EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2002

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* ${}^{n}C_{r}$ *and that* ln *stands for* $\log_e$.

This examination paper consists of 9 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. Three diets D_1, D_2 and D_3 are to be studied to assess their effect on the growth of mice. Six litters, each containing three mice, are available.

The experimenter has a tidy mind and so suggests the following experimental layout.

Litter					
1	2	3	4	5	6
D_1	D_2	D_3	D_1	D_2	D_3
D_1	D_2	D_3	D_1	D_2	D_3
D_1	D_2	D_3	D_1	D_2	D_3

He shows this to you and asks for your advice.

(i) Explain to the experimenter what is wrong with his layout and why a randomised (complete) block design would be better. Include in your advice explanations of what a randomised block design is, and why blocking and randomisation are useful.

(5)

(ii) Using a table of random digits, construct a randomised block design for allocating the mice to the diets. Explain your construction carefully.

(3)

In the event, a randomised complete block design was planned, but two litters contained only two mice that could be used, and for those litters one diet, selected at random, was omitted. The data given below are the increases in weight, in grams, of the mice over the experimental period; m and n denote the missing items of data.

LITTER:	1	2	3	4	5	6	Totals
DIET:							
D_1	152	93	110	m	143	104	$602+m$
D_2	106	n	72	92	127	93	$490+n$
D_3	109	55	58	59	112	81	474
Totals	367	$148+n$	240	$151+m$	382	278	$1566+m+n$

(iii) Based on the model for analysing a randomised block design, obtain formulae for the residuals for the observations in the positions m and n. Hence show that the estimates of m and n which make the corresponding residuals zero are 107.76 and 68.42 respectively.

(5)

(iv) Given that the residual (error) mean square in an analysis using the estimates of m and n is 92.846, find approximate standard errors for the differences between pairs of diet means. Hence construct 95% confidence intervals for the difference in mean increases in weight between diets D_1 and D_2, and between diets D_2 and D_3. Comment on whether the diets appear to result in significantly different increases.

(7)

2. (a) The yields of two aubergine varieties, V_1 and V_2, were examined in two soil types, S_1 and S_2, and at two levels of moisture, M (low, M_1, and high, M_2). Four complete replicates of a 2^3 factorial design were run in 4 greenhouses. The yields y (coded in suitable units) and treatment combinations are given below.

In the coding of the treatment combinations,
presence of v indicates V_2 was used, otherwise V_1 was used;
presence of s indicates S_2 was used, otherwise S_1 was used;
presence of m indicates M_2 was used, otherwise M_1 was used.

	Greenhouse				
	I	**II**	**III**	**IV**	Total
(1)	7	19	13	11	50
v	30	33	28	31	122
s	24	30	19	25	98
vs	39	36	35	43	153
m	21	30	24	21	96
vm	31	36	31	33	131
sm	27	31	26	29	113
vsm	39	41	36	66	182
Total	218	256	212	259	

Effect estimates	
V	14.44
S	9.19
M	6.19
VS	1.06
VM	–1.44
SM	–0.69
VSM	3.19

$\Sigma y = 945 \qquad \Sigma y^2 = 31513$

(i) Explain briefly why the above is called a factorial layout. What are the advantages of this type of design?

(3)

(ii) Show how the value 3.19 for the *VSM* interaction estimate was obtained.

(2)

(iii) Construct the analysis of variance. Carry out any further significance tests that you consider appropriate, and report on the results.

(8)

(b) Suppose now that only four aubergine plants can be grown in each greenhouse, and that the greenhouse used may influence the yield. However, 8 greenhouses of this size are available.

(i) Explain briefly the importance of *confounding* in 2^k factorial experiments.

(3)

(ii) Write down an appropriate design for an experiment using the same treatment combinations as in part (a), in which each of these 8 greenhouses is treated as a block.

(4)

3. Ten examiners are to be compared in an inter-examiner reliability study of the use of a rating scale for depression. Up to 45 patients are available. Patients cannot tolerate more than five examinations. There is time for up to 90 assessments in the whole study.

Three possible balanced incomplete block designs have been suggested for comparing the ratings of the ten examiners:

A 18 patients each assessed by five examiners;

B 30 patients each assessed by three examiners;

C 45 patients each assessed by two examiners.

Earlier, similar inter-examiner rating studies in depression indicate that the within-patient variance (i.e. the residual variance) for patients undergoing three examinations is 20% greater than the within-patient variance for patients undergoing two examinations. Similarly, the within-patient variance for patients undergoing five examinations is 40% greater than the within-patient variance for patients undergoing two examinations.

(i) Explain what is meant by a *balanced incomplete block* design with patients as blocks and examiners as treatments.

(2)

(ii) Write down the conditions necessary for a balanced incomplete block design to exist, defining all the symbols you use.

(3)

(iii) Verify that balanced incomplete block designs may exist for comparing 10 treatments in blocks of 2, 3 or 5 units using a total of 45, 30 or 18 blocks respectively. Does a balanced incomplete block design exist for 10 treatments in blocks of 4 units, using not more than 90 units in all?

(5)

(iv) For each scheme *A*, *B*, *C*, write down the variance of the difference between two treatment means and the degrees of freedom for the residual error. Which of these schemes would you recommend, and why?

(10)

4. A manufacturer of cutting tools is conducting a series of experiments to determine the manufacturing time and hardness of steel to be used to maximise blade life.

The first of these experiments produces the following results.

Steel hardness, x_1	Manufacturing time (minutes), x_2	Blade life (hours), y
75	10	3.5
75	10	4.0
75	20	7.0
75	20	6.5
85	10	12.0
85	10	11.5
85	20	16.5
85	20	16.0

For analysis, the values of $x_1 = 75, 85$ are coded $-1, +1$ respectively; and the values of $x_2 = 10, 20$ are also coded $-1, +1$ respectively.

(i) Name and describe the type of design used here, and discuss its use in the method of steepest ascent for locating optimum operating conditions.

(4)

(ii) Produce an interaction diagram to show the effect of manufacturing time and steel hardness on blade life. Estimate the two main effects and the interaction between x_1 and x_2.

(6)

(iii) What settings would you recommend for the next experiment in this series given that a convenient step size for steel hardness is 2.5 units? Explain why you recommend these settings.

(5)

(iv) Suppose now that the unit blade cost z (dollars) in terms of steel hardness (coded x_1) and manufacturing time (coded x_2) is given by the first-order model

$$z = 23 + 3x_1 + 4x_2 \qquad \text{where } -1.5 \leq x_1, x_2 \leq 1.5 .$$

Unit blade cost must be kept below 27.5 dollars and blade life must exceed 12 hours for the product to be competitive. Using the first-order model for blade life, assess whether there is a feasible set of operating conditions for this process. At what settings would you recommend that the process be run?

(5)

5. (i) Explain what is meant by a *quota sample*, a (linear) *systematic random sample* and a *two-stage cluster sample*. Are these equal probability selection methods? Why or why not?

(5)

(ii) Discuss any practical difficulties that might arise in planning a survey to study family income in a mixed urban and rural population. Explain how and why stratification and clustering might be used in such a survey.

(5)

(iii) What is meant by *post-stratification* (i.e. post-hoc stratification)? Explain the main distinction between this and ordinary stratified sampling. What are the main consequences of using post-stratification?

(5)

(iv) In a district containing 4000 houses, the percentage of homes owned by the occupier, thought to be between 45% and 65%, is to be estimated with a standard error of not more than 2%. From the same survey, the percentage of households running two (or more) cars, thought to lie between 5% and 10%, is to be estimated with a standard error of not more than 1%. How large a sample is necessary to satisfy both aims?

(5)

6. A region has 3510 farms which cluster naturally into 90 different "villages". In any village, x denotes the total number of farms and y the total number of cattle. A simple random sample of 15 villages (clusters) was selected, giving the data shown in the table.

Village (cluster) i	Number of farms x_i	Number of cattle y_i	Mean number of cattle per farm $\bar{z}_i = y_i / x_i$
1	35	418	11.94
2	25	402	16.08
3	48	360	7.50
4	30	394	13.13
5	70	515	7.36
6	55	910	16.55
7	66	600	9.09
8	18	316	17.56
9	30	288	9.60
10	32	350	10.94
11	64	784	12.25
12	24	290	12.08
13	48	795	16.56
14	40	478	11.95
15	82	906	11.05
Total	667	7806	

$\Sigma x_i y_i = 393716; \quad \Sigma x_i^2 = 34883; \quad \Sigma y_i^2 = 4759890; \quad \Sigma \bar{z}_i = 183.64; \quad \Sigma \bar{z}_i^2 = 2395.8018$

(i) Estimate the mean number of cattle per farm in the region as a whole in three different ways:

(a) using $\bar{z}$, the simple mean of the cluster means;

(b) using the cluster sample ratio estimate of y to x;

(c) using the cluster sample total.

(8)

(ii) Estimate the variance of each of the three estimators. Comment on the properties of the estimators.

(12)

Turn over

7. (a) A large student population consists of 10000 males and 5000 females aged 15 to 20. A sample survey is to be carried out to investigate the prevalence of smoking in the student population, using a sample size of 75. Two sample designs are proposed:

A a simple random sample of 75 students is drawn from the total population, and the overall proportion of students who smoke is estimated from it;

B the population is stratified by sex and a stratified random sample of 50 male and 25 female students is drawn. The proportions of male and of female students who smoke are estimated separately, and the overall proportion who smoke is estimated by combining these.

(i) Suppose that design *B* is adopted. Defining suitable notation, write down an expression for an appropriate estimator of the proportion of the population who smoke.

(2)

(ii) Evaluate the variances of the estimators for designs *A* and *B* when the true proportions p_m and p_f of males and females in the population who smoke are

(a) $p_m = p_f = 0.5$,

(b) $p_m = 0.5$ and $p_f = 0.1$.

(4)

(iii) What conclusions do you draw about which sample design is preferable in each case, and why? Explain the purpose of stratification in design *B*.

(4)

(b) Design a suitable set of questions to study the smoking habits of students.

(10)

8. (i) Explain why simple *crude* rates may not be appropriate for comparing mortality across different subgroups. Distinguish between *direct* and *indirect* methods of standardisation, stating any underlying assumptions.

(7)

(ii) The data below summarise the maternal age distribution for all infants born in the state of Michigan from 1950 to 1964, and the prevalence of Down's Syndrome among live-born infants. The figures are broken down according to the age of the mother at the birth of the child; data specific to first-born infants are picked out, as are those for any children who were born fifth or later in the family.

	Entire State		**Birth Order**			
			First-born		**Fifth or later**	
Maternal age (years)	*Live births*	*With Down's Syndrome*	*Live Births*	*With Down's Syndrome*	*Live Births*	*With Down's Syndrome*
< 20	319933	136	230061	107	327	0
20 – 24	931318	396	329449	141	30666	8
25 – 29	786511	411	114920	60	123419	63
30 – 34	488235	428	39487	40	149919	112
35 – 39	237863	628	14208	39	104088	262
≥ 40	61313	530	3052	25	34392	295
Total	2825173	2529	731177	412	442811	740

(a) Calculate the crude rates for the prevalence of Down's Syndrome among infants born first, and born fifth or later in birth order.

(2)

Using the entire state of Michigan as the standard:

(b) Calculate the *indirect adjusted rate* for the prevalence of Down's Syndrome among infants born first in birth order.

(4)

(c) Calculate the *direct adjusted rate* for the prevalence of Down's Syndrome among infants born first in birth order.

(4)

(d) You may assume that the indirect and direct adjusted rates for those born fifth or later in birth order are respectively 84.8 and 75.5 infants per 100 000.

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

(3)