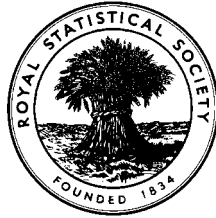


EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



GRADUATE DIPLOMA, 1999

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. There are many ways to bake cakes. The purpose of this experiment was to determine how the pan material, the brand of cake mix and the stirring method affect the taste of cakes. The factor levels were:

Factor	Low (-)	High (+)
A = pan material	glass	aluminium
B = stirring method	spoon	mixer
C = brand mix	expensive	cheap

The response variable was taste, a subjective measure derived from a questionnaire given to the subjects who sampled each batch of cakes. An eight-person test panel sampled each batch and filled out the questionnaire. The complete set of data is shown below.

Cake Batch	A	B	C	Test Panel Results								Total
				1	2	3	4	5	6	7	8	
1	-	-	-	11	9	10	10	11	10	8	9	78
2	+	-	-	15	10	16	14	12	9	6	15	97
3	-	+	-	9	12	11	11	11	11	11	12	88
4	+	+	-	16	17	15	12	13	13	11	11	108
5	-	-	+	10	11	15	8	6	8	9	14	81
6	+	-	+	12	13	14	13	9	13	14	9	97
7	-	+	+	10	12	13	10	7	7	17	13	89
8	+	+	+	15	12	15	13	12	12	9	14	102
<i>Total</i>				98	96	109	91	81	83	85	97	740

$$\Sigma x = 740 \quad \Sigma x^2 = 8988$$

- (i) Analyse the data from this experiment as if there were eight replicate blocks of a 2^3 factorial design. Comment briefly on the results. (8)
- (ii) Is the analysis in part (i) the correct approach? There are only eight batches. Do we really have eight replicate blocks of a 2^3 factorial? (3)
- (iii) Analyse the average taste ratings of the eight test panel results. Plot the effect estimates on Normal probability paper. Which factors appear to have large effects? (6)
- (iv) Is this analysis more appropriate than the one in part (i)? Why or why not? (3)

2. (a) Explain fully the circumstances in which a *balanced incomplete block* design is appropriate. In the usual notation, let v represent the number of treatments, b the number of blocks, r the number of replicates of each treatment, k the number of units in each block and λ the number of times each pair of treatments occurs together in a block. Derive two equations connecting the parameters v, b, r, k and λ .

(4)

- (b) In an industrial experiment, 5 batches of metal ingots were selected at random from a production process where each batch comprised 4 ingots. Each ingot was melted and mixed with an amount of cadmium (Cd) and tin (Sn), then allowed to cool. When the treated ingot was reheated, its melting point ($^{\circ}\text{C}$) was recorded, as shown.

Batch 1	A: 194	D: 205	E: 250	B: 214
Batch 2	B: 204	E: 243	D: 198	C: 238
Batch 3	D: 206	B: 205	C: 238	A: 186
Batch 4	A: 183	E: 247	B: 202	C: 229
Batch 5	E: 255	C: 244	D: 209	A: 198

$$\Sigma x = 4348 \quad \Sigma x^2 = 955360$$

The treatments were described as follows:

- A: 10% Cd, no Sn
- B: 20% Cd, no Sn
- C: 30% Cd, no Sn
- D: 10% Cd, 10% Sn
- E: 30% Cd, 10% Sn

- (i) State the values of the parameters v, b, r, k and λ of the above design. (2)

- (ii) Obtain the analysis of variance for the data. Briefly state your conclusions. (8)

- (iii) Estimate the adjusted treatment means. Interpret your results. (6)

3. The effect of fertiliser on the yield of potatoes was investigated in a 2^3 factorial experiment, involving 32 plots. There were 4 replications of each treatment combination, allocated at random to the plots.

The factors were three fertiliser constituents: nitrogen (*A*), potassium (*B*) and dung (*C*). Each was either at a single level or absent. The yields of potatoes (kg) are displayed in the accompanying computer output.

- (i) Interpret the results from the computer output (you may assume the usual statistical assumptions are reasonable). Where appropriate, undertake any additional statistical analyses to investigate significant treatment effects. (12)
- (ii) Explain why the treatment combinations were randomly allocated to the experimental plots. (3)
- (iii) Describe how to use random numbers to allocate the treatments to experimental plots. (3)
- (iv) Suppose that the field is divided into a 4×8 rectangular grid with plots running in an East direction within rows, and in a North direction within columns of the field. Suppose that there is a soil fertility gradient running in a North direction across the field from one of its boundaries to the opposite boundary. Describe a suitable experimental design which still has 4 replicates of each treatment combination. (2)

Rows: replicates Columns: treatment combinations

	(-)	<i>a</i>	<i>b</i>	<i>ab</i>	<i>c</i>	<i>ac</i>	<i>bc</i>	<i>abc</i>
1	101	106	265	291	312	373	398	450
2	106	89	272	306	324	338	407	449
3	87	128	279	334	323	324	423	471
4	131	103	302	272	324	361	445	437

Analysis of Variance Procedure

Source	DF	Sum of Squares	Mean Square	F-Value	Pr > F
Model	7	458717.969	65531.138	195.09	0.0001
Error	24	8061.750	335.906		
Corrected Total	31	466779.719			

Source	DF	Anova SS	Mean Square	F Value	Pr > F
<i>A</i>	1	3465.281	3465.281	10.32	0.0037
<i>B</i>	1	161170.031	161170.031	479.81	0.0001
<i>A*B</i>	1	344.531	344.531	1.03	0.3213
<i>C</i>	1	278817.781	278817.781	830.05	0.0001
<i>A*C</i>	1	810.031	810.031	2.41	0.1335
<i>B*C</i>	1	13986.281	13986.281	41.64	0.0001
<i>A*B*C</i>	1	124.031	124.031	0.37	0.5491

4. The 2026 households in a city are divided up into four strata. Simple random samples of households are selected from within strata and the proportion of households renting the house they live in is found by interview. It is required to estimate the proportion of households living in rented houses.

<i>Stratum based on income</i>	<i>Stratum population size</i> N_h	<i>Stratum sample size</i> n_h	<i>Number renting</i>
< 50	1190	40	30
50-100	523	35	18
100-200	215	35	7
>200	98	40	5

Let p_h be the sample proportion for the h th stratum who are renting, N_h the number of units in stratum h , and n_h the number of units sampled in stratum h . Define N and n as the total population size and total sample size respectively, and L as the number of strata.

- (i) Show that the sample proportion $p_{st} = \frac{1}{N} \sum_{h=1}^L N_h p_h$ is an unbiased estimator for the proportion of households living in rented houses. (5)

- (ii) Find the variance of p_{st} and show that

$$\frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{p_h(1-p_h)}{(n_h - 1)}$$

is an unbiased estimator of this variance. (7)

- (iii) Estimate the proportion of households living in rented houses and its standard error. (5)
- (iv) Is the sample allocation reasonable? If not, suggest a more suitable allocation of a total sample of 150 households. (3)

5. A survey is to be conducted to find out how much adults resident in London exercise.

(i) Criticise the following questions:

Do you *never* exercise, *occasionally* exercise, or *frequently* exercise?

Which of the following sports do you play: squash, tennis, football, cycling, running?

(6)

(ii) Suggest your own questions, setting your version in the style that you would present in a self-completion questionnaire.

(7)

(iii) Discuss the merits of personal interviews, telephone interviews and postal questionnaires as methods of carrying out this survey.

(7)

6. National income from manufacturing industries is to be estimated for 1989 from a sample of 6 of the 19 industrial categories that reported figures early for that year. Incomes from all 19 industries are known for 1980 and the total is \$674 billion.

<i>Industry</i>	<i>\$billion</i>	
	1980	1989
Lumber and wood products	21	26
Electric and electronic equipment	63	91
Motor vehicles and equipment	35	47
Food and kindred products	60	70
Textile mill products	16	17
Chemical and allied products	50	76

- (i) Estimate *total* national income from manufacturing in 1989 using the following:
- (a) a simple random sample estimator,
 - (b) a ratio estimator,
 - (c) a regression estimator.
- (6)
- (ii) Which of the three methods (a), (b) or (c) above do you consider to be most appropriate in this case? Why?
- (5)
- (iii) Estimate and compare the relative efficiencies of your estimators. Do the results support your answer to part (ii) above?
- (9)

7. (i) Explain briefly the purpose of a *first-order* model in the context of *response surface* methodology. Discuss the main criteria for comparing different first-order designs. (4)

(ii) A laboratory experiment is to be designed to examine the relationship between growth (y) of a particular organism and the percentage of glucose (x_1), concentration of yeast extract (x_2), and the time in hours (x_3) allowed for organism growth. A first-order model is postulated though it is not quite clear that this is the correct model. Eight design runs are to be used.

It has been suggested that one of the following designs may be useful:

Design 1: 2^3 factorial

Design 2: $\frac{1}{2}$ fraction of 2^3 factorial augmented with 4 centre runs

Design 3: $\frac{1}{2}$ fraction of 2^3 factorial with replicate runs at each design point

Compare the above 3 designs. Your answer should include reference to criteria such as *orthogonality*, *variance optimal* and *predicted variance*. Marks will be awarded for supporting calculations.

(10)

(iii) Discuss the merits of each design for fitting a first-order model. Explain the circumstances in which you would use each of the 3 designs.

(6)

8. (i) Explain the notation n_i , \hat{q}_i , \hat{p}_i and $\hat{S}(t_i)$ used in the context of clinical life tables. How would you calculate each in terms of the number lost to follow-up, l_i , number withdrawn alive, w_i , number dying, d_i , and number entering the i th interval, n_i' ?

(6)

(ii) Survival data for 2418 males with angina pectoris for the first 8 years after diagnosis are shown below. With the exception of 86 who were lost to follow-up (shown in l_i column below) all were followed either to death (d_i column) or were known to be alive at the end of the 8th year (w_i column).

Year after diagnosis	Number lost to follow-up l_i	Number withdrawn alive w_i	Number dying d_i	Number entering interval n_i'
0	0	0	456	2418
1	9	30	226	1962
2	10	12	152	1697
3	0	23	171	1523
4	9	15	135	1329
5	10	97	125	1170
6	25	108	83	938
7	15	87	74	722
8	8	60	51	546

Complete this life-table and determine $\hat{S}(t_i)$, the estimate of the probability of survival to each year after diagnosis.

(4)

(iii) Another important concept in survival is the hazard function $\hat{h}(t_{mi})$ defined as the probability of failure (usually instantaneous) given survival to date. The hazard function for the i th interval, estimated at the midpoint, is :

$$\hat{h}(t_{mi}) = \frac{2\hat{q}_i}{(1+\hat{p}_i)} \quad i = 1, \dots, s-1$$

where s refers to the number of intervals.

Calculate $\hat{h}(t_{mi})$ for these data.

(2)

(iv) Plot $\hat{S}(t_i)$ against time t . Also plot $\hat{h}(t_{mi})$ against the midpoint of the interval. Comment.

(8)