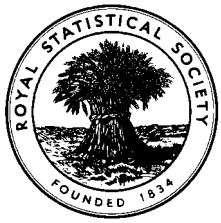
EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



GRADUATE DIPLOMA IN STATISTICS, 1996

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. In a social investigation in Sheffield, 291 male and 291 females selected from two areas were each asked whether they were satisfied or not with their present employment. The results were:

	Male		**Female**	
	Area 1	*Area 2*	*Area 1*	*Area 2*
Satisfied	123	55	90	63
Dissatisfied	92	21	89	49

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

$$\log_e \lambda_{ijk} = \mu + S_i + M_j + A_k + (SM)_{ij} + (SA)_{ik} + (MA)_{jk} + (SMA)_{ijk}$$

where *S* is the satisfaction factor, *M* is the sex factor, *A* is the area factor and λ_{ijk} is the expected frequency in cell *ijk*.

(i) Interpret, in terms understandable to a non-statistician, the terms $(SM)_{ij}$ and $(SMA)_{ijk}$.

(ii) Part of the output for fitting the model with first order effects only gave

Factor	*Level*	*Coefficient*
Constant		4.211
S	Satisfied	0.138
	Dissatisfied	-0.138
M	Male	0.000
	Female	0.000
A	Area 1	0.370
	Area 2	-0.370

Explain why the coefficients for M_j must be 0 for this data set. State, giving your reasons, whether this means that sex may be ignored and that the data may be collapsed into a single 2 x 2 table.

Using this model, calculate the expected number of male respondents in area 1 who are satisfied with their employment.

(Question continued on next page)

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

Terms in model	*Deviance*
μ	106.8
μ , *S, M, A*	21.2
μ , *S, M, A, SM*	16.8
μ , *S, M,,A, SA*	17.2
μ , *S, M, A, AM*	11.0
μ , *S, M, A, SM, SA*	12.9
μ , *S, M, A, SM, AM*	6.6
μ , *S, M, A, AM, SA*	7.0
μ , *S, M, A, SM, SA, AM*	1.3

Write a short report on the factors appearing to affect whether a respondent is satisfied or not with their present employment.

2. A random variable *X* has mean μ and variance σ^2 and *f(X)* is a function of *X*. Show that to a first approximation *f(X)* has mean *f(* μ *)* and variance $\sigma^2 (f'(\mu))^2$ where $f'(\mu) = \left.\frac{df(x)}{dx}\right|_{x=\mu}$. What is a necessary condition for the approximations to be reasonable?

Show that if *X* has a Poisson distribution with parameter λ then $\sqrt{X}$ has approximately a mean of $\sqrt{\lambda}$ and a variance of 1/4.

In an attempt to model traffic flow, the number of vehicles passing a given point in a 10 minute interval was noted on 5 occasions on each of 3 different days with the following results:

	Number of vehicles passing in a 10 minute interval				
Day 1	5	5	4	1	3
Day 2	14	9	5	12	10
Day 3	5	4	4	3	8

The analyst wishes to use the Poisson distribution to model these data and has performed a one factor analysis of variance after using the square root transformation, the result being

Source	*Degrees of freedom*	*Sums of squares*
Between days	2	4.440
Within days	12	3.046
Total	14	7.486

(Question continued on next page)

Using the analysis of variance table, confirm that the variance of the transformed variable is approximately 1/4. Complete the analysis to test whether there is a difference in the mean rate of traffic passing the point on different days.

What conditions must be satisfied for the Poisson distribution to be an appropriate model for these data?

3. In the context of stationary time series, define an autoregressive series of order *p*, AR(*p*), a moving average series of order *q* , MA(*q*), and a mixed autoregressive moving average series of orders *p, q*, ARMA (*p, q).*

Show that a stationary autoregressive series of order 1 may be represented as an infinite-order moving average series and obtain the autocorrelation function of this series.

The weight of a subject has been recorded weekly for a period of two years. Several models have been fitted to these data, an extract of the computer output being given below.

Model Fitted	*Parameter Estimates*					*Residual*
	AR(1)	AR(2)	MA(1)	MA(2)	Constant	*Mean Square*
AR(1)	0.646				50.12	2.33
	(0.08)				(0.15)	
AR(2)	0.506	0.224			38.18	2.23
	(0.10)	(0.10)			(0.15)	
MA(1)			-0.48		141.6	2.84
			(0.09)		(0.25)	
MA(2)			-0.56	-0.27	141.6	2.60
			(0.10)	(0.10)	(0.29)	
ARMA(1,1)	0.934		0.558		9.32	2.08
	(0.05)		(0.11)		(0.06)	
ARMA (2,1)	1.005	-0.06	0.600		8.29	2.10
	(0.22)	(0.18)	(0.19)		(0.06)	
ARMA (1,2)	0.942		0.522	0.067	8.29	2.10
	(0.05)		(0.11)	(0.11)	(0.06)	
ARMA (2,2)	0.538	0.376	0.111	0.294	12.23	2.11
	(1.09)	(1.00)	(1.07)	(0.54)	(0.09)	

(*Note that the figures in parentheses are standard errors of the estimates.)*

Which of the models do you consider the most suitable for these data? Give reasons for your choice. How could you check on the fit of the model you have selected?

Turn over

4. State the Gauss-Markov theorem and explain its importance in the context of estimation of parameters in the general linear model.

A general linear model relating a response variable y to p-1 predictor variables $X_1,\ldots,X_{p-1}$ is given by

$$E(y)=X\beta \quad \text{Var}\,(y) = I\sigma^2$$

where X is the design matrix, β is a px1 vector of parameters and I is the identity matrix.

Four observations y_1, y_2, y_3 and y_4 are taken with expectations $\beta_1+\beta_2, \beta_1-\beta_2, \beta_1+\beta_3$ and $\beta_1-\beta_3$ respectively. Write down the design matrix X, obtain $\hat{\beta}$, the least squares estimator of β, in terms of the y_i $(i=1,\ldots,4)$ and also obtain the dispersion (variance/covariance) matrix of $\hat{\beta}$. What are the advantages of the dispersion matrix being diagonal?

If $y_1=7, y_2=5, y_3=11$ and $y_4=5$, obtain an estimate of σ^2.

5. The data in the table show the lengths of time (in seconds) taken by rats to find their way out of a maze. Three strains of rat with differing degrees of intelligence were used, the rats being kept under one of two conditions for a period of three months before being placed in the maze. Four different rats were used for each possible combination of strain and condition.

		Intelligence					
		Bright		*Average*		*Dull*	
Condition	*Free*	34	16	107	101	130	110
		39	33	81	98	107	102
	Restricted	125	93	121	132	95	108
		127	89	156	138	98	134

Note that: $\sum_i \sum_j x_{ij} = 2374, \qquad \sum_i \sum_j x_{ij}^2 = 264708$

(i) Give a suitable model for these data, explaining carefully the meaning of each of the terms.

(ii) Complete the analysis of variance and interpret the results, including graphical representations where appropriate.

(iii) By referring to the original data, explain to the experimenter the variability represented by the residual (error) mean square.

6. Distinguish between discriminant analysis and cluster analysis and give, for each, an example of a situation where it would be the appropriate analysis to use. You should not use an example similar to that given below.

A market researcher wishes to determine whether six products fall into distinct categories with respect to the perception the public have of them. She has conducted a survey which has enabled her to obtain a meaningful measure of distance between each pair of products. The distances are given in the form of the matrix below.

	Product					
	A	*B*	*C*	*D*	*E*	*F*
A	0	12	8	11	3	14
B	12	0	12	3	10	4
C	8	12	0	13	7	15
D	11	3	13	0	10	2
E	3	10	7	10	0	13
F	14	4	15	2	13	0

Describe an hierarchical clustering method of your choice and apply it to these data. Comment on your results.

7. Data concerning road usage have been collected in 46 areas of the USA, the variables being

No. dr	-	number of drivers x 10^{-4}
Pop/sm	-	number of persons per square mile
Rd. mil	-	miles of road x 10^{-3}
Fl.con	-	fuel consumption (gallons x 10^{-6})
Deaths	-	number of road deaths in one year

A principal component analysis of the correlation matrix of the first four variables (excluding "Deaths") yielded

	Principal Component			
	1	2	3	4
No. dr	-0.61	-0.14	0.59	-0.50
Pop/sm	-0.10	-0.82	-0.50	-0.24
Rd. mil	-0.48	0.53	-0.63	-0.31
Fl. con	-0.62	-0.14	-0.02	0.77
Eigenvalue	2.23	1.33	0.25	0.19
Cumulative proportion	0.56	0.89	0.95	1.00

(Question continued on next page)

Comment on the results of the analysis including, where possible, an interpretation of the components.

In a follow-up analysis, a stepwise regression has been performed using 'Deaths' as the dependent variable and the principal component scores as the regressor variables. The computer output is as follows:

STEP	1	2	3
CONSTANT	968.5	968.5	968.5
PC1	-545	-545	-545
T-RATIO	-13.29	-17.08	-24.64
PC3		519	519
T-RATIO		5.45	7.86
PC4			-517
T-RATIO			-6.89
S	411	319	221
R-SQUARED	80.1	88.2	94.5

where PC1, PC3, PC4 are the principal component scores from components 1, 3 and 4 respectively. Note that the package stops when the introduction of further variables is not significant at the nominal 5% significance level.

Interpret the output in terms of the factors which appear to affect the number of road deaths.

Discuss the advantages and disadvantages of using principal component scores in a regression rather than the original variables.

8. An experiment has been conducted to determine the length of time a subject takes to complete a task. The factor of interest is the effect of the number of previous practice trials on the time taken. A subject was allowed a number of practice trials (x) and then the time taken to complete the task (y) was determined. For each value of x there were three different subjects and x varied between one and six.

	Number of practice trials (x)					
	1	2	3	4	5	6
Time (y)	172	114	110	78	54	44
in seconds	148	113	96	64	63	43
	157	132	82	92	53	51

It is suspected that the relationship between x and y is

$$E(y) = Ae^{bx}$$

where A and b are constants.

(a) Plot a suitable graph to verify this relationship.

(b) Two methods have been suggested for obtaining estimates of A and b.

Method 1 consists of transforming the data to obtain a linear relationship between the variables and applying least squares to these transformed variables.

Method 2 consists of applying least squares directly to the untransformed variables.

For each method state the assumptions being made about the model.

(c) Using method 1 obtain estimates for A and b and outline, without performing any calculations, how you could test the model for lack of fit.

(d) In the case of method 2, obtain the normal equations which could provide least squares estimates for A and b. (You are <u>not</u> expected to solve the equations.)

(e) Explain briefly why you would expect your estimates from the two methods to differ.