

**THE ROYAL STATISTICAL SOCIETY**

**GRADUATE DIPLOMA EXAMINATION**

**NEW MODULAR SCHEME**

**introduced from the examinations in 2009**

**MODULE 5**

**SPECIMEN PAPER B**

**SOLUTIONS ARE CONTAINED IN A SEPARATE FILE**

The time for the examination is 3 hours. The paper contains eight questions, of which candidates are to attempt **five**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

Note. In accordance with the convention used in all the Society's examination papers, the notation  $\log$  denotes logarithm to base  $e$ . Logarithms to any other base are explicitly identified, e.g.  $\log_{10}$ .

1. An economist wants to develop a scale to measure doctors' attitudes to the cost of health care. She designs a questionnaire comprising 15 questions about attitude. Each of the responses is coded on a 5-point scale, where 1 means "strongly agree" and 5 means "strongly disagree". Completed questionnaires have been returned by 149 doctors.

(i) The economist wishes to devise subscales consisting of combinations of answers from subsets of the questions, and intends to use principal component analysis on the correlation matrix. Explain why this technique might be useful, and comment on the suitability of the method for these data. (5)

(ii) Describe another multivariate method that could be used to explore relationships between questions on the questionnaire. (2)

(iii) Unfortunately there are many missing values in the data. In fact nine questions have a response rate of less than 50%. The economist decides to restrict her analysis to the remaining six questions, which have been answered by all 149 doctors. Comment on whether you think this is a sensible strategy. (3)

(iv) The six questions are as follows.

- Q1 Best health care is expensive
- Q2 Cost is a major consideration
- Q3 I would determine the cost of tests before referral
- Q4 I would monitor likely complications only
- Q5 I would use all means irrespective of cost
- Q6 I prefer unnecessary tests to omitting tests.

[You are reminded that low values indicate agreement with the statements given in the questions.]

A summary of the first three eigenvalues and associated principal components using the correlation matrix is given in the table below.

<i>Eigenvalue</i>	3.24	1.38	0.52
Q1	0.27	0.50	-0.29
Q2	-0.37	0.45	-0.51
Q3	-0.41	0.51	0.13
Q4	-0.35	0.26	0.68
Q5	0.49	0.34	0.42
Q6	0.52	0.32	-0.04

Interpret those principal components that you consider to be meaningful, justifying your answer. Would you want to see any further principal components from this analysis? (6)

(v) What advice would you give to the economist about the scores she should use? Justify your answer. (4)

2. A group of archaeologists are studying remains found at a number of sites. They have measurements on a set of 32 skulls, 17 of which were found at one archaeological site and the other 15 at another site. They believe that each of these two sites was inhabited by a different tribe of people. They are now working on other sites in the same region and wish to decide which tribe the skulls they are finding belong to.

The measurements comprised 5 different dimensions of the skulls, all in mm.

- (i) Explain how linear *discriminant analysis* might be useful in studying these data, and describe the way in which the results may be applied to data from the new sites. (3)
- (ii) State the assumptions that would need to be made about the data in order for linear discriminant analysis to be valid. Describe the checks that you could do to investigate these assumptions, stating any limitations on the methods. (4)
- (iii) Summary statistics for each of the two groups of skulls are shown below. Describe the main features in relation to your answers to (i) and (ii). (4)

Means for the 5 variables

<i>Variable</i>	<i>Site 1 (n = 17)</i>	<i>Site 2 (n = 15)</i>
$x_1$	174.82	185.73
$x_2$	139.35	138.73
$x_3$	132.00	134.77
$x_4$	69.82	76.47
$x_5$	130.55	137.50

Variance-covariance matrices:

Site 1

$$\begin{pmatrix} 45.53 & 25.22 & 12.39 & 22.15 & 27.97 \\ 25.22 & 57.81 & 11.88 & 7.52 & 48.06 \\ 12.39 & 11.88 & 36.09 & -0.31 & 1.41 \\ 22.15 & 7.52 & -0.31 & 20.94 & 16.77 \\ 27.97 & 48.06 & 1.41 & 16.77 & 66.21 \end{pmatrix}$$

**Question continued on the next page**

Site 2

$$\begin{pmatrix} 74.42 & -9.52 & 22.74 & 17.79 & 11.13 \\ -9.52 & 37.75 & -11.26 & 0.70 & 9.46 \\ 22.74 & -11.26 & 36.32 & 10.72 & 7.20 \\ 17.79 & 0.70 & 10.72 & 15.30 & 8.66 \\ 11.13 & 9.46 & 7.20 & 8.66 & 17.96 \end{pmatrix}$$

- (iv) Below is shown a summary of two discriminant analyses on these data. Describe the differences between the two sets of results. Which model do you prefer and why?

(9)

OUTPUT FROM LINEAR DISCRIMINANT ANALYSIS

**METHOD 1** – including all 5 variables

Linear discriminant function is:  $-0.09x_1 + 0.16x_2 + 0.01x_3 - 0.18x_4 - 0.18x_5$

Classification Results

			Predicted group membership		Total
			1	2	
Original	Count	SITE 1	14	3	17
		2	3	12	15
	%	1	82.4	17.6	100.0
		2	20.0	80.0	100.0
Cross-validated	Count	SITE 1	12	5	17
		2	6	9	15
	%	1	70.6	29.4	100.0
		2	40.0	60.0	100.0

**METHOD 2** – stepwise, using forward selection

Discriminant function:  $-0.37x_4$

Classification Results

			Predicted group membership		Total
			1	2	
Original	Count	SITE 1	12	5	17
		2	3	12	15
	%	1	70.6	29.4	100.0
		2	20.0	80.0	100.0
Cross-validated	Count	SITE 1	12	5	17
		2	3	12	15
	%	1	70.6	29.4	100.0
		2	20.0	80.0	100.0

3. Define the *hazard function* used in the analysis of data describing time to failure. (1)

Write down the hazard function that is modelled in Cox proportional hazards regression, for the case where there are two predictor variables,  $x_1$  (a continuous variable) and  $x_2$  (a factor with two levels, coded 0 and 1), and the model contains  $x_1$ ,  $x_2$  and the interaction between  $x_1$  and  $x_2$ . (4)

A company is trying to improve the reliability of one of its products. Scientists working for the company know that the level of a particular chemical affects the lifetime of the product. During the production process some of the items develop a slight fault.

The scientists have collected data on a random sample of 36 items. For each item they have recorded the values of  $x_1$  and  $x_2$ , where  $x_1$  is the level of the chemical and  $x_2$  indicates the presence or absence of the fault (0 = absent, 1 = present). They tested each item for 23 days or until it failed, whichever was the earlier. There were 15 items without the fault, of which 9 failed within the 23 days. There were 21 with the fault, all of which failed within the 23 days.

The manager wants to know how the presence of the fault affects the lifetime of the product, once the possible confounding effect of  $x_1$  is taken into account. He is also interested in whether there is an interaction between  $x_1$  and  $x_2$ .

The table below summarises models from Cox proportional hazards regression analyses.  $x_1 * x_2$  denotes the interaction between  $x_1$  and  $x_2$ . Some entries have been omitted.

Model	Variables in model	Coefficient	Standard error	Hazard ratio	Log likelihood
A	$x_1$	1.465	0.296	4.328	-75.640
B	$x_1$	1.528	0.327		-71.518
	$x_2$	1.172	0.429		
C	$x_1$	2.059	1.684		-71.369
	$x_2$	1.695	0.448		
	$x_1 * x_2$	-0.284	0.517		

- (i) Do you think that  $x_1$  is an important predictor of the lifetime? Do you think that there is an interaction effect? Select the model that you think best models the data. Justify your answers. (5)
- (ii) Show how the hazard ratio is derived from the coefficient in the output. Use the output for the model you chose in part (i) to estimate a 95% confidence interval for the coefficient of  $x_2$  and the associated 95% confidence interval for the hazard ratio corresponding to  $x_2$ . (3)

**Question continued on next page**

(iii) Write a brief explanation for the manager of the extent to which the presence of the fault affects the lifetime of the product. (2)

(iv) Consider two items:

item 1, for which  $x_1 = 1.4$  and  $x_2 = 1$ ;

item 2, for which  $x_1 = 2.9$  and  $x_2 = 0$ .

Which of the two would you expect to fail first? Explain your answer. (3)

(v) Someone suggests that the above results are unreliable, because the proportional hazards assumption has not been checked. Comment briefly on the possible consequences of this assumption not being true. (2)

4. The survival times from diagnosis (in months) for a random sample of 24 patients with colorectal cancer from one centre are given below.

3\*, 6, 6, 6, 6, 8, 8, 12, 12, 12\*, 15\*, 16\*, 18\*, 18\*, 20, 22\*, 24, 28\*, 28\*, 28\*, 30, 30\*, 33\*, 42 (\* Indicates a right-censored observation.)

- (i) Explain what is meant by a *right-censored observation*. (2)
- (ii) Compute the Kaplan-Meier estimate of the survival curve and plot it. Using Greenwood's formula, calculate the associated standard error for the Kaplan-Meier survival function estimate at 12 months follow-up. (14)
- (iii) Find an approximate 95% confidence interval for the one-year survival rate for colorectal cancer patients from this centre. (2)
- (iv) Use your graph to estimate the median survival time for colorectal cancer patients from this centre. (1)
- (v) The 24 patients were part of a larger randomised controlled clinical trial to compare two treatments, control and  $\gamma$ -linolenic acid, for colorectal cancer. If the data from all patients in the study were available, give the name of an analysis which could be used to compare the survival times for these two treatments. (1)

5. (a) Explain clearly the differences between infant, neonatal, perinatal and maternal mortality rates and show how they are calculated. (6)

- (b) The age-sex structure of the population of *U*, a developing country, is given (in thousands) below. Draw an age pyramid to illustrate these data.

<i>Age</i>	<i>Males</i>	<i>Females</i>
0	35.7	34.8
1 – 4	143.9	140.0
5 – 14	328.1	320.6
15 – 24	202.4	216.2
25 – 34	120.7	142.2
35 – 44	114.7	123.3
45 – 54	93.6	87.2
55 – 64	63.5	60.4
65 – 74	50.2	47.6
75 – 84	9.0	12.8
85 and over	0.9	1.7

(6)

- (c) (i) The sex-age-specific death rates (per thousand) for *U* and for the population of *D*, a developed country, are given below. Explain clearly how these rates have been calculated.

<i>Age</i>	<b>Country D</b>		<b>Country U</b>	
	<i>Males</i>	<i>Females</i>	<i>Males</i>	<i>Females</i>
0	33.2	25.5	54.2	41.1
1 – 4	1.2	1.0	3.3	3.5
5 – 14	0.6	0.4	0.9	0.6
15 – 24	1.5	0.6	1.3	0.9
25 – 34	1.9	1.1	2.7	1.7
35 – 44	3.7	2.2	4.1	3.1
45 – 54	9.7	5.1	7.3	5.0
55 – 64	22.9	11.8	15.8	9.9
65 – 74	51.6	30.7	34.5	24.5
75 – 84	101.3	75.1	69.7	55.4
85 and over	202.6	202.5	198.5	161.8

(4)

- (ii) Outline the similarities and differences in mortality levels between the two countries, and between males and females.

(4)

6. The data in the table below describe an unmatched case-control study of lung cancer as related to tobacco consumption. The 483 cases and 447 controls were cross-classified according to smoking status and sex, with the following results.

**Results of an unmatched case-control study of smoking status and lung cancer**

<b>MALES</b>	<b>Smoking status</b>		
	<i>Smokers</i>	<i>Non-smokers</i>	<i>Total</i>
<i>Cases</i>	58	245	303
<i>Controls</i>	6	271	277
<i>Total</i>	64	516	

<b>FEMALES</b>	<b>Smoking status</b>		
	<i>Smokers</i>	<i>Non-smokers</i>	<i>Total</i>
<i>Cases</i>	31	149	180
<i>Controls</i>	7	163	170
<i>Total</i>	38	312	

We are interested in estimating the relative risk of lung cancer in smokers compared with non-smokers.

- (i) Explain why we cannot estimate the relative risk directly in a case-control study. (4)
- (ii) What is the *odds* of an event? What is the *odds ratio* for exposure and disease? (2)
- (iii) Ignoring sex, what is the odds ratio for the occurrence of lung cancer for smokers, relative to non-smokers? Comment on the result. (2)
- (iv) Calculate the Mantel-Haenszel estimate of the odds ratio for occurrence of lung cancer for smokers relative to non-smokers, allowing for sex. Calculate a 95% confidence interval for this odds ratio. (8)
- (v) Perform a test of the null hypothesis that smoking status is unrelated to occurrence of lung cancer. Comment on the results of all your calculations. (4)

7. (i) A simple random sample of size  $n$  is selected from a population of  $N$  units. The response of interest,  $y$ , and an auxiliary variable,  $x$ , are measured on each unit in the sample. The population mean of the auxiliary variable is  $\bar{X}$ .

The sample estimator,  $\hat{R}$ , of a population ratio is given by  $\hat{R} = \frac{\bar{y}}{\bar{x}}$ . Show that, approximately,

(a) the bias of  $\hat{R}$  is  $-\left\{\text{Cov}(\hat{R}, \bar{x})\right\} / \bar{X}$ , (2)

- (b) the variance of  $\hat{R}$  may be estimated by

$$\frac{1-f}{n} \frac{1}{\bar{x}^2} \left( s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\hat{\rho}s_y s_x \right),$$

where  $f$  is the sampling fraction and  $s_y^2$ ,  $s_x^2$  and  $\hat{\rho}$  are to be defined. (4)

[You may assume that (if  $\bar{X}$  is not known) the variance of  $\hat{R}$  may be estimated (approximately) from a sample as

$$\frac{N-n}{Nn\bar{x}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 .]$$

- (ii) Explain briefly the circumstances under which the ratio estimator of a population mean will be less precise than the sample mean of a simple random sample of the same total size. (3)
- (iii) The wholesale price for oranges in large shipments is based on the sugar content of the load. The exact sugar content cannot be determined prior to the purchase and extraction of the juice from the load.

The data below show the sugar content in grams ( $y$ ) and weight in grams ( $x$ ) of a random sample of ten oranges from a consignment with total weight 820 kg.

Sugar content (gm) $y$	9.5	13.6	11.4	9.1	15.0	12.3	8.6	9.5	10.5	11.4
Weight (gm) $x$	182	218	195	191	227	209	177	186	190	200

$$\Sigma xy = 22\,194.80, \quad \Sigma y^2 = 1268.69, \quad \Sigma x^2 = 392\,389.$$

- (a) Explain why a ratio or regression estimator is appropriate for these data. (2)
- (b) Estimate the total sugar content for the oranges and give a standard error of your estimate, giving reasons for your choice of estimator. (6)
- (c) Show that about 25 oranges must be sampled from a consignment with total weight 820 kg so that the half-width of the 95% confidence interval for the total sugar content is less than 2 kg. (3)

8. Wildland managers want to estimate the total number of caribou in the Nelchina herd located in south central Alaska. The density of caribou differs dramatically in different types of habitat. A preliminary aerial survey has identified the area used by the herd, and divided it into six strata based on habitat type.

The organiser has decided to divide the area into sub-areas called quadrats, each  $4 \text{ km}^2$ . The main survey will be conducted by selecting a simple random sample of quadrats from each stratum; the number of caribou,  $y$ , in the quadrats will be counted from an aerial photograph.

Estimates of the means and standard deviations of the measurements,  $y$ , in each stratum based on the preliminary survey of 211 quadrats are as follows.

Stratum ( $h$ )	$N_h$	$n_h$	$\bar{y}_h$	$s_h$	$N_h s_h$
1	400	98	24.1	74.7	29880
2	40	10	25.6	63.7	2548
3	100	37	267.6	589.5	58950
4	40	6	179.0	151.0	6040
5	70	39	293.7	351.5	24605
6	120	21	33.2	99.0	11880
Total	770	211			133903

$$\sum N_h s_h^2 = 47882186$$

- (i) Discuss briefly the merits of using stratified sampling for this survey. (4)
- (ii) Based on the results of the preliminary aerial survey, estimate the total number of caribou in the herd and obtain an estimate of the standard error for your estimator. (5)
- (iii) For the main survey, the managers wish to estimate the total number of caribou to within  $d$  animals with 95% probability (i.e. the width of the interval is  $2d$ ). You may assume that the formula for the total sample size  $n$  is

$$n = \frac{\sum_h N_h^2 S_h^2 / w_h}{V + \sum_h N_h S_h^2}.$$

Define  $N_h$ ,  $S_h$ ,  $w_h$  and  $V$  as used in this formula. (2)

- (iv) Define *optimal* allocation. Discuss briefly why you would choose an optimal allocation rather than proportional allocation for this survey. You may assume that the cost of sampling any unit is constant. (3)
- (v) Use optimal allocation to calculate the total sample size and the allocations  $n_h$  needed to estimate the total population of caribou to within 8000 animals with 95% probability. Calculate the standard error for your estimator of the population total. (6)