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

GRADUATE DIPLOMA, 2007

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 12 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. A conservationist wishes to assess the effectiveness of four management treatments for promoting meadow brome *Bromus commatatus* in a hay meadow nature reserve. The treatments are as follows.

A: Hay cut and harvested (removed from site)
B: Hay flail cut and left on ground in windrows
C: Hay scythed and left *in situ*
D: Hay not cut

The meadow slopes in one direction towards a river. On another side, at right angles, it is bounded by a motorway. The experiment was arranged as a Latin square design with the columns representing distance from the river and the rows representing distance from the motorway.

(i) Explain why this arrangement would be better than complete randomisation of treatments to plots. Describe in detail how you would choose at random a 4×4 Latin square design for the layout of this experiment. (4)

(ii) The treatment allocation and the data obtained during a fixed period after the intended harvest date are shown in the following table. The data were transformed into units such that high values show a more effective treatment.

		(Coded) distance from river				
		1	2	3	4	*Total*
(Coded) distance from motorway	1	D 34.45	A 64.16	B 53.13	C 49.02	200.76
	2	A 66.42	D 31.31	C 49.60	B 51.94	199.27
	3	C 52.54	B 54.94	A 62.73	D 25.84	196.05
	4	B 58.05	C 53.73	D 29.33	A 56.79	197.90
	Total	211.46	204.14	194.79	183.59	

$\sum y = 793.98$ $\sum y^2 = 41802.6036$

Totals for treatments are: A 250.10; B 218.06; C 204.89; D 120.93.

(a) Analyse the data to determine the significance of the effects due to distance from motorway, distance from river and hay management treatment, and briefly report your results. (5)

(b) Write down a set of orthogonal contrasts which assess the following comparisons between treatments.

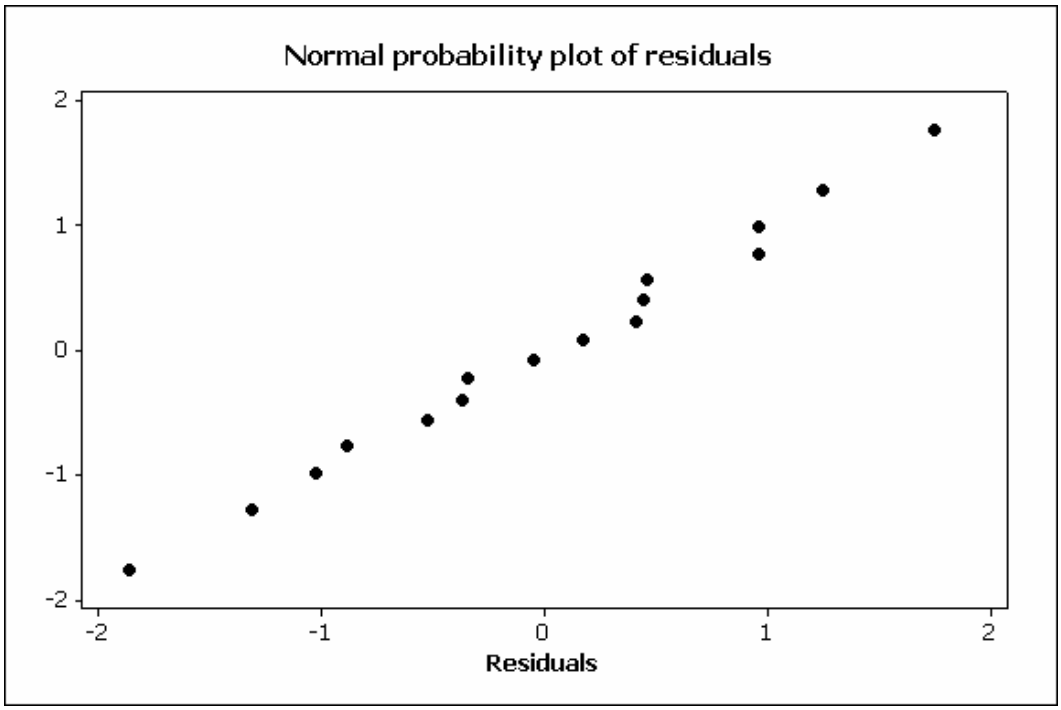
(1) Hay cut vs not cut.
(2) Hay removed from site vs hay not completely removed after cutting.
(3) Hay left *in situ* vs hay left in windrows.

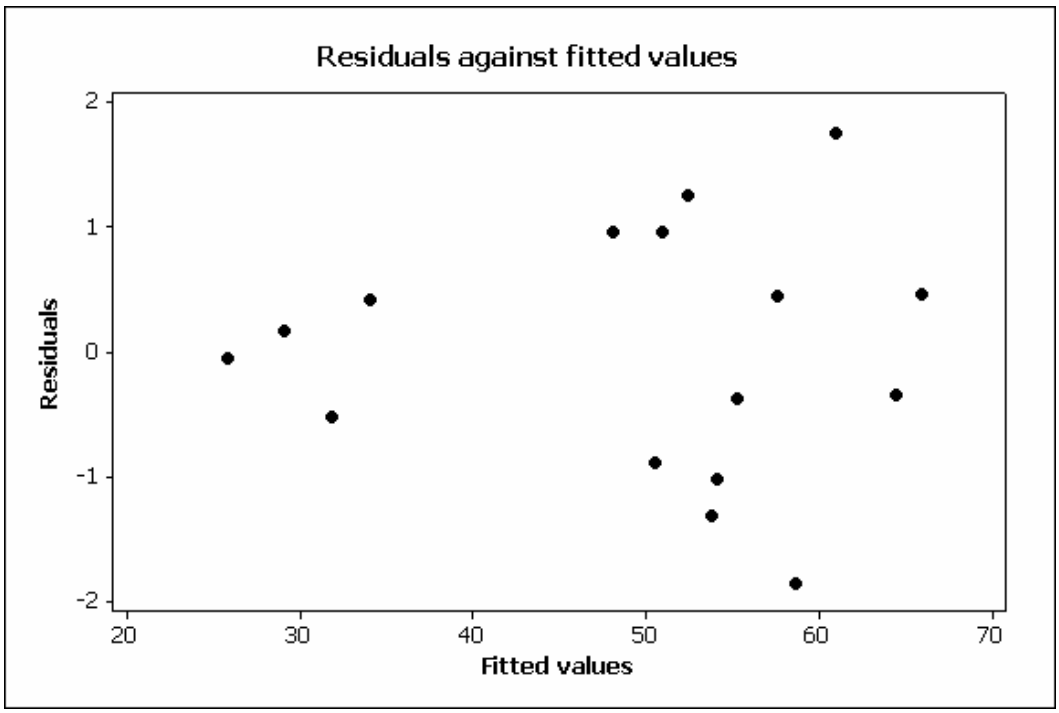
(2)

(c) Partition the treatment sum of squares into the single degree of freedom contrasts specified in (b). Summarise your conclusions. (6)

Question 1 is continued on the next page

(d) State the assumptions needed for the validity of the analysis of variance in (a). The graphs below show a Normal probability plot of the residuals and a plot of residuals against fitted values. Based on these graphs, explain whether you consider any of the assumptions for the analysis of variance to be invalid. (3)





Turn over

2. (a) Explain briefly what is meant by *blocking* in experimental design and why it is used. Illustrate your comments with a simple example.

(3)

Four different hardwood concentrations are to be studied to determine their effect on the strength of paper produced from them. Only one hardwood concentration can be used in each run of a pilot plant.

(i) Suppose that the pilot plant can carry out four runs on each day. Name and describe a suitable design with days as blocks.

(3)

(ii) Now suppose that the pilot plant can only carry out three runs per day. Explain what is meant by a balanced incomplete block design with days as blocks. Suggest one of these designs if three runs for each hardwood concentration are to be made. State the number of degrees of freedom there will be in the analysis, and comment on this.

(4)

(b) Under certain conditions, bacteria grown on cultural plates in an incubator produce circular colonies. Four different growth supplements (treatments *A*, *B*, *C*, *D*) were compared in a randomised block design, to study their effect on the diameter (mm) of these colonies. Six blocks (replicates) were used. The data on diameters were transformed by taking logarithms, and the transformed data were used in the analysis of variance.

(i) Outline conditions under which a logarithmic transformation of the raw data might be appropriate in the context of this type of experiment.

(2)

(ii) The treatment means of the log-transformed data (logs to base *e*) were as follows.

A	*B*	*C*	*D*
3.34	2.72	2.45	2.88

The residual sum of squares from the analysis of variance was 0.496.

What conclusions can be drawn about the performances of the supplements?

(5)

(iii) Construct a 95% confidence interval for the difference in mean log-transformed colony diameters between growth supplements *A* and *B*. How would you interpret this interval in terms of colony diameter?

(3)

3. Explain what is meant by the main effect of a factor and the interaction between two factors in a 2^2 factorial experiment. Illustrate your explanation with a suitable diagram.

(4)

A study of people with psychiatric illness assessed the effect of two different drugs (A and B) and psychotherapy (C) on health-related quality of life. These three factors could each be given, or not given, to any individual patient. There were therefore 8 possible factor combinations which formed the set of treatments for the study: for example *abc* indicates that all factors A, B, C were given; *ab* that A and B were given but C was not; *a* that only A was given; and (1) that none of the factors was given to a patient. Two hospitals participated in the study; each enrolled 8 individuals who were randomised to one of the 8 treatments.

The data below are the total scores y on a questionnaire, obtained at 12 weeks after starting study treatment. High values indicate better quality of life.

Treatment	*Hospital 1*	*Hospital 2*	*Total*
(1)	2	3	5
a	6	14	20
b	10	15	25
ab	4	6	10
c	6	9	15
ac	15	25	40
bc	18	22	40
abc	8	12	20
Total	69	106	

$\sum y = 175$ $\sum y^2 = 2605$

(i) Copy and complete the analysis of variance table **shown on the next page**.

(4)

(ii) Draw a suitable diagram showing the relationship between the three factors A, B and C. Comment on any apparent main effects of factors or interactions between them.

(4)

(iii) Carry out any significance tests that you consider necessary, and state your conclusions.

(5)

(iv) Discuss briefly any concerns you have about the design of this study.

(3)

The analysis of variance table is on the next page

Analysis of variance table for question 3

Source of variation	DF	Sum of squares
Hospitals	1	*****
A	1	*****
B	**	14.0625
C	**	189.0625
AB	**	351.5625
AC	**	1.5625
BC	**	1.5625
ABC	**	*****
Treatments	**	573.4375
Residual	**	*****
TOTAL	15	690.9375

Turn over

4. The temperature, X_1, and the length of reaction time, X_2, of a chemical reaction are known to affect the reaction rate and hence the percent yield Y from the reaction. An experimenter is interested in finding the combination of levels of these factors which will produce the highest yield. He intends to begin his investigation, using response surface methods, by fitting a first-order model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where ε is a Normally distributed random variable with mean 0 and variance σ^2.

From existing knowledge, he decides on experimental regions for X_1 and X_2, chooses two values for each of X_1 and X_2 which he codes ±1, and plans to base his experimental design on these. In discussion with colleagues, he receives three suggestions for possible designs. In these designs, values (x_1, x_2) of X_1 and X_2 are as follows.

A: a 2^2 factorial design, using (1, 1), (1, –1), (–1, 1) and (–1, –1), together with five observations taken at (0, 0).

B: the design $(-\sqrt{3/2}, -\sqrt{1/2})$, (0, √2), $(\sqrt{3/2}, -\sqrt{1/2})$, with six observations also taken at (0,0).

C: the design (1,1), (–√2, 0), (0, √2), with six observations also taken at (0, 0).

(i) Explain clearly the reason for taking more than one observation at (0, 0).

(2)

(ii) (a) For design A, show that when the model is written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ the design matrix $\mathbf{X}$ is the transpose of

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Using the result $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X'X}\right)^{-1}\sigma^2$, find the variance $\mathrm{Var}\left(\hat{Y}\right)$ of the estimated response at a point (x_1, x_2).

(6)

(b) For design B, the matrix $\mathbf{X'X}$ is $\begin{pmatrix} 9 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}$. Find $\mathrm{Var}\left(\hat{Y}\right)$ for this design.

(2)

(c) For design C, $\mathbf{X'X}$ is not a diagonal matrix. Explain briefly (no extensive calculations are required) the effect this has on the formula for $\mathrm{Var}\left(\hat{Y}\right)$.

(2)

Question 4 is continued on the next page

(iii) Draw a graph to show the points (x_1, x_2) used in each design A, B, C. Which of these designs are rotatable?

(3)

(iv) The experimenter decides to use the rotatable design with the smallest value of $\mathrm{Var}\left(\hat{Y}\right)$. Indicate briefly how he should proceed if the first-order model proves inadequate.

(5)

5. (i) Define *stratified random sampling*. Explain what is meant by *stratification with proportional allocation* and *stratification with optimal allocation.*

Why should stratified sampling be expected to be an improvement on simple random sampling? When would this improvement be greatest? Illustrate your answer with reference to a practical survey with which you are familiar.

(5)

(ii) A population consists of 20 000 large companies and 80 000 small companies. A pilot survey suggests that approximately 90% of large companies recognise trade unions and approximately 50% of small companies do so. It is intended to take a stratified random sample of size 1000 in order to estimate the overall proportion of companies which recognise trade unions. It may be assumed that the cost of sampling a unit (company) is the same in the two strata.

(a) Write down the formula for the variance of the stratified sampling estimator of a proportion, ignoring the finite population correction.

(1)

(b) If the sample is stratified by size of company, calculate the required number of companies in the sample for each of the two strata, using

(1) proportional allocation,

(2) optimal allocation.

(4)

(c) Suppose now that the percentage estimates from the preliminary survey are close to the true values P_h. Calculate and compare the relative efficiencies of these two methods of allocation with that of a simple random sample of 1000 companies. You may ignore the finite population correction.

(6)

(d) What does your analysis suggest about the usefulness of stratified sampling?

(4)

6. (i) A survey is being planned to study agricultural land use in a large region in a developing country. Explain how and why stratification and clustering might be useful in such a survey, and what practical problems they could help to overcome.

(5)

(ii) A simple random sample of 2055 farms was selected from the 75 308 farms in a region, and the number of cattle (y) and the total area (x) were recorded for each farm. The results were as follows.

Sample total number of cattle, Σy_i	25 751
Sample total area (hectares), Σx_i	62 989

The sums of the squares are $\sum y_i^2 = 596\,737$ and $\sum x_i^2 = 2\,937\,851$, and the sum of the products is $\sum x_i y_i = 1\,146\,391$. The total area under cattle in this region is 2 353 365 hectares.

(a) Estimate the total number of cattle in the region, and the standard error of your estimate, using

(1) the simple random sample mean,

(2) the ratio estimator.

(9)

[Note. An estimate of the variance of a ratio estimator of the total is given by

$$\frac{N(N-n)}{n(n-1)}\left\{\sum y_i^2 - 2r\sum x_i y_i + r^2 \sum x_i^2\right\},$$

where $r = \dfrac{\Sigma y_i}{\Sigma x_i}$ and the other symbols have their usual meanings.]

(b) You may assume that the regression estimate of the total number of cattle is 959 651.6 and the estimated standard error of this estimate is 13 881.9. What conclusions can you draw about the relative merits of the three estimators?

(3)

(c) Discuss briefly to what extent these are general results, which would happen in any survey, and to what extent they depend on the particular data obtained.

(3)

7. A simple random sample of 10 hospitals was selected from a population of 33 hospitals that had received state funding to upgrade their emergency medical services. Within each of the selected hospitals, the records of all patients hospitalised in the past 12 months for traumatic injuries (i.e. accidents, poisonings, violence, burns, etc) were examined. The numbers of patients hospitalised for trauma conditions and the numbers discharged dead for the selected hospitals are given below.

Hospital	*Number of patients hospitalised for trauma conditions*	*Number with trauma conditions discharged dead*
1	560	4
2	190	4
3	260	2
4	370	4
5	190	4
6	130	0
7	170	9
8	170	2
9	60	0
10	110	1

(i) Explain why this design may be considered as a cluster sample. What are the first-stage and second-stage units?

(2)

(ii) (a) Obtain a point estimate and an approximate 95% confidence interval for the total number of persons hospitalised for trauma conditions for the 33 hospitals. State the properties of your estimator.

(b) Obtain a point estimate of the proportion of persons discharged dead among those hospitalised for trauma conditions for the 33 hospitals, using the cluster totals. Hence calculate an approximate 95% confidence interval for this proportion, and comment on the validity of the assumptions necessary for this calculation.

(12)

(iii) Give reasons why, for this survey, cluster sampling might be preferred to stratified random sampling. What might be the drawbacks of cluster sampling?

Discuss, with reasons, any improvements you might make if another survey was being planned on the same topic.

(6)

8. Explain clearly the purpose of standardisation with respect to death rates, and distinguish between *direct* and *indirect* standardisation. (5)

Death rates from coronary heart disease (CHD) in the United Kingdom are among the highest in the world. The data below summarise the age distributions, and the associated prevalences of CHD mortality, for males and females in the United Kingdom in 2004. Also given is the age distribution for Scotland, a subpopulation of the United Kingdom. In 2004, CHD accounted for 10 778 deaths in Scotland (5814 males and 4964 females).

(i) Calculate the crude rates for the prevalence of CHD mortality separately for males and females in the United Kingdom in 2004. (2)

(ii) Calculate the direct adjusted rate for the prevalence of CHD mortality among males in the United Kingdom in 2004 using the *European Standard Population* as the standard population. (5)

> [Note. The hypothetical *European Standard Population* shown in the table is the same for both sexes, and is used for comparisons between countries and over time.]

(iii) Calculate the standardised mortality ratio for males in Scotland in 2004, using the male age distribution for the United Kingdom in 2004 as the standard population. (5)

(iv) You may assume that the direct adjusted rate for prevalence of CHD mortality among females in the United Kingdom is 80.13 deaths per 100 000, and the standardised mortality ratio for the prevalence of CHD mortality for females in Scotland is 1.23.

Comment on the conclusions that may be drawn from comparisons among all these rates.

(3)

	Population and CHD deaths, United Kingdom, 2004				**European Standard Population**	**Population, Scotland, 2004**	
	Males		**Females**				
Age (years)	*Population ('000)*	*CHD deaths*	*Population ('000)*	*CHD deaths*		*Males ('000)*	*Females ('000)*
< 35	13457	136	13067	36	50 000	1088	1072
35 – 44	4552	850	4639	196	14 000	384	412
45 – 54	3780	3041	3859	610	14 000	339	353
55 – 64	3390	7369	3509	2006	11 000	293	310
65 – 74	2374	14149	2659	6634	7 000	207	248
75 +	1717	33010	2830	37805	4 000	134	237
Total	29270	58555	30563	47287	100 000	2445	2632

Sources: *Mid-2004 population estimates, United Kingdom* — *Office for National Statistics*
Mid-2004 population estimates, Scotland — *Office for National Statistics*
Deaths by cause, sex and age, 2004, United Kingdom — *British Heart Foundation*