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

GRADUATE DIPLOMA, 2001

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.
Normal probability graph paper is available for use in question 4.

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* $\log_e$.

This examination paper consists of 10 printed pages. This front cover is page 1. The reverse of the front cover, which is intentionally left blank, is page 2. Question 1 starts on page 3.

There are 8 questions altogether in the paper.

1. (a) A baking technologist wishes to compare 4 different biscuit recipes. Mixing, forming and baking the biscuits from one recipe takes approximately 1½ hours.

(i) Suppose that the pilot plant can produce four bakes in a working day. Name and describe a suitable design with days as blocks.

(2)

(ii) One of the response variables to be measured is average biscuit weight. Now suppose in part (i) that the technologist suspects that there may be a trend in the weights throughout the day due to the oven getting progressively hotter. Write down an appropriate *Latin square design* and explain its relevance.

(4)

(iii) Explain what is meant by a *balanced incomplete block design* with days as blocks. Suppose that the pilot plant can only produce three bakes in a working day. Suggest one of these designs if three bakes for each biscuit recipe are to be made.

(4)

(b) An experiment is performed to determine the effect of two temperatures and three heat treatment times on the strength of normalised steel. Four shifts were used to collect the data. In each shift, the various combinations of time and temperature were used in a random order. The results are shown below where the data on the response (strength) are coded.

		Heat treatment times (minutes)			
Shift	**Temp (^{o}F)**	**10**	**20**	**30**	**Total**
1	**1500** **1600**	63 89	54 91	61 62	420
2	**1500** **1600**	50 80	52 72	59 69	382
3	**1500** **1600**	48 73	74 81	71 69	416
4	**1500** **1600**	54 88	48 92	59 64	405
Total	**1500** **1600**	215 330	228 336	250 264	693 930

$\sum y = 1623$ $\sum y^2 = 114159$

(i) Describe the type of design used here. Explain why the experimental design used here is not completely randomised.

(4)

(ii) Use the above data to construct the analysis of variance. Interpret the results of the analysis of variance.

(6)

2. (i) An experiment was conducted to compare several treatments, each of which was replicated r times during the experiment. Explain what is meant by a *contrast* in the comparison of treatment means and derive its standard error. Define any notation. State the condition under which two contrasts are *orthogonal* and explain its relevance.

(6)

(ii) In a randomised block experiment on potato scab, 8 treatments were replicated 4 times. Two treatments were identical controls, the other six consisted of 3 amounts of sulphur (3, 6 and 12 units) applied either in spring (S) or autumn (A). The totals for the eight treatments were as follows.

C_1	C_2	S_3	S_6	S_{12}	A_3	A_6	A_{12}
75	106	38	67	62	73	23	57

(a) Write down a set of meaningful orthogonal contrasts which assess the following types of treatment differences:

- sulphur versus no sulphur
- spring versus autumn application
- response to increasing sulphur levels
- interaction of sulphur response with application timing
- control 1 versus control 2.

(6)

(b) Calculate the value of each contrast. If the error mean square is 30, test the statistical significance of each contrast, stating any assumptions required for the validity of the test. Summarise the results.

(8)

3. In a randomised block experiment on poppy plants in oats, 5 treatments (*A*, *B*, *C*, *D*, *E*) were compared. The numbers of poppy plants per unit are shown below:

Block	*A*	*B*	*C*	*D*	*E*	*Total*
1	438	538	77	17	18	1088
2	442	422	61	31	26	982
3	319	377	157	87	77	1017
4	380	315	52	16	20	783
Total	1579	1652	347	151	141	3870

$\sum y = 3870$ $\sum y^2 = 1395658$

(i) Calculate the means and standard deviations for the 5 treatment groups and examine the relationship between these two statistics.

(2)

(ii) Using the above information, construct the analysis of variance. State any assumptions required for the validity of this analysis, and (without doing further calculations) explain how you would check whether these were reasonable.

(5)

(iii) Calculate the standard error for a treatment difference. Hence identify pairs of treatments whose effects you would consider to be different.

(3)

(iv) Explain why a square root transformation might be considered appropriate for these data. Transform the data for treatment *A* using a square root transformation and calculate the treatment mean.

(3)

The other treatment means for the square root transformed data are

B	*C*	*D*	*E*
20.2	9.1	5.8	5.6

(v) If the mean square error for the transformed data is 4.066, identify pairs of treatments whose effects you would consider to be different.

(3)

(vi) Compare the results for the raw and transformed data. What are the implications of the assumptions in part (ii) on the results? Explain.

(4)

4. (a) Explain what is meant by a *central composite design* and mention briefly the advantages of building up the design sequentially.

(3)

(b) In a process development study on optimising the strength of bread wrapper stock (grams per square inch), three factors were studied: sealing temperature (x_1), cooling bar temperature (x_2) and percent polyethylene additive (x_3). A single replicate of a 2^3 factorial design was run with the following results, where −1 and +1 indicate low and high levels of each factor:

	A	*B*	*C*	*Strength*
(1)	−1	−1	−1	10.6
a	+1	−1	−1	11.1
b	−1	+1	−1	11.9
ab	+1	+1	−1	13.0
c	−1	−1	+1	19.7
ac	+1	−1	+1	18.1
bc	−1	+1	+1	18.7
abc	+1	+1	+1	22.1

Effect estimates	
A	0.85
B	1.55
C	8.00
AB	1.40
AC	0.05
BC	−0.05
ABC	1.10

(i) Verify the *ABC* interaction effect. (2)

(ii) Plot the effect estimates and interactions on Normal probability graph paper. Explain how a Normal probability plot works in this situation. What should the plot look like if none of the factors affect the response? Why? What if some factors do affect the response? Interpret your plot.

(4)

(iii) Write down a first-order model and fit it to these data. Is this an adequate approximation to the data? Why or why not?

(2)

(iv) Four additional runs were made at the centre of the region of experimentation with response values 16.0, 15.5, 15.6 and 16.2. Use this information to test for curvature and lack of fit in the first-order response function. Interpret your results.

(5)

(v) Now suppose that a maximum of 7 runs could be carried out in one day, and that it is decided to run the experiment over two days, with days being used as blocks. Suggest an appropriate design for the 2^k factorial design plus four centre points and give reasons for your choice of blocking.

(2)

(vi) Augment the design in part (v) to produce a central composite design. How can this be made rotatable?

(2)

5. (a) Give an example of a survey in which you would choose to use each of the following methods as the main method of data collection:

(i) face-to-face interviews,

(ii) telephone interviews,

(iii) postal questionnaires.

In each case, explain the reasons for your choice and the limitations of the method for the survey in question.

(7)

(b) In a study of HIV-related risk groups, as part of a face-to-face interview, people in the survey were asked intimate questions about their personal lives. What problems are likely to be encountered? Discuss the nature and limitations of the so-called *randomised response method* that may be used in this situation.

(6)

(c) A survey of household income was conducted in which no time was available for further attempts to contact those who were not at home when the first call was made. All interviews were conducted in the evenings. Married women were asked the following questions about their husband's income:

Do you know how much your husband earns? Y/N

If yes, could you indicate how much your husband's income is each week

____ pounds

Comment on the weaknesses of the above questions and suggest alternative wordings.

(7)

6. A simple random sample of 1 in 20 households in a small town provided the following data about the availability of cars and the number of adults in households.

		Adults in household (x_i)					
		1	2	3	4	5	Total
Number of cars (y_i) in household	0	58	127	9	6	0	200
	1	68	140	27	4	1	240
	2	4	30	5	8	3	50
	3	0	3	4	2	1	10
Total		130	300	45	20	5	500

Note: summing over all 500 households in the sample, $\sum x_i y_i = 795$.

(i) Obtain point estimates and approximate 95% confidence intervals for the following:

(a) the total number of cars in the town's households,

(b) the ratio of cars per adult in the town's households,

(c) the proportion of households with 1 or more cars per adult.

(14)

(ii) Additional information is available that the mean number of adults per household based on all town households is 1.8. Use this to provide a revised estimate of the total number of cars in the town's households and construct an approximate 95% confidence interval for the population total. Comment briefly on the two estimators.

(6)

[You may use without proof the formula

$$\hat{V}\left(\hat{R}\right) = (1-f)\left(\hat{R}^2 s_x^2 - 2\hat{R}s_{xy} + s_y^2\right)/\left(n\bar{x}^2\right)$$

in standard notation.]

7. A wholesale food distributor in a large city wants to assess the demand for a new product based on mean monthly sales. He plans to sell this product in a sample of stores he services. He only services four large chains in the city. Hence, for administrative convenience, he decides to use stratified random sampling with each chain as a stratum. A stratified random sample of 20 stores yields the following sales figures after a month:

Stratum (Chain)			
1	2	3	4
$N_1 = 24$	$N_2 = 36$	$N_3 = 30$	$N_4 = 30$
$n_1 = 4$	$n_2 = 6$	$n_3 = 5$	$n_4 = 5$
$\bar{y}_1 = 99.3$	$\bar{y}_2 = 100.0$	$\bar{y}_3 = 98.0$	$\bar{y}_4 = 100.0$
$s_1 = 9.00$	$s_2 = 7.46$	$s_3 = 6.28$	$s_4 = 10.61$
94	91	108	92
90	99	96	110
103	93	100	94
110	105	93	91
	111	93	113
	101		

(i) Explain what is meant by *stratification with proportional allocation*, and verify that this has been used to construct the above stratum sample sizes.

(4)

(ii) Write down an expression for the unbiased estimator of the population mean in terms of the stratum means and derive its standard error.

(5)

(iii) Estimate the mean monthly sales and obtain an estimate of the standard error of your estimator. Construct an approximate 95% confidence interval for the population mean.

(3)

(iv) Suppose instead that a simple random sample of 20 stores from the population of 120 stores had been selected, with the same responses as given in the table. Had this been done, the estimate for the population standard deviation would have been 7.75. Construct an approximate 95% confidence interval for the population mean.

(2)

(v) Compare the efficiencies of your estimators in parts (iii) and (iv). Suggest why stratified random sampling gives a less precise estimate of the population mean than simple random sampling in this case.

(3)

(vi) Advise the wholesaler on how to select appropriate strata for a stratified random sample when the objective of stratification is to produce estimators with small variance.

(3)

Turn over

8. (a) Explain the role of *life tables* in population studies. Distinguish between *current* and *cohort* life tables, and state when each would be used. Explain briefly the relationship between life tables and *age-specific death rates*.

(4)

(b) The mortality rates for a certain stationary population, A, with 1000 births per year are given in the following table, where ${}_{10}q_x$ is the probability that a person aged x years dies within the next 10 years.

Age	${}_{10}q_x$
0	0.250
10	0.024
20	0.040
30	0.051
40	0.062
50	0.091
60	0.172
70	0.335
80	0.624
90	1.000

Using only this information construct a life table and estimate

(i) the age distribution in 10 year class intervals,

(5)

(ii) the expected age at death of groups now aged 20,

(3)

(iii) the life-expectancy of people in this population.

(2)

Define any notation used in parts (i) to (iii).

(iv) State any assumptions required for the validity of the calculations in parts (i) to (iii) and comment on whether they are appropriate.

(3)

(c) A different stable population, B, experiences the same mortality rates as population A and an annual growth rate of 1%. Without doing further calculations, explain how you would find the age distribution of population B in a form that is suitable for comparison with the distribution obtained for population A. How would you expect these distributions to differ?

(3)