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

GRADUATE DIPLOMA, 2006

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 calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

The notation log *denotes logarithm to base* ***e****.*
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^nC_r$.

This examination paper consists of 9 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. An experiment was carried out to investigate both the effect of rate of seeding and the effect of spatial arrangement on the yield of turnips. Five seeding rates (0.5, 2, 8, 20, 32 lb/acre) and four row widths (4, 8, 16, 32 inches) were tested. The experiment was laid out in three blocks, each with 20 equal-sized plots. Within each block, the 20 treatment combinations were allocated at random to the plots.

The table below summarises the total yields for the three plots for each treatment combination (in coded units), obtained during a fixed period in the growing season.

		Row width (inches)				
		4	8	16	32	*Total*
Seed rate (lb/acre)	0.5	1.87	3.07	3.13	2.85	10.92
	2	5.40	6.28	6.08	5.73	23.49
	8	7.67	7.94	6.94	6.58	29.13
	20	8.16	8.46	8.43	6.80	31.85
	32	8.38	8.40	7.81	7.05	31.64
	Total	31.48	34.15	32.39	29.01	127.03

Block totals (of 20 plots each) are: I, 36.09; II, 43.27; III, 47.67.

The sum of the squares of all 60 observations is 301.4107.

You may also use the fact that $1.87^2 + 5.40^2 + \ldots + 7.05^2 = 889.5165$.

(i) Carry out an analysis of variance to examine the effects of seed rate, row width, and their interaction, on the yield of turnips.

(6)

(ii) Partition the sum of squares for the row width main effect into single-degree-of-freedom components. Examine and comment on these.

[Note. The row width levels used were equally spaced on the logarithmic scale. The coefficients of linear, quadratic and cubic components for four equally-spaced levels of a factor are, respectively, (–3, –1, 1, 3), (1, –1, –1, 1) and (–1, 3, –3, 1).]

(5)

(iii) Draw a diagram showing all 20 totals of seed rate and row width combinations.

(4)

(iv) Using the diagram and the analysis of variance, explain the results found by this experiment, including mention of any interaction between seed rate and row width.

(5)

Turn over

2. 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 you use in your answers. State the conditions under which two contrasts are *orthogonal* and explain the relevance of orthogonality.

(6)

An experiment was conducted on the effect of inoculating *Phaseolus vulgaris* (bean) seeds with nitrifying *Rhizobium* bacteria. The aim of the experiment was to investigate the effect of a liquid fertiliser when used with different strains of Rhizobium. Eight treatments were used, labelled A – H, and defined below. The fertiliser was used at two levels. Treatments G and H were "controls", not inoculated with Rhizobium.

Treatment	*Strain of Rhizobium*	*Cultural history*	*Fertiliser level*
A	R 3644	Newly-cultured	Low
B	R 3644	Newly-cultured	High
C	R 3644	Repeatedly subcultured	Low
D	R 3644	Repeatedly subcultured	High
E	CC 511	Peat-based	Low
F	CC 511	Peat-based	High
G	Non-inoculated		Low
H	Non-inoculated		High

The experiment consisted of 5 replicates in a completely randomised design. The response was total root nodule weight after a fixed period of time.

The mean responses (in coded units) for the 5 replicate samples were:

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
88	198	66	235	265	233	40	41

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

(a) effect of fertiliser;

(b) difference between the two cultures of R 3644;

(c) effect of inoculation;

(d) difference between the two strains of Rhizobium;

(e) interactions of (b), (c) and (d) with fertiliser.

(6)

(ii) Calculate the value of each contrast. Given that the residual (error) mean square is 3265.8, test the statistical significance of each contrast, stating any assumptions required for the validity of the test. Summarise the results found by this experiment, including mention of any fertiliser/*Rhizobium* interactions.

(8)

3. Explain clearly what is meant by a *balanced incomplete block* design. When is this design useful?

(4)

Seven different confectionery products, A to G, made from the same ingredients but using slightly different recipes, were examined by a panel of experts. There were 7 panel sessions; at each session, 3 recipes were tested, the order of testing being random. The panel assessed the recipes blind, and gave a total score to each recipe based on a variety of characteristics. A higher score indicated a better perceived quality.

The scheme for the experiment and the results (y) were as shown in the table.

Block (panel session)	I	II	III	IV	V	VI	VII
Product tested and response y	A: 20	B: 25	C: 24	A: 16	B: 20	C: 19	A: 19
	B: 23	D: 21	E: 18	D: 14	E: 16	D: 17	F: 20
	C: 16	F: 20	F: 19	E: 15	G: 25	G: 22	G: 24
Session total	59	66	61	45	61	58	63

$\sum y = 413$, $\sum y^2 = 8341$.

The estimated treatment effects adjusted for blocks ($\hat{\tau}_i$) were:

(Adjusted) estimated treatment effects						
A	B	C	D	E	F	G
–0.2857	2.5714	–0.1429	–1.8571	–2.8571	–1.8571	4.4290

(i) Verify that this is a balanced incomplete block design with panel sessions forming the blocks, and find the values of its structural parameters N, b, k, v, r and λ. Define all the notation that you use.

(5)

(ii) Use the estimated treatment effects (adjusted for blocks) given above to construct the analysis of variance.

[You may use the fact that the treatment sum of squares, adjusted for blocks, is $\frac{v\lambda}{k}\sum \hat{\tau}_i^2$.]

(6)

(iii) Carry out any tests of significance which you consider necessary to investigate differences between treatments, and state your conclusions.

[Note. $\mathrm{Var}(\hat{\tau}_i - \hat{\tau}_j)$ can be estimated by $2k\hat{\sigma}^2/(v\lambda)$ for any $i \neq j$, where $\hat{\sigma}^2$ is suitably defined.]

(5)

Turn over

4. (a) In an investigation into the failure stress of sand aggregate pavements, the effects of the amounts of sulphur (% S) and asphalt (% A) were studied. Each factor had 3 levels. All nine possible combinations of these factor levels were used as the treatments, and there were two replicates of each combination. These 18 runs were carried out in a completely randomised order. It was assumed that experimental conditions would not change during this experiment.

The data given below are the failure stresses y (in pounds per square inch) for the two runs for each treatment combination.

%A (Asphalt)	*%S (Sulphur)*	*Failure stresses*	*Total*
2	14	338, 344	682
4	14	258, 272	530
6	14	320, 334	654
2	16	264, 290	554
4	16	242, 207	449
6	16	308, 310	618
2	18	332, 325	657
4	18	258, 233	491
6	18	336, 350	686

$\sum y = 5321$, $\sum y^2 = 1\,605\,355$; also, $682^2 + 530^2 + \ldots + 686^2 = 3\,207\,507$.

(i) Briefly describe how this form of experiment is useful for determining operating conditions that minimise failure stress.

(4)

(ii) Copy and complete the following analysis of variance table for these data, using the fact that the coefficients of linear and quadratic components for three equally-spaced levels of a factor are, respectively, (–1, 0, 1) and (1, –2, 1).

(5)

Source of variation	DF	Sum of squares	MS	*F* ratio
A_{linear}				
$A_{quadratic}$		23053.361		
S_{linear}		85.333		
$S_{quadratic}$				
$A_{linear} \times S_{linear}$				
Other AS components		1083.098		
Treatments	8	30806.778		
Residual				
TOTAL				

(iii) Write down the second-order linear model that would be fitted to these data. Explain which terms in the analysis of variance are used up by fitting this model. (2)

(iv) Obtain sums of squares for *lack of fit* and *pure error*. Hence complete the analysis and comment on the adequacy of the fit. (4)

(b) 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 where a mixture experiment would be used. (5)

Turn over

5. (i) A simple random sample of 500 households was selected from the 10 000 households in a town. The number of adults in the household (x) and the total number of cars owned by these adults (y) were recorded.

		Number of adults (x)					
		1	2	3	4	5	*Total*
Number of cars (y)	0	18	119	28	11	0	176
	1	21	130	84	8	0	243
	2	1	28	16	15	1	61
	3	0	3	12	4	1	20
	Total	40	280	140	38	2	500

In the most recent census, the proportions of households in the town with 1, 2, 3, 4 and 5 adults were 0.10, 0.50, 0.30, 0.09 and 0.01 respectively.

(a) Calculate the ratio and regression estimates of the mean number of cars per household in this town.

(7)

(b) Estimate the relative efficiency of the ratio and regression estimators, and comment.

(4)

(c) Giving your reasons, say which you think is the better of these two estimators, and use it to construct an approximate 95% confidence interval for the total number of cars owned by adults living in this town.

[You may assume that $\text{Var}(\hat{y}_{LR}) = (1-f)(1-r^2)S_y^2/n$, using standard notation.]

(4)

(ii) Discuss briefly the advantages and disadvantages of using a telephone directory as compared with random digit dialling as methods of data collection for this survey.

(5)

6. A regional council wishes to assess the amount of hazardous waste produced by the 6231 manufacturing companies in its area. They are split into three strata:

(1) basic metal industries;

(2) food, textiles and mineral products;

(3) other manufacturing.

A simple random sample of companies was taken in each stratum, and for each company the total quantity of hazardous waste (in thousands of tonnes) produced in 2003 was measured.

	Hazardous waste ('000 tonnes)			
Stratum	N_h	n_h	$\bar{y}_h$	s_h
1	92	11	166.6	207.7
2	1612	61	7.7	14.7
3	4527	292	0.3	4.5
Total	6231	364		

(i) Define the symbols N_h, n_h, $\bar{y}_h$, s_h as used above.

Show that $\bar{y}_{st} = \sum \frac{N_h}{N} \bar{y}_h$ is an unbiased estimator for the mean hazardous waste produced per company, and find the variance of $\bar{y}_{st}$. [Results from simple random sampling may be assumed without proof.]

(7)

(ii) Estimate the mean hazardous waste produced per company and obtain an estimate of the standard error of your estimator. Give an approximate 95% confidence interval for the mean hazardous waste per company.

(7)

(iii) Compute the sample sizes in the strata if proportional allocation had been used for this survey. Give brief reasons why a stratified sample using proportional allocation would be more efficient than a simple random sample of 364 units. Explain briefly whether the allocation actually used has been effective in improving precision compared with a proportional allocation.

(6)

7. An orange grower is to sell a truckload of oranges. The oranges are packed into 140 crates containing 120 oranges each. Before striking the deal, the buyer wants to estimate the quantity of juice in the oranges, and proposes to inspect a sample of oranges.

(a) *Convenience sampling* chooses the items which are most accessible while sampling is in progress. Suggest reasons why *cluster sampling* might be preferred to convenience sampling for this sampling inspection. How do *one*- and *two-stage* cluster sampling differ in the context of this example? Mention any practical difficulties that might arise in choosing genuinely random samples in this study.

(6)

(b) A simple random sample of 10 crates was selected from the truck, and then 5 oranges from each chosen crate were randomly sampled. The quantity of juice (y) obtained from squeezing each of the selected oranges was recorded, as shown below.

Crate (i)	*Juice (in ml)*					$\overline{y}_i$	s_i^2
1	90	103	76	84	89	88.4	97.3
2	107	80	72	110	70	87.8	372.2
3	104	93	83	76	91	89.4	112.3
4	101	81	77	99	109	93.4	188.8
5	97	85	110	101	80	94.6	147.3
6	84	99	106	92	78	91.8	126.2
7	101	100	91	113	108	102.6	70.3
8	109	78	89	91	90	91.4	124.3
9	114	96	108	80	103	100.2	171.2
10	94	90	109	102	84	95.8	97.2

(i) Treating the crates as clusters and the oranges as the individual sampling units, estimate the mean quantity of juice per orange. Explain whether this estimator is unbiased.

(4)

(ii) In a two-stage sampling procedure with equal-sized clusters, and with equal-sized simple random samples selected from the chosen clusters, an unbiased estimator of the variance of the sample mean can be written as

$$\frac{1-f_1}{n}s_b^2 + f_1\frac{1-f_2}{nm}s_w^2 .$$

Define n, m, f_1, f_2, s_b and s_w as used in this formula.

(3)

(iii) Use the formula in part (ii) to obtain an estimate of the variance of the sample mean quantity of juice per orange. Give an approximate 95% confidence interval for the true mean quantity of juice per orange.

(4)

(iv) The buyer decides to use a two-stage sample, and wishes to sample no more than 50 oranges in total. Advise the buyer how to choose the number of crates to sample, and how to choose the number of oranges to sample from each of those crates.

(3)

Turn over

8. A personnel manager wishes to compare the number of employees leaving one company (A) within the last three years with the number of employees leaving another company (B) in the same period. The data available are given below.

Company A

Age (years)	*Mean number employed*	*Number leaving*	*Duration of service (years)*	*Mean number employed*	*Number leaving*
16 – 24	115	23	0 – 4	485	102
25 – 34	585	117	5 – 9	280	45
35 – 44	234	16	10 – 14	153	9
45 – 54	187	3	15 +	367	5
55 +	164	2			
Total	1285	161	*Total*	1285	161

Company B

Age (years)	*Mean number employed*	*Number leaving*	*Duration of service (years)*	*Mean number employed*	*Number leaving*
16 – 24	18	2	0 – 4	118	20
25 – 34	35	6	5 – 9	47	6
35 – 44	75	11	10 – 14	29	2
45 – 54	70	7	15 +	58	1
55 +	54	3			
Total	252	29	*Total*	252	29

Three indices have been suggested as providing suitable methods for making the comparison:

(1) the crude "death" (i.e. leaving) rate;

(2) the age-adjusted "death" rate;

(3) the duration-of-service-adjusted "death" rate.

Calculate these rates for company A and for company B and comment on the information given by them.

(11)

Explain clearly the differences observed between the crude and the adjusted rates.

(4)

What further enquiries and calculations, if any, would you make before coming to any conclusions about the differences between the losses of employees from companies A and B?

(5)