EXAMINATIONS

18 September 2000 (pm)

Subject 101 — Statistical Modelling

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Write your surname in full, the initials of your other names and your Candidate's Number on the front of the answer booklet.*

2. *Mark allocations are shown in brackets.*

3. *Attempt all 16 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 and this question paper.

In addition to this paper you should have available graph paper, Actuarial Tables and an electronic calculator.

1 A random sample of fifty claim amounts (£) arising in a particular section of an insurance company's business are displayed below in a stem and leaf plot:

15	14678
16	0233368889
17	0000001233457888
18	3456779
19	0257
20	0
21	3
22	07
23	
24	
25	3
26	
27	3
28	8
29	
30	
31	2

Stem unit = 100
Leaf unit = 10

The sum of the fifty amounts (before rounding) is £92780.

Calculate the mean and median claim amounts. [2]

2 Consider a random sample of 47 white-collar workers and a random sample of 24 blue-collar workers from the workforce of a large company. The mean salary for the sample of white-collar workers is £28,470 and the standard deviation is £4,270; whereas the mean salary for the sample of blue-collar workers is £21,420 and the standard deviation is £3,020.

Calculate the mean and the standard deviation of the salaries in the combined sample of 71 employees. [4]

3 The number of claims arising in a period of one month from a group of policies can be modelled by a Poisson distribution with mean 24.

Determine the probability that fewer than 20 claims arise in a particular month. [2]

4 Suppose that a random sample of nine observations is taken from a normal distribution with mean $\mu = 0$. Let $\overline{X}$ and S^2 denote the sample mean and variance respectively.

Determine (to 2 decimal places) the probability that the value of $\overline{X}$ exceeds that of S, i.e. determine $P(\overline{X} > S)$. [3]

5 The number of claims which arise under a policy of a particular type in a year is to be modelled as a Poisson(λ) random variable. A random sample of 500 such policies gave rise to a total of 84 claims in 1999.

Calculate a 95% confidence interval for λ. [2]

6 Suppose that the linear regression model

$$Y = \alpha + \beta x + e$$

is fitted to data $\{(y_i, x_i) : i = 1, 2, \ldots, n\}$, where y is the salary (£) of a company manager and x (years) is the number of years of relevant experience of that manager.

State the units of measurement (if any) of

(a) $\hat{\alpha}$, the estimate of α ,
(b) $\hat{\beta}$, the estimate of β ,
(c) R^2 , the coefficient of determination of the fit. [2]

7 In a correlation analysis based on a random sample of 10 values from a bivariate normal distribution, a t-test of

$$H_0 : \rho = 0 \text{ v. } H_1 : \rho > 0$$

results in a probability-value of 0.025.

Calculate the value of the sample correlation coefficient. [3]

8 Claims on a certain class of policy are classified as being of two types, I and II. Past experience has shown that:

25% of claims are of type I and 75% are of type II;
Type I claim amounts have mean £500 and standard deviation £100;
Type II claim amounts have mean £300 and standard deviation £70.

Calculate the mean and the standard deviation of the claim amounts on this class of policy. [6]

9 The size of a claim, X, which arises under a certain type of insurance contract, is to be modelled using a gamma random variable with parameters α and θ (both > 0) such that the moment generating function of X is given by

$$M(t) = (1 - \theta t)^{-\alpha} \text{ , } t < 1/\theta \text{ .}$$

By using the cumulant generating function of X, or otherwise, show that the coefficient of skewness of the distribution of X is given by $2/\sqrt{\alpha}$. [5]

10 Let Z be a random variable with mean 0 and variance 1, and let X be a random variable independent of Z with mean 5 and variance 4. Let $Y = X - Z$.

Calculate the correlation coefficient between X and Y. [5]

11 Let X_1 , X_2 ,..., X_n be a random sample of size n from a distribution with density

$$f(x;\theta) = \frac{2x}{\theta^2}, 0 < x < \theta \quad : \theta > 0 .$$

(i) Write down the likelihood function clearly and hence by drawing a rough sketch of the likelihood function, show that the maximum likelihood estimator of θ is given by $\hat{\theta} = \max\{X_1, X_2, ..., X_n\}$. [4]

(ii) Without calculating $E(\hat{\theta})$, explain why $\hat{\theta}$ is a biased estimator of θ. [2]

[Total 6]

12 A market research company intends to estimate the proportion of the population, θ, who support a certain political party. They intend to poll a sample large enough so that a 95% confidence interval for θ has a width of 0.03 or less. It is thought that θ is approximately equal to 0.4.

Assuming that everyone questioned will respond to the poll, calculate the minimum size of sample which the company should take. [4]

13 Consider a one-way analysis of variance for comparing k treatments using the same number $n_i = r$ responses for each treatment. The model is

$$Y_{ij} = \mu + \tau_i + e_{ij} : i = 1, 2, ..., k; j = 1,2, ..., r$$

where the errors e_{ij} are independent $N(0, \sigma^2)$ random variables and where $\Sigma\tau_i = 0$. Show that the parameter estimates $\hat{\mu} = \bar{Y}_{..}$ and $\hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}$ are unbiased and that their variances are given by

$$V(\hat{\mu}) = \frac{\sigma^2}{kr} \text{ and } V(\hat{\tau}_i) = \frac{(k-1)\sigma^2}{kr}.$$ [6]

14 Let X be a random variable with cumulative distribution function

$$F_X(x) = P(X < x) = 1 - \exp(-x^2/\theta) \ , x > 0 \quad , \ F_X(x) = 0 \ , x \leq 0$$

and let $(X_1, X_2, \ldots, X_n)$ be a random sample from X.

(i) By considering $P(Y < y)$, show that $Y = X^2$ has an exponential distribution. [3]

(ii) (a) Show that the maximum likelihood estimator of θ is given by

$$\hat{\theta} = \frac{1}{n}\Sigma X_i^2$$

(b) Show that $\hat{\theta}$ is an unbiased estimator of θ which attains the Cramer-Rao lower bound on variance.

(c) Using moment generating functions, show that $\frac{2n}{\theta}\hat{\theta} \sim \chi^2_{2n}$. [11]

(iii) The above distribution of X is to be used as a model for claim amounts in a particular situation. A random sample of 50 such claim amounts (in appropriate units) gives the following summary:

$$\Sigma x^2 = 485.7518$$

(a) Calculate the maximum likelihood estimate of θ.

(b) Using the result of (ii)(c) above, calculate an exact 95% confidence interval for θ. [4]

[Total 18]

15 A manufacturer uses a certain type of electrical component from supplier A in high quality computing equipment and uses similar components from supplier B in inexpensive playstations. Previous investigations have shown that supplier A produces components whose resistances are normally distributed about a mean of 100 units and with a standard deviation of 0.1 units. Similarly, the resistances of supplier B's components are normally distributed about a mean of 100.5 units and with a standard deviation of 0.5. Components are supplied in large batches and are externally identical.

A batch known to come entirely from one supplier has no labels. The value of $\bar{X}$, the mean resistance from a random sample of components, is to be used to decide whether the batch has come from supplier A.

The hypotheses to be tested are:

H_0 : components come from supplier A i.e. $X \sim N(100, 0.1^2)$

v. H_1 : components come from supplier B i.e. $X \sim N(100.5, 0.5^2)$.

(i) (a) If the manufacturer insists that the probabilities of the type I and type II errors are each to be restricted to at most 5%, show that at least 4 components from the batch have to be examined.

(b) Using a sample size of 4, calculate the power of the test if the probability of the type I error is reduced to 1%. [11]

(ii) Suppose that a sample of 10 components from a batch has mean resistance $\bar{x} = 100.1$.

Calculate the probability-value of this observed mean. [4]

[Total 15]

16 At the end of the skiing season the tourist board in a mountain region examines the records of ten ski resorts. For each one it obtains the total number (y, thousands) of visitor-days during the season as a measure of the resort's popularity, and the ski-lift capacity (x, thousands), being the maximum number of skiers that can be transported per hour. The resulting data are given in the following table:

Resort:	A	B	C	D	E	F	G	H	I	J
Lift capacity x:	1.9	3.3	1.2	4.2	1.5	2.2	1.0	5.6	1.9	3.8
Visitor-days y:	15.1	22.6	9.2	37.5	8.9	21.1	5.8	41.0	9.2	32.4

$\Sigma x = 26.6$, $\Sigma x^2 = 91.08$, $\Sigma y = 202.8$, $\Sigma y^2 = 5603.12$, $\Sigma xy = 707.58$

(i) Draw a scatterplot of y against x and comment briefly on any relationship between a resort's popularity and its ski-lift capacity. [2]

(ii) Calculate the correlation coefficient between x and y and comment briefly in the light of your comment in part (i). [3]

(iii) Calculate the fitted linear regression equation of y on x. [2]

(iv) (a) Calculate the "total sum of squares" together with its partition into the "regression sum of squares" and the "residual sum of squares".

(b) Use the values in part (iv)(a) above to calculate the coefficient of determination R^2 and comment briefly on its relationship with the correlation coefficient calculated in part (ii).

(c) Use the values in part (iv)(a) above to calculate an estimate of the error variance σ^2 in the usual linear regression model. [6]

(v) Suppose that a resort can increase its ski-lift capacity by 500 skiers per hour.

Estimate the increase in the number of visitor-days it can expect in a season, and specify a standard error for this estimate. [4]

[Total 17]