EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2012

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. The organisers of a large marathon selected a simple random sample of 1000 athletes from the 10 000 who entered. Athletes were notified as a condition of entry to the marathon that if they were selected they would have to provide a urine sample and answer questions about their training.

(i) The primary purpose of the sampling was to estimate the proportion *p* of all athletes using steroids and performance-enhancing drugs. Of the 1000 athletes tested, 35 were found to be positive.

(a) Give approximate 95% and 99% confidence intervals for the proportion of all athletes using steroids and performance-enhancing drugs.

(4)

(b) For future marathons, the organisers' aim is to sample a sufficient number so that the half-width of the 95% confidence interval for *p* is less than 0.01. If the organisers were to use simple random sampling show that this aim should be achieved with a sample of about 1300.

(4)

(ii) The 1000 sampled athletes were asked to give the average weekly mileage *X* they ran during the 8 weeks preceding the marathon. The mean of *X* in this sample was 46.8 and the sample standard deviation was 6.2 miles.

Calculate a 95% confidence interval for the mean of *X* for the 10 000 runners who entered the marathon. Explain what this confidence interval shows.

(4)

(iii) The organisers would, for practical reasons, much prefer to use a systematic sample. Explain briefly what assumptions this would require, and what advice you might give them.

(5)

(iv) If any athletes had refused to be tested for steroids and performance-enhancing drugs when selected, how might this have led to bias in the survey results?

(3)

Turn over

2. A regional council wishes to assess the amount of hazardous waste produced by the 6231 manufacturing companies in its area. The companies are split into three strata:

(1) basic metal industries;

(2) food, textiles and mineral products;

(3) other manufacturing.

A simple random sample of companies was taken in each stratum, and for each company the total amount of hazardous waste produced in 2003 was measured.

	Hazardous waste ('000 tonnes)			
Stratum (h)	N_h	n_h	$\bar{y}_h$	s_h
1	92	11	166.6	207.7
2	1612	61	7.7	14.7
3	4527	292	0.3	4.5
Total	6231	364		

(i) Define N_h, n_h, $\bar{y}_h$ and s_h.

Estimate the mean amount of hazardous waste produced per company and obtain an estimate of the standard error of your estimator. Give an approximate 95% confidence interval for the mean amount of hazardous waste per company.

(8)

(ii) Compute the sample sizes in the strata if proportional allocation had been used for this survey. Give brief reasons why a stratified sample using proportional allocation should give much more precise results than a simple random sample of 364 units. Explain briefly whether the allocation actually used has been effective in improving precision compared with a proportional allocation.

(8)

(iii) The council wishes to report estimates of the mean amount of hazardous waste produced per company for each of the three strata, as supporting information. Obtain a point estimate and an approximate 95% confidence interval for the mean amount of hazardous waste produced per company for the basic metal industries.

(4)

3. A researcher selects a simple random sample of 2055 farms from the 75 308 farms in a large region in a developing country, and the number of cattle (y) and the total area under cattle (x) were recorded for each farm. The results were as follows.

Sample total number of cattle, Σy_i	25 751
Sample total area (hectares), Σx_i	62 989

The sum of the squares is $\Sigma y_i^2 = 596\,737$. The total area under cattle in this region is 2 353 365.

(i) Using the mean of the simple random sample, estimate the total number of cattle in the region, and the standard error of your estimator.

(4)

(ii) The researcher seeks your advice on how the supplementary information on the area under cattle in the region might be used to estimate the total number of cattle in the region.

(a) Discuss briefly why either a ratio or regression estimator could be appropriate for these data. Explain how you would decide whether to use a ratio or regression estimator.

(3)

(b) The researcher decides to use a ratio estimator, and asks you to comment on his results compared with those obtained in part (i). You may assume that the ratio estimate of the total number of cattle in the region is 962 055, and its estimated standard error is 14 020.7. Comment on the relative standard errors. If it was suggested to you that the ratio estimate should not be used because it is biased, how would you reply?

(5)

(iii) Explain how and why *stratification* and *clustering* might be useful in such a survey, and what practical problems they could help to overcome.

(8)

4. A careers advisor at a university is interested in finding out details relating to occupations of the university's graduates one, three and seven years after graduation.

The university administration holds student records on a computer database and could easily provide a list of all students who graduated in a given year, with the following information for each student.

- Identification number
- Name
- Sex
- Mode of study (full-time or part-time)
- Level of qualification (post-graduate, first degree, other undergraduate)
- Faculty/subject
- Date awarded
- Current address (e-mail or home) or telephone number
- Domicile prior to year of entry

(i) The careers advisor is wondering whether to use a cross-sectional sample survey of those students who graduated one, three and seven years ago, or a longitudinal study of a sample of last year's graduates.

Explain how a cross-sectional sample survey and a longitudinal study differ in their construction and use.

(7)

(ii) The careers advisor decides to use a longitudinal study of a sample of last year's graduates.

(a) Discuss the potential advantages and disadvantages of a longitudinal study for the proposed survey.

(5)

(b) Outline a possible survey methodology that the careers advisor could use to follow up on the occupations of graduates. You should discuss factors such as identification of the population, constructing a sampling frame, the sampling methodology, and how the study should be conducted, mentioning any difficulties the university is likely to encounter in attempting to contact students.

(8)