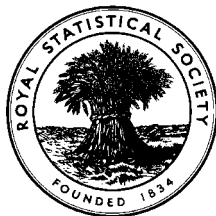


EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



GRADUATE DIPLOMA IN STATISTICS, 1998

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) (i) State the assumptions underlying the analysis of variance.
- (ii) What steps may be taken to overcome the problems caused by departures from these assumptions.
- (b) A bacteriologist is interested in the effects of two different culture media and two different times on the growth of a particular virus. She performs six replicates of a 2^2 design with the following results, for growth rates

<i>Time</i>	<i>Culture Medium</i>			
	<i>1</i>		<i>2</i>	
12 hr	21	22	25	26
	23	28	24	25
	20	26	29	27
18 hr	37	39	31	34
	38	38	29	33
	35	36	30	35

- (i) State an appropriate model for these data. Calculate the residuals, and the predicted values.
- (ii) Plot the residuals on normal probability paper and also plot the residuals versus the predicted growth rates. Comment on the model's adequacy.

2. (a) Sixteen animals, four from each of four litters, were available for an experiment to compare the effect on growth rate of four additives to a basic diet. The following three different designs are suggested:
- I Four pens each containing the four animals of one litter. One animal in each pen is given one additive at random, all additives being used in each pen.
 - II Each litter in a single pen, and given one additive (group feeding) chosen at random.
 - III Four pens each containing one animal from each litter. Additives are allocated so that each pen has one animal on each additive and also every litter has one animal on each additive.

For each design write out an outline ANOVA table, listing the items and their degrees of freedom.

- (b) The following comments were made by statisticians consulted by the experimenter. For each comment, explain for which design it correctly applies and why the comment is correct for that design and not for the others. Give an outline ANOVA table (items plus degrees of freedom) in each case.
- (A) There is no true replication in this since a pen is your experimental unit;
 - (B) This design provides more residual degrees of freedom than the other two;
 - (C) If you cannot feed animals individually then I suppose you are forced to do something like this. How about having one animal from each litter in each pen; then if you can ignore the effects of pens you could analyse it as a randomised block design with litters as blocks;
 - (D) You could analyse this as a Latin Square;
 - (E) You cannot distinguish the effects of litters and additives in this design;
 - (F) This is the only design which separates the effects of pens from that of litters;
 - (G) You will only have six degrees of freedom for error if you do it this way.

3. An experiment was conducted to investigate the effects of fertilizers and watering on the growth of carrots. Two fertilizers were used, ammonium sulphate (A.S.) and monocalcium phosphate (M.P.) as well as a control (no fertilizer). The carrots were also either given a heavy or light watering during growth.

The six treatments were applied at random to twelve plots of carrots and the average weight (in hundreds of grams) of the plant roots within each plot were as follows:

Heavy Watering			Light Watering		
<i>Control</i>	<i>A.S.</i>	<i>M.P.</i>	<i>Control</i>	<i>A.S.</i>	<i>M.P.</i>
72	89	84	61	56	85
82	110	89	40	54	81

- (i) Calculate the between and within treatment sums of squares and test whether there are differences between treatments.
- (ii) Give orthogonal linear contrasts between the treatment means which measure the following types of treatment differences, and give an estimate of the standard error of each contrast.
- heavy versus light watering
 - A.S. versus M.P.
 - A.S. versus M.P. at heavy watering against A.S. versus M.P. at light watering
 - fertilised versus non-fertilised
 - fertilised versus non-fertilised at heavy watering against fertilised versus non-fertilised at light watering.
- (iii) Split the between treatment sum of squares into components for each of the contrasts in (ii) and test their significance. Interpret your results and write a summary of your conclusions.

4. A new product was considered by the bakery products research division of a large corporation. Of paramount concern was the maximum peak height obtained from a standard container of mixed dough just prior to baking. Five major factors were thought to be important in affecting peak height: percentage of fat, percentage of water, amount of flour in the brew, the speed of the mixer in rpm, and mixing time in minutes.

It was proposed to use a composite design with three components, I, II and III, as follows:

- I a 2^{5-1} fractional design using values $x_A = \pm 1, \dots, x_E = \pm 1$;
- II $k (\geq 1)$ centre points i.e. $x = (0, 0, 0, 0, 0)$;
- III *ten* points on the axes i.e. $(\pm\alpha, 0, 0, 0, 0), \dots, (0, 0, 0, 0, \pm\alpha)$.

- (i) Discuss the purpose of the different components of this design in the context of a second order model for predicting maximum peak height.
- (ii) Write down a suitable defining relation and the aliases for this design. Explain carefully why the use of a quarter fraction in place of the half fraction to save experimental runs, would not be sensible. Your answer should include the alias structure of a 2^{5-2} factorial design.

An experiment of this form was performed, and no main effects or first order interactions were aliased with each other. The resulting data were analysed by fitting a second order model to the data. This gave the following analysis of variance table:

Source of variation	df	MS
Constant term	1	108.17
First -order terms	5	84.15
Interaction terms	*	131.80
Second order terms	*	70.91
Lack of fit	*	38.12
Pure error	4	26.31
Total	31	459.46

- (iii) Complete the table and summarise the results using appropriate statistical tests. Briefly suggest further analyses of the data that you consider appropriate, giving your reasons.

5. (a) Define the demographic usage of the term *fertility*, and distinguish between *period* and *cohort* analysis of fertility.
- (b) **Mid-year Female Population and Live Births by Maternal Age, 1961**

<i>Age</i>	<i>Female Population</i>	<i>Live Births</i>
15-19	18 000	299
20-24	20 000	3 008
25-29	21 000	2 814
30-34	19 000	1 938
35-39	27 000	1 485
40-44	24 000	456
15-44	129 000	10 000
All ages	315 000	10 122

Mid-year male population, 285 000

Maternal and Infant Deaths and Stillbirths, 1961

Maternal deaths	3
Infant deaths (1st year of life)	210
Neonatal deaths (1st four weeks of life)	126
Early neonatal deaths (1st week of life)	106
Stillbirths	200

Using the above data on population, births and deaths, calculate the following:

- (i) Birth rate
- (ii) General fertility rate
- (iii) Fertility rate at ages 20 to 24
- (iv) Infant mortality rate
- (v) Neonatal mortality rate
- (vi) Postneonatal mortality rate
- (vii) Stillbirth rate
- (viii) Perinatal mortality rate
- (ix) Maternal mortality rate

6. The wholesale price paid 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 entire load. You may assume that (a) the sugar content of an individual orange, y is closely related to its weight, x ; and (b) the ratio of the total sugar content τ_y to the total weight of the truckload τ_x is equal to the ratio of the mean sugar content per orange, μ_y to the mean weight, μ_x .

(i) How can the *total* sugar content of the load be estimated from a random sample n oranges from the load

(a) if the total number of oranges, N in the load is known?

(b) if only the total weight of the oranges, τ_x in the truck is known?

In each case, what measurements must be made on the sample of n oranges?

(ii) Variates y_i and x_i are measured on each unit of a simple random sample of size n , assumed large. Show that the variance of $r = \frac{\bar{y}}{\bar{x}}$ is

$$V(r) \approx \frac{1-f}{n\mu_x^2} \sum_{i=1}^N \frac{(y_i - Rx_i)^2}{N-1}$$

where $R = \mu_y/\mu_x$ is the ratio of the population means and $f = n/N$.

(c) Hence derive an approximate expression for the variance of your estimator in (b) of part (i) above. State the condition under which the use of your estimator is better than the use of the estimator in (a) of part (i).

(d) Roughly how many oranges must be sampled from a very large truckload of oranges weighing 1800 pounds in order for the standard error of the estimator in (b) of part (i) to be about 3 pounds, where $\sum_{i=1}^N \frac{(y_i - Rx_i)^2}{N-1} = (0.0030)^2$. You may assume that the mean weight of an orange is 0.4 pounds.

7. (a) Define the term *stratified random sampling*. Explain what is meant by the expressions *stratification with proportional allocation* and *stratification with optimal allocation*.
- (b) A population of size N is divided into L strata, the stratum h being of size N_h . The mean of the elements in the h^{th} stratum is Y_h and the variance is S_h^2 . The population mean is \bar{Y} and the variance is S^2 . A sample of size n is selected by taking independent random samples of size n_h from stratum h .

The population mean \bar{Y} is to be estimated, either using a stratified random sample with proportional allocation or using a simple random sample of the same size. Let V_{prop} and V_{ran} be the variances of the two estimators. Show that

$$V_{ran} = V_{prop} + \frac{N-n}{nN(N-1)} \left[\sum^L N_h (\bar{Y}_h - \bar{Y})^2 - \frac{1}{N} \sum^L (N - N_h) S_h^2 \right]$$

- (c) The following data show the stratification of all farms in a country by farm size showing summary information about the area devoted to corn (maize) per farm in each stratum.

<i>Farm Size (acres)</i>	<i>Number of Farms N_h</i>	<i>Average Corn Acres \bar{Y}_h</i>	<i>Standard Deviation S_h</i>
≤ 40	394	5.4	8.3
41- 80	461	16.3	13.3
81- 120	391	24.3	15.1
121- 160	334	34.5	19.8
161- 200	169	42.1	24.5
201- 240	113	50.1	26.0
≥ 241	148	63.8	35.2
Overall	2010	26.3	

$$\sum^L W_h S_h^2 = 343.2788; \quad \sum^L W_h S_h = 17.0183, \text{ where } W_h = \frac{N_h}{N} \text{ for } h = 1, 2, \dots, L$$

For a sample of 100 farms, compute the sample sizes in each stratum under:

- (i) Optimum allocation
(ii) Proportional allocation.

Compare the precisions of these methods with that of simple random sampling.

8. At an experimental station 100 fields (each of area one hectare) were sown with wheat. Each field was divided into sixteen equal plots. A simple random sample of 10 fields was selected, and from each of these fields a simple random sample of 4 plots was selected. The yields in kg/plot are given below:

<i>Field</i>	<i>Plot within field</i>			
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
1	4.28	4.36	3.00	3.52
2	4.20	4.66	3.64	5.00
3	4.40	4.72	4.04	3.98
4	5.16	4.24	4.96	3.84
5	4.08	3.96	3.42	3.08
6	4.12	4.68	3.46	4.02
7	4.00	4.84	4.32	3.72
8	3.06	4.24	4.76	3.12
9	4.16	4.36	3.50	5.00
10	4.32	4.84	3.96	4.04

- (i) Estimate the wheat yield per hectare for the experimental station and its standard error (note: you may assume that, in standard notation, that

$$\hat{V}(\bar{y}) = \frac{(1-f_1)}{n} S_b^2 + f_1 \frac{(1-f_2)}{nm} S_w^2).$$

- (ii) Estimate the relative efficiency of this estimator to that obtained from a simple random sample of 40 plots.
- (iii) If the cost of including a field in the sample is four times the cost of including an extra plot, and total cost (excluding overheads) must not exceed 100 units, use the method of Lagrange multipliers to derive the optimum number of fields and the optimum number of plots per field for the sample.