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

GRADUATE DIPLOMA IN STATISTICS, 1997

Applied Statistics II

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. A 4×4 Latin square design was used to compare four varieties of Brussels sprouts. The yields, shown in row and column pattern with the variety letter beside each yield, are given below.

   Unfortunately, during the course of the trial, several plots in the bottom left corner of the experimental area were partially waterlogged. As a result, it was decided to ignore the yield from the plot in row 4 column 1 (variety A) which was clearly atypical compared to the yields from other plots.

B	98	D	100	A	127	C	142
C	141	A	91	D	110	B	124
D	97	C	102	B	103	A	127
A	34	B	71	C	119	D	118

   (i) The missing value was estimated to be 103 (after rounding). Give a brief explanation of how this estimate could have been derived (no calculations are required).

   (ii) Analyse these data to determine whether the varieties produce different mean yields.

   (iii) Using a suitable approximation, find the standard errors of the differences in mean yields.

   (iv) Construct approximate 95% confidence intervals for the difference in mean yields between varieties A and B, and between C and D. Comment on the results.

Turn over

2. A randomised block experiment was planned to compare four drugs A, B, C and D using three patients as blocks. At the time of administration of the drugs, it was found that there was a shortage of C and a surplus of D. The following design was used instead of that originally planned.

Patient Number	*Drugs administered*			
1	B	A	D	C
2	C	D	B	A
3	D	A	D	B

It was known that carryover effects could be discounted.

Write down the usual additive model for the response to the drugs and obtain the normal equations. [*Hint.* Modify the model for a randomised block design to allow for the fact that patient 3 was administered drug D twice.] Show that the least squares estimator of the difference between drugs C and D is

$$\frac{4\,(2T_C - T_D) + 3P_3 - G}{14}$$

where G is the grand total of all the observations, P_3 is the total for the 3rd patient and T_C and T_D are the totals for drugs C and D respectively.

Indicate how you would complete the analysis of variance.

3. An experiment was conducted to investigate the water uptake of amphibia. Frogs and toads were kept in dry or moist conditions prior to the experiment. Half of the animals were injected with a mammalian water balance hormone. Thus there were three treatment factors: species (S), pre-experiment moisture condition (M) and hormone (H) - each with two levels

S = species: toad or frog;
M = moisture: wet or dry;
H = hormone: control or hormone.

Two animals were observed for each of the eight treatment combinations, but there was no blocking of the 16 animals. The variable measured was the percentage increase in weight after immersion in water for two hours.

Treatment	*Results*		*Total*
Toad wet control	2.31	-1.59	0.72
Toad dry control	17.68	25.23	42.91
Toad wet hormone	28.37	14.16	42.53
Toad dry hormone	28.39	27.94	56.33
Frog wet control	0.85	2.90	3.75
Frog dry control	3.47	16.72	20.19
Frog wet hormone	3.82	2.86	6.68
Frog dry hormone	13.71	7.38	21.09

Note: Total sum of squares = 1757.827

Investigate the effects of moisture, species and hormone treatment on the water uptake of amphibia, using the analysis of variance.

Interpret your results, including a line plot to illustrate the significant interaction(s).

Turn over

4. A corporation desires to estimate the *total* number of man-hours lost for a given month, because of accidents among all employees. The company has 132 labourers, 92 technicians and 27 administrators. At random, 8 labourers, 10 technicians and 2 administrators were chosen, and the following data were collected from them:

Sample numbers of man-hours lost during the month

Labourers	8	0	6	7	24	16	0	1		
Technicians	4	0	8	3	1	5	24	12	2	8
Administrators	1	8								

Let $\bar{y}_h$ be the usual estimate of *mean* response for the *h*th stratum, N_h the number of units in stratum *h* .

(i) Explain why in the above example, stratified random sampling is preferable to simple random sampling.

(ii) Show that $\hat{Y} = \sum N_h \bar{y}_h$ is an unbiased estimator for the *total* number of man-hours lost during a given month, and find the variance of $\hat{Y}$. Note that results from simple random sampling can be assumed without proof.

(iii) Estimate the total number of man hours lost and its standard error.

(iv) Without computing the optimal allocation of the sample into the 3 groups, give an intuitive reason why a total sample of 20 employees could be allocated in a different way such that the variance of the estimator, $\hat{Y}$, is smaller.

5. (a) Discuss the use of sampling frames in a sample survey design. Your discussion should include a description of two sampling frames which would be suitable for sampling households in a country of your choice, comparing the units of the two frames.

(b) Consider a survey in which the sampling units are households and the only available sampling frame is a list of addresses.

(i) Under what circumstances would it be sensible to choose the next address in the frame should there be no reply from a selected address ?

(ii) Discuss the relative merits of choosing *one* household or choosing *all* households at a multi-household address.

6. A simple random sample of 33 low-income families yielded the following information on family size, weekly income (£) and weekly expenditure (£) on food.

Family Number	*Size* x_1	*Income* x_2	*Food Cost* y	*Family Number*	*Size* x_1	*Income* x_2	*Food Cost* y
1	2	62	14.3	18	4	83	36.0
2	3	62	20.8	19	2	85	20.6
3	3	87	22.7	20	4	73	27.7
4	5	65	30.5	21	2	66	25.9
5	4	58	41.2	22	5	58	23.3
6	7	92	28.2	23	3	77	39.8
7	2	88	24.2	24	4	69	16.8
8	4	79	30.0	25	7	65	37.8
9	2	83	24.2	26	3	77	34.8
10	5	62	44.4	27	3	69	28.7
11	3	63	13.4	28	6	95	63.0
12	6	62	19.8	29	2	77	19.5
13	4	60	29.4	30	2	69	21.6
14	4	75	27.1	31	6	69	18.2
15	2	90	22.2	32	4	67	20.1
16	5	75	37.7	33	2	63	20.7
17	3	69	22.6				

$\Sigma x_1 = 123$ $\Sigma x_2 = 2\,394$ $\Sigma y = 907.2$

$\Sigma x_1^2 = 533$ $\Sigma x_2^2 = 177\,254$ $\Sigma y^2 = 28\,224$

The sampling fraction for the survey was 2%.

From the sample, estimate each of the following together with its estimated standard error:

(i) the mean weekly expenditure on food per family.

(ii) the mean weekly expenditure on food per person.

(iii) the percentage of the income that is spent on food.

(iv) the mean weekly expenditure on food of 2-member families.

Turn over

7. (a) In the context of response surface methodology, explain carefully what is meant by the *method of steepest ascent* for finding the optimum value of a response variable y which depends on the values of quantitative factors $x_1, x_2, ...$ Your answer should include explanations of the terms *response surface*, *first- and second-order designs*, and *composite design.*

(b) A series of experiments is to be conducted to determine the temperature, ξ_1 , and the pressure ratio, ξ_2 , of a chemical reaction which produces oxygen of maximum purity. The first of these experiments was centered about the current operating conditions temperature (ξ_1) = -220°C and pressure ratio (ξ_2) = 1.2. The settings used, and the yields obtained are:

Temperature (ξ_1)	*Pressure Ratio* (ξ_2)	*Purity* (y)
-225	1.1	82.8
-225	1.3	83.5
-215	1.1	84.7
-215	1.3	85.0
-220	1.2	84.1
-220	1.2	84.5
-220	1.2	83.9
-220	1.2	84.3

Comment on the structure of this design, and the purpose of the replicated points.

Fit a first-order model to the above data. Construct the analysis of variance table, and hence test for the lack of fit of the model. Find the slope of the path of steepest ascent.

8. (a) Explain clearly the purpose of standardisation with respect to death rates, and distinguish between *direct* and *indirect* standardisation.

(b) **Population and Deaths, 1992**

	Northern Ireland		**United Kingdom**	
	Population ('000)	*Deaths*	*Population* ('000)	*Deaths*
Under 1	26	153	787	5 141
1 - 4	104	39	3 124	1 014
5 - 14	262	55	7 289	1 200
15 - 44	703	582	24 986	21 749
45 - 64	317	2 321	12 682	87 987
65 - 74	121	3 726	5 104	148 710
75 - 84	67	5 001	3 100	217 380
85 +	17	3 111	936	151 057
Total	1 617	14 988	58 008	634 238

(Source: Population derived from Population Trends, Spring 1994.
Deaths derived from Annual Abstract of Statistics, 1993.)

Calculate the crude death rates per thousand in Northern Ireland and the United Kingdom. Using the data for the United Kingdom as the standard population

(i) calculate the standardised death rate for Northern Ireland,
(ii) calculate the standardised mortality ratio for Northern Ireland,
(iii) calculate the indirect standardised death rate for Northern Ireland.