ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2017

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 8 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 4 questions altogether in the paper.

1. (a) Discuss the reasons for using stratification in a sample survey.

(4)

(b) Dairy farms in a certain region have been divided into four categories, depending on their size and range of produce. In a survey to estimate the total number of milk-producing cows in the region, a stratified random sample of 28 farms is chosen with (roughly) proportional allocation. The results are as follows.

Category (h)	*Total number of farms*	*Number of milk-producing cows (y) in farms selected*	*Sample mean*	*Sample standard deviation*
1	72	61, 47, 44, 70, 28, 39, 51, 52, 101, 49, 54, 71	55.6	18.73
2	37	160, 148, 89, 139, 142, 93	128.5	29.95
3	50	26, 19, 21, 34, 28, 15, 20, 24	23.4	5.95
4	11	17, 11	14.0	4.24

(i) The formula for the variance of the estimator of the population total based on a stratified random sample, where there are L strata, is

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

Define the terms N_h, S_h and n_h in the formula above.

(2)

(ii) Using the data above, estimate the total number of milk-producing cows on dairy farms in this region and construct an approximate 95% confidence interval for this total. Explain what this confidence interval shows.

(7)

(iii) Given that stratified random sampling is used, explain what is meant by *proportional allocation* and verify that this has been used to construct the above stratum sample sizes. Give reasons why, in practice, proportional allocation might be preferred to optimal allocation. What might be a drawback?

(7)

2. A researcher is designing a postal survey of full-time undergraduate students at universities throughout the United Kingdom. The purpose is to estimate the proportion of such students spending more than 10 hours per week in paid employment during term time.

The researcher is seeking your advice in developing a sampling methodology.

Some of the information below might be of assistance in developing the sampling scheme.

- The United Kingdom has approximately 140 universities (as of December 2015).
- The numbers of students registered at each university are held centrally, but there is no central list of names and addresses.
- Each university holds student records and could provide a list of all full-time undergraduate students, with the following information on each student: name, current address and telephone number, gender, faculty/subject, start date.

(i) Outline the meaning of the terms *population*, *sample*, *sampling frame* and *random sampling* in the context of this survey.
(5)

(ii) Suggest reasons why it might be preferable to use cluster sampling, rather than simple random sampling, in this survey. What might be a drawback? How do one-stage and two-stage cluster sampling differ in the context of this example?
(6)

(iii) The researcher likes the idea of using cluster sampling, but is wondering if every student will have an equal probability of selection, given that universities (clusters) vary substantially in size. How would you respond?
(3)

(iv) Briefly contrast the merits of using a postal survey as compared with a telephone survey. Suggest strategies to increase the response rate for a postal survey here.
(6)

3. In a particular sector of industry a survey is conducted in an attempt to investigate the extent of absenteeism from *casual holidays*, that is, not connected with illness or official holidays. A simple random sample of employees, from the total workforce of 36 000, were asked how many days they have taken off work, in the previous six months, as such casual holidays. The results for the 1000 employees who responded were as follows.

	Number of days of casual holiday in past 6 months										
	0	1	2	3	4	5	6	7	8	9 or more	Total
Number of employees	451	162	187	112	49	21	5	11	2	0	1000

(i) Obtain point estimates, and approximate 95% confidence intervals, for

(a) the mean number of days of casual holiday taken by employees in the industry in the past six months,

(b) the proportion of employees in the industry who have taken at least one day of casual holiday in the past six months.

(12)

(ii) The survey was conducted through the use of an *anonymous* questionnaire. Comment on whether you think this is sensible or not. What might be a drawback?

(5)

(iii) The organisers would, for practical reasons, much prefer to use a *systematic* sample for future surveys. Briefly describe systematic sampling, and explain what assumptions it would require.

(3)

4. (a) A city directory, 4 years old, lists the addresses in order along each street, and gives the names of persons living at each address. The city has undergone a good deal of regeneration in the past 4 years.

For a current interview survey of the people in the city, briefly discuss the deficiencies of this frame. How could they be remedied by the interviewers during the course of the field work? In using a directory, would you draw a list of addresses (dwelling places) or a list of persons?

(7)

(b) A survey exploring smoking, drinking and drug use behaviour of young people aged 11–15 is being planned. One objective is to estimate the proportion of young people, p, who have taken illegal drugs in the past year. There are 8000 such young people altogether in a certain local authority area.

The organizer requires the standard error of the estimator of p not to exceed 0.01. The value of p is thought to be between 0.05 and 0.20. If the organizer were to use simple random sampling, show that this aim would be achieved with a sample of about 1334.

Briefly discuss the practical difficulties that might arise in planning this survey.

(7)

(c) A train company is to ask a sample of passengers on a *monthly* basis whether they think the train service offers value for money.

Briefly compare the advantages and disadvantages of selecting a new sample of passengers on each occasion over randomly selecting a panel of passengers and then re-interviewing the same panel on subsequent occasions.

(6)

BLANK PAGE

BLANK PAGE

BLANK PAGE