EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2009
(Modular format)

MODULE 5 : Topics in Applied Statistics

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 calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

The notation log *denotes logarithm to base* ***e****.*
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^nC_r$.

This examination paper consists of 13 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. (i) State briefly three reasons why an analyst may wish to perform a principal component analysis.

(3)

(ii) Under what circumstances would it be sensible to use the variance-covariance matrix instead of the correlation matrix in principal component analysis?

(2)

(iii) Describe how the principal components of a data set can be derived from the variance-covariance or correlation matrix.

(2)

(iv) Five variables have been recorded for each of fifty households selected at random among those with at least two wage earners within an area. The variables, all in pounds sterling, are as follows.

X_1 monthly income of main wage-earner in household
X_2 monthly income of second wage-earner in household
X_3 total debts of household excluding mortgage
X_4 monthly mortgage payment
X_5 monthly payment for gas, electricity and water

An analyst has extracted principal components from the correlation matrix for these data, the coefficients for the first three being given in the table below.

		Component		
	Variable	*1*	*2*	*3*
	X_1	–0.46	–0.18	+0.67
	X_2	–0.24	–0.55	–0.68
	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

(a) Comment on the apparent dimensionality of these data.

(2)

(b) Interpret the first and third principal components.

(2)

Question continued on next page

(c) A manager has queried the coefficients for the second principal component, because he thinks that this component is difficult to interpret. He says that the coefficient of X_3 should be +0.48 rather than −0.48. What would your response be to this query?

(3)

(d) The analyst has suggested that for these data it would be valid to do a principal component analysis on the variance-covariance matrix. The manager argues that the results would be very similar, and so there is no point. How would you advise the analyst and manager?

(3)

(e) The analyst reports that there were a few missing values in the dataset, but that this should not have affected the results. What would your response be?

(3)

2. (i) Explain the purpose of *linear discriminant analysis*.

(1)

(ii) State the distributional assumptions that are desirable for the method to be useful and briefly discuss how these can be checked in practice.

(4)

(iii) A clinical psychologist is testing people to determine whether they tend to exaggerate their hearing loss (defined as exaggerators) or not (defined as honest). The clinical opinion is that people in the study population do not understate their hearing loss.

Each person's result comprises two responses, X_1 and X_2, and past experience has shown that these responses are consistent with bivariate Normal populations having the following parameters.

Exaggerators $\boldsymbol{\mu} = \begin{pmatrix} 20 \\ 19 \end{pmatrix}$

Honest $\boldsymbol{\mu} = \begin{pmatrix} 11 \\ 11 \end{pmatrix}$

Common variance-covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 98 & 57 \\ 57 & 92 \end{pmatrix}$

(a) State the criterion that is optimised to determine Fisher's linear discriminant function for discriminating between these two groups.

(2)

(b) Confirm that the inverse of the matrix $\boldsymbol{\Sigma}$ is $\frac{1}{5767}\begin{pmatrix} 92 & -57 \\ -57 & 98 \end{pmatrix}$ and hence show that Fisher's linear discriminant function is $0.0645x_1 + 0.0470x_2$.

(3)

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

(5)

(d) In fact the assumption in part (c) is unreasonable. Past records show that 90% of the population of interest are honest. If a new rule

"If discriminant function < 2, subject is honest"

is applied, what proportion of people would be expected to be incorrectly classified?

(5)

Turn over

3. Six years ago a clinician collected data on the following measures from 207 patients who had a terminal illness.

X a continuous clinical outcome measuring disease severity, where higher values correspond to better health

QL1 a measure of patient-reported quality of life, where higher values correspond to better quality of life

QL2 a measure of the patient's quality of life reported by the nurse, where higher values correspond to better quality of life

The table below shows correlations between the three measures.

	X	*QL1*	*QL2*
X	1.00	0.07	0.38
QL1		1.00	0.49
QL2			1.00

Previous studies suggest that *X* predicts survival for this population of patients. The clinician is interested in whether, having taken account of *X*, either or both of the quality of life measures can add any predictive information about survival for these patients. A total of 62 patients have died since the measures were taken, and the clinician has survival times for these patients.

The intention is to use Cox's proportional hazards regression to answer the clinician's question, and the table below summarises results from a series of Cox regression analyses.

Model	*Variables present*	*Beta coefficient (SE)*	*–2 log likelihood*
A	None	—	639.19
B	*X*	–0.046 (0.008)	591.81
C	*X* *QL1*	–0.048 (0.008) –0.024 (0.006)	576.74
D	*X* *QL2*	–0.036 (0.008) –0.022 (0.007)	581.62
E	*X* *QL1* *QL2*	–0.043 (0.008) –0.018 (0.007) –0.011 (0.008)	574.76

Question continued on next page

(i) Is *X* a statistically significant predictor of survival? Justify your answer.

(2)

(ii) Calculate the hazard ratio, and a corresponding 95% confidence interval, for *QL1* from model C. Explain in a manner suitable for the clinician what these quantities mean, and discuss their implication if the clinician believes that a change of 10 points on a quality of life scale is meaningful.

(6)

(iii) After taking account of *X* does *QL1* add any information about survival? After taking account of *X* does *QL2* add any information about survival? Justify your answers.

(2)

(iv) Which of the models in the table do you consider to be the best? Justify your choice.

(2)

(v) Consider the responses of the following two patients.

Patient 1 *QL1* = 60, *QL2* = 50

Patient 2 *QL1* = 50, *QL2* = 50

Using your chosen model, discuss how each patient's *X* value will determine which of the two survive longer. Justify your answer.

(3)

(vi) Explain briefly what is meant by *checking the proportional hazards assumption*.

Discuss the validity of your answers to each of parts (i) to (v) if the assumption is found to be valid for *X* but invalid for *QL1* and *QL2* in these data.

(5)

Turn over

4. (i) The lifetime of a product can often be modelled by a Weibull distribution. The probability density function of a Weibull distribution is

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}); \quad t > 0; \quad \lambda > 0, \gamma > 0.$$

(a) Derive the hazard function for a Weibull distribution.

(2)

(b) In order to show the versatility of the Weibull distribution, sketch the form of the hazard function for each of a suitably chosen set of values for γ.

(4)

(c) Briefly describe scenarios in which each of your chosen hazard functions in part (b) could be useful in lifetime analysis.

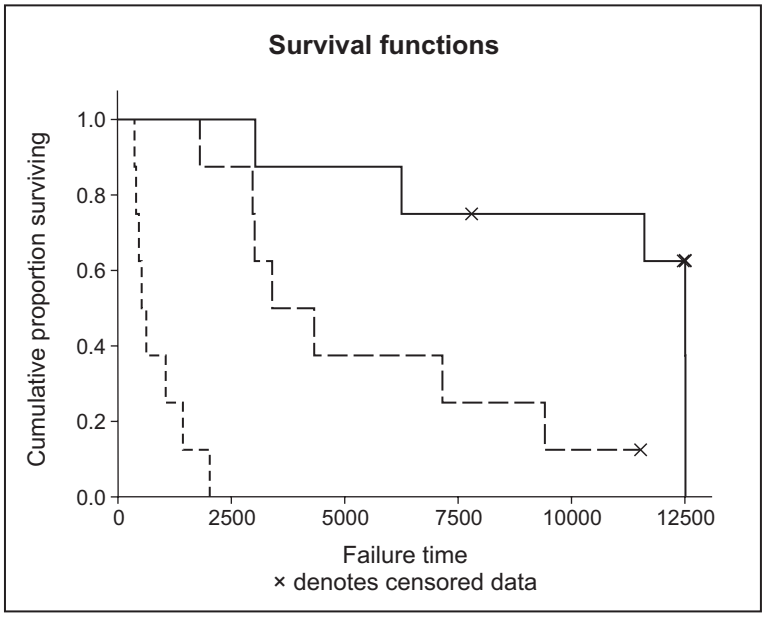
(4)

(ii) A quality manager is investigating the lifetimes of springs produced in his factory. He has randomly selected 24 springs and subjected them to 3 different stress levels, each stress level being applied to 8 of the springs. The table below shows the results of his experiment. The results are the numbers of cycles to failure (in units of 10^3 cycles).

Stress (N/mm^2)	*Failure times*
800	625, 1053, 1432, 2024, 523, 400, 462, 365
750	3400, 9413, 1806, 4327, 11524*, 7154, 2969, 3014
700	12513, 12507*, 3028, 12508*, 6253, 11607, 12475*, 7798*

* denotes data that are right-censored

Kaplan-Meier curves derived from the data are shown below.



Question continued on next page

Turn over

(a) State which survival curve corresponds to each of the three stress levels.

(1)

(b) Derive the survival function for stress level 800 N/mm^2.

(3)

(c) The manager wants to know whether the lifetimes of the springs can be modelled by Weibull distributions. Use your result in part (b) to derive a function which can be used to investigate this request. Plot a suitable graph and then explain whether you think that the lifetime of the springs at stress level 800 N/mm^2 can be modelled with a Weibull distribution.

(6)

5. The data in the table below are from a case-control study, showing numbers of patients attending a particular ante-natal clinic in the last 5 years. The aim of the study was to examine risk factors associated with low birthweight of babies in the population of women who used the centre. Cases were defined as women giving birth to a baby weighing less than 2500 g, and were a random sample from all such women attending in the past 5 years. Controls were a similarly drawn random sample of women whose baby weighed at least 2500 g.

The risk factors of interest were as follows.

Ethnic origin

Number of ante-natal visits in the first trimester of pregnancy

Smoking status during pregnancy

Exposure to the risk factors was derived from case notes.

		Ethnic origin					
		White		*Black*		*Other*	
	Birthweight (g)	< 2500	≥ 2500	< 2500	≥ 2500	< 2500	≥ 2500
Smoked during pregnancy	*Ante-natal visits in first trimester*						
No	0	2	11	2	6	12	22
	1	0	20	1	3	5	6
	2 or more	2	9	2	2	3	7
Yes	0	11	19	4	2	5	4
	1	4	5	1	1	0	1
	2 or more	4	9	1	1	0	2

(i) (a) Provide a table to summarise the association of smoking with low birthweight, and present your conclusions.

(3)

(b) Use your table in part (a) to calculate the odds ratio, and a corresponding 95% confidence interval, for the risk factor smoking status.

(3)

(c) Under what circumstances would the odds ratio be a good approximation to the relative risk?

(1)

Question continued on next page

(ii) The odds ratios and 95% confidence intervals for the other risk factors are given below.

Black *vs* White	2.33	(0.94 to 5.77)
Other *vs* White	1.89	(0.96 to 3.74)
0 *vs* 1 ante-natal visit	1.84	(0.84 to 4.05)
2 or more *vs* 1 ante-natal visit	1.31	(0.51 to 3.39)

Interpret the results for all of the risk factors.

(5)

(iii) Use the Mantel-Haenszel method to calculate an adjusted odds ratio for the comparison of 2 or more *vs* 1 ante-natal visit.

Interpret this statistic, comparing it with the value in part (ii), and suggest explanations for any differences between the unadjusted and adjusted odds ratios.

(6)

(iv) Briefly describe a different technique that could be used to calculate adjusted odds ratios for these data.

(2)

Turn over

6. (a) In the context of clinical life tables, let l_i be the number lost to follow-up, w_i the number withdrawn alive, d_i the number dying, and n_i' the number entering the ith interval.

Define in terms of l_i, w_i, d_i and n_i'

(i) n_i, the adjusted number at risk during the interval $I_i = [t_i, t_{i+1})$,

(ii) $\hat{q}_i$, the estimated probability of dying in the interval I_i, conditional on being alive at t_i,

(iii) $\hat{p}_i$, the estimated probability of surviving the interval I_i,

(iv) $\hat{S}(t_i)$, the estimate of the probability of survival to time t_i.

(6)

(b) Breast cancer is the leading cause of death from cancer among women worldwide, accounting for more than 400000 deaths per year. In a clinical trial in women with advanced or metastatic breast cancer, participants were randomised into one of two groups: treatment and control. An interim analysis was conducted on 163 women randomised to the treatment group and 161 to the control group. The following (incomplete) table shows the survival times in months after randomisation (t_i) for women in the treatment group.

Months after randomisation	$[t_i, t_{i+1})$	n_i'	$w_i + l_i$	d_i	n_i	$\hat{q}_i$	$\hat{p}_i$	$\hat{S}(t_i)$
0		163	43	7	141.5			
2		113	19	11	103.5			
4		83	11	5	77.5			
6		67	18	10	58.0			
8		39	14	7	32.0			
10		18	4	7	16.0			
12		7	2	2	6.0			
14+		3	3	0	1.5			

(i) Copy and complete this life-table. State any assumptions underlying your calculations. (6)

(ii) Plot $\hat{S}(t_i)$ against time t for the treatment group, and use the plot to estimate the median overall survival time of women on treatment. If the median survival time for women in the control group was approximately 6 months, what conclusions would you draw? (5)

(iii) Discuss briefly why it would be considered inappropriate to summarise these data by the proportion of women still disease-free or alive (out of those randomised) at some suitable time after randomisation. (3)

Turn over

7. (a) The variance of a stratified random sample mean can be written as

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{N_i^2}{N^2}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 \ .$$

Explain the (conventional) notation used here.

(3)

Suppose the total cost of sampling is

$$C = c_0 + \sum_{i=1}^{k} c_i n_i \ ,$$

where $c_0, c_1, \ldots, c_k$ are positive constants.

If $\mathrm{Var}(\bar{y}_{st})$ is fixed and the stratum sample sizes are chosen to minimise the total cost of sampling, show that the ith stratum sample size n_i is proportional to

$$\frac{N_i S_i \,/\, \sqrt{c_i}}{\sum_{i=1}^{k} N_i S_i \,/\, \sqrt{c_i}} .$$

(5)

(b) In one region there are 227 orchards located in three different districts of the region. The number of orchards, an estimate of the standard deviation of the areas of the orchards (in hectares), and the sampling cost per orchard are given below.

District	*Total number of orchards*	*Estimated standard deviation (ha)*	*Sampling cost per orchard*
1	68	34	10
2	143	20	14
3	16	59	18

Find the stratum sample sizes required to estimate the total area of orchards in the region to within 300 ha, given that the total cost of sampling is to be minimised and also that we are prepared to accept a one in twenty chance that the estimate will be more than 300 ha from the true total area.

(10)

Find the total cost of sampling in this case, if the overhead cost is 200 units.

(2)

Turn over

8. Monthly magazines are often directed at special interests of their readers, such as the countryside, wildlife, sport, motor cars. The marketing manager for one of these magazines is trying to encourage advertisers to place advertisements in her magazine. She therefore wishes to collect information on the number and types of people who read the magazine, regularly or occasionally, and their interests.

(i) Paying particular attention to the problems of constructing a sampling frame, undercoverage, non-response and likely requirements of advertisers, recommend a method for selecting a sample from the readership.

(10)

(ii) Comment on the types of information the manager should aim to collect, in addition to standard questions about name, address, age, occupation, etc.

(5)

(iii) How could the frame be used, and how could the results of the exercise be kept up to date, to satisfy the likely needs of the advertisers?

(5)