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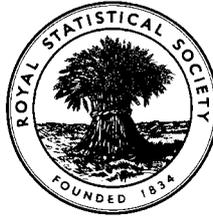
(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA IN STATISTICS, 1996

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. Describe some circumstances in which data transformation may be useful in statistical analysis. Discuss how you would examine your data to decide whether a transformation was needed. Discuss briefly any implications the use of transformations may have on the presentation of conclusions.

Two insecticides, *A* and *B*, have been developed for controlling boll weevils when growing cotton. It is thought that the insecticides might be most effective when used together. Four powders were prepared, the first containing no insecticide, the second containing *A*, the third *B*, and the fourth with equal quantities of *A* and *B*.

A field trial of the four powders was carried out using a randomised block design. Each block was divided into eight plots, so that each powder was sprayed on two of the plots within each block, the same amount of powder being sprayed on all 32 plots used in the trial. One hundred buds were examined from each plot, selected from the centre rows of each plot, and the numbers showing attack by boll weevils were as follows.

| | Insecticides | | | | | | | |
Block	None		*A*		*B*		*A* and *B*	
I	9	13	5	16	6	4	3	6
II	16	11	8	14	12	7	5	5
III	33	20	17	15	13	18	13	7
IV	15	13	5	6	10	6	7	4

Explain why the arc sine square root transform is appropriate for these data. Transform the data using an arc sine square root transformation. Construct a two-way analysis of variance and, carrying out any tests which you feel are necessary, comment on the effectiveness of the insecticides in this experiment.

2. Describe the purpose and use of stratification in sample surveys, distinguishing between proportional and optimal allocation.

An investigator proposes to take a stratified random sample with two strata. He expects that his total field cost will be of the form $c_1 n_1 + c_2 n_2$, where c_h is the cost of sampling one unit from stratum h and n_h is the number of units randomly selected from stratum h. His advance estimates of relevant quantities for the two strata are as follows

Stratum	W_h	S_h	C_h ($)
1	0.4	10	4
2	0.6	20	9

where $W_h = N_h/N$ is the stratum weight of stratum h, and S_h^2 is the population variance of y, the item of interest, in stratum h.

Find the values of n_1/n and n_2/n that minimise the total field cost for a given value of variance of $\bar{y}_{st}$, where n is the total number of units selected and $\bar{y}_{st}$ is the usual estimate of the population mean from a stratified sample. Find the total number of units, n, required, under this optimum allocation, to make $V(\bar{y}_{st}) = 1$. Assume the finite population correction is negligible.

How much will the total field cost be ?

After the sample is taken, the investigator finds that his field costs were actually $2 per unit in stratum 1 and $12 per unit in stratum 2.

(a) How much greater is the field cost than anticipated ?

(b) If he had known the correct field costs in advance, could he have attained $V(\bar{y}_{st}) = 1$ for the original estimated field cost ?

3. (a) Define *confounding* and *partial confounding*, and explain briefly their importance in 2^n factorial experiments.

The following treatments were applied to 1/60 acre plots of potatoes:

N = Sulphate of ammonia at 0 and 0.45 cwt N per acre
K = Sulphate of potash at 0 and 1.12 cwt K_2O per acre
D = Dung at 0 and 8 tons per acre.

Each block available for the experiment contains 4 plots. The treatments are to be applied in all factorial combinations to plots.

Produce a design for four replicates of this experiment with NKD confounded in replicate I, NK in replicate II, ND in replicate III, and KD in replicate IV. Outline the analysis of variance and comment on the information obtained.

(b) Discuss the problems associated with using only single replicates in a factorial experiment. Explain how you would proceed to assess the significance of the main effects and interactions in such experiments.

4. Explain briefly how you would design sample surveys in *each* of the following cases, giving reasons for your choice of design in each case. What problems of data collection might you encounter, and how would you deal with them ?

(i) A survey of political attitudes amongst voters.
(ii) A survey of a country's annual production of food crops.
(iii) A survey of food consumption by householders.

Turn over

5. Four different food enzymes are being studied to determine their effect on bread quality, and on the volume of bread produced. The experimental procedure consists of selecting a batch of raw material, adding each enzyme to the ingredients in a separate run of the process, baking the dough and measuring loaf volume. Since variations in the batches of raw material may affect the quality of bread, the technologist decided to use batches of raw material as blocks. However, each batch is only large enough to test three enzymes. The design for this experiment, along with the loaf volumes (ml) recorded, is shown below.

Enzyme	Batch of raw material			
	1	2	3	4
A	238	196	254	-
B	238	213	-	312
C	279	-	334	421
D	-	308	367	412

State the properties of the above experimental design. Explain how the estimation of the treatment effects differs from that for the randomised block design.

Perform a within-block analysis of this experiment. Comment on the results.

Define the term *inter-block* information. Comment on when it is worthwhile to obtain inter-block information.

6. Define the term *systematic sampling*. Explain the circumstances under which the procedure might be regarded as equivalent to (i) simple random sampling, (ii) cluster sampling, and (iii) stratified random sampling. In what circumstances would you consider it unwise to use systematic sampling ?

The total weight (kg) of catch landed by fishing boats at a harbour over a 12-hour period is to be estimated by observing the weight landed during a sample of complete hours. Three methods are being considered:

(i) a systematic sample of 4 hours

(ii) a simple random sample of 4 hours

(iii) a cluster sample of two clusters. Treat hours 1 and 2 as cluster 1, hours 3 and 4 as cluster 2, hours 5 and 6 as cluster 3, hours 7 and 8 as cluster 4, hours 9 and 10 as cluster 5 and hours 11 and 12 as cluster 6.

Using the table of random numbers, employ each of these methods to select a sample of 4 hours from the 12 hours in the period, showing in detail how the samples were selected.

Illustrate your answer by finding, for each method, an estimate of the *total* catch landed in the 12-hour period if, in fact, the weights landed were as shown in the table below.

Hour	1	2	3	4	5	6	7	8	9	10	11	12
Weight (kg)	567	861	231	92	347	1117	946	1301	465	444	96	0

Compare the relative efficiencies of your estimators. Comment on your results.

5

7. In the context of response surface methodology, explain what is meant by a *mixture experiment*, and how it differs from a *factorial experiment*. Give an example of an experimental situation which falls under the heading of a mixture experiment.

Three chemical pesticides - Vendex (x_1), Omite (x_2) and Kelthane (x_3) - were sprayed on strawberry plants in an attempt to control the mite population. Each chemical was applied individually and in combination with each of the others to comprise three pure blends and three binary blends. Altogether, 24 plants were used, four allocated to each of the six blends. For each plant, the response measured was the number of mites seven days after spraying, expressed as a percentage of the number of mites just prior to spraying. This was averaged over the four plants allocated to each blend.

It is proposed to use a $\{q, m\}$ simplex-lattice design for this experiment, which will allow a polynomial function of degree m in q components to be fitted to the data. A $\{q, m\}$ simplex-lattice design for q-components consists of points defined by the following coordinate system: the proportions assumed by each component take the $m+1$ equally spaced values from 0 to 1

$$x_i = 0, 1/m, 2/m, ..., 1 \qquad\qquad i = 1,2,...,q \qquad (1)$$

and all possible combinations (mixtures) of the proportions from equation (1) are used.

Produce a $\{3, 2\}$ simplex-lattice design for a single replicate of this experiment, and construct a two-dimensional simplex to represent the factor space.

Show that the $\{3, 2\}$ polynomial function associated with this design can be written in the form

$$\eta = \sum_{i=1}^{3} \beta_i x_i + \sum_{j=1}^{3} \sum_{\substack{i=1 \\ i \neq j}}^{3} \beta_{ij} x_i x_j \qquad (2)$$

and deduce an association between the points of the $\{3, 2\}$ simplex-lattice design and the parameters of the polynomial function.

The following least squares estimates of the parameters in (2) were calculated from the multiple observations collected at the points of the $\{3, 2\}$ simplex-lattice design

$$b_1 = 1.8, \qquad b_2 = 28.6, \qquad b_3 = 38.5$$
$$b_{12} = -48.4, \qquad b_{13} = 34.2, \qquad b_{23} = -91.4 .$$

The estimated standard errors of the parameter estimates were 7.4 for b_i and 36.4 for b_{ij}.

What would you conclude about the chemical blends used in this experiment?

8. Explain the notation $_nq_x$, l_x, $_nL_x$, T_x and $\overset{o}{e}_x$ in the context of life tables.

The data below show the population size and number of deaths of females in England and Wales in 1982. Calculate the columns $_nq_x$, l_x, $_nL_x$, T_x and $\overset{o}{e}_x$ of an abridged life table for females in England and Wales. Assume that $l_0 = 100\,000$. Indicate how you have calculated each column.

Age (yrs)	Mid-year Population	Registered Deaths
0	300 800	2 861
1 - 4	1 190 600	483
5 - 9	1 473 800	253
10 - 19	3 861 500	941
20 - 39	7 035 900	4 029
40 - 59	5 592 500	22 976
60 - 79	4 958 000	128 984
80 +	1 048 300	131 166
Total	25 461 400	291 693

[Source: Office of Population Censuses and Surveys.   Death Statistics 1982]

A stationary population is supported by 10 000 female births per annum and experiences the mortality of females in England and Wales in 1982. Use your life table to determine the following:

If the female working population is considered to be females aged 20 to 60

(a) what is the number of females of working age ?
(b) how many deaths of females of working age occur each year?
(c) what is the average death rate for this particular age group ?

What are the expected ages at death of three groups of females now aged 20, 40 and 60 respectively ?

State any assumptions you have made in these calculations.