EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2008

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.

1. Explain what is meant by a *randomised complete blocks design* for a set of t experimental treatments. Why is this design useful?

(3)

A field experiment to study the response of a cereal crop to the addition of six new fertilisers A to F was laid out as four randomised complete blocks.

The data below are the crop yields (in coded units) for each plot at the end of the growing season.

		Fertiliser						
		A	*B*	*C*	*D*	*E*	*F*	*Total*
Block	I	3.21	3.79	3.44	3.62	3.72	3.44	21.22
	II	3.53	4.32	3.66	3.83	4.23	3.76	23.33
	III	3.58	3.52	3.31	3.24	3.69	3.90	21.24
	IV	3.44	3.81	3.40	3.42	4.30	3.64	22.01
	Total	13.76	15.44	13.81	14.11	15.94	14.74	87.80

$\sum y = 87.80, \ \sum y^2 = 323.3288$

(i) An 80 × 120 metre field divided into 20 × 20 metre plots was available for the experiment and was known to have a North-South fertility gradient. With the aid of a sketch, show how the blocks of plots for this experiment should be arranged in the field, and how the treatments could be allocated to a typical block.

(3)

(ii) Write down the linear model appropriate for a randomised complete blocks design, and state what assumptions are required for the analysis of variance.

(3)

(iii) Construct the analysis of variance for these data. Carry out further significance tests for comparing pairs of treatment effects, and report on the results.

(6)

(iv) Discuss any concerns you have about carrying out the analysis in part (iii), and about the use of new fertilisers only.

(2)

(v) Suppose now that the treatments had been six combinations of three equally spaced levels of nitrogenous fertiliser at each of two dates of application, with the lowest level of application being no fertiliser at all. Suggest how the treatment sum of squares would be partitioned in this situation.

(3)

Turn over

2. Explain the principles of confounding and fractional replication for 2^k factorial experiments.

(4)

In a preliminary investigation into the compressive strength of cylinders of concrete, five factors, each at two levels, were studied: type of sand (*S*), type of cement (*C*), amount of water (*W*), time to mix (*M*), and time in mould (*T*).

The results below are from a ½-replicate of a 2^5 factorial experiment in which the defining contrast was $I = -SCWMT$. The order of the 16 treatment combinations was completely randomised. The table shows the treatment combinations that were run and the responses (compressive strengths in coded units) obtained. Also given are the estimates of main effects and two-factor interactions calculated from these data.

Treatment	Response	Treatment	Response
cw	10.2	*scwm*	15.1
sm	17.5	*cm*	10.6
wm	13.0	*swmt*	19.1
cwmt	17.3	*ct*	16.9
(1)	13.4	*sc*	14.8
wt	20.2	*mt*	19.5
sw	18.1	*scwt*	15.7
scmt	15.7	*st*	19.2

Effect Estimates			
S	1.76	*ST*	–2.81
C	–2.96	*CW*	–0.06
W	0.14	*CM*	0.36
M	–0.09	*CT*	–0.14
T	3.86	*WM*	0.16
SC	–0.19	*WT*	0.11
SW	0.06	*MT*	–0.01
SM	–0.01		

(i) Explain why the treatment combination *st* is part of this scheme but the combination *scwmt* is not. What are the consequences of performing only 16 of the 32 possible treatment combinations?

(4)

(ii) Show how the value of –2.81 for the *ST* interaction estimate was obtained.

(2)

(iii) One method of analysing small 2^k experiments with several factors is to make a Normal probability plot of the estimable main effects and two-factor interactions. Points lying off the line at either end may indicate important main effects and interactions. Such a plot for this experiment is shown **on the next page**. Explain why it may be useful for analysing data from this experiment, and interpret the plot. Give reasons to justify any conclusions that you make.

(4)

(iv) It is desired to set the levels of the factors *S, C, W, M* and *T* to achieve as large a value of the response as possible. Suggest levels of *S, C, W, M* and *T* which would achieve this objective.

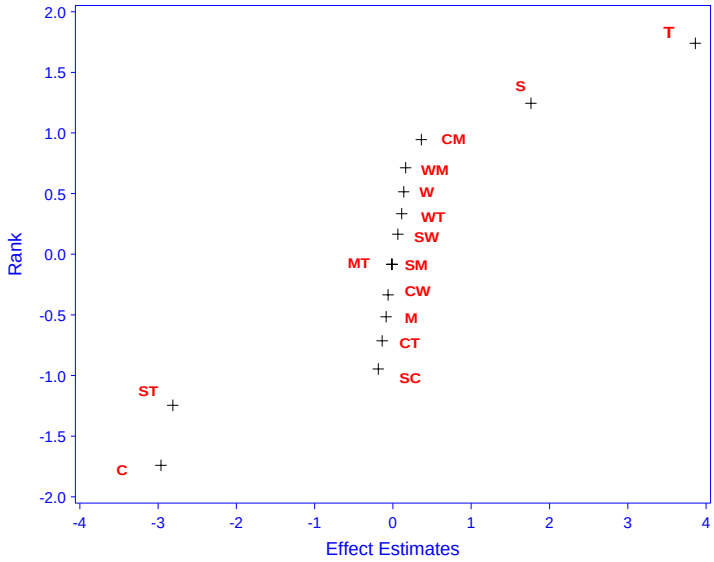
(4)

(v) It is sometimes argued that one should pool effects which appear small in order to obtain more degrees of freedom for the residual sum of squares. Comment on the advisability of doing this in fractional factorial designs.

(2)

Turn over

Normal probability plot of estimates of main effects and two-factor interactions **for question 2**.



Turn over

3. Write down the linear model for a Latin square design and the assumptions underlying its analysis of variance.

(2)

Give two examples of situations where analysis of variance of the raw data might not be valid. Explain the use of transformations prior to analysis of variance and give details of how an appropriate transformation could be chosen in practice.

(4)

An experiment has been conducted using a 5×5 Latin square design to compare treatments *A*, *B*, *C*, *D* and *E*. The table below gives the fitted values and the residuals (in brackets) for this experiment.

B:	19.6 (–1.6)	*A*:	24.8 (0.2)	*D*:	24.2 (3.8)	*E*:	24.8 (–0.8)	*C*:	26.6 (–1.6)
D:	25.2 (–2.2)	*E*:	27.0 (0.0)	*A*:	29.2 (–3.2)	*C:*	31.0 (2.0)	*B*:	27.6 (3.4)
C:	19.6 (0.4)	*D*:	20.8 (–1.8)	*E*:	20.4 (1.6)	*B*:	21.0 (–2.0)	*A*:	23.2 (1.8)
E:	22.4 (1.6)	*C*:	27.2 (0.8)	*B*:	24.6 (–0.6)	*A*:	28.6 (–0.6)	*D*:	27.2 (–1.2)
A:	23.2 (1.8)	*B*:	23.2 (0.8)	*C*:	26.6 (–1.6)	*D*:	26.6 (1.4)	*E*:	25.4 (–2.4)

(i) Explain in detail how the fitted values and residuals shown above were calculated.

(4)

(ii) Plot the residuals against the fitted values. What conclusions do you draw from your plot?

(6)

(iii) Without additional calculation, describe any further checks you would carry out in an analysis of these residuals, identifying for each the assumptions being investigated and the steps that could be taken if these assumptions are violated.

(4)

4. In a pilot plant, an experimenter is interested in determining how the time (X_1) and temperature (X_2) affect the build-up of an unwanted by-product (Y) in a chemical process. It is decided to explore the region of interest in time and temperature by a series of experiments. The data below are from an initial experiment consisting of 12 runs performed in random order.

Time (min.), X_1	*Temperature (°C), X_2*	*Response, y*
30	240	2.5
40	240	4.0
30	250	0.7
40	250	2.2
35	245	2.5
35	245	2.3
30	240	2.7
40	240	4.3
30	250	0.3
40	250	2.5
35	245	3.0
35	245	2.4

$\Sigma y = 29.4$, $\Sigma y^2 = 86.00$.

(i) Comment on the structure of this design, and the purpose of the replicated points.

(3)

For analysis, the factor levels are coded: $X_1 = 40$ becomes $x_1 = +1$, $X_1 = 35$ becomes $x_1 = 0$, and $X_1 = 30$ becomes $x_1 = -1$; also $X_2 = 250$ becomes $x_2 = +1$, $X_2 = 245$ becomes $x_2 = 0$, and $X_2 = 240$ becomes $x_2 = -1$.

(ii) The response y is regressed on the coded values of time (x_1) and temperature (x_2), using the results of the 12 runs. The fitted regression model is

$$\hat{y} = 2.450 + 0.850x_1 - 0.975x_2,$$

and the residual sum of squares from the analysis of variance table is 0.5850. Use this information to test for lack-of-fit for the first-order response function. Does the model provide an adequate fit to the data?

(6)

(iii) Assuming this model is adequate, give the direction of steepest **descent** in terms of the changes in temperature per minute change in time.

(2)

(iv) Draw an (x_1, x_2) graph which shows the five treatment combinations actually used. Plot the fitted model in part (ii) over the experimental region for $\hat{y} = 1$, 2, 3 and 4, and also the line of steepest descent.

(6)

(v) Briefly describe how you would continue with the series of experiments if the objective is to try to find optimum operating conditions by the method of steepest descent.

(3)

Turn over

5. (i) The formula for the variance of the estimator of a population mean based on a stratified (random) sample is

$$V = \sum_{h=1}^{L} W_h^{\,2} \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^{\,2}}{n_h}.$$

Define the symbols in the above formula. Explain the conditions under which stratified sampling may be superior to simple random sampling.

(5)

(ii) A region aims to estimate the total number of children who have played truant from school in the past week (that is, who have been absent from lessons without good reason). The region is divided into four education authorities (strata) and a sample of ten schools is taken from each education authority. The results are as follows.

Education authority, h	*Total number of schools*	*Number of truants (y) in schools selected*										$\bar{y}_h$	s_h
		1	2	3	4	5	6	7	8	9	10		
1	141	43	84	98	0	10	44	0	124	13	0	41.6	45.69
2	471	50	147	62	87	84	150	170	104	56	160	107.0	46.00
3	256	228	262	110	232	139	170	334	0	63	220	175.8	99.80
4	1499	17	34	25	34	36	0	25	7	15	31	22.4	12.31

(a) Obtain a point estimate and a 95% confidence interval for the total number of children in the region who have played truant from school in the past week.

(5)

(b) Suppose that the cost of sampling in different strata varies and let c_h be the cost of sampling an element from stratum h. If c_1 = £25, c_2 = £10, c_3 = £10, c_4 = £5, and the total sample size is 40, use the results of this survey to compute the sample sizes in each stratum under

(*A*) optimal allocation,

(*B*) proportional allocation.

For each method, state the total cost of sampling, and obtain an estimate of the variance of the estimated total number of children who have played truant in this region.

(7)

(c) The region is planning a new survey, and is intending to sample an equal number of schools from each authority in the region. Write a short report on the merits of using optimal, proportional or equal allocation for this survey.

(3)

Turn over

6. (a) Define what is meant by a *sampling frame*. Discuss whether a national set of telephone directories would provide a good sampling frame for adults/ households. Suggest two suitable sampling frames for adults/ households that you are familiar with and compare their advantages and disadvantages. In what ways do these frames differ from the target population? Your discussion should distinguish between the cases where the target population consists of households and where it consists of adults.

(7)

(b) An official from a local authority has asked you to design a questionnaire for a large self-completion survey of households in its area. One aim is to estimate how many households have been victims of crime. The following is suggested as a question that any household member should be able to answer on behalf of the household.

Over the past 12 months, have you or your household been victims of any crimes? Yes/No

Comment on the strengths and weaknesses of the above question as a way of providing reliable and useful information about crime victimisation.

Draft a set of questions that overcome any weaknesses you have identified.

(7)

(c) In a city, the percentage of adults in favour of the proposed road user charging scheme, thought to be between 35% and 55%, is required to be estimated with a standard error of not more than 2%. If the city has 15 000 adult residents, how large a sample is necessary to meet this objective?

Give reasons why, for this survey, quota sampling might be preferred to simple random sampling. What might be its drawbacks?

(6)

7. A small survey was carried out in a Middle Eastern country to estimate the total number of bunches of bananas produced in a district during a given growing period. The district was divided into 289 primary units such that each unit had about 500–1000 banana pits. Each pit may produce 0, 1 or more bunches of bananas. The total number of banana pits for the whole district was known to be 181 336. A simple random sample of 20 primary units was selected from the 289 units, and for each unit the number of banana pits (x) and the total number of banana bunches (y) were obtained. The results are summarised below.

	Sample (n = 20)	
	Mean	SD
Number of banana pits per unit (x)	644.35	115.9025
Total number of banana bunches per unit (y)	901.70	221.8112

The correlation is 0.7737 between the number of banana pits and the total number of banana bunches.

(i) Explain why a ratio estimator is appropriate for these data.

(3)

(ii) Estimate the **total** number of banana bunches for the district, and obtain a standard error of your estimate, using

(a) the simple random sample mean,

(b) the ratio estimator.

(9)

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

$$\frac{1-f}{n}\frac{1}{\bar{X}^2}\left\{s_y{}^2 + r^2 s_x{}^2 - 2r\hat{\rho}s_x s_y\right\},$$

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

(iii) Comment on the results. If it was suggested to you that the ratio estimate should not be used because it is biased, how would you reply?

(2)

(iv) Show that about 225 primary units must be sampled from the 181 336 banana pits in the district to make the half-width of the 95% confidence interval for the total number of banana bunches about 2500 (using a ratio estimate).

(3)

(v) Discuss briefly an alternative sampling scheme that might be suitable for such a survey and the practical difficulties that might arise in carrying it out.

(3)

8. Demographic data for China in 2005 (source: U.S. Census Bureau) are shown below.

Mid-year population (in millions)

Age	*Males*	*Females*
0 – 4	43.8	38.4
5 – 14	104.3	92.7
15 – 24	117.3	109.2
25 – 34	108.5	103.3
35 – 44	115.2	109.7
45 – 54	83.6	79.1
55 – 64	52.5	49.4
65 – 74	32.9	33.0
75 – 84	12.7	15.6
85+	1.8	3.3

Births per 1000 population	13
Deaths per 1000 population	7
Life expectancy at birth (years)	72.3
Infant deaths per 1000 live births	24
Total fertility rate (per woman)	1.7

(i) Explain clearly the terms death rate, life expectancy, infant death and total fertility rate, and show how they are calculated.

(6)

(ii) Draw an age pyramid to illustrate the age-sex structure of the population of China in 2005. Comment briefly on the main features of the pyramid, and what may have caused them, setting your comments in the current world population context.

(8)

(iii) The age-specific death rates and sex-ratios of these death rates for children aged up to 5 years in China, 1999–2000 census, are given below. Explain clearly how these rates have been calculated.

Age	*Males*	*Females*	*Sex Ratio*
0	22.56	32.10	0.70
1	2.37	2.64	0.90
2	1.59	1.61	0.99
3	1.19	1.15	1.03
4	0.92	0.80	1.15
5	0.77	0.63	1.22

(3)

(iv) Comment on the similarities and differences in mortality levels between girls and boys in China, 1999–2000. Discuss briefly what other demographic analyses you might consider to explore the age-sex structure in China.

(3)