ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2016

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* $^{n}C_{r}$.

1. Each year, a company surveys the level of employee stress in its four offices. This year, the company proposes to select a stratified random sample from its 5400 employees. Each of the sampled employees will be interviewed and subjected to a physical examination by trained investigators, and awarded a stress "score" of between 0 and 100.

The company is seeking your advice in determining the number of employees to interview.

Results from previous years suggest that there are differences between offices in the standard deviation of stress scores. The estimated standard deviations and the total numbers of employees in the offices are shown below.

Office number	Total number of employees	Estimated standard deviation
1	1400	25
2	500	30
3	2500	12
4	1000	20

(i) You may assume that the formula for the total sample size n given a specified variance, V, of the estimator of the population mean is

$$n = \frac{\sum_h W_h^2 S_h^2 / w_h}{V + \frac{1}{N} \sum_h W_h S_h^2} \ .$$

Define the terms N, W_h, S_h and w_h as used in this formula.

(3)

(ii) The company wishes to estimate the overall mean employee stress score to within d units of the true value with 95% probability (i.e. the width of the interval is $2d$). Show that this requires the variance, V, of the estimator of the population mean to be given by $V = \left(\frac{d}{1.96}\right)^2$.

(4)

(iii) Define *optimal allocation*. Briefly discuss why you might choose optimal allocation rather than proportional allocation for this survey.

(4)

(iv) Using the results from previous years, apply optimal allocation to calculate the sample size, n, and the allocation n_h needed to estimate the overall mean employee stress score to within ±2.5 units with 95% probability. The cost of sampling is the same at each office.

(9)

2.   In 2014, the Scottish Government held a referendum on Scottish independence. People in Scotland voted on whether Scotland should remain as part of the United Kingdom. The electoral register for the referendum covered everyone from age 16 upwards who was resident in Scotland and had appropriate citizenship, and everyone on this register was entitled to vote; this was the first time, anywhere in the United Kingdom, that people as young as 16 or 17 were able to vote.

   (i)   Throughout the referendum campaign, various polling organisations carried out public opinion polls on behalf of national newspapers and broadcasters, to gauge the voting intention. Discuss how *selection bias* might relate to each of the following methods used by polling organisations in collection of their data. Briefly discuss the overall relative merits of using these three methods.

      (a)   Face to face quota sampling.

      (b)   Random digit dialling of landline telephones, to obtain interviews over the phone.

      (c)   An internet sample derived from people who have previously agreed to join a panel of potential interviewees.

                                                                              (14)

   (ii)   Suggest a possible question to find out about people's voting intention for the referendum, with suitable tick-boxes for respondents' replies.

                                                                              (3)

   (iii)   Briefly contrast the difficulties of predicting outcomes of referendums as compared with parliamentary elections.

                                                                              (3)

3.   A local authority is investigating various aspects of the usage of its public libraries.

(a)   A survey of a simple random sample of students in secondary education was undertaken, to determine their use of library services. There are 12 000 such students altogether in this local authority area. One of the survey questions was "How many times have you visited a public library in the past four weeks?" The results were as follows.

Number of visits in past four weeks	Number of students
0	68
1	99
2	50
3	45
4	130
>4	58
Total	450

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

(i)   the proportion of students in this local authority area who visited a public library in the past four weeks,

(ii)   the total number of visits made to a public library in the past four weeks by students in this local authority area. You should use a value of 5 for calculating the mean and standard deviation where the number of visits in the past four weeks is 5 or more.

(12)

(b)   The proportion of adults visiting a public library in this area has declined over the past few years. The local authority is designing a survey of adults and is considering a longitudinal survey. Adults would be selected from the electoral register to take part in the survey by systematic sampling.

(i)   Briefly describe *systematic sampling*, and discuss whether it should be treated as random or non-random.

(ii)   Explain what is meant by a *longitudinal survey* and how it might be useful in this situation. What might be its main drawback?

(8)

4. 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

(i) Using the mean of a simple random sample, estimate the total number of banana bunches for the district, and estimate the standard error of your estimator.

(5)

(ii) The researcher seeks your advice on how he might use the supplementary data on the numbers of banana pits per unit to estimate the total number of banana bunches in the district. He has estimated the correlation between the number of banana pits per unit and the total number of banana bunches per unit (i.e. between x and y above) and thinks a ratio estimator might be suitable.

(a) Explain what is meant by *correlation*. Given that the estimated correlation is 0.7737, how would you respond? Briefly discuss the properties of the ratio estimator and the estimator based on the sample mean in part (i).

(5)

(b) You are given that the ratio estimate of the total number of banana bunches in the district is 253 760.64 and its estimated standard error is 8550.947. Hence, giving a reason, say whether this estimate is better than that calculated in part (i), and use it to construct an approximate 95% confidence interval for the true total number of banana bunches in the district. Explain what this confidence interval shows.

(4)

(iii) Briefly discuss how cluster sampling might be useful in such a survey and what practical difficulties might arise in carrying it out. What might be a drawback of cluster sampling? Why might cluster sampling nevertheless be useful?

(6)

BLANK PAGE

BLANK PAGE

BLANK PAGE