

EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY



GRADUATE DIPLOMA, 2003

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 .

1. An experiment was carried out to study brake wear (y) on cars of similar size made by four different manufacturers F , V , R and P . Four different test roads (1, 2, 3, 4) were used and there were four drivers (A , B , C , D). The scheme for the experiment, and the results y , are shown in the table.

		Driver and response (y)				Total
		Road 1	Road 2	Road 3	Road 4	
Car manufacturer	F	B : 44	A : 46	D : 39	C : 52	181
	V	C : 51	B : 37	A : 43	D : 40	171
	R	A : 42	D : 39	C : 46	B : 34	161
	P	D : 45	C : 52	B : 36	A : 42	175
Total		182	174	164	168	

$$\Sigma y = 688, \quad \Sigma y^2 = 30042.$$

- (i) Analyse the data to extract effects due to manufacturers, roads and drivers, and report briefly on the results. (6)
- (ii) Discuss briefly the advantages and disadvantages of a Latin square design. (4)
- (iii) Explain the principles that should be followed when selecting a particular Latin square of the appropriate size for use in an experiment. (4)

You are now told that drivers A and C always drove in the morning, while B and D always drove in the afternoon. Also, A and B drove at the weekend, while C and D drove on weekdays.

- (iv) Investigate the data further, using linear contrasts to examine differences between times of day, differences between times in the week, and any other relevant comparisons. Discuss your results, and explain whether it is reasonable to ascribe these differences to time. (6)

2. The data below show the time to death by cyanide in *Phoxinas laevis*, a European minnow (small fish), using 4 concentrations of cyanide ion (mg CN/litre) and 3 different types of gas (denoted by G_1, G_2, G_3).

Three replicates of the 3×4 factorial design were performed, and the order of the 36 runs was completely random. For each run there were 10 fishes, and the logarithm of the survival time, in minutes, of each fish was calculated. These values were coded and summed over the 10 fishes in each run to give the values y in the table below.

	Cyanide concentration (mg CN/litre)				Total
	0.16	0.80	4	20	
Gas G_1	161 119 124	103 110 104	91 90 86	57 60 72	
Total for G_1	404	317	267	189	1177
Gas G_2	206 158 169	111 124 94	99 91 98	62 86 74	
Total for G_2	533	329	288	222	1372
Gas G_3	182 231 207	130 133 117	81 109 102	72 59 92	
Total for G_3	620	380	292	223	1515
Total	1557	1026	847	634	

$$\Sigma y = 4064, \quad \Sigma y^2 = 525276, \quad \frac{1}{3}(404^2 + 317^2 + \dots + 292^2 + 223^2) = 519688.667.$$

- (i) What characteristic of survival time distributions suggests that a logarithmic transformation of raw data might often be appropriate? (1)
- (ii) Carry out an analysis of variance to examine the effects of cyanide concentration and type of gas, and their interaction, on survival time. (4)
- (iii) Partition the sum of squares for the cyanide main effect into single-degree-of-freedom components.
 [NOTE that the cyanide levels used were equally spaced on the logarithmic scale. The coefficients of linear, quadratic and cubic components for four equally spaced levels of a factor are, respectively, $(-3, -1, 1, 3)$, $(1, -1, -1, 1)$ and $(-1, 3, -3, 1)$.] (5)
- (iv) Draw a diagram showing all the 12 means of the gas/cyanide combinations. (4)
- (v) Using the diagram and the analysis of variance, explain the results found by this experiment, including mention of any gas/cyanide interaction. (6)

3. (i) A randomised block experiment using b blocks is to be conducted using a standard treatment S and v other treatments A, B, \dots . Each block is to contain $(v + c)$ plots, consisting of c replicates of S and one of each of A, B, \dots .
- (a) Write down a suitable mathematical model, stating any assumptions to be made. (4)
- (b) Derive the least squares estimator of the difference between the effect of treatment S and that of any other treatment. (4)
- (c) Find the variance of the difference between the effect of treatment S and that of any other treatment. (3)
- (ii) Four new (and possibly improved) methods of controlling a crop pest were applied to plots within a randomised block experiment. Each block contained 6 plots, consisting of 2 replicates of a standard treatment, S , and one of each new treatment, A, B, C and D . The percentage damage to the crop (y) from each plot is given below.

Block	Treatment					Total
	S	A	B	C	D	
I	22, 30	14	16	8	10	100
II	24, 26	11	16	5	6	88
III	18, 14	9	12	8	6	67
IV	16, 12	8	6	4	5	51
Total	162	42	50	25	27	

$$\Sigma y = 306, \quad \Sigma y^2 = 5076.$$

The researcher is interested only in comparing each new treatment with the standard treatment, S .

- (a) Construct an analysis of variance for these data. Construct 95% confidence intervals for the difference in mean percentage crop damage between S and each new treatment. Comment on which of the new methods, if any, appear to result in a significant improvement over the standard method. (6)
- (b) Discuss briefly any concerns you have about carrying out the analysis specified in part (ii)(a). (3)

4. (i) Two controllable factors which influence process yield y are reactant concentration (A) and feed rate (B). The current operating conditions use a reactant concentration of 20% and a feed rate of 25 kg/hr, which result in yields of approximately 145 units.

The plant manager is interested in determining operating conditions that maximise yield, and suggests an experimental plan in which the factors are varied one at a time.

He proposes to use current operating conditions as a baseline, and to include as the other treatment combinations [i] A at 20%, B at 30 kg/hr, [ii] A at 25%, B at 25 kg/hr.

- (a) Explain to the plant manager, with the aid of a diagram, why this experiment may not be appropriate for locating the operating conditions that give the highest yield. (3)
- (b) The plant manager now agrees to run four replicates of a 2^2 experiment, centred on the current operating settings. Discuss the advantages and disadvantages of this proposal. (3)
- (ii) Four replicates of a 2^2 factorial design were performed, with the four replicates of the four treatment combinations run in random order. The factors A and B were varied ± 5 units from their present settings. The yield y was regressed on reactant concentration (x_1) and feed rate (x_2), using the results of these 16 runs. The regression equation obtained was

$$\hat{y} = 146.6875 + 1.0125(x_1 - 20) - 0.5875(x_2 - 25).$$

- (a) Explain how an appropriate test for lack of fit of the above model can be constructed. (3)
- (b) Calculate the path of steepest ascent. Draw an (x_1, x_2) graph which shows the four treatment combinations actually used and also the line of steepest ascent. Explain how this may be used to choose treatment combinations for a follow-up experiment. (5)
- (c) Suppose the reactant concentration cannot exceed 37.5% and the feed rate cannot go below 11.5 kg/hr. Show the "practical" path of steepest ascent on your plot – that is, the path that accounts for these constraints. What settings would you recommend for the follow-up experiment? (4)
- (d) If the manager had asked you at the outset to design the experiment so that it would also enable you to test the hypothesis that the mean yield at $x_1 = 20$, $x_2 = 25$ was significantly different from 145, would you have recommended any change to the experimental design? Give reasons for your answer. (2)

5. (i) Explain the difference between *random* and *non-random* methods of sampling, discussing both the construction of samples and the methods of analysing data collected by them. Suggest reasons why non-random samples may sometimes be preferred. Include an explanation of *systematic sampling*, and whether it should be treated as random or non-random.

(6)

- (ii) Write down the formula for calculating an unbiased estimate, s^2 , of the variance of a large (but finite) population, based on a simple random sample of n items. Define any symbols you use. Show also that, for a binary variable, $s^2 = np(1 - p)/(n - 1)$, where n and p are to be defined.

A pilot survey has given rough estimates of the mean and variance of a measurement x , and of the proportion p of a special type of member, in the population being studied. The main sample survey will be required to estimate, at the 95% confidence level, the population mean of x within ± 1.5 units, and also the proportion p within ± 0.04 . If the pilot survey value of the variance of x was 168.33 and the value of the required proportion was 0.36, find the minimum sample size that should be used to meet requirements.

(7)

- (iii) Define *one-stage* and *two-stage cluster sampling*. How do cluster sampling and *stratified sampling* differ, both in their construction and in their use? Give an example of a survey in a country of your choice that uses both stratification and clustering in the sample design.

(7)

6. (i) A simple random sample of size n is selected from a population of N units. The response of interest, y , and an auxiliary variable, x , are measured on each unit in the sample. The population mean of the auxiliary variable is \bar{X} .

The sample estimator, \hat{R} , of a population ratio is given by $\hat{R} = \frac{\bar{y}}{\bar{x}}$.

Show that, approximately,

(a) the bias of \hat{R} is $-\left\{\text{Cov}(\hat{R}, \bar{x})\right\} / \bar{X}$, (2)

- (b) the variance of \hat{R} may be estimated by

$$\frac{1-f}{n} \frac{1}{\bar{x}^2} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\hat{\rho}s_y s_x),$$

where f is the sampling fraction and s_y^2 , s_x^2 and $\hat{\rho}$ are to be defined. (4)

[You may assume that (if \bar{X} is not known) the variance of \hat{R} may be estimated (approximately) from a sample as

$$\frac{N-n}{Nn\bar{x}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2.]$$

- (ii) Explain briefly the circumstances under which the ratio estimator of a population mean will be less precise than the sample mean of a simple random sample of the same total size. (3)

- (iii) The wholesale price for oranges in large shipments is based on the sugar content of the load. The exact sugar content cannot be determined prior to the purchase and extraction of the juice from the load.

The data below show the sugar content in grams (y) and weight in grams (x) of a random sample of ten oranges from a consignment with total weight 820 kg.

Sugar content (gm) y	9.5	13.6	11.4	9.1	15.0	12.3	8.6	9.5	10.5	11.4
Weight (gm) x	182	218	195	191	227	209	177	186	190	200

$$\Sigma xy = 22\,194.80, \quad \Sigma y^2 = 1268.69, \quad \Sigma x^2 = 392\,389.$$

- (a) Explain why a ratio or regression estimator is appropriate for these data. (2)
- (b) Estimate the total sugar content for the oranges and give a standard error of your estimate, giving reasons for your choice of estimator. (6)
- (c) Show that about 25 oranges must be sampled from a consignment with total weight 820 kg so that the half-width of the 95% confidence interval for the total sugar content is less than 2 kg. (3)

7. (a) You are about to design a questionnaire. Discuss considerations which will affect your choice of wording and ordering of questions, including examples of:

(i) a question that could bias the responses because of its strong wording,

(ii) two questions whose responses are likely to depend on the order in which they are asked,

(iii) open and closed forms of questions,

(iv) questions with and without a balanced alternative.

(12)

(b) A Teenager Attitudes and Practices Survey obtained completed questionnaires, either by telephone or mail, from some members of a randomly selected group aged 12–18 years living in households across the country. One type of question asked the teenagers about the perceived behaviour of members of their age-group (referred to as their "peers"). One specific question asked was:

Do your peers care about staying away from drugs?

- A lot
- Somewhat
- A little
- Not at all

Why would a question be formed in terms of "peer" behaviour rather than as a direct question to the person being interviewed? Why might respondents who failed to answer the question give rise to bias in the results of the survey? How could you use questions about the respondent's background and attitudes to investigate this possible bias?

Mention other sources of errors that could arise in surveys like this.

(8)

8. (a) Define the demographic usage of the term *fertility*, and distinguish between *period* and *cohort* analysis of fertility. (5)

- (b) The total numbers of births, and the numbers of these which were stillbirths (fetal deaths), in each of three health districts have been recorded. These are tabulated by the age of the mother.

	Health district					
	A		B		C	
<i>Age of mother</i>	<i>Total births</i>	<i>Still-births</i>	<i>Total births</i>	<i>Still-births</i>	<i>Total births</i>	<i>Still-births</i>
< 20	3721	31	6083	57	3326	24
20 – 24	12803	89	19420	123	9951	52
25 – 29	14032	97	19318	120	10628	60
30 – 34	8260	61	12144	83	6642	41
35 +	2493	30	3745	32	2670	32
Total	41309	308	60710	415	33217	209

For each of the three health districts, calculate:

- (i) the crude fetal death rate per 1000 births,
- (ii) the age-specific fetal death rates per 1000 births,
- (iii) the age-adjusted fetal death rate per 1000 births using health district A as the standard.

Comment on the information conveyed by these death rates, explaining clearly the differences observed between the crude and age-adjusted fetal death rates.

(15)