E(Y) = e^{\mu+\frac{\sigma^2}{2}}$ and $V(Y) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.]

(i) Using the above lognormal model with parameters $\mu = 0.0125$ and $\sigma = 0.055$, determine the probability that:

(a) the price of the security decreases over the next week

(b) the price of the security decreases over each of the next two weeks

(c) the price at the end of two weeks is greater than it is at present

(d) the price at the end of 20 weeks is less than it is at present

[11]

(ii) At the start of a period the price is £1,245 and it is then observed for 10 weeks. The resulting 10 prices (£) and ratios are given in the following table:

Week	*Price*	*Ratio(y)*
0	1,245	-
1	1,230	0.988
2	1,280	1.041
3	1,392	1.088
4	1,431	1.028
5	1,428	0.998
6	1,439	1.008
7	1,346	0.935
8	1,265	0.940
9	1,513	1.196
10	1,468	0.970

$\Sigma y = 10.192 \quad \Sigma y^2 = 10.441562$

(a) Based on the specified model, explain why the 10 ratios constitute a random sample from a lognormal distribution.

(b) Calculate the mean and the standard deviation for the random sample of 10 ratios.

(c) Hence determine the method of moments estimates of the parameters μ and σ for the lognormal model. [6]

[Total 17]