EXAMINATIONS

27 September 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 Claim sizes are normally distributed about a mean μ = £6,000 and with standard deviation σ = £1000.

Calculate the probability that a claim is for more than £7,500, given that it is for more than £6,000. [3]

2 Let X be a random variable which has a Poisson distribution with parameter μ.

(i) Write down the cumulant generating function $C_X(t)$. [1]

(ii) By differentiation of $C_X(t)$ show that the mean and variance of X are both equal to μ. [2]

[Total 3]

3 A random sample of 25 observations from $X \sim N(\mu, 4)$ has sample mean $\bar{x} = 15.6$.

Calculate a symmetrical 90% confidence interval for μ. [3]

4 The following test concerning the mean claim amount (μ) for a certain class of policy

$$H_0: \mu = £200 \quad \text{v.} \quad H_1: \mu \neq £200$$

is to be performed. A random sample of 50 claims is examined and yields a mean amount of £207 and standard deviation £42.

Calculate the approximate probability-value for the test. [3]

5 When comparing the mean premiums for policies issued by two companies, a two-sample t test is performed assuming equal population variances. The sample sizes and sample variances are given by

$$n_1 = 25, \quad s_1^2 = 139.7$$
$$n_2 = 30, \quad s_2^2 = 76.6$$

Perform an appropriate F test at the 5% level to investigate the validity of the equal variance assumption. [3]

6 A 2×2 contingency table was set up to investigate whether or not two classification criteria are independent and resulted in the following data:

	I	*II*	
A	22	28	50
B	28	22	50
	50	50	100

Calculate the observed χ^2 test statistic and state an appropriate conclusion concerning the independence of the two criteria. [3]

7 A claim size distribution is modelled using a simple distribution with density of the form

$$f(x) = \begin{cases} k(100-x), & 0 \le x \le 100 \\ 0, & \text{otherwise} \end{cases}$$

(i) Verify that $k = 0.0002$. [1]

(ii) Determine the mean of this claim size distribution. [2]

(iii) Calculate the probability that an individual claim size is greater than 50. [1]

(iv) Calculate the probability that an individual claim size is less than 60 given that it is greater than 50. [3]

[Total 7]

8 Let $X_1, X_2, \ldots, X_{100}$ be independent random variables, each having a gamma(4,1) distribution (and hence with mean 4 and variance 4).

Calculate an approximate value for the probability that the sum of the variables assumes a value which exceeds 425. [3]

9 A statistician suggests that, since a t variable with k degrees of freedom is symmetrical with mean 0 and variance $\frac{k}{k-2}$ for $k > 2$, one can approximate the distribution using the normal variable $N\left(0, \frac{k}{k-2}\right)$.

(i) Use this to obtain an approximation for the upper 5% percentage points for a t variable with:

(a) 4 degrees of freedom, and
(b) 40 degrees of freedom

[2]

(ii) Compare your answers with the exact values from tables and comment briefly on the result. [2]

[Total 4]

10 In order to estimate a certain probability of success a single observation is taken from the binomial random variable $X \sim \text{binomial}(20, p)$.

(i) Write down an expression for the mean square error of the maximum likelihood estimator, $\hat{p} = \frac{X}{20}$, and evaluate this mean square error at $p = 0.5$.

[2]

(ii) Determine an expression for the mean square error of the estimator, $\tilde{p} = \frac{X+1}{21}$, and evaluate this mean square error at $p = 0.5$. [4]

(iii) Comment briefly on the comparison of $\hat{p}$ and $\tilde{p}$ as estimators of p in the case $p = 0.5$. [1]

[Total 7]

11 The continuous random variables X and Y have a bivariate probability density function

$$f(x, y) = 2 \quad \text{for } 0 < x + y < 1,\ x > 0, y > 0.$$

The conditional distribution of X given $Y = y$ is a uniform distribution with probability density function

$$f(x|y) = \frac{1}{1-y} \quad 0 < x < 1 - y$$

and the marginal distribution of Y is a beta distribution with probability density function

$$f(y) = 2(1 - y) \quad 0 < y < 1.$$

(i) Show that the conditional expectation of X given $Y = y$ is

$$E(X|Y = y) = \frac{1-y}{2},$$

and obtain the conditional variance of X given $Y = y$. [3]

(ii) Verify in this case that $\text{var}(X) = \text{var}(E(X|Y)) + E(\text{var}(X|Y))$. [3]

[Total 6]

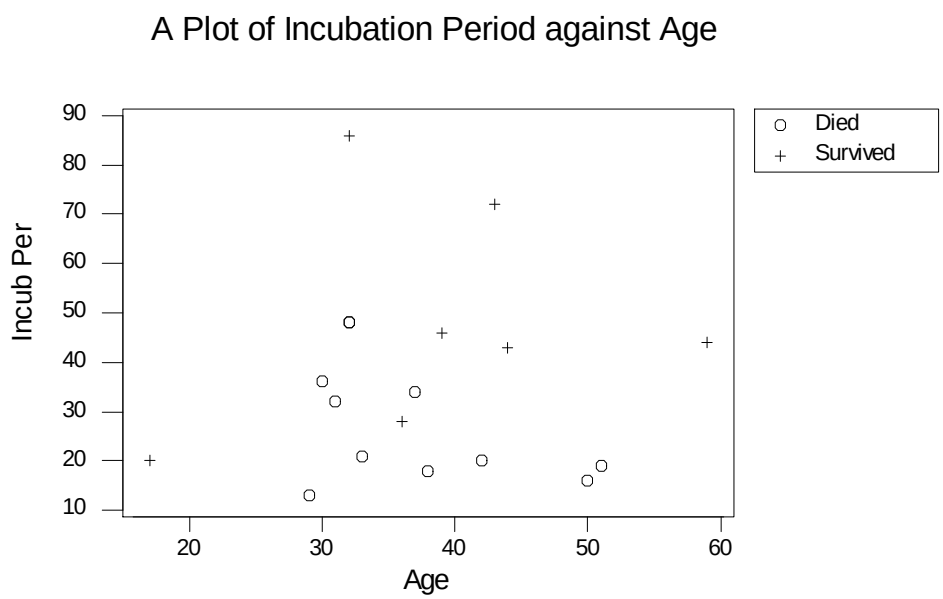
12 The following data refer to an outbreak of botulism, a form of food poisoning that may be fatal. Each subject is a person who contracted botulism in the outbreak. The variables recorded are the subject's age in years, the time in hours between eating the infected food and the first signs of illness (incubation period) and whether the subject survived (denoted by survival category Y) or died (denoted by survival category N).

Subject	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
Age (x)	29	39	44	37	42	17	38	43	51	30	32	59	33	31	32	32	36	50
Incubation period (y)	13	46	43	34	20	20	18	72	19	36	48	44	21	32	86	48	28	16
Survival	N	Y	Y	N	N	Y	N	Y	N	N	N	Y	N	N	Y	N	Y	N

Died: $\sum x = 405;\ \sum y = 305;\ \sum x^2 = 15517;\ \sum y^2 = 10035$

Survived: $\sum x = 270;\ \sum y = 339;\ \sum x^2 = 11396;\ \sum y^2 = 19665$

(i) A scatterplot of incubation period against age is given below, in which different symbols are used for subjects who died and for subjects who survived.



Comment briefly on any relationships between age and incubation period for those subjects who died and for those subjects who survived.

[2]

(ii) Construct suitable dotplots to investigate any relationship between:

(a) age and survival, and

(b) incubation period and survival

and make a brief informal comparison of the died and survived groups based on these dotplots. [3]

(iii) Construct 95% and 99% confidence intervals for the mean difference between the incubation period for subjects who survived and subjects who died (i.e. take the mean incubation period for subjects who survived minus the mean incubation period for subjects who died).

Comment briefly on these confidence intervals. [6]

(iv) (a) Conduct a test to investigate whether the variances of the incubation periods for subjects who died and subjects who survived are equal.

(b) Comment on the validity of the assumptions that are required for the confidence intervals given in part (iii) to be appropriate.

[4]

[Total 15]

13 For each of a group of policyholders the number of claims, Y, occurring in a period of one year is modelled by the following modified Poisson random variable, which incorporates a reluctance to claim:

there is a probability of α that Y equals zero and a probability of $(1 - \alpha)$ that Y comes from a Poisson distribution with mean μ, so that

$$P(Y = 0) = \alpha + (1-\alpha)P(X = 0)$$
$$P(Y = r) = (1-\alpha)P(X = r), r = 1, 2, 3, \ldots$$

where $X \sim$ Poisson(μ).

(i) Show that the mean and variance of Y are given by

$$E(Y) = (1-\alpha)\mu$$
$$V(Y) = (1-\alpha)\mu(1+\alpha\mu)$$

and comment briefly on these values in comparison to those for the unmodified Poisson variable, X. [5]

(ii) A random sample of such policyholders resulted in a distribution of numbers of claims with sample mean $\bar{y}$ and sample standard deviation s. Use the method of moments to determine estimators for α and μ in terms of $\bar{y}$ and s^2. [4]

(iii) Data on the numbers of claims from a sample of 200 policyholders resulted in the following frequency distribution.

number of claims:	0	1	2	3	4	5	>5
frequency:	90	56	37	12	4	1	0

(a) Calculate the mean and variance for this sample and hence calculate the method of moments estimates of α and μ.

(b) Using these estimates the expected frequencies under the fitted modified Poisson model were calculated for y = 0, 1, 2, 3 and are given in the table below.

y	*exp. freq.*
0	88.9
1	59.2
2	34.0
3	13.1

Calculate the expected frequencies for y = 4, 5 and y > 5 and comment briefly on the suitability of the model for these data.

[8]

[Total 17]

14 Consider the following data, which comprise four groups of claim sizes (y), each comprising four observations. In scenario I, information is also given on the sum assured under the policy concerned — the sum assured is the same for all four policies in a group. In scenario II, we regard the policies in the different groups as having been issued by four different companies — the policies in a group are all issued by the same company.

All monetary amounts are in units of £10,000. Summaries of the claim sizes in each group are given in a second table.

Group	*1*	*2*	*3*	*4*
Claim sizes y	0.11 0.46	0.52 1.43	1.48 2.05	1.52 2.36
	0.71 1.45	1.84 2.47	2.38 3.31	2.95 4.08
I: Sum assured x	1	2	3	4
II: Company	A	B	C	D

Summaries of claim sizes:

Group	*1*	*2*	*3*	*4*
Σy	2.73	6.26	9.22	10.91
Σy^2	2.8303	11.8018	23.0134	33.2289

(i) In scenario I, suppose we adopt the linear regression model

$$Y_i = \alpha + \beta x_i + e_i$$

where Y_i is the i^{th} claim size and x_i is the corresponding sum assured, $i = 1, \ldots, 16$.

(a) Calculate the total sum of squares and its partition into the regression (model) sum of squares and the residual (error) sum of squares.

(b) Fit the model and calculate the fitted values for the first claim size of group 1 (namely 0.11) and the last claim size of group 4 (namely 4.08).

(c) Consider a test of the hypothesis H_0: $\beta = 0$ against a two-sided alternative. By performing appropriate calculations, assess the strength of the evidence against this "no linear relationship" hypothesis.

[13]

(ii) In scenario II, suppose we adopt the analysis of variance model

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

where Y_{ij} is the j^{th} claim size for company i and τ_i is the i^{th} company effect, $j = 1, 2, 3, 4$ and i = A, B, C, D.

(a) Calculate the partition of the total sum of squares into the "between companies" (model) sum of squares and the "within companies" (residual/error) sum of squares.

(b) Fit the model.

(c) Calculate the fitted values for the first claim size of group 1 and the last claim size of group 4.

(d) Consider a test of the hypothesis H_0: $\tau_i = 0$, i = A, B, C, D against a general alternative. By performing appropriate calculations, assess the strength of the evidence against this "no company effects" hypothesis.

[10]

[Total 23]

END OF PAPER