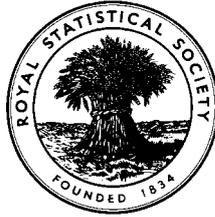# EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA IN STATISTICS, 1998

Applied Statistics I

Time Allowed: Three Hours

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

All questions carry equal marks.

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* $\log_{e}$.

1. (a) Under what circumstances might it be desirable to transform data before performing a statistical analysis when using a linear model?

    (b) A manufacturer of dyestuffs has two possible sources of a raw material. Six estimates of the percentage impurity in the raw material from each source gave:

	% Impurity						*Mean*	*Standard Deviation*
Source A	2.5	0.8	1.0	15.3	5.1	4.0	4.78	5.42
Source B	7.5	17.7	19.3	50.8	5.4	2.1	17.13	17.86

    (i)   The manufacturer is interested in whether the percentage impurity differs between the two suppliers. Explain why it might be appropriate to apply a logarithmic transformation to these data before further analysis.

    (ii) By using this transformation, obtain a 95% confidence interval for the ratio of the mean percentage impurities for the two suppliers.

Turn over

1

2. As part of an investigation into the public perception of the quality of meat, three members of the public were asked to judge the quality of each of nine pieces of beef. The table gives the marks awarded.

		Judge 1	Judge 2	Judge 3	$\sum x$
	1	47	31	43	121
	2	72	30	53	155
	3	61	27	22	110
Beef	4	66	48	28	142
piece	5	37	20	75	132
	6	76	21	53	150
	7	64	30	15	109
	8	21	5	5	31
	9	71	55	27	153
	$\sum x$	515	267	321	
	$\sum x^2$	32193	9685	15339	

(i) Under what circumstances is a random effects (type 2) model appropriate for these data?

(ii) Assuming the random effects model to be appropriate, report on the relative importance of the factors affecting the marks awarded. You should include estimates of all components of variance.

(iii) Explain, in terms understandable to a non-statistician, the meaning of your estimates of the components of variance. You may assume the non-statistician understands the meaning of "variance".

(iv) Obtain a 95% confidence interval for the residual (error) variance.

3. (a) (i) Define the autocorrelation function and partial autocorrelation function of a stationary time series and explain how these functions can be used to identify suitable models for these series.

(ii) The non-stationary series $\{X_t\}$ is generated by

$$X_t = X_{t-1} + a_t - \theta a_{t-1} \qquad\qquad |\theta| < 1$$

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

Show that by considering first differences of this series the problem of non-stationarity can be overcome.

(b) The price of a share, X_t, on 201 successive trading days has been recorded. The first 10 values of the sample autocorrelation function $(\hat{r}_k)$ and the sample partial autocorrelation function $(\hat{\phi}_{kk})$ for X_t and for first differences $\nabla_t = X_t - X_{t-1}$ are as follows:

lag

		1	2	3	4	5	6	7	8	9	10
X_t	$\hat{r}_k$	0.36	0.35	0.41	0.36	0.34	0.30	0.25	0.28	0.30	0.23
	$\hat{\phi}_{kk}$	0.36	0.26	0.27	0.16	0.11	0.04	−0.03	0.04	0.09	0.00
∇_t	$\hat{r}_k$	−0.49	−0.07	0.10	−0.03	0.02	0.02	−0.07	−0.01	0.09	−0.03
	$\hat{\phi}_{kk}$	−0.49	−0.42	−0.25	−0.19	−0.11	−0.03	−0.08	−0.16	−0.06	−0.02

Select a suitable model for the series, giving reasons for your choice. Describe briefly how you would check on the fit of your chosen model.

Turn over

4. A large tank contains β_1 litres of liquid and from it β_2 litres are removed. Neither β_1 nor β_2 are known exactly but they have been estimated by the measurements y_1 and y_2. An estimate of the volume of liquid remaining in the tank is given by y_3. All measurements are known to be unbiased and uncorrelated with one another.

   (i) Assuming equal variances, obtain the least squares estimators for β_1 and β_2 in terms of y_1, y_2 and y_3.

   (ii) If $y_1 = 1000$, $y_2 = 305$ and $y_3 = 690$ obtain a 95% confidence interval for the amount of liquid remaining in the container. You may assume normality.

   (iii) Due to differing measurement techniques it has been suggested that y_1 and y_3 have the same variance whereas y_2 has a variance half this value. Re-estimate β_1 and β_2 in (i) using this extra information.

5. As part of an investigation into the use of the Gaelic language amongst adults in the outer Hebrides, data have been collected on 257 subjects. The data include proficiency in the language of the subject (y), of the father (x_1), of the mother (x_2) and the age of the subject in years (x_3). Various linear regression models relating y to the other variables have been fitted with the following results:

Variables in Model	Regression Coefficients			Regression sum of Squares
	x_1	x_2	x_3	
x_1	4.577			2461.8
x_2		2.508		2115.3
x_3			−0.0037	1.0
x_1, x_2	1.705	1.015		2598.3
x_1, x_3	2.458		−0.0261	2519.5
x_2, x_3		2.507	−0.0032	2115.7
x_1, x_2, x_3	1.799	0.932	−0.0199	2629.7

The corrected total sum of squares is 5249.0

   (i) Explain why the partial regression coefficient for x_1 changes according to the other variables included in the model.

   (ii) The computer output for each model identifies certain subjects as having <u>high influence</u> and/or <u>large standardised residuals</u>. Explain the meaning of each of these terms and how their detection might affect the conclusions you draw from the analysis.

   (iii) On the basis of the information supplied, determine which of the variables should be included in the model. Use any appropriate method with a nominal significance level of 1%.

6. An organisation is investigating the effects of age and sex on whether an individual is prepared to take part in a survey. The results of a pilot investigation are given in the table in the form r/n where r is the number prepared to take part out of n asked.

	Age < 30	30≤Age≤59	Age ≥ 60	Total
Male	$30/52$	$18/42$	$8/19$	$56/113$
Female	$27/56$	$28/51$	$19/37$	$74/144$

A logistic regression model has been proposed for these data:

$$\text{Logit}\,(\Pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where Π is the proportion of individuals prepared to take part in the survey, x_1 is the sex of the individual (0 = male, 1 = female) and x_2, x_3 are dummy variables representing the three age groups with coding

Age group	x_2	x_3
Age < 30	0	0
30≤Age≤59	1	0
Age ≥ 60	0	1

A statistical package provides the following estimates for the model:

Coefficient	Estimate
β_0	0.0655
β_1	0.0884
β_2	−0.1354
β_3	−0.1953

(i) Define the logit transformation for proportions.

(ii) What are the advantages and disadvantages of using this method of coding the age groups, as opposed to using a single variable which takes the values 0, 1, and 2?

(iii) Extend the given model to allow for the effect of age on the logit to differ between the sexes.

(iv) Explain why, when fitting another logistic regression model using the row margins (i.e. the Total column in the table above) only, the deviance is zero with zero degrees of freedom.

(v) Estimate from the model the proportion of females aged 60 or over who would be prepared to take part in the survey.

Turn over

5

7. A clinical scientist is involved in testing subjects to determine whether they are exaggerating their hearing loss (exaggerators) or not (honest). Two tests are given and past experience has shown that the results, x_1 and x_2, have a multivariate normal distribution with the following parameters:

Exaggerators: $\quad \mathbf{\mu}^T = [20 \ 19] \qquad \Sigma = \begin{bmatrix} 98 & 57 \\ 57 & 92 \end{bmatrix}$

Honest: $\qquad \mathbf{\mu}^T = [11 \ 11] \qquad \Sigma = \begin{bmatrix} 98 & 57 \\ 57 & 92 \end{bmatrix}$

(i) State the criterion used to determine Fisher's linear discriminant function for discriminating between these two groups.

(ii) Show that the function $0.0645 \, x_1 + 0.0470 \, x_2$ is a suitable discriminant function.

(iii) Assuming a subject is equally likely to be honest or an exaggerator show that use of the function will result in the overall proportion of subjects classified incorrectly being 0.312.

(iv) From past records it is found that 90% of subjects tested are honest. If a new decision rule
    "If discriminant function < 2, subject is honest"
is applied, obtain the overall proportion of subjects who will be incorrectly classified.

8. An organisation wishes to investigate the effect of speed driven on the fuel consumption of cars. As it is known that engine size can also affect fuel consumption, a series of trials has been conducted using similar cars but with different engine sizes driven at various speeds. As far as was possible the conditions for each trial were the same. Three readings were obtained on fuel consumption (x) for each combination of engine size and speed with results as given in the following table

Size of engine (cc)

		1100	1500	1800	Σx	Σx^2
	30	43.8, 45.2, 44.9	38.0, 38.6, 37.4	37.1, 35.5, 35.2	355.7	14185.91
Speed (mph)	50	43.0, 42.1, 43.4	41.7, 41.6, 39.9	41.0, 40.0, 41.6	374.3	15577.99
	70	32.6, 31.4, 32.4	29.9, 28.0, 29.0	29.9, 31.1, 31.8	276.1	8489.95
	Σx	358.8	324.1	323.2		
	Σx^2	14580.94	11913.19	11759.72		

The figures in the table are fuel consumption given as miles per gallon.

(i) Specify a suitable model for these data, carefully explaining the meaning of all the terms. State any assumptions that you have made.

(ii) Complete the analysis and report on the effect of speed driven on fuel consumption.