EXAMINATIONS

5 September 2001 (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 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 and this question paper.

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

1 Data were collected on 100 consecutive days for the number of claims, x, arising from a group of policies. This resulted in the following frequency distribution

x:	0	1	2	3	4	≥5
f:	14	25	26	18	12	5

Calculate the median and interquartile range for these data. [2]

2 Let X_1 and X_2 be independent Poisson random variables with respective means μ_1 and μ_2.

Assuming the moment generating function of a Poisson random variable, determine the moment generating function of $X_1 + X_2$ and hence state the distribution of $X_1 + X_2$. [2]

3 Let $X_1, X_2, \ldots, X_5$ be independent $N(0,1)$ random variables and let

$$\bar{X} = \frac{1}{5}\sum_{i=1}^{5} X_i \quad \text{and} \quad S^2 = \frac{1}{4}\sum_{i=1}^{5}\left(X_i - \bar{X}\right)^2 .$$

Calculate $P\left[\bar{X} > 0 \quad \text{and} \quad \sum_{i=1}^{5}\left(X_i - \bar{X}\right)^2 < 9.488\right]$. [3]

4 Let $\{(x_i, y_i); i = 1, \ldots, n\}$ denote a set of n pairs of points with

$$\bar{x} = \sum_{i=1}^{n} x_i / n \quad \text{and} \quad \bar{y} = \sum_{i=1}^{n} y_i / n .$$

Assuming the usual expressions for the estimated coefficients, verify that the least squares fitted regression line of y on x passes through the point $(\bar{x}, \bar{y})$. [2]

5 The number of claims, X, to be processed in a day by an employee of an insurance company is modelled as $X \sim$ Poisson with mean 10. The time (minutes) the employee takes, Y, to process x claims is modelled as having a distribution with conditional mean and variance given by

$$E(Y \mid X = x) = 15x + 20 , \; V(Y \mid X = x) = x + 12.$$

Calculate the unconditional variance of the time the employee takes to process claims in a day. [3]

6 (i) The occurrence of claims in a group of 200 policies is modelled such that the probability of a claim arising in the next year is 0.015 independently for each policy. Each policy can give rise to a maximum of one claim.

Calculate an approximate value for the probability that more than 10 claims arise from this group of policies in the next year. [2]

(ii) The occurrence of claims in a group of 2000 policies is modelled such that the probability of a claim arising in the next year is 0.015 independently for each policy. Each policy can give rise to a maximum of one claim.

Calculate an approximate value for the probability that more than 40 claims arise from this group of policies in the next year. [3]

[Total 5]

7 The probability density function of a random variable X is given by

$$f(x) = \begin{cases} kx(1-ax^2)\,, & 0 \le x \le 1 \\ 0\,, & \text{otherwise} \end{cases}$$

where k and a are positive constants.

(i) Show that $a \le 1$, and determine the value of k in terms of a. [3]

(ii) For the case $a = 1$, determine the mean of X. [2]

[Total 5]

8 A job takes X minutes to complete, where X is modelled as a $N(28,2^2)$ random variable. Another job, independent of the first, takes Y minutes to complete, and begins 5 minutes after the first job begins. Y is modelled as a $N(25,1^2)$ random variable.

Calculate the probability that the job that was begun last is first to be completed. [4]

9 Two independent random samples of sizes n_1 and n_2 are selected from a normal population with variance σ^2. The sample variances are denoted by S_1^2 and S_2^2 respectively. Let S_W^2 denote a weighted average of the sample variances given by

$$S_W^2 = \alpha S_1^2 + (1 - \alpha) S_2^2 .$$

where α is a constant such that $0 \leq \alpha \leq 1$.

(i) Show that S_W^2 is an unbiased estimator of σ^2, and obtain an expression for the mean square error of S_W^2.

(You may use $\text{Var}(S_i^2) = \dfrac{2\sigma^4}{n_i - 1}$, $i = 1, 2$.) [3]

(ii) Show that S_W^2 has minimum mean square error if $\alpha = \dfrac{n_1 - 1}{n_1 + n_2 - 2}$. [2]

[Total 5]

10 The number of incomplete insurance proposals Y, in a batch of x proposals, is to be modelled as a Poisson random variable with mean λx, where λ is unknown. Data are available from n independent batches of proposals as follows: batch number i contains x_i proposals of which y_i are incomplete, $i = 1,2, \ldots, n$.

The least squares estimator of λ is that value of λ for which

$$\sum_{i=1}^{n} \left(Y_i - E(Y_i)\right)^2$$

is minimised.

(i) Show that the least squares estimator of λ is given by:

$$\tilde{\lambda} = \frac{\sum x_i Y_i}{\sum x_i^{\,2}} .$$ [3]

(ii) Determine $\hat{\lambda}$, the maximum likelihood estimator of λ. [3]

(iii) Determine whether neither, one, or both of $\tilde{\lambda}$ and $\hat{\lambda}$ provide unbiased estimators of λ. [2]

[Total 8]

11 A random sample of 11 policies on the contents of private houses was examined for each of three insurance companies and the sum insured under each policy noted. The observations were rounded to the nearest £100 and expressed in units of £1,000.

The sums and sums of squares of the observations are as follows:

	Sum	*Sum of squares*
Company 1	129.1	1,534.37
Company 2	109.8	1,109.88
Company 3	123.5	1,401.73

The data are to be analysed under the one-way analysis of variance model to examine whether company effects are present.

The ANOVA table is given below, with three entries deleted.

Source of variation	d.f.	SS	MSS
Between companies	2	***	***
Residual	30	48.24	1.61
	32	***	

(i) Copy the table into your answer book, filling in the three values which have been deleted. [2]

(ii) Test the null hypothesis of no company effects and state your conclusion. [2]

[Total 4]

12 Claims are classified on inception into one of three categories, "simple", "standard" and "complex". Last year the percentage of all claims classified in each of these categories was 18.4%, 70.3% and 11.3% respectively.

A random sample of 120 of this year's claims to date shows that the numbers classified in each category are 15, 87 and 18 respectively.

Perform a goodness-of-fit test to investigate whether this year's pattern to date differs from that of last year, and state your conclusion. [5]

13 Twenty overweight executives take part in an experiment to compare the effectiveness of two exercise methods, A (isometric), and B (isotonic). They are allocated at random to the two methods, ten to isometric, ten to isotonic methods. After several weeks, the reductions in abdomen measurements are recorded in centimetres with the following results:

A (isometric method)	3.1	2.1	3.3	2.7	3.4	2.7	2.7	3.0	3.0	1.6
B (isotonic method)	4.5	4.1	2.7	2.2	4.7	2.2	3.6	3.0	3.3	3.4

(i) (a) Plot the data for the two exercise methods on a single diagram. Comment on whether the response values for each exercise method are well modelled by normal random variables.

(b) Perform a test to investigate whether the assumption of equal variability for the responses for the two exercise methods is reasonable.

(c) Perform a *t*-test to investigate whether these data support the claim that the isotonic method is more effective than the other method. [9]

(ii) (a) Determine a two-sided 95% confidence interval for the difference in the means for the two exercise methods.

(b) Assuming that the two sets of 10 measurements are taken from normal populations with the same variance, determine a 95% confidence interval for the common standard deviation. [7]

[Total 16]

14 In a medical study on hypertension amongst young male athletes the researchers were interested in the effects of the use of a particular (legal) stimulant on systolic blood pressure (*bp*).

Ten young male athletes from a larger group who had agreed to take part in the study were selected at random — none of those in the group were currently users of the stimulant. The *initial bp* of each of the sample was measured in controlled conditions. Each sample member was then exposed to the use of the stimulant in a controlled manner for a fixed period of time. Each sample member was subject to a similar exercise regime, and at the end of the period the *bp* of each of the sample was again measured in the same controlled conditions as initially, giving the *follow-up bp*.

The data obtained were as follows:

Athlete	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Initial bp	116	107	129	119	116	113	135	121	112	123
Follow-up bp	123	111	140	129	130	118	143	128	110	132

Summaries: *Initial* $\Sigma x = 1191$, $\Sigma x^2 = 142471$ *Follow-up* $\Sigma x = 1264$, $\Sigma x^2 = 160872$

The following models are proposed as possible bases for the analysis:

(M1) The *initial bp* of the ith athlete, X_i, is distributed as $X_i \sim N(\mu, \sigma_1^2)$ and the *follow-up bp*, Y_i, is distributed such that $Y_i \mid X_i = x \sim N(x + \alpha, \sigma_2^2)$.

(M2) The *initial bp* of the ith athlete, X_i, is distributed as $X_i \sim N(\mu_i, \sigma_1^2)$ and the *follow-up bp*, Y_i, is distributed such that $Y_i \mid X_i = x \sim N(x + \alpha, \sigma_2^2)$.

(M3) The *initial bp* of the ith athlete, X_i, is distributed as $X_i \sim N(\mu_i, \sigma_1^2)$ and the *follow-up bp*, Y_i, is distributed such that $Y_i \mid X_i = x \sim N(x + \alpha_i, \sigma_2^2)$.

(i) (a) Explain briefly the differences between the physical assumptions underlying the three models proposed above.

(b) Explain briefly why model (M3) above could not be used in the analysis of the data as given. [7]

(ii) Adopting model (M1) above

(a) Using the *initial bp* data, calculate a symmetrical, two-sided, 95% confidence interval for μ, and

(b) Calculate a point estimate of α. [6]

(iii) Adopting model (M2) above, calculate a one-sided 95% confidence interval for α which brings out the minimum realistic value for the mean increase in *bp* attributable to taking the stimulant. [6]

[Total 19]

15 In a study into employee share ownership plans, data were obtained from ten large insurance companies on the following two variables:

employee satisfaction with the plan (x);
employee commitment to the company (y).

For each company a random sample (of the same size) of employees completed questionnaires in which satisfaction and commitment were recorded on a 1–10 scale, with 1 representing low satisfaction/commitment and 10 representing high satisfaction/commitment. The resulting means provide each company's employees' satisfaction and commitment score. These scores are given in the following table:

Co.	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
x	5.05	4.12	5.38	4.17	3.81	4.47	5.41	4.88	4.64	5.19
y	5.36	4.59	5.42	4.35	4.03	5.34	5.64	4.89	4.52	5.88

$\sum x = 47.12$, $\sum x^2 = 224.8554$, $\sum y = 50.02$, $\sum y^2 = 253.5796$, $\sum xy = 238.3676$

(i) Draw a scatterplot of y against x and comment briefly on any relationship between employee satisfaction and commitment. [2]

(ii) Calculate the fitted linear regression equation of y on x. [3]

(iii) Calculate the coefficient of determination R^2 and relate its value to your comment in part (i). [2]

(iv) Assuming the full normal model, calculate an estimate of the error variance σ^2 and obtain a 95% confidence interval for σ^2. [3]

(v) Calculate a 95% confidence interval for the true underlying slope coefficient. [3]

(vi) For companies with an employees' satisfaction score of 5.0, calculate an estimate of the expected employees' commitment score together with 95% confidence limits. [4]

[Total 17]