EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2004

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.*

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. An experiment on the growth of tomato plants is carried out in a glasshouse. The plants are grown in commercial "grow-bags" instead of in pots. Each grow-bag contains a standard growing medium and will hold four plants, all of which must receive the same treatment.

There is sufficient space in the glasshouse to place on benches two rows of grow-bags, side by side with 16 in each row. The experiment includes three factors *A*, *B*, *C* that are different nutrients, with each factor used at two levels (low and high). The glasshouse length runs north to south and there are doors at each end.

(i) At the planning stage it was decided to include blocking in the design, with the northernmost eight bags forming block I and the remaining three blocks constructed in a similar way. With the aid of a sketch, explain why this decision would be better than complete randomisation. Show in your sketch how the treatments could be allocated in a typical block, using the coding in the following table of data.

(5)

(ii) The data below are the total yields (kg) of the four plants in each unit plot, obtained during a fixed period in the growing season. The treatment combinations are coded in the usual way; for example the letter *a* is included in the combination when factor *A* is present at its high level, and not otherwise.

		Treatment combination								*Block total*
		(1)	*a*	*b*	*ab*	*c*	*ac*	*bc*	*abc*	
Block	I	4.4	10.7	3.9	13.4	5.0	10.3	6.2	14.9	68.8
	II	7.6	10.8	5.0	19.0	6.4	13.4	4.0	15.6	81.8
	III	7.1	11.6	5.6	18.0	7.0	14.0	3.6	16.4	83.3
	IV	5.2	9.3	4.6	12.9	5.6	11.0	5.7	15.8	70.1
Treatment total		24.3	42.4	19.1	63.3	24.0	48.7	19.5	62.7	

The sum of squares of all 32 observations is 3563.28

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

(6)

(b) Carry out such tests of significance as you consider necessary, and construct any diagrams that will help in interpreting the results.

(6)

(c) Write a brief report for the scientist who commissioned the experiment.

(3)

The analysis of variance table for part (ii)(a) is shown on the next page

Turn over

Analysis of variance table for question 1 part (ii)(a)

Analysis of Variance for total yield of 4 tomato plants per unit plot

Source of variation	DF	Sum of squares
Blocks	3	*******
A	1	*******
B	**	19.84500
C	**	1.05125
AB	**	62.16125
AC	**	0.98000
BC	**	1.20125
ABC	**	*******
Treatments	**	616.795
Residual	**	*******
Total	31	675.280

Turn over

2. (i) Write down a balanced incomplete block design to compare five treatments in 10 blocks, using a total of 30 experimental units. Explain how you would randomise this design.

(5)

(ii) Five different treatments for refining raw rubber were compared to determine their effect on the elongation of rubber under a stress test. The raw rubber was received in batches, each of which contained sufficient rubber to test two treatments. Ten batches were selected at random. The elongation of rubber under each test was recorded, as shown below.

	Treatment					
Batch	*A*	*B*	*C*	*D*	*E*	*Total*
1	35	16				51
2	20		10			30
3	13			26		39
4	25				21	46
5		16	5			21
6		21		31		52
7		27			16	43
8			18	37		55
9			12		20	32
10				31	17	48
Total	93	80	45	125	74	417

In order to construct an analysis of variance table for the data, a computer package was used to calculate the following residual (error) sums of squares.

After fitting the grand mean only (i.e. the total corrected sum of squares)	1352.55
After fitting batches only	804.50
After fitting treatments only	503.25
After fitting batches and treatments	171.50

(a) Calculate the adjusted treatment means.

(5)

(b) Construct the analysis of variance.

(5)

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

(5)

3. The germination rate of a particular species of plant is studied under four different growing conditions. The four conditions are the factorial arrangements of two storage methods, stratified and unstratified, and two storage temperatures, called "spring" and "summer".

Forty batches of seeds were prepared, each containing 20 seeds, and 10 batches were allocated at random to each of the four treatments. First the seeds were stored for two weeks according to their pre-assigned storage method. Then the seeds were placed on dishes with 5 ml of water, and stored for two weeks according to their pre-assigned storage temperature.

The table below shows the numbers of germinated seeds (out of 20) for each batch of seeds, and also (in brackets) the residuals obtained from a one-way analysis of variance.

Spring/ Stratified	12	13	2	7	19	0	0	3	17	11
	(3.6)	(4.6)	(–6.4)	(–1.4)	(10.6)	(–8.4)	(–8.4)	(–5.4)	(8.6)	(2.6)
Spring/ Unstratified	6	2	0	2	4	1	0	10	0	0
	(3.5)	(–0.5)	(–2.5)	(–0.5)	(1.5)	(–1.5)	(–2.5)	(7.5)	(–2.5)	(–2.5)
Summer/ Stratified	6	4	5	7	6	5	7	5	2	3
	(1.0)	(–1.0)	(0.0)	(2.0)	(1.0)	(0.0)	(2.0)	(0.0)	(–3.0)	(–2.0)
Summer/ Unstratified	0	6	2	5	1	5	2	3	6	6
	(–3.6)	(2.4)	(–1.6)	(1.4)	(–2.6)	(1.4)	(–1.6)	(–0.6)	(2.4)	(2.4)

(i) A one-way analysis of variance was conducted, to compare the four treatments. Explain in detail how the residuals shown above were calculated. Explain how you would analyse the residuals to examine the usual assumptions concerning Normality and constant variance.

(14)

(ii) Write down all the assumptions required for this analysis of variance. Discuss whether you would expect these data to satisfy all the assumptions, and explain any concerns you have about the validity of the analysis. State the steps which could be taken to overcome any concerns you have.

(6)

4. In a pilot plant, experimenters were interested in improving percent conversion of a chemical process. An initial screening experiment showed that temperature and concentration were the two most important variables, and a further experiment was planned to relate the yield to these factors. The experimenters thought the process was operating near the optimum values, and so they decided to use a design for the second experiment which would allow them to fit a second-order model relating yield to temperature and concentration.

It was proposed to use a central composite design with three components, I, II and III, as follows:

I a 2^2 factorial design using values $x_1 = \pm 1$ and $x_2 = \pm 1$;

II four centre points, i.e. points at $\mathbf{x} = (0, 0)$;

III four points on the axes, i.e. at $\mathbf{x} = (\pm\alpha, 0)$ and $\mathbf{x} = (0, \pm\alpha)$.

(i) Sketch the experimental design. Discuss briefly the purposes of the different components of the design in the context of fitting a second-order model.

(5)

(ii) Define the terms *rotatability* and *orthogonality*, discussing the advantages and disadvantages of each design property. Explain how you would choose α to make the above design rotatable. Comment on whether the above design is orthogonal.

(5)

(iii) The data **shown on the next page** are from the above rotatable experiment consisting of 12 runs performed in a random order. The response represents the percent conversion of the chemical process. In order to construct an analysis of variance table for the data, a computer package was used to calculate the following residual (error) sums of squares.

After fitting the intercept only (i.e. the total corrected sum of squares)	1768.917
After fitting linear terms only	854.510
After fitting linear and interaction terms	614.260
After fitting linear, quadratic and interaction terms	35.344

(a) Construct the analysis of variance for fitting the second-order model. You should include the sums of squares, mean squares and F statistics for linear terms, for quadratic terms and for the interaction between temperature and concentration.

(4)

(b) Partition the residual sum of squares into pure error and lack of fit components, and test the lack of fit for the second-order model. Does the model provide an adequate fit to the data?

(3)

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

(3)

The data for part (iii) are shown on the next page

Data for question 4 part (iii)

Variables in coded units		**Percent conversion**
x_1	x_2	y
–1	–1	43
1	–1	78
–1	1	69
1	1	73
$-\alpha$	0	48
α	0	76
0	$-\alpha$	65
0	α	74
0	0	76
0	0	79
0	0	83
0	0	81

5. Wildland managers want to estimate the total number of caribou in the Nelchina herd located in south central Alaska. The density of caribou differs dramatically in different types of habitat. A preliminary aerial survey has identified the area used by the herd, and divided it into six strata based on habitat type.

The organiser has decided to divide the area into sub-areas called quadrats, each 4 km^2. The main survey will be conducted by selecting a simple random sample of quadrats from each stratum; the number of caribou, y, in the quadrats will be counted from an aerial photograph.

Estimates of the means and standard deviations of the measurements, y, in each stratum based on the preliminary survey of 211 quadrats are as follows.

Stratum (h)	N_h	n_h	$\bar{y}_h$	s_h	$N_h s_h$
1	400	98	24.1	74.7	29880
2	40	10	25.6	63.7	2548
3	100	37	267.6	589.5	58950
4	40	6	179.0	151.0	6040
5	70	39	293.7	351.5	24605
6	120	21	33.2	99.0	11880
Total	770	211			133903

$\sum N_h s_h^2 = 47882186$

(i) Discuss briefly the merits of using stratified sampling for this survey.

(4)

(ii) Based on the results of the preliminary aerial survey, estimate the total number of caribou in the herd and obtain an estimate of the standard error for your estimator.

(5)

(iii) For the main survey, the managers wish to estimate the total number of caribou to within d animals with 95% probability (i.e. the width of the interval is $2d$). You may assume that the formula for the total sample size n is

$$n = \frac{\sum_h N_h^2 S_h^2 / w_h}{V + \sum_h N_h S_h^2} .$$

Define N_h, S_h, w_h and V as used in this formula.

(2)

(iv) Define *optimal* allocation. Discuss briefly why you would choose an optimal allocation rather than proportional allocation for this survey. You may assume that the cost of sampling any unit is constant.

(3)

(v) Use optimal allocation to calculate the total sample size and the allocations n_h needed to estimate the total population of caribou to within 8000 animals with 95% probability. Calculate the standard error for your estimator of the population total.

(6)

Turn over

6. A National Shoppers' Survey is to be conducted in the UK on behalf of manufacturers. A total of 5000 households will be selected to participate in the survey. Each selected household will be sent a printed questionnaire and asked to complete questions on a range of topics such as shopping, hobbies and activities, computer use and motoring.

(a) In order to select the sample of households, a sampling plan based on the UK postcode scheme is proposed. The postal areas in the UK will first be stratified into 50 geographical regions. In each of these strata, the postcode sectors (combinations of letters and numbers) will be listed, together with the population density (people per hectare) in each. Postcode sectors will then be selected for the sample with probability proportional to size. Within each selected postcode sector, a simple random sample of households will be selected.

(i) Explain why this sample of households is a *cluster sample*, and define clearly the sampling units and sampling plan at each stage. What are the merits of sampling with *probability proportional to size* (PPS)?

(3)

(ii) Street directories are available, which list all the buildings (by name or number) in each street of the postcode sectors that might be chosen. It is proposed to use a street directory as a sampling frame for obtaining a sample of households in selected postcode sectors. Discuss briefly the typical problems associated with using a street directory as a sampling frame. Suggest a suitable sampling frame of households in a country of your own choice.

(3)

(b) Part of the printed questionnaire to be sent to households is shown **on the next page.**

Comment on the strengths and weaknesses of this questionnaire. Suggest alternative phrasing for those questions which you think could be improved.

(14)

The questionnaire for part (b) is shown on the next page

Turn over

Questionnaire for question 6 part (b)

NATIONAL SHOPPERS SURVEY

Instructions

1) This questionnaire should be completed by the main shopper in your household
2) Please give the answers that apply to you by putting an "X" in the appropriate boxes
3) Please return your completed survey in the reply-paid envelope

1. ABOUT YOU AND YOUR FAMILY

1. What is your date of birth? |__|__| |__|__| |__|__|__|__|
   Day Month Year

2. Are you? 1 □ Married 2 □ Single

3. How many persons are living at home? _____

4. Which group best describes your combined household income?

   1 □ Up to £5 000
   2 □ £5 000 – £10 000
   3 □ £10 000 – £20 000
   4 □ £20 000 – £30 000
   5 □ £30 000 – £50 000
   6 □ £50 000 plus

2. YOUR SHOPPING

When doing your main shopping:

1. Which stores do you use for food and grocery shopping? (please specify) _____

2. Why do you buy where you do?

   1 □ Distance
   2 □ Convenience
   3 □ Quality of Products
   4 □ Parking Facilities
   5 □ Prices
   6 □ Food Range

3. What do you spend on groceries?

   1 □ Under £15
   2 □ £15 – £24.99
   3 □ £25 – £34.99
   4 □ £35 – £44.99
   5 □ £45 – £59.99
   6 □ £60 – £74.99
   7 □ £75 – £89.99
   8 □ £90 – £104.99
   9 □ £105+

4. How often have you bought the following products in the past 3 months?

   Environmentally Friendly Products
   1 □ Very Often 2 □ Somewhat often 3 □ Not too often 4 □ Never
   Recycled Products
   1 □ Very Often 2 □ Somewhat often 3 □ Not too often 4 □ Never

5. How many packets of breakfast cereals do you buy?

   □ 1 □ 2 □ 3 4 □

6. What breakfast cereal brands do you buy regularly? (please tick all that apply)

   1 □ Kellogg
   2 □ Nabisco
   3 □ Jordan
   4 □ Shops Own
   5 □ Other (please list) _____

Turn over

7. (i) Give three situations in which a cluster sampling design might be used rather than simple random sampling. What might be the drawbacks of cluster sampling in such cases?

(4)

(ii) In the usual notation for one-stage cluster sampling (using a simple random sample of clusters),

$$\hat{Y}_{ub} = N\bar{y} \quad \text{and} \quad \bar{y}_{cl} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

are estimators of the population total Y and the population mean $\bar{Y}_{cl}$ respectively.

(a) Explain the notation N, n, m_i, y_i and $\bar{y}$. (3)

(b) Show that $\hat{Y}_{ub}$ is an unbiased estimator of the population total Y. (3)

(c) Why is $\bar{y}_{cl}$ a biased estimator of the population mean $\bar{Y}_{cl}$ and when is its bias small? (3)

(iii) An accountancy firm was studying the error rate in one of its compliance audits. The population consisted of 828 claim forms. A claim form contained an entry in each of 215 fields.

A simple random sample of 85 claim forms was selected. Each form was checked and the total number of fields in error, y_{iT}, was recorded. The maximum number of fields in error on any form was 4. The table below summarises the number of forms which had 0, 1, 2, 3 or 4 erroneous fields.

Number of errors	*Number of forms*
4	1
3	1
2	4
1	22
0	57
	Total 85

(a) Treating the claim forms as clusters and the fields as the individual sampling units, estimate the total number of errors in the population of 828 claim forms. Calculate the standard error for your estimator.

(4)

(b) Suppose instead that these data came from a simple random sample of 18275 fields. What would now be the estimated standard error of the estimator of the total number of errors in the population of 828 claim forms? Comment on why the firm chose to carry out a cluster sample of whole claims.

(3)

8. The mortality rates for a certain stationary population, *A*, with 1000 births per year are given in the following table, where ${}_{10}q_x$ is the probability that a person aged x years dies within the next 10 years.

Age	${}_{10}q_x$
0	0.029
10	0.009
20	0.015
30	0.020
40	0.047
50	0.125
60	0.291
70	0.551
80	0.846
90	0.979
100	1.000

Using only this information, estimate:

(i) the age distribution in 10 year class intervals,

(ii) the expected ages at death of groups of people now aged 20, 40, 60, 90 and 100,

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

(15)

Using the unabridged life table, the life-expectancy of people in this population is 67.92 years, and the expected ages at death of people now aged 20, 40, 60, 90 and 100 are 70.40, 71.84, 75.61, 93.05 and 101.80 respectively. Explain why these figures differ from those obtained in parts (ii) and (iii).

(2)

A different stable population, *B*, experiences the same mortality rates as population *A* and an annual growth rate of 1%. Without doing any additional calculations, explain how you would find the age distribution of population *B* in a form which is suitable for comparison with the distribution obtained for population *A*. How would you expect these distributions to differ?

(3)