EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2009

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* ${}^{n}C_{r}$.

This examination paper consists of 10 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.

© RSS 2009

1. An experiment to investigate the effect of winter feeding on milk production of dairy cows used a Latin square design. Four diets (*A, B, C, D*), in order of increasing starch equivalent, were each fed for 3-week periods to each cow, and the total milk yield in the third week of each period was recorded.

The yields *y* (in pounds of milk) are given in the table below. Unfortunately one record was lost, and *x* denotes this missing value.

		Cow								
		1		2		3		4		Total
Period	1	*A*	192	*B*	195	*C*	292	*D*	249	928
	2	*B*	190	*D*	203	*A*	*x*	*C*	210	603 + *x*
	3	*C*	214	*A*	139	*D*	245	*B*	163	761
	4	*D*	221	*C*	152	*B*	204	*A*	134	711
	Total		817		689		741 + *x*		756	3003 + *x*

(i) Explain briefly why a missing observation in a Latin square experiment introduces complications into the analysis of the data.

(3)

(ii) A general formula which can be used to estimate a single missing value *x* in treatment *i*, row *j* and column *k* of a Latin square design is

$$x = \frac{vT_i' + vR_j' + vC_k' - 2G'}{(v-1)(v-2)}.$$

Define v, T_i', R_j', C_k' and G'.

(1)

(iii) Apply this formula to find an estimate of the missing value *x* in this experiment. State what effect using this estimate will have on the rest of the analysis of these data. What assumptions about the lost observation must be made if this analysis is to be valid?

(5)

(iv) Given that the residual (error) mean square in an analysis using the estimate of *x* is 149.600, find approximate standard errors for each of the following contrasts:

$$\bar{y}_A - \tfrac{1}{3}(\bar{y}_B + \bar{y}_C + \bar{y}_D), \quad \bar{y}_B - \tfrac{1}{2}(\bar{y}_C + \bar{y}_D) \quad \text{and} \quad \bar{y}_C - \bar{y}_D.$$

Hence test whether each of these contrasts is significantly different from zero. Summarise the results found by this experiment, identifying the best diet(s) among the four.

(8)

(v) An alternative design that was considered for this experiment was to use a completely randomised design with 16 cows, four cows randomly assigned to each of the four diets, and one observation per cow. Discuss briefly why the Latin square was likely to have been the better design. What might be its disadvantages?

(3)

2. Explain what is meant by a *factorial* experiment. Describe the advantages of using a single factorial experiment to investigate two treatment factors simultaneously, rather than investigating each factor in a separate experiment.

(4)

A company manufacturing a foodstuff is interested in how the pH (acidity) of the final product depends on the temperature at which production takes place and on the strain of yeast used. Four strains of yeast, which is an important ingredient in the process, can be used. These are included in an experiment, together with three production temperatures. Each combination of strain and temperature is tested twice, and the 24 observations are obtained in random order. The following table gives the acidity level of the final product, in coded units.

		Strain of yeast				
		I	II	III	IV	Total
Temperature (°C)	12	113, 130	115, 100	52, 43	61, 74	688
	16	118, 105	106, 116	80, 70	101, 92	788
	20	105, 88	131, 121	93, 107	96, 108	849
	Total	659	689	445	532	2325

The sum of the squares of all 24 observations is 237 799. You may also use the fact that $(113 + 130)^2 + \ldots + (96 + 108)^2 = 473\,655$.

(i) Write down the linear model for a 3×4 factorial design, and the assumptions underlying its analysis of variance.

(3)

(ii) Analyse the data to estimate the effects attributable to yeast strain and temperature, and their interaction.

(4)

(iii) The company wishes to identify one or more of the yeasts as suitable for general production. Carry out any significance tests that you consider necessary, and construct any diagrams that will help in interpreting the results.

Use these tests and diagrams to write a brief report for the company. You should keep in mind that it will not be as easy to control the temperature in general production as under experimental conditions. Also, the target acidity level should not be too far from 100 units.

(9)

3. Explain what is meant by a *balanced incomplete block* design (BIBD). When is this design useful? Write down the conditions necessary for a BIBD to exist, defining all the symbols you use. Use these conditions to investigate whether a BIBD exists for 10 treatments in blocks of 4 units, using not more than 90 units in all.

(6)

A glasshouse experiment is to be conducted to compare the yields of 5 new varieties of a species of vegetable. The plants are to be grown in tubs, and only one size of tub may be used in the experiment.

Three possible balanced incomplete block designs have been suggested for comparing the 5 new varieties of vegetable:

A 10 tubs which can hold 2 plants each;

B 10 tubs which can hold 3 plants each;

C 5 tubs which can hold 4 plants each.

In earlier, similar experiments it was found that the within-tub variance (i.e. the error variance) for the 3-plant tubs is 20% greater than the within-tub variance for the 2-plant tubs. Similarly the within-tub variance for the 4-plant tubs is 40% greater than the within-tub variance for the 2-plant tubs.

(i) For each of these three designs, write down the variance of the difference between two means and the degrees of freedom for the residual error. Which design would you recommend for this experiment and why?

(9)

[**Note**. The variance of a difference between any two means is $\frac{2k\sigma^2}{v\lambda}$ in the usual notation.]

(ii) Write down a plan for your recommended design. Explain how you would randomise this design.

(5)

4. In an investigation into the volume of bread, the effects of flour protein and proof time (the length of time the dough is left to rise between forming and baking) were studied. A central composite design was used, consisting of 13 runs performed in a random order. The resulting volumes of loaves, y, in ml, are shown in the table below.

	Variables in coded units		*Loaf volume (ml)*
Run	x_1	x_2	y
1	0	0	1866
2	1	1	2048
3	–1	–1	1827
4	–1.414	0	1825
5	0	0	1970
6	0	0	2003
7	0	–1.414	1800
8	0	0	1960
9	0	0	1988
10	1	–1	1746
11	1.414	0	1939
12	0	1.414	1919
13	–1	1	1849

The coded values of x_1 = (–1, +1) correspond to actual values (20, 30) minutes for proof time respectively; and the coded values of x_2 = (–1, +1) correspond to (11%, 13%) for flour protein.

(i) Sketch the experimental design. Briefly discuss how this form of experiment is useful for determining operating conditions that maximise loaf volume. Comment on the choice of ± 1.414 as levels for x_1 and x_2.

(8)

(ii) Given the computer output below, write down the second-order model that has been fitted to the data. Use the fitted equation to predict loaf volume when the proof time is 27.5 minutes and the protein is 11.5%.

(3)

(iii) Copy and complete the analysis of variance table given below for these data. Does the second-order model provide an adequate fit?

(4)

(iv) Briefly suggest any further analyses of the data that you consider desirable, giving your reasons.

(3)

(v) Now suppose that a maximum of 8 runs could be carried out in one day, and that it is decided to run the experiment over two days, with days being used as blocks. Suggest an appropriate design and give reasons for your choice of blocking.

(2)

The computer output is on the next page

Computer output for question 4

Estimated regression coefficients

Term	*Coefficient*	*Standard error*
Constant	1957.40	20.577
Linear x_1	34.90	16.268
Linear x_2	61.54	16.268
Quadratic x_1	–38.52	17.448
Quadratic x_2	–49.77	17.448
Linear x_1 × Linear x_2	70.00	23.005

Analysis of variance for y

Source of variation	*DF*	*Sum of squares*	*MS*	*F ratio*
Linear terms		40 041		
Quadratic terms		24 483		
Interaction term		19 600		
Residual				
Lack of Fit				
Pure Error				
Total	12	98 943		

Turn over

5. Outline the main advantages and disadvantages of cluster sampling as compared to simple random sampling. Give an example of a survey in a country of your own choice that uses clustering in the sample design.

(6)

A human population is divided into 150 area segments, each containing 5 households. A simple random sample of 20 segments is selected from the population, all households in each segment are visited and the number of people in each of the sampled households is recorded. Let y_{ij} be the number of people in the jth household of the ith sampled segment. Then

$$\sum_{i=1}^{20}\sum_{j=1}^{5} y_{ij} = 482, \quad \sum_{i=1}^{20}\sum_{j=1}^{5} y_{ij}^2 = 2708, \quad \sum_{i=1}^{20}\left(\tfrac{1}{5}\sum_{j=1}^{5} y_{ij}\right)^2 = 470.32.$$

(i) Estimate the total number of people in the population and the variance of your estimate.

(5)

(ii) Suppose instead that these data had come from a simple random sample. What would now be the estimated variance of the estimator of the total number of people in the population?

(3)

(iii) Hence estimate the intra-cluster correlation coefficient.

(3)

(iv) Comment on your results.

(3)

6. A wildlife researcher wishes to estimate the otter population along the coast of Shetland, United Kingdom. As a direct count of otters would be time-consuming, the researcher decides to count the number of "active holts" being used by otters, as an index for the number of otters.

The coastline is divided into 242 sub-areas called sections, each 5 km^2, based on an Ordnance Survey map. The 237 sections that are not urban are assigned to one of four strata based on predominant terrain type. The survey is conducted by selecting a simple random sample of sections from each stratum and counting the number y of active holts in the sections.

Estimates of the means and standard deviations of the measurements, y, in each stratum based on a survey of 83 sections are as follows.

Stratum (h)	N_h	n_h	$\bar{y}_h$	s_h	$N_h s_h$
1 (Cliff)	89	19	1.79	2.30	204.70
2 (Agriculture)	61	21	1.57	2.60	158.60
3 (Peatland)	40	22	13.27	7.67	306.80
4 (Others)	47	21	4.10	3.95	185.65
Total	237	83			855.75

$\sum N_h s_h^2 = 3969.644$

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

(2)

(ii) Define the notation N_h, n_h, $\bar{y}_h$ and s_h.

Show that $\hat{Y} = \sum N_h \bar{y}_h$ is an unbiased estimator for the total number of holts in Shetland, and find the variance of $\hat{Y}$. [Results from simple random sampling may be assumed without proof.]

(8)

(iii) Estimate the total number of otter holts in Shetland and obtain an estimate of the standard error of your estimator. Give an approximate 95% confidence interval for the total number of holts.

(6)

(iv) Without computing the optimal allocation of the sample into 4 groups, give an intuitive reason why a sample of 83 sections could be allocated in a different way such that the variance of the estimator $\hat{Y}$ is smaller.

(4)

7. (i) Explain what is meant by a *quota sample*, a *systematic random sample* and a *two-stage cluster sample*. For each case, state with reasons whether or not it is an equal probability selection method.

(7)

(ii) In the context of sampling from a finite population, explain the term *finite population correction factor* and illustrate its use in estimating the population mean based on simple random sampling. Describe the conditions under which its importance is slight.

(5)

(iii) It is desired to estimate the amount overdue (per account) in the accounts of a certain company. A simple random sample of accounts will yield the current amount overdue y_i (in pounds) for the ith account in the sample. A check on records will give the overdue amount x_i for that account at the same time last year. Records will also give the average amount μ_x overdue for all accounts at this time last year.

Discuss briefly, giving reasons, how you would decide whether to use (a) a method based on a simple random sample mean of the y_i, (b) a ratio estimator, or (c) a regression estimator, to estimate the total amount currently overdue for all accounts. Give a formula for each estimator, defining any notation you use.

(8)

8. Explain the role of life tables in population studies. Distinguish between *current* and *cohort* life tables and state when each would be used. Explain briefly the relationship between life tables and age-specific death rates.

(6)

The data below show the population size and number of deaths of males in England and Wales in 2006. Calculate the columns ${}_nq_x$, l_x, ${}_nL_x$ and T_x of an abridged life table for males in England and Wales. Assume that $l_0 = 10\,000$. Indicate how you have calculated each column.

(6)

Age (years)	*Mid-year population ('000)*	*Registered deaths ('000)*
0	1595.4	2.15
5	3284.4	0.45
15	3636.1	2.05
25	3516.9	3.13
35	4079.6	6.32
45	3419.1	12.30
55	3118.0	27.60
65	2110.8	48.90
75	1267.0	81.90
85+	343.9	56.20
Total	26 371.2	241.00

[**Note**. Assuming that deaths occur uniformly throughout the age intervals, ${}_nq_x$ is given by

$$\frac{n \times {}_nM_x}{1+(0.5\,n \times {}_nM_x)}$$

where ${}_nM_x$ is the age-specific death rate for the age-interval x to $x + n$.]

A stationary population is supported by 10 000 male births per annum and experiences the mortality of people in England and Wales in 2006. Use your life table to estimate

(i) the age distribution for the above age intervals,

(ii) the expected ages at death of groups of males now aged 25, 45, 65 and 85,

(iii) the life-expectancy of males in this population.

(6)

State any assumptions required for the validity of the calculations in parts (i) to (iii) and comment on whether they are appropriate.

(2)