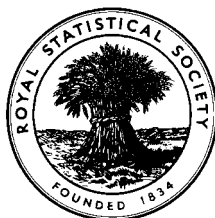


EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



GRADUATE DIPLOMA, 2000

Applied Statistics II

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 nC_r and that \ln stands for \log_e .

1. (a) Explain what is meant by a *Latin square* design. Briefly discuss the disadvantages of small Latin square designs (e.g. 3×3 or 4×4). (3)
- (b) A biological assay of insulin from the blood sugar response of rabbits was performed with two linked 4×4 Latin squares. Each rabbit was injected with four doses of insulin (*A*, *B*, *C* and *D*). The four doses were the factorial arrangement of two preparations, standard and test, each used at two doses, 0.6 and 1.2.

<i>Treatment</i>	<i>Preparation</i>	<i>Dose</i>
<i>A</i>	Standard	0.6
<i>B</i>	Standard	1.2
<i>C</i>	Test	0.6
<i>D</i>	Test	1.2

Eight rabbits (columns) were injected on each of 4 days (rows) with the treatment *A* to *D* of the Latin square (squares). The measured response is mg % blood sugar 50 minutes after injection.

<i>Treatment and Response in Rabbit Number</i>									
<i>Day</i>	1	2	3	4	5	6	7	8	<i>Total</i>
1	<i>B</i> 24	<i>C</i> 46	<i>D</i> 34	<i>A</i> 48	<i>C</i> 61	<i>D</i> 72	<i>A</i> 68	<i>B</i> 28	381
2	<i>D</i> 33	<i>A</i> 58	<i>B</i> 57	<i>C</i> 60	<i>D</i> 58	<i>C</i> 83	<i>B</i> 62	<i>A</i> 65	476
3	<i>A</i> 57	<i>D</i> 26	<i>C</i> 60	<i>B</i> 45	<i>B</i> 46	<i>A</i> 75	<i>C</i> 68	<i>D</i> 54	431
4	<i>C</i> 46	<i>B</i> 34	<i>A</i> 61	<i>D</i> 47	<i>A</i> 54	<i>B</i> 62	<i>D</i> 50	<i>C</i> 56	410
<i>Total</i>	160	164	212	200	219	292	248	203	

$$\sum x = 1698 \quad \sum x^2 = 96\,518$$

- (i) Compute a combined analysis of variance for the two replicated 4×4 Latin squares. Comment. (6)
- (ii) Suggest a set of mutually orthogonal linear contrasts for subdividing the sum of squares for treatments, and explain what each is comparing. (3)
- (iii) Split the sum of squares for treatments into components for each of the contrasts in part (ii) and test their significance. Interpret your results and write a summary of your conclusions. (8)

2. A greenhouse test was conducted to determine the effect of five soil treatments on the growth of barley plants. The experiment tested the effect of soil treatments on barley grown in volcanic ash (*a*) and in mineral acid soil, of high (*b*) pH and low pH, in the presence (*c*) and absence of potassium, presence (*d*) and absence of magnesium and presence (*e*) and absence of calcium oxide.

It was decided to use a 2^{5-2} fractional factorial design consisting of 8 treatment combinations. One pot per soil combination was prepared. Seed-bearing spikes of naked barley were planted in each pot. The mean plant heights in centimetres for each treatment combination were as follows:

Treatment combination	Growth (cm)
<i>e</i>	8.7
<i>ad</i>	12.0
<i>bde</i>	17.5
<i>ab</i>	11.0
<i>cd</i>	9.0
<i>ace</i>	13.0
<i>bc</i>	16.1
<i>abcde</i>	17.7

- (i) Verify that the design generators used were $I = ACE$ and $I = BDE$. (4)
- (ii) Write down the complete defining relation and the aliases for this design. (4)
- (iii) Estimate the main effects and interactions and comment on any effects that appear to be substantial. (8)
- (iv) It is sometimes argued that one should pool effects which appear small in order to obtain more degrees of freedom for the residual sum of squares. Comment on the advisability of doing this in fractional factorial designs. (4)

3. As part of a rehabilitation programme during early recovery in subjects who had had myocardial infarction or cardiac surgery, subjects were randomly assigned to one of three groups:

Group A received teaching, treadmill exercise testing and exercise training 3 times per week;
 Group B received only teaching and treadmill exercise testing;
 Group C received only routine care without supervised exercise or teaching.

Subjects were assessed on self-efficacy. Total efficacy scores were to be recorded as follows:

<i>Group</i>	<i>Subject</i>				
<i>A</i>	y_{A1}	y_{A2}	y_{A3}	y_{An_A}
<i>B</i>	y_{B1}	y_{B2}	y_{B3}	y_{Bn_B}
<i>C</i>	y_{C1}	y_{C2}	y_{C3}	y_{Cn_C}

- (i) The linear expression

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{for } i = A, B, C \text{ and } j = 1, 2, \dots, n_i \text{ ,}$$

where $E(\varepsilon_{ij}) = 0$ and $V(\varepsilon_{ij}) = \sigma^2$ for all i, j and the ε_{ij} are uncorrelated, is being considered as a model for the data. Define clearly each term in the model. (2)

- (ii) Show that the expected value of the sum of squares for residuals is $(N - 3)\sigma^2$ where $N = n_A + n_B + n_C$. (6)

- (iii) Using Cochran's theorem or otherwise, deduce the distribution of the test statistic appropriate for testing the null hypothesis of no differences in treatment means. State any assumptions. (5)

- (iv) Results for the above study at 4 weeks after cardiac event were:

	Group		
	<i>A</i>	<i>B</i>	<i>C</i>
<i>n</i>	11	13	13
<i>Mean</i>	126.8	129.0	103.9
<i>St.dev</i>	24.25	24.07	17.71

Find 95% confidence intervals for each of the following:

$$\alpha_A - \alpha_B \text{ , } \frac{\alpha_A + \alpha_B}{2} - \alpha_C \text{ .}$$

Summarise any conclusions from the confidence intervals, stating clearly any assumptions. (7)

4. An experimenter begins a steepest ascent procedure on two variables (x_1, x_2) at the current central point (90, 20) and performs five runs with the following response results.

x_1	80	100	80	100	90
x_2	10	10	30	30	20
y	11	0	29	6	12

- (i) Explain what is meant by a *steepest ascent* procedure in the context of searching for the optimum value of a response variable. (4)
- (ii) Code the factor levels (x_1, x_2). Use the method of least squares to fit an appropriate first-order model to the data. (6)
- (iii) Determine the direction of steepest ascent. (2)
- (iv) The experimenter now performs six more runs:

x_1	64.5	47.5	39	30.5	43.25	34.75
x_2	38	50	56	62	53	59
y	43	58	72	62	65	68

- Which of these runs lie on the path of steepest ascent you determined in part (iii)? (2)
- (v) Plot, as points on a diagram, all the runs performed so far, with their y values attached. (4)
- (vi) What briefly should the experimenter do next? Go off in a new direction of steepest ascent? Fit a second-order surface? Or what? (2)

5. (a) For stratified random sampling (without replacement), the variance of the estimated proportion, p_{st} , of units in a population possessing a certain attribute is

$$\sum_{h=1}^L \frac{W_h^2 P_h (1-P_h)}{n_h} \left(1 - \frac{n_h}{N_h}\right).$$

Explain the notation W_h , N_h , P_h , n_h and L . (3)

- (b) The cost (in suitable units) of data collection in a stratified sample survey is

$$C = c_0 + \sum_{h=1}^L c_h n_h$$

where c_0 is the overhead cost and c_h is the cost per individual observation in stratum h .

- (i) Show that the sample size allocation that minimises $V + \lambda C$, where V is the variance of the estimated proportion (p_{st}) and λ is a positive constant, is given by $n_h = W_h \sqrt{\frac{P_h(1-P_h)}{\lambda c_h}}$. (4)
- (ii) Show how to choose λ so that the optimal allocation minimises the total cost of sampling for fixed variance V . (4)
- (c) An interview survey is to be conducted to determine the proportion of households in a city living in rented houses. The 2026 households in the city are divided up into four strata based on location. The number of households, the sampling costs per household and the initial estimates of proportions living in rented houses are as follows:

<i>Stratum based on city areas</i>	<i>Stratum population size, N_h</i>	<i>Estimated proportion renting</i>	<i>Sampling cost per household</i>
Area 1	1190	0.75	9
Area 2	523	0.50	9
Area 3	215	0.20	16
Area 4	98	0.12	16

- (i) For the above data, evaluate $n_h = W_h \sqrt{\frac{P_h(1-P_h)}{\lambda c_h}}$ for $h = 1, 2, 3, 4$ in terms of λ . (4)
- (ii) Assume it is required to estimate the proportion of households living in rented houses to within 0.1 of the true value with 95% confidence. Determine the appropriate value for λ . Hence find the total sample size and optimal strata allocations that minimise the total cost of sampling. (3)
- (iii) How much will the total survey cost be? (2)

6. (a) Define the terms *target population*, *study population*, *sampling frame*, *simple random sampling*, *equal probability selection method*. (5)
- (b) In a company survey to investigate alcohol consumption of employees, an anonymous questionnaire asked respondents about their background and social activities before asking the following question:
- How much alcohol do you drink in an average week?*
- Criticise this question. Suggest your own question(s). (5)
- (c) What are the advantages and disadvantages of *quota* sampling? (5)
- (d) A community consisting of $N = 10\,000$ households is to be monitored to estimate the average daily water consumption used during a specified dry spell. What size of simple random sample should be taken to estimate the average daily consumption within 1.5 gallons of the true value with 95% confidence? Use 35.38 gallons as the population standard deviation. (5)

7. A political scientist has developed a test designed to measure the degree of awareness of current events. She wants to estimate the average score per person that would be achieved on this test by all students in a certain school. She selects 10 classes at random from a total of 108 classes and gives the test to each member of the sampled classes, with the following results.

<i>Class</i>	<i>Number of students</i>	<i>Total score</i>
1	29	1510
2	38	1990
3	22	1080
4	30	1620
5	18	710
6	40	1980
7	22	1310
8	19	860
9	31	1590
10	21	1140
<i>Total</i>	270	13790

- (i) Explain why the above sample is a cluster sample and what its advantages are in this context. (3)
- (ii) Explain why the average score per student based on the above sample is a ratio estimator. (2)
- (iii) Variates y_i and m_i are measured on each unit of a simple random sample of size n , assumed large. Show that the variance of $r = \frac{\bar{y}}{\bar{m}}$ is approximately

$$\frac{1-f}{n\bar{M}^2} \sum_{i=1}^N \frac{(y_i - Rm_i)^2}{N-1}$$

where $R = \bar{Y} / \bar{M}$ is the ratio of the population means and $f = n/N$.

- (iv) Estimate the average score per student that would be achieved in this test by all students in the school, and give a 95% confidence interval for this average. (5)
- (v) Suppose the scientist would like to repeat the survey for a similar high school which has 100 classes. She wants to choose a sample size so that the half-width of the 95% confidence interval for the average score per student is less than 1. How many classes should she sample? (4)

8. Mortality data for coronary heart disease during 1991 in the United Kingdom and in Scotland alone were as follows:

Age	Scotland		UK	
	Population (thousands)	Deaths	Population (thousands)	Deaths
Under 35	2513	18	28226	192
35 – 44	701	185	7932	1595
45 – 54	580	751	6593	6035
55 – 64	537	2346	5814	19515
65 – 74	441	4886	5075	46369
75 and over	328	8680	4009	97473
Total	5100	16866	57649	171179

(Source: *Coronary Heart Disease Statistics, British Heart Foundation, June 1992. Annual Abstract of Statistics, 1993*)

- (i) Explain why you would consider it important to standardise death rate to compare the numbers of deaths due to coronary heart disease in different geographical regions in the United Kingdom. Distinguish between direct and indirect standardisation. (5)
- (ii) Calculate the crude death rates per thousand in Scotland and the United Kingdom. (2)

Using the data for the United Kingdom as the standard population:

- (iii) Calculate the age standardised mortality rate for coronary heart disease in Scotland. Compare this with the crude death rate. (5)
- (iv) Calculate the standardised mortality ratio (SMR) for coronary heart disease in Scotland. What information is provided by the calculation? (5)
- (v) Calculate the indirect standardised death rate for coronary heart disease in Scotland. Comment. (3)