THE ROYAL STATISTICAL SOCIETY

2010 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, 2010. Question 1

(i) In units of £000, for the "fewer than 5" businesses we are given $\bar{y}$ = 3.5 and SE($\bar{y}$) = $s/\sqrt{n}$ = $3/\sqrt{100}$ = 0.3.

Hence a 95% confidence interval for the true mean advertising expenditure is $3.5 \pm (1.96 \times 0.3) = 3.5 \pm 0.588$, i.e. 2.912 to 4.088 thousand pounds.

(ii) We can be 95% certain that this interval does contain the true value of the mean advertising expenditure for the whole population of small businesses in Kingstown that have 5 or fewer employees.

(iii) Again using units of £000, a confidence interval of width no greater than 0.5 requires that the achieved sample size, n, satisfies $1.96 \times 3/\sqrt{n} \leq 0.25$.

Thus $\sqrt{n} \geq 1.96 \times 3/0.25 = 23.52$, i.e. $n \geq 553.19$.

Rounding up so as to be sure that the interval is of width no greater than 0.5, we take n = 554.

(iv) We again work in units of £000.

The overall mean is

$$\bar{\bar{y}} = \frac{1}{18}\left((10 \times 3.5) + (6 \times 10.0) + (2 \times 35.0)\right) = \frac{165}{18} = 9.167 .$$

The estimated variance of $\bar{\bar{y}}$ is

$$\sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{s_h^{\,2}}{n_h}\right) = \left(\frac{10}{18}\right)^2 \left(\frac{9}{100}\right) + \left(\frac{6}{18}\right)^2 \left(\frac{25}{60}\right) + \left(\frac{2}{18}\right)^2 \left(\frac{225}{15}\right)$$

$$= 0.02777 + 0.04630 + 0.18519$$

$$= 0.25926,$$

and so SE($\bar{\bar{y}}$) = $\sqrt{0.25926}$ = 0.5092.

Hence a 95% confidence interval for the true mean advertising expenditure for all small businesses in Kingstown is $9.167 \pm (1.96 \times 0.5092) = 9.167 \pm 0.998$, i.e. 8.169 to 10.165 thousand pounds.

Solution continued on next page

(v) A supplementary variable must be measurable for each business. It should be a variable that is likely to be quite closely related to (correlated with) y. Possible variables include the company's profits in the previous year or its turnover – both these are likely to be useful. The number of employees is also likely to be a useful variable as (for example) an increase in this is likely to indicate increased activity and therefore a need for more advertising. Wholly new lines of work might also be modelled, perhaps using dummy (0, 1) variables, as these would also be likely to lead to increased advertising. Any special promotions might also be modelled in a similar way.

[In the examination, credit was given for all valid and relevant comments and suggestions for possible variables.]

Higher Certificate, Module 8, 2010. Question 2

[Solution continues on next two pages]

Part (a)

(i) In a survey, a measurement is made on each unit in the sample. In the example in this question, it is a Yes/No vote on the question asked. The totals of "Yes" and "No" votes in the sample are found and thus the proportion of "Yes" votes. This proportion is used to estimate the same proportion in the whole population that is being studied. A good sampling method should lead to an estimate that is (in some sense) near to the (unknown) true population value. If a sampling method leads to consistently poor estimates – too high or too low systematically – every time it is used, it is said to be biased.

(ii) Many criticisms can be made. Key ones concern the lack of definition of the population under study, the lack of any sensible sampling frame, the very low level of response, the leading and over-simplistic nature of the question and the inability to capture important supplementary information. All in all, there is likely to be heavy upward bias in the reported estimate of the "Yes" proportion.

The lack of definition of the population under study is a serious deficiency. Is it the local population only? If this is the intention, exactly what does "local" mean? How far does the local newspaper's circulation reach? Even with a very limited definition of the population, the response rate is extremely low (an achieved sample size of 150 whereas the population of the town is 25 000). Little reliance can be placed on the resulting estimate.

Alternatively, is it intended to include the drivers (and passengers?) of vehicles entering or passing through the town? If so, there is even more concern about the small achieved sample size. This is perhaps unsurprising, as such people, particularly in through vehicles, are not likely to see the local newspaper. They may perhaps be expected to be in favour of the by-pass road so as to reduce or avoid possible traffic problems in the town, but the survey cannot claim to have revealed that.

Thus the sampling frame for the survey is seriously incomplete – in fact it can hardly be said to exist at all. Respondents are limited to those who happened to read the newspaper and saw the question – and then further limited to those who are willing and able to send a text message to the newspaper.

Another key point concerns the nature of the question asked. It is highly "leading". It begs the answer "Yes" – how can a respondent possible think that dangerous traffic should be allowed to remain in the town? It is even worse in that the possible consequential damage to the area of unspoilt beauty is not even mentioned – some respondents may be unaware of this and feel that they are simply voting for an uncontroversial by-pass. There is also a more subtle, but still important, point arising from the implicit labelling of all

traffic as "dangerous". A genuine response that is intended to imply that only the dangerous traffic should be kept out of the town would be misinterpreted as agreeing that all traffic should be kept out.

On the other hand, there may well be respondents who are in favour of a road to divert (dangerous) traffic away from Chiptown, but not the particular road that has been proposed. The question does not give any scope for a proper reply from such respondents.

There is also the point that there is no facility for an intermediate "don't know" response, so only those who feel very strongly about the issues of freeing-up the town centre and/or passing through an area of unspoilt beauty are likely to vote. This would be likely to lead to the familiar problem of capturing extreme opinions that may not be representative of majority views.

There is no facility for indicating age group – different age groups may well have markedly different views – or any other accompanying information that might affect opinions. Examples are how near to the town a respondent lives, does the respondent usually move around the town on foot, does the respondent own a car, and many other possibilities. The apparent simplicity of the unelaborated Yes/No response is far too simple to be a worthwhile indicator.

Further important supplementary questions that should be considered depend on the nature of the town. For example, does it have a railway station, shopping centre, business park? If people come to the town for these reasons, car park space is likely to be all they are interested in, by-pass or not. Does it have a rush-hour, which may be made worse by through traffic? How much through traffic is there? How much damage would the road do to the area of unspoilt beauty?

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

Part (b)

In cluster sampling, the whole of Chiptown is divided into groups, called clusters, each of which is expected to be roughly similar to the whole town in respect of the opinion being surveyed. In particular, each group should as far as possible capture all the variability that is represented in the whole town. The areas described in the question may be suitable to be used as these groups, if it is felt that each area is

roughly similar to the whole town.

Assuming suitable clusters have been defined, cluster sampling recognises that only a limited number of them need to be surveyed. So a sample of clusters is selected for study. This sample is itself usually selected by simple random sampling from all the clusters. Typically it will consist of only a few clusters, and it is quite common in practice for it to consist of only one. The survey is then carried out in these selected clusters. This is usually administratively convenient, especially if the survey is to be conducted by face-to-face interviewing; much less time is likely to be needed for this. There is also the point that the sampling frame is limited to the selected clusters; an up-to-date list of sampling units in other clusters is not needed.

Cluster sampling thus lends itself to direct interviewing. In the present example, this would be likely to give very much better results as a proper questionnaire could be developed.

Various extensions of this basic concept of cluster sampling may be appropriate in this example. For example, it may be thought that areas near the centre of Chiptown (perhaps being older?) exhibit differences from areas in inner suburbs which are themselves different from areas further away in the outer suburbs. If so, it might be appropriate to select a few clusters in the centre, a few in the inner suburbs and a few in the outer suburbs and then combine the results in the eventual analysis – in effect, a combination of cluster and stratified sampling.

Higher Certificate, Module 8, 2010. Question 3

(i) Stratification enables us to study each of the sub-populations separately as well as studying the population as a whole. If the strata have been chosen sensibly, the variances of estimators for the whole population will be less under stratified sampling than they would have been under simple random sampling.

In the present survey, it seems reasonable to take urban and rural areas as strata. It might be thought that the proportions could differ considerably between the two and, if this is so, reduction in variance of estimators compared with simple random sampling should be achieved. Whether it is so or not, it seems useful to obtain information about urban and rural areas separately as well as about the country as a whole.

(ii) Using the notation in the question, the standard error of each estimated proportion p_h is estimated by $\mathrm{SE}(p_h) = \sqrt{\dfrac{p_h(1-p_h)}{n_h}}$ and the required 95% confidence interval is given by $p_h \pm 1.96\,\mathrm{SE}(p_h)$.

Thus for urban areas the interval is $0.3 \pm (1.96 \times \sqrt{(0.3 \times 0.7/300)}) = 0.3 \pm 0.052$, i.e. 0.248 to 0.352.

For rural areas the interval is $0.6 \pm (1.96 \times \sqrt{(0.6 \times 0.4/150)}) = 0.6 \pm 0.078$, i.e. 0.522 to 0.678.

(iii) We have $p_r - p_u = 0.3$ with estimated underlying variance

$$\frac{p_r(1-p_r)}{n_r} + \frac{p_u(1-p_u)}{n_u} = \frac{0.6 \times 0.4}{150} + \frac{0.3 \times 0.7}{300} = 0.0023.$$

Thus the test statistic for examining whether the true difference in proportions is zero is

$$\frac{0.3}{\sqrt{0.0023}} = 6.255.$$

This is extremely highly significant as an observation from N(0, 1), so we have overwhelming evidence against the null hypothesis of equal proportions.

Solution continued on next page

(iv) The estimated total number is $(8000 \times 0.3) + (4000 \times 0.6) = 4800$.

The estimated underlying variance for the overall proportion is

$$\sum_h \left(\frac{N_h}{N}\right)^2 \mathrm{Var}(p_h) = \left(\frac{8000}{12000}\right)^2 \frac{0.3\times 0.7}{300} + \left(\frac{4000}{12000}\right)^2 \frac{0.6\times 0.4}{150} = 0.0004889.$$

Thus the estimated underlying variance for the overall total is

$$(12000)^2 \times 0.0004889 = 70400.$$

Thus the required 95% confidence interval for the overall total is $4800 \pm (1.96 \times \sqrt{70400}) = 4800 \pm 520(.05)$, i.e. 4280 to 5320.

(v) 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 500 with the population strata sizes in the ratio 2:1, the n_h will be 333.33 [= (2/3) × 500] and 166.67 [= (1/3) × 500] respectively, so we take stratum sample sizes 333 and 167 respectively.

Optimal allocation (in the simplified present situation where costs are not considered, often referred to as Neyman allocation) aims to minimise, for a given total sample size n, the variance of an overall population estimate, in this case the estimate of the overall total. Thus it takes into account the variability in each stratum, increasing the sample size in more variable strata. For this, using the estimates from 2009 and noting that the sample sizes are large, n_h is taken as proportional to $N_h s_h$ where $s_h = \sqrt{(n_h p_h (1 - p_h))}$.

The values of $N_h s_h$ are $8000\sqrt{(300 \times 0.3 \times 0.7)} = 63498.03$ for urban areas and $4000\sqrt{(150 \times 0.6 \times 0.4)} = 24000.00$ for rural areas.

Thus for a sample of total size 500, the n_h will be 362.85 and 137.15 respectively, so we take stratum sample sizes 363 and 137 respectively.

(i) The headline is wrong.

It is incorrect, alarmist even, in giving the distinct impression that this is a statement about the entire population of businesses in the region covered by the Southern England Chamber of Commerce. In fact it refers only to a sample survey of those businesses that are members of the Chamber. There is no information as to whether or not it reflects the business community as a whole, including businesses that are not members. Further, there are only 100 respondents, which seems a very small achieved sample size for what is presumably a large region with a very large number of businesses that are members of the Chamber. So there are multiple problems of non-coverage and non-response.

It is also incorrect and alarmist in reporting that "most businesses will be shedding jobs" whereas in fact the information is that 40% anticipate increasing their workforce. Those that are not increasing are not necessarily decreasing – they may anticipate remaining stable in terms of employment.

There are further issues regarding general lack of detailed information.

The 40% figure is itself subject to considerable inaccuracy as an estimate of the population proportion, because of the small achieved sample size as well as the problems of non-response. There should be some indication of its accuracy, for example by giving a confidence interval (which may well be found to be so wide as to span 50%).

There are also issues concerned with the sizes of the businesses. For example, a large employer moving its production elsewhere could well have a very serious impact on the region as a whole; whereas a few small employers closing down might have little impact beyond their immediate workforces. A related but separate point is that businesses, even of the same size, may have very different plans – a large contraction (or expansion) could have regional consequences, whereas very small changes might have little impact.

In addition, there may be various amounts of part-time working and this may or may not be expected to change during the next 12 months. There is no information as to whether the survey has captured this.

[In the examination, credit was again given for all valid and relevant comments.]

Solution continued on next page

(ii) It is first necessary to define whether the population consists of all businesses within the region or only those that are members of the Chamber of Commerce. The latter cannot be criticised as such provided the results are properly reported as only pertaining to the membership. There may be issues as to exactly what constitutes a "business" and as to whether a business is regarded as in the region (for example, it may be merely a very small department of a large national or even multinational company).

If the survey is being limited to Chamber members, a sampling frame is immediately available as there will be a list of them. If a wider population is being studied, it may be possible to use an official (e.g. government) list of some kind.

Stratification will almost certainly be sensible, so as to obtain information about distinct strata and, hopefully, to obtain more precise estimates for the population as a whole. Several criteria for stratification suggest themselves; examples are size, sector, geographical location.

A postal survey is a possibility, but sometimes it is hard to find the right person to send it to. Reminders are also often needed; even so, response rates may be poor. The questionnaire would need to be pre-designed and, if at all possible, should be piloted (perhaps using a small number of Chamber members).

An alternative is a face-to-face interview with an appropriate person in each chosen business. Though more resource-intensive, this may give better results. More questions can be asked and there is more scope for opinions to be sought. Making the initial contact with each chosen business may be difficult and is probably best attempted by telephone.

The analysis should produce results for each of the strata as well as for the population as a whole, with a careful definition of what "the population" is. The results should not be limited to point estimates but also include confidence intervals. There should be proper discussion of any problems of non-response. The analysis should include the detailed information and also suitable summaries, and an accompanying report should highlight the salient outcomes and suggest possible explanations.