ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

THE ROYAL STATISTICAL SOCIETY

2016 EXAMINATIONS − SOLUTIONS

HIGHER CERTIFICATE – MODULE 8

The Society is providing these solutions to assist candidates preparing for the examinations in 2017.

The solutions are intended as learning aids and should not be seen as "model answers".

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

Question 1

(i) N = total number of units in population
$W_h = N_h / N$, the population weight in stratum h
S_h = population standard deviation in stratum h
$w_h = n_h / n$, the proportion of the whole sample that comes from stratum h

[3 marks for all four definitions correct, 2 for three or two correct definitions, 1 for one or two correct definitions, and 0 for none]

(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 $2d$).

Require $Pr\{|\bar{y} - \bar{Y}| > d\} = 0.05$. [1 mark]

Assume $\bar{y}$ is normally distributed about the true value $\bar{Y}$. Then $Z = \dfrac{|\bar{y} - \bar{Y}|}{\sqrt{\text{Var}(\bar{y})}} \sim N(0, 1)$. [1 mark]

$$Pr\{Z > \frac{d}{\sqrt{\text{Var}(\bar{y})}}\} = 0.05$$ [1 mark]

This would require

$$\frac{d}{\sqrt{\text{Var}(\bar{y})}} = 1.96 \quad \therefore \text{ Var}(\bar{y}) = (\frac{d}{1.96})^2$$ [1 mark]

(iii) Optimal allocation aims to minimise the variance Var $(\bar{y}_{st})$ for a fixed total sample size, n [1 mark]. As well as allocating more sample to strata with higher density (as in proportional allocation), it allocates more to those with larger standard deviations [1 mark], so the precision is comparable with those having lower variability. [1 mark] In this present survey there are wide variations among stratum sizes and standard deviations [1 mark]; proportional allocation with the same sample size as optimal allocation is likely to lead to a considerably larger value of Var $(\bar{y}_{st})$.

(iv) Use the formula quoted in part (i) of the question.

Optimal allocation with constant cost of sampling any unit has $w_h = \dfrac{W_h S_h}{\sum W_h S_h}$. [1 mark]

We have $\sum W_h S_h = 18.519$, using the results from previous years. [1 mark]

Further, $\dfrac{W_h^2 S_h^2}{w_h} = (W_h S_h)(\sum W_h S_h)$, so that $\sum \dfrac{W_h^2 S_h^2}{w_h} = (\sum W_h S_h)^2$. [1 mark]

Also $\sum W_h S_h^2 = 386.111$ (this appears in the denominator of the formula). [1 mark]

Finally, we need V. From part (ii), the criterion $d = 2.5$ with (one-sided) tail probability 0.025 gives $V = (\frac{2.5}{1.96})^2$. [1 mark]

$$\therefore \quad \frac{(18.519)^2}{(\frac{2.5}{1.96})^2 + 0.0715} = 201.92.$$ [1 mark]

So we take $n = 202$. The allocation in each stratum is then given by

$$n_h = 202\, w_h = 202 \times \frac{W_h S_h}{\sum W_h S_h}$$ [1 mark]

which gives $n_1 = 70.7$, $n_2 = 30.3$, $n_3 = 60.6$, $n_4 = 40.4$ [1 mark] i.e., (71, 30, 61, 40). [1 mark]

Note, the last mark requires the answers to be rounded to integers (and that add up to 202).

Question 2

(i) (a) A quota sample is a non-random selection, by trained observers, of individuals from a population which has been stratified by characteristics such as age, sex, social class [1 mark]. A specified number from each stratum has to be interviewed but the observer/ interviewer is free to select the actual individuals. Selection bias is introduced by the high refusal rates and easy availability of persons interviewed, [1 mark] and by the subjective selection of the interviewer. [1 mark]

(b) The sample includes those who are accessible by a landline telephone and ready and willing to respond. [1 mark] There is, by definition, no direct way of covering households having no phone, particularly poorer families, or those with mobile access only. [1 mark] In a large household there might be biases in who answers the telephone and if the decision as to whom to interview is left to the interviewer, there could be selection bias due to the interviewer. [1 mark]

(c) The sampling frame is respondents who have previously agreed to join a panel of potential interviewees. Voluntary samples cannot be representative of the overall population—or even people with common characteristics—because of the self-selection bias**.** [1 mark] Internet samples exclude large portions of the overall *adult* population (in this case, all people who do not use the Internet). [1 mark] 16 and 17 year olds may be excluded, i.e. too young to register as a panel member. [1 mark]

Marks will be awarded for other sensible suggestions.

None of the proposed methods gives a representative sample of adults and young people (16 or 17 year olds). [1 mark]

On cost and ease of administration, on-line polls are better than face-to-face quota surveys and telephone surveys, i.e. low cost and simple administration. [1 mark]

On response rate, face-to-face quota surveys are better than telephone surveys and telephone surveys are better than online polls (which are often self-administered). [1 mark]

On coverage of young people, face-to-face quota surveys are likely to be best, although identifying 16 and 17 year olds in a group of children in the street might be difficult. [1 mark]

On overall accuracy, telephone surveys using random digit dialling are likely to be best, i.e. built on randomisation/have the advantage of not depending on the underlying data on social factors needed for setting quotas. [1 mark]

Note, marks will be awarded for other sensible comments.

(ii) A possible question that might be asked by a polling organisation in conducting its enquiry during the campaign is:

"How do you intend to vote in the coming referendum?" [1 mark]

Tick boxes such as

- ☐ Scottish independence
- ☐ Stay as part of the United Kingdom
- ☐ Don't know
- ☐ Refusal [2 marks: 1 balanced set of alternatives; 1 choice]

Note, marks will be awarded for other sensible comments.

(iii) The relative infrequency of referendums [1 mark] and the fact that they typically ask quite different questions each time [1 mark] make it difficult to estimate voters' intentions compared with a parliamentary election. The minimum age is 18 for parliamentary election. In this referendum, all polls, whether conducted online or by phone, will struggle to interview a representative number of 16 or 17 year olds. [1 mark]

Note, marks will be awarded for other sensible comments.

Question 3

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

For simple random sampling, the sample estimate is $\hat{p} = \frac{382}{450} = 0.849$ [1 mark]

The variance of $\hat{p}$ is estimated as $(1- f)\,\frac{\hat{p}\,(1 - \hat{p})}{n - 1} = \frac{0.9625 \times 0.849 \times 0.151}{449} = 0.00027$ [1 mark],

where $f = \frac{n}{N} = \frac{450}{12000} = 0.0375$ and $(1 - f) = 0.9625$ is the finite population correction.

Assume $\hat{p}$ is approximately normally distributed about the population value. Approximate 95% limits for p are $\hat{p} \pm 1.96\ se\ (\hat{p})$. [1 mark] Limits are 0.849 ± 0.0322, i.e. 0.817 to 0.881. [1 mark]

Note, accept a divisor 'n' (=450) in the formula for the estimated variance of $\hat{p}$.

Accept answers ignoring finite population correction here and in part (ii) (has minimal effect on standard error/CI).

(ii) Let y_i denote number of visits made to a public library in past four weeks for student i.

$\sum y_i = (68\times0) + (99\times1) + (50\times2) + (45\times3) + (130\times4) + (58\times5) = 1144$

$\sum y_i^2 = (68\times0) + (99\times1) + (50\times4) + (45\times9) + (130\times16) + (58\times25) = 4234$ [2 marks]

Note, if the sum and/or sum of squares is incorrect, follow-through with incorrect values.

Population size $N = 12000$. The simple random sample estimate of the total number of visits made to the library in the past four week $\hat{Y}_{srs} = N\,\bar{y}$ [1 mark] $= 12000 \times \frac{1144}{450} = 30506.6$ [1 mark]

Its estimated variance is $N^2(1 - f)\,\frac{s^2}{n}$. To find the estimated variance underlying this, first calculate

$s_y^2 = \frac{1}{449}\ (4234 - \frac{1144^2}{450}) = \frac{1325.70}{449} = 2.95256$ [1 mark]

Thus the estimated variance of $\hat{Y}_{srs}$ is

$N^2(1 - f)\,\frac{s^2}{n} = 12000^2 \times 0.9625 \times \frac{2.95256}{450}$ [1 mark]

So, the estimated standard error is 953.6186

Assume $\hat{Y}_{srs}$ is normally distributed around population value, Y. An approximate 95% confidence interval for Y is given by $30506.6 \pm 1.96\ se\ (\hat{Y}_R)$. Limits are 28637.5, 32375.7 [2 marks: 1 method 1 answer]

Note, in part (a), small numerical variations in answers may arise depending on the detailed accuracy of a candidate's intermediate working.

(b) (i) Systematic sampling is done from a population whose members are listed in some standard order (such as alphabetical). [1 mark] It consists of choosing a random starting point at the beginning of the list followed by a regular selection of every k th item, where $k = N/n$ = (population size)/(sample size). [1 mark] If there are no trends, a systematic sample might behave as if it were a simple random sample, though strictly speaking it is not. [1 mark]

(ii) A longitudinal survey follows a group ("cohort") or the target population over a period of time. [1 mark] In this survey, a longitudinal survey would be done at fixed points of time for the same respondents. [1 mark]

By visiting the same respondents, the local authority would be able to capture change over time in respondents' attitude to services provided, and to understand how changes in circumstances and life events might impact on participation levels. [1 mark]

Drawback: there is a potential risk of bias due to incomplete follow up or "drop-out" of study participants. [1 mark] Participants who complete to the planned end of study may not be representative of the original target population. [1 mark]

Note, marks will be awarded for other sensible comments.

Question 4

(i) The simple random sample estimate of the total number of banana bunches in the district is $\hat{Y} = N\bar{y} = 289 \times 901.70 = 260591.30$ [2 marks: 1 formula 1 answer]

The standard deviation in this sample is 221.8112. The estimated variance of $\hat{Y}$ is

$$N^2(1-f)\frac{s^2}{n} = 289^2\left(1 - \frac{20}{289}\right)\frac{221.8112^2}{20}$$ [2 marks: 1 formula 1 answer]

So, the estimated standard error is 13829.088. [1 mark]

(ii) (a) The correlation measures the mutual linear association between two quantities. [1 mark]. A necessary condition for the ratio estimator to be useful is that the population correlation coefficient between the two quantities must be 'large' and positive. [1 mark] A value of 0.7737 suggests the number of banana pits and the total number of banana bunches per unit, are highly correlated, making the ratio estimator a reasonable choice. [1 mark]

$\hat{Y} = N\bar{y}$ is an unbiased estimate of the population total Y. [1 mark] $\hat{Y}_R$ is a biased estimator. The bias is likely to be negligible in large samples. [1 mark]

(b) We are given the ratio estimate of the total number of banana bunches in the district is 253760.64 and its estimated standard error is 8550.947. Prefer the ratio estimator, as it gives a considerably smaller standard error. [1 mark]

Let $\hat{Y}_R$ be the ratio estimator of the population total Y. Assume $\hat{Y}_R$ is approximately normally distributed around the population value, Y. An approximate 95% confidence interval for Y is given by $253760.64 \pm 1.96\ se\ (\hat{Y}_R)$. [1 mark] So, the interval is (237000.78, 270520.50) [1 mark]

Note, accept a critical value of $t_{19,5\%}$ in formula for 95% confidence interval.

Were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter 95% of the time. [1 mark] Note that this does not refer to repeated measurements on the same sample, but repeated sampling.

Note, candidates will also get marks if they interpret the confidence interval in terms of a single sample. "There is a 95% probability that the calculated confidence interval (237000.78, 270520.50) encompasses the true value of the population parameter."

(iii) Clustering would seek to identify parts of the district that exhibit most of the characteristics of the whole district (i.e. sub districts). [1 mark] Survey work could then be restricted to a few clusters (maybe only one), instead of having to try to cover the whole region. [1 mark]

Practical issues: how to define the clusters, choosing an appropriate cluster size, number of clusters to be sampled, locating clusters in the district may be time-consuming, etc. [2 marks for covering a range of reasons]

Cluster sampling may require a larger sample than SRS to achieve the same level of accuracy [1 mark] – but cost savings from clustering might still make this a cheaper option. [1 mark]

Note, marks will be awarded for other sensible comments.

Question 4 (supporting information)

$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{901.70}{644.35}$. Hence, $\hat{Y}_R = X\,\hat{R} = 253760.6$

From Cochran, chapter 6, equation 6.13 *, the estimated variance, $V(\hat{R})$, is

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

Re-write $s_{yx} = \hat{\rho}\,s_x\,s_y$.

$$V(\hat{R}) = \frac{(1-\frac{20}{289})}{20 \times 644.35^2}\,\{221.8112^2 + \left(\frac{901.70}{644.35}\right)^2 \times 115.9025^2 - 2 \times \left(\frac{901.70}{644.35}\right) \times 0.7737 \times 115.9025 \times 221.8112\}$$

The estimated variance of $\hat{Y}_R$ is $X^2\,V(\hat{R})$.
Hence estimated SE $(\hat{Y}_R)$ is 8550.9465.

* Cochran and others (later) have suggested that it may be better to use the alternative form of the variance formula based on the sample mean $\bar{x}$, 'to compensate for a tendency for the sample estimate to be too large when $\bar{x}$ is much larger than $\bar{X}$'.