EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2011

MODULE 8 : Survey sampling and estimation

Time allowed: One and a half hours

Candidates should answer **THREE** *questions.*

Each question carries 20 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 5 printed pages **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 4 questions altogether in the paper.

1. In an opinion poll, a simple random sample of 1600 voters in a town was selected from the electoral roll (list of voters) and interviewed, in order to estimate support in the town for the candidates from two parties, A and B, in a forthcoming election. The electoral roll lists 160 000 persons, divided geographically into seven wards (areas) that have distinct characteristics in terms of economic well-being.

(i) State the advantages and disadvantages of using simple random sampling to select this sample of 1600 voters from the electoral roll. Suggest another sampling method that could be used with advantage to select the voters, and explain what benefits this method would have over simple random sampling.

It is nearly a year since the electoral roll was compiled. Explain what consequences this fact may have for the survey.

(8)

(ii) Voters were asked to say which of the two parties they would support if there were an election tomorrow. Of the sample of 1600, 720 voters said party A and 880 said party B. Using this information, obtain an approximate 95% confidence interval for p, the current proportion of party A supporters. Explain what this confidence interval shows.

State why the finite population correction need not be used in this calculation.

(7)

(iii) For future surveys, the organisers wish to estimate p to within 3% of the true value. How large a (simple random) sample is needed to meet this objective?

(5)

2. A water company wishes to estimate the total amount of water used by its domestic consumers. Its domestic business serves 400 000 households altogether and is divided into four regions.

As an initial exercise, before the main survey, a simple random sample of 50 metered households was selected in each region, and for each household the amount y of water used (in megalitres) over the past 12 months was determined by inspecting its water bill. Estimates of the means and standard deviations of y in each region, based on this sample of 200 households, are as follows.

Region	1	2	3	4
Households (thousands) in whole region	160	110	80	50
Sample mean	0.131	0.095	0.112	0.172
Sample standard deviation	0.022	0.015	0.032	0.064

(i) Comment briefly on the merits of using stratified random sampling for this survey.

(3)

(ii) Based on the results given above, estimate the total domestic water consumption over the past 12 months, and obtain an estimate of the standard error of your estimator.

(7)

(iii) Define *optimal allocation*. For the main survey, discuss briefly why you might choose optimal allocation rather than proportional allocation. (You may assume that the cost of sampling any unit is constant.)

Use optimal allocation to calculate the total sample size and the allocation needed to estimate total domestic water consumption over the past 12 months, to within 600 megalitres with 95% confidence.

(10)

3. A government agency wishes to conduct a survey of 400 schools to collect data on alcohol and drug use among 15–16 year olds.

In order to select the sample of schools, a sampling plan based on local administrative areas is proposed. These areas will first be stratified into 50 geographical regions. In each of the strata the schools will be listed, and schools will be selected for the sample. Within each chosen school, the classes in the chosen age-group will be listed and a simple random sample of classes selected. It is assumed that class sizes in this age-group are similar in all the schools.

All pupils in the chosen classes within each sampled school will be given a printed questionnaire and asked to complete questions on family background and demographics, and self-reported substance use.

(i) Explain why this sample of pupils is a *cluster sample*, and define clearly the sampling units and sampling plan at each stage.

(6)

(ii) Give reasons why, for this survey, sampling whole classes is preferred to sampling pupils individually from all classes. Discuss other practical difficulties that might arise in planning and operating a school survey such as the one described.

(6)

(iii) Why might pupils who fail to answer the questions on substance use give rise to bias in the results of the survey? How could you use the questions about a pupil's background and attitudes to investigate the bias?

Discuss other sources of errors that could arise in surveys such as the one described.

(8)

4. A simple random sample of 33 low-income families in a large city yielded the following information on family size, weekly household income (£) and weekly expenditure on food (£).

Family number	*Size* x_1	*Income* x_2	*Food cost* y	*Family number*	*Size* x_1	*Income* x_2	*Food cost* y
1	2	62	14.3	18	4	83	36.0
2	3	62	20.8	19	2	85	20.6
3	3	87	22.7	20	4	73	27.7
4	5	65	30.5	21	2	66	25.9
5	4	58	41.2	22	5	58	23.3
6	7	92	28.2	23	3	77	39.8
7	2	88	24.2	24	4	69	16.8
8	4	79	30.0	25	7	65	37.8
9	2	83	24.2	26	3	77	34.8
10	5	62	44.4	27	3	69	28.7
11	3	63	13.4	28	6	95	63.0
12	6	62	19.8	29	2	77	19.5
13	4	60	29.4	30	2	69	21.6
14	4	75	27.1	31	6	69	18.2
15	2	90	22.2	32	4	67	20.1
16	5	75	37.7	33	2	63	20.7
17	3	69	22.6				

The totals are: $\Sigma x_1 = 123$, $\Sigma x_2 = 2394$, $\Sigma y = 907.2$

Sums of squares are: $\Sigma x_1^2 = 533$ $\Sigma x_2^2 = 177\,254$, $\Sigma y^2 = 28\,224$

The sampling fraction for the survey was 2%.

(i) Obtain a point estimate and an approximate 95% confidence interval for

(a) mean weekly expenditure on food per family,

(b) mean weekly expenditure on food per 2-member family.

(10)

(ii) A researcher wishes to estimate two further quantities: the proportion of family income spent on food, and the mean weekly expenditure on food per person, in all families. Explain how you would estimate each of these quantities from the above information. Discuss the properties of these estimators.

(5)

(iii) The city council wishes to improve its services to low-income families in the city. Briefly explain the difference between longitudinal and cross-sectional surveys, and how the distinction might be relevant in this situation.

(5)