THE ROYAL STATISTICAL SOCIETY

2009 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

MODULE 8

SURVEY SAMPLING AND ESTIMATION

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

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.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Higher Certificate, Module 8, 2009. Question 1

(i) For the small projects, $SE(p_h) = \sqrt{\frac{0.7 \times 0.3}{80}} = 0.0512$.

(ii) (a) The formula used in part (i) is appropriate for sampling from a very large ("infinite") population. In practice, for sampling fractions less than about 5% the effect of ignoring the finite population correction factor $(1 - f)$ is negligible. But for larger sampling fractions, it should be used. In this example, the sampling fractions in the three strata Large, Medium, Small are 50%, 20%, 20%. Thus ignoring the finite population correction factor would be far from negligible. As can be seen, the corrected standard errors are substantially lower than the uncorrected ones.

(b) For the large projects, the corrected standard error is

$$\sqrt{\left(1 - \frac{30}{60}\right)\left(\frac{0.4 \times 0.6}{30}\right)} = 0.0632.$$

(c) The approximate 95% confidence interval is given by

$$\hat{p} \pm 2\,SE_c(\hat{p}) = 0.4 \pm (2 \times 0.0632) = 0.4 \pm 0.1264,$$

i.e. it is (0.274, 0.526). [Note: 2 is used as an approximation to 1.96, the double-tailed 5% point of N(0, 1). 1.96 could of course be used.]

Solution continued on next page

(d) We first need the stratified sampling estimate of the overall proportion, which is

$$\hat{p}_{st} = \frac{1}{N}\sum_{h=1}^{3} N_h \hat{p}_h$$

$$= \frac{1}{660}\left((60\times 0.4) + (200\times 0.6) + (400\times 0.7)\right) = \frac{424}{660} = 0.642\,.$$

The (estimated) variance of this is given by

$$\mathrm{Var}\left(\hat{p}_{st}\right) = \frac{1}{N^2}\sum N_i^{\,2}\mathrm{Var}\left(\hat{p}_h\right)$$

$$= \frac{1}{660^2}\left\{\left(60^2\times(0.0632)^2\right) + \left(200^2\times(0.0693)^2\right) + \left(400^2\times(0.0458)^2\right)\right\}$$

$$= \frac{542.101264}{660^2} = 0.0012445\,,$$

so the (estimated) standard error is 0.035277.

So the approximate 95% confidence interval is given by

$$0.642 \pm (2\times 0.035277) = 0.642 \pm 0.0706\,,$$

i.e. it is (0.571, 0.713).

(iii) Stratification divides the population into groups which may differ from one another but are homogeneous within each group. If it is thought that the population does, or might, divide in this way, stratification is likely to be sensible. It leads to better precision for overall population estimates, and also enables the groups to be studied separately. But a possible disadvantage is that all strata need to be visited as part of the survey and this may be expensive (for example if geographical regions are the strata) and run up against constraints of total cost.

In the present study, it might perhaps be, say, that the larger projects are more difficult to budget for; so stratification would be useful both for the overall examination of the projects and for gaining information on the three strata (sizes). There is no obvious drawback in terms of the overall cost of studying all of the strata.

Higher Certificate, Module 8, 2009. Question 2

[Solution continues on next page]

A target population is the collection of units/people/businesses/hospitals to which we want the results of a survey to apply. In theory we must sample from that population. In practice some parts of it may be very difficult to reach, and the study population that we actually use leaves those parts out. As examples, people who work "unsocial hours" are commonly not available for interviews; and remotely situated agricultural areas are often extremely expensive, in terms of both time and money, to reach, and this can be a major problem for government departments undertaking agricultural surveys.

A sampling frame may be constructed from lists of names/organisations etc that do include any elusive or difficult to reach ones that there may be. Either we use such a frame and select a few extra units to replace members of the original sample that turn out not to be able to be reached, or we use additional sources of information to indicate in advance which are the units that would be difficult to reach and then exclude them from the frame. The "extra units" approach can cope with problems such as "unsocial hours", while a case such as remote farms can be coped with by identifying them through additional information (in this case, geographical information).

Random sampling can be tedious to carry out, especially if refusal rates of selected members are high and/or selected members are not easy to contact. If results are required quickly, for example for a routine requirement such as an opinion poll or a shopping-centre survey of residents' reactions to a topic of local interest, a quota sample may be adequate. When surveys are done regularly, any apparent very sharp changes in opinion (e.g. support for political parties) can either be explained or checked against similar surveys in other places. Quota samples do aim to cover the whole range of social/age/occupation(/etc) categories – which cannot be guaranteed in simple random sampling – even though the members of the quota sample are not chosen randomly from a well-defined population. Interviewers in quota sampling should be properly trained and their results scrutinised regularly.

Any item which can be measured (or, sometimes, estimated) and which may help to explain the quantity of interest, y, that is to be analysed is potentially useful. For example, in a social survey taking place every year, the previous year's value x of y on the same unit is sometimes a good predictor of y. As another example, consumption of food items or use of power (gas, electricity) or leisure spending (sport, holidays) or simply current income can all be useful in household surveys.

It is often reasonable to suppose that y, the quantity of interest, varies roughly in proportion to x, the ancillary variable. An exact relationship of this kind would give $y_i = Rx_i$ for each pair of observations (y_i, x_i) in the bivariate population, where R is the

population ratio of the totals (Y and X) or of the means ($\overline{Y}$ and $\overline{X}$), i.e. $R = Y / X = \overline{Y} / \overline{X}$. R can be estimated by the sample ratio $r = \Sigma y / \Sigma x = \overline{y} / \overline{x}$ calculated using the observations in a random sample.

Suppose now that the object of the survey is to obtain an estimate of the population total for the y variable, i.e. of Y. Assuming that the population total X for the ancillary variable is known, we can use the "ratio estimator" $\hat{Y}_{ratio} = rX = X.\Sigma y / \Sigma x$.

Similarly, if we want to estimate the population mean for y, i.e. $\overline{Y}$, we can use $\overline{Y}_{ratio} = r\overline{X} = \left(\overline{X} / \overline{x} \right) \overline{y}$.

[Notice that this illustrates an intuitive appeal of ratio estimators: if it happens that the sample mean $\overline{x}$ is (say) smaller than the known value $\overline{X}$, and bearing in mind the approximate proportionality between y and x, the estimator gives an upward adjustment of $\overline{y}$ in estimating $\overline{Y}$.]

The proportionality relationship leading to use of ratio estimators is, in effect, linear regression through the origin. It often happens that there is a roughly linear relationship between y and x which does not pass through the origin. Similar arguments lead to regression estimators for the population total and mean of the y variable. A simple example often arises in agricultural surveys: it is often the case that there are established well-known relationships between a measurement that can be taken early in the season and the amount of the crop that can be expected at harvest (e.g. length of raspberry cane and the eventual crop).

Particularly in medical work, but also elsewhere (e.g. educational studies), it is sometimes necessary to follow the effect of a treatment or a drug on the *same* group of patients for an extended period. This is a longitudinal survey and enables us to study the dynamic development of members of the population. In a cross-sectional survey, a study is made using a single sample at a fixed time point (or perhaps over a fixed and fairly short time period) with the aim of investigating the characteristics of the population at that particular time. Often cross-sectional surveys cover widely defined populations (men, women, different ages, different ethnic backgrounds, etc), whereas longitudinal surveys may perhaps follow a more narrowly defined population over an extended time period.

[There are many relevant points that could be made in response to this question, in addition to those set out briefly above. Other examples are also obviously possible. In the examination, credit was given for any relevant points and for good examples.]

Higher Certificate, Module 8, 2009. Question 3

We ignore finite population corrections in this question as the sampling fractions are small (see solution to question 1(ii)(a) above). We use N(0, 1) rather than a *t* distribution in forming the confidence intervals as the sample sizes are reasonably large. For this question, we use 1.96 rather than the approximation of 2 as used in question 1, but there can be no great objection to use of 2.

(i) (a) The required 95% confidence interval is given by $40 \pm (1.96 \times 10/\sqrt{50})$, i.e. it is (37.2, 42.8) (hours).

(b) The half-width of this interval is 2.77 (to 2 decimal places) hours, but it was required to be 2 hours. To achieve a half-width of 2 hours, we require a sample of size n where n is given by $1.96 \times 10/\sqrt{n} \leq 2$. This gives $n \geq (5 \times 1.96)^2 = 96.04$. So we would need an achieved sample size of 97.

(c) For all the students, the sample mean is

$$\bar{y} = \{(800 \times 45) + (2000 \times 40) + (600 \times 55)\} / 3400 = 43.82 \text{ hours}.$$

The underlying variance is

$$\text{Var}(\bar{y}) = \sum_h \left(\frac{N_h}{N}\right)^2 \text{Var}(\bar{y}_h)$$

which we estimate by

$$\text{Var}(\bar{y}) = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{{s_h}^2}{n_h}$$

$$= \left(\frac{800}{3400}\right)^2 \frac{15^2}{40} + \left(\frac{2000}{3400}\right)^2 \frac{10^2}{50} + \left(\frac{600}{3400}\right)^2 \frac{10^2}{40} = 1.0813$$

and so $\text{SE}(\bar{y}) = 1.04$ (hours). So the required 95% confidence interval is given by $43.82 \pm (1.96 \times 1.04)$, i.e. it is 41.8 to 45.9 hours.

Solution continued on next page

(ii) Proportional allocation chooses the stratum sample sizes n_h in the same ratio as the stratum population sizes N_h. For a sample of total size 200 with the population strata sizes in the ratio 8:20:6, the n_h will be 47.06 [= (8/34) × 200], 117.65 and 35.29 respectively, so we take stratum sample sizes 47, 118, 35 respectively.

Optimal allocation aims to minimise, for given total sample size n, the variance of an overall population estimate. For this, n_h has to be proportional to $N_h s_h$. The values of this are 12000, 20000 and 6000 respectively. So for a sample of total size 200, the n_h will be 63.16 [= (12000/38000) × 200], 105.26 and 31.58 respectively, so we take stratum sample sizes 63, 105, 32 respectively.

As stated, optimum allocation improves the precision of overall estimates of population parameters. This will be important here, as the standard deviation for the Faculty of Technology is considerably more than for the other two faculties – which optimum allocation has allowed for by specifying a larger sample in that faculty than would have been the case under proportional allocation.

Higher Certificate, Module 8, 2009. Question 4

(i) Bias is very likely. There are many groups in the local population who would be unlikely to look at a local government website often, or indeed would be unlikely to look at it at all. For instance, older people do not usually spend much (or any) time at a computer, but these are likely to use public transport and quite unlikely to cycle. It is also possible for groups, either in favour or against, to organise either a "yes" or a "no" vote in these circumstances, and the total response is unlikely to be large enough to withstand this. In short, very little reliance, if any at all, can be placed on the figure.

(ii) First, it is necessary to decide on the population of interest. It might be all the residents in the area; or only those subgroups who use the present cycle track; or existing users of public transport who might want an improvement on the present service; and there are many other possibilities. It would also be necessary to consider whether there might be different views according to what time of day people wish to travel; whether a factor might be how far they need to travel; and issues such as whether cyclists might use the new tram service instead of cycling to work.

A sampling frame could probably be constructed from the electors lists (but some people may have exercised their option of not having their names on the version of the list that is available to the general public), or, since it is an official body carrying out the survey, a list of households may be used.

Any information known to the City Council about the characteristics of the populations of various areas in the city should be considered as bases for stratification – for example, the age distributions (are the residents in an area predominantly retired, predominantly young with families, predominantly middle-aged, etc), and various indicators of socio-economic status (e.g. are they predominantly owner-occupiers). If any such known characteristics are not used in constructing the sampling frame, they should be verified by questions on the response form. Standard questions on age etc should in any case be included; post-hoc stratification may be found to be necessary when the results are being studied.

A questionnaire to be used by interviewers visiting people at home might be the best method, if resources allow. Revisits would be necessary for those who could not be contacted on an interviewer's first visit; the budget for the survey must allow for these. Refusals should be minimised by careful training of the interviewers so that they can explain the project fully; for example, they should be able to answer questions such as the anticipated frequency and hours of operation of the tram service and how long it would take to reach key destinations. It must be recognised that some respondents may say that the change will not affect them, and they should not be pressed to say yes or no to the project. Careful attention to these matters should minimise bias.

[As with question 2, credit was given in the examination for all relevant points and justifications of viewpoint. But a quota sample does not seem a proper basis here, and answers based on quota samples gained less credit.]