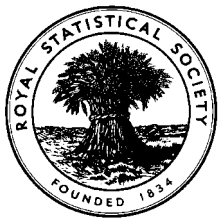
EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



GRADUATE DIPLOMA, 1999

Applied Statistics I

Time Allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use silent, cordless, non-programmable electronic calculators.

Where a calculator is used the ***method*** *of calculation should be stated in full.*

Note that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$ *and that* ln *stands for* $\log_e$.

This examination paper consists of 12 printed pages. This front cover is page 1. The reverse of the front cover, which is intentionally left blank, is page 2. Question 1 starts on page 3.

There are 8 questions altogether in the paper.

1. Fee paying schools in a country have been classified by region (South, Midlands and North) and also by gender balance (mainly boys, mixed, and mainly girls). Within each of these nine categories, five schools were selected at random and the percentage of students entering higher education in the academic year 1997-8 noted, the results being given in the table below.

Percentage of students entering higher education (1997-8)

	Region															
Gender balance	*South*					*Midlands*					*North*					
B	93	86	91	92	97	95	92	100	90	95	89	89	80	62	91	
M	90	91	96	72	80	93	60	90	80	83	90	91	90	91	82	
G	80	80	56	70	94	100	77	86	83	100	92	95	96	96	87	

B = mainly boys, M = mixed, G = mainly girls

Interest centres on whether the percentage varies with either of the factors and, if so, how.

(i) It has been suggested that these data could be analysed by a two factor analysis of variance. Explain why it may be considered desirable to apply the inverse sine transformation before continuing with the analysis. What requirement should be satisfied in this case for the transformation to be appropriate?

(4)

(ii) The following table gives certain statistics for each category where $x = \sin^{-1}\sqrt{p}$, where p is the proportion of students entering higher education.

	South	*Midlands*	*North*
Mainly boys	$\Sigma x = 6.4370$ $\Sigma x^2 = 8.3102$	6.7945 9.2958	5.7455 6.6899
Mixed	6.0050 7.2908	5.6910 6.5818	6.1630 7.6092
Mainly girls	5.3745 5.9001	6.5455 8.8036	6.5700 8.6539

Complete the analysis on x to examine which factors (if any) affect the percentage of students entering higher education. Produce a suitable diagram to illustrate your conclusions.

(16)

Turn over

2. (i) Let X be a random variable with mean μ and variance σ^2 where $\sigma^2 \ll \mu$. Obtain first approximations for the mean and variance of $1/X$ in terms of μ and σ^2.

(4)

(ii) From a population of wild animals of size N, a random sample of m is selected and each animal marked. These animals are then released and, after a period of time, a second random sample of n is selected. Let the random variable X denote the number found to be marked. Show that the probability mass function of X is

$$P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

where $\max(0, n-N+m) \le x \le \min(n, m)$.

(4)

(iii) It is required to estimate N from a knowledge of m, n and x. Explain why $\hat{N} = \frac{mn}{x}$ is an intuitively acceptable estimator for N and show that, to a first approximation, it has mean N and variance

$$\frac{(N-m)(N-n)N}{mn}.$$

You may assume that X has mean $\frac{nm}{N}$ and variance

$$\frac{nm(N-m)(N-n)}{N^2(N-1)}.$$

(7)

(iv) If $m = n = 100$ and $x = 20$, obtain an approximate 95% confidence interval for N using a Normal approximation for the distribution of $\hat{N}$.

(5)

3. (i) In many regressions, the regressor variables are highly correlated. What are the problems which may be encountered as a result of this? Suggest a way in which you could attempt to overcome these problems.

(5)

(ii) An investigation into the effect of temperature on electrical resistivity has used three different alloys *A*, *B* and *C*. Measurements were taken at nine different temperatures for each of the alloys.

Various regression models have been fitted to these data with x_1 as temperature and x_2 and x_3 as dummy variables representing the three alloys with coding

Alloy	x_2	x_3
A	0	0
B	1	0
C	0	1

The regression sums of squares for some of the fitted models are

Variables in model	*Regression sum of squares*
x_1	2847
$x_1 \quad x_2 \quad x_3$	35232
$x_1 \quad x_1x_2 \quad x_1x_3$	33800
$x_1 \quad x_2 \quad x_3 \quad x_1x_2 \quad x_1x_3$	35517
$x_1 \quad x_1^2$	2947
$x_1 \quad x_2 \quad x_3 \quad x_1x_2 \quad x_1x_3 \quad x_1^2x_2 \quad x_1^2x_3 \quad x_1^2$	35669

The corrected total sum of squares is 35673.

An engineer is considering the following models for these data:

(a) A common straight line for all three alloys.
(b) Parallel straight lines with a different intercept for each alloy.
(c) Every alloy represented by a straight line with different slopes and intercepts.
(d) Every alloy represented by a unique quadratic curve.

Write down the regression sum of squares corresponding to each model.

Using the information given in the table above, select the simplest model appropriate for the data. Justify your selection. What other information might help in your decision?

(12)

How could you test in (b) whether a different intercept is necessary for each alloy?

(3)

Turn over

4. (i) State the *Gauss-Markov theorem* and explain its importance to estimation in the general linear model.

(4)

(ii) (a) A general linear model relating n observations of a response variable to $p-1$ predictor variables is given by

$$E(Y) = X\beta \qquad \text{Var}(Y) = I_n\sigma^2$$

where X is an $n \times p$ design matrix of full rank and β is a $p \times 1$ vector of parameters, the first parameter being an intercept.

Derive the least squares estimator $\hat{\beta}$ of β and obtain its expectation and dispersion (variance-covariance) matrix.

(7)

(b) A data set to be modelled using the general linear model with $n = 50$ and $p = 3$ gives

$$X^T X = \begin{bmatrix} 50 & 69 & 12423 \\ 69 & 113 & 17098 \\ 12423 & 17098 & 3199537 \end{bmatrix} \qquad X^T Y = \begin{bmatrix} 604 \\ 791 \\ 146578 \end{bmatrix}$$

$$\left[X^T X\right]^{-1} = \begin{bmatrix} 0.690129 & -0.083363 & -0.002234 \\ -0.083363 & 0.056302 & 0.000023 \\ -0.002234 & 0.000023 & 0.000009 \end{bmatrix}$$

$Y^T Y = 11194$.

Calculate $\hat{\beta}$ and obtain a 95% confidence interval for β_0, the intercept term.

Calculate the (uncorrected) regression sum of squares for the model involving β_0 only and hence test the hypothesis that simultaneously $\beta_1 = \beta_2 = 0$. You may assume Normality.

(9)

5. Six batches of antibiotic were randomly selected from a large number of batches. Each batch was then stored for a period of six months, three batches under a set of conditions *A* and three under a different set *B*. At the end of the period, four independent estimates of the potency (in coded units) of each batch were made with the following results.

Conditions *A*			**Conditions *B***		
Batch			*Batch*		
1	*2*	*3*	*4*	*5*	*6*
23	32	28	14	7	13
20	35	28	12	9	17
21	28	30	15	10	18
19	26	31	10	10	15

A model suggested for these data is

$$y_{ijk} = \mu + C_i + b_{j(i)} + \varepsilon_{ijk} \qquad \begin{matrix} i = 1,2 \\ j = 1,2,3 \\ k = 1,2,3,4 \end{matrix}$$

(i) Interpret each of the terms in this model.

(4)

(ii) A partially completed analysis of variance using this model gave

Source	*Sum of squares*	*Expected mean square*
Between conditions		$\sigma^2 + 4\sigma_b^2 + 12\sum_{i=1}^{2} C_i^2$
Within conditions, between batches		
Within batches	99.75	σ^2
Total	1627.625	

Derive the expected mean square for within conditions, between batches.

(4)

Complete the analysis of variance, reporting on the relative effectiveness of the two methods of storage and the variation in potency between batches.

(10)

What assumptions have you made?

(2)

Turn over

6. (i) Define a moving average series of order q, $MA(q)$, and an autoregressive series of order p, $AR(p)$, in the context of a stationary time series.

(4)

(ii) A stationary time series $\{X_t\}$ is given by

$$X_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2}$$

where a_t is white noise with zero mean and variance σ^2.

(a) Let ρ_k denote the autocorrelation function of $\{X_t\}$. Determine ρ_k in terms of θ_1, θ_2 and σ^2 for all k.

(8)

(b) A new series $\{Z_t\}$ is generated from $\{X_t\}$ by defining

$$Z_t = (X_t + X_{t-1})/2 .$$

Obtain the variance of Z_t and hence determine the values of θ_1 and θ_2 which minimise this variance.

(8)

7. (i) (a) Define the first and second principal components of a data set. Why do principal components extracted from a dispersion (variance-covariance) matrix differ from those extracted from a correlation matrix? Discuss the relative advantages of using each of these two matrices to perform a principal components analysis.

(8)

(b) Describe briefly three reasons why an analyst may wish to perform a principal components analysis.

(4)

(ii) Five variables have been recorded for each of fifty randomly selected households within an area. The variables are

x_1 - first monthly income in household (£)
x_2 - second monthly income in household (£)
x_3 - total debts of household excluding mortgage (£)
x_4 - monthly mortgage payment (£)
x_5 - monthly payment for gas, electricity and water (£).

Principal components have been extracted from the correlation matrix for these data, the first three being given in the table below.

		Component		
		1	*2*	*3*
	x_1	-0.46	-0.18	+0.67
	x_2	-0.24	-0.55	-0.68
Variable	x_3	-0.50	-0.48	+0.11
	x_4	-0.48	+0.47	-0.22
	x_5	-0.49	+0.47	-0.20
Eigenvalue		2.59	1.47	0.89

Comment on the apparent dimensionality of these data. Briefly attempt an interpretation of these components.

(8)

8. One hundred and seventy investment trusts have been classified according to whether they are invested mainly in the United Kingdom or not, whether they have been in existence for at least seven years or not and whether they have given an above average return in the last year or not (source: Money Observer). The results were as follows.

		Invested mainly in the UK			
		YES		**NO**	
		In existence for at least 7 years		*In existence for at least 7 years*	
		YES	*NO*	*YES*	*NO*
Above average return	*YES*	30	25	18	9
	NO	11	19	17	41

A log linear model has been proposed for these data, the full model being

$$\log_e \lambda_{ijk} = \mu + R_i + E_j + I_k + (RE)_{ij} + (RI)_{ik} + (EI)_{jk} + (REI)_{ijk}$$

where R is the return factor, E the existence factor, I the investment factor and λ_{ijk} is the expected frequency in cell ijk.

(i) Using this example, describe the main features of a generalised linear model. (5)

(ii) Part of the output from a statistical package used to fit the proposed log linear model containing first and second order effects only gave

Term	*Coefficient*
μ	+2.953
R_1	-0.011
I_1	+0.011
E_1	-0.104
$(RI)_{11}$	+0.348
$(RE)_{11}$	+0.284
$(IE)_{11}$	-0.021

Note that level 1 for all factors is the YES level and that all factors are subject to the sum to zero constraint, e.g. $R_1 + R_2 = 0$.

Use the model to calculate the expected number of investment trusts producing an above average return last year which invest mainly in the UK and have been in existence for less than seven years. (6)

(question 8 continued on next page)

(iii) Various log linear models have been fitted with the following results:

Terms in model	*Deviance*
μ	34.94
μ, R, E, I	32.82
μ, R, E, I, RE	20.42
μ, R, E, I, RI	14.00
μ, R, E, I, EI	31.96
μ, R, E, I, RE, RI	1.60
μ, R, E, I, RE, EI	19.56
μ, R, E, I, RI, EI	13.14
μ, R, E, I, RE, RI, EI	1.54

Use these results to show that both E and I appear to affect the chance of a trust having given an above average return last year.

(9)

BLANK PAGE