THE ROYAL STATISTICAL SOCIETY

2007 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

(MODULAR FORMAT)

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, 2007. Question 1

Part (a)

We have population size N = 800 and sample size n = 25. $\bar{x}$ = 8000, s = 3000. An underlying Normal distribution for the amount of loan is assumed.

(i) Without a finite population correction, a 95% confidence interval for the true population amount of loan $\bar{X}$ is given by $\bar{x} \pm ts/\sqrt{25}$ where t is the double-tailed 5% point of t_{24}, i.e. 2.064.

So the interval is given by 8000 ± (2.064×3000)/5, i.e. 8000 ± 1238.4, i.e. it is (6762, 9238).

The sampling fraction is $f = n/N$ = 25/800 = 0.03125. It is usual to ignore the finite population correction when f is less than 5%, as is the case here. Using it would only change the result marginally, for we would then have that Var($\bar{x}$) is estimated by $(1-f)s^2/n$ [instead of just s^2/n]. Thus, as $1 - f$ = 0.96875, the interval becomes 8000 ± (√0.96875 × 1238.4), i.e. 8000 ± 1218.9, i.e. (6781, 9219).

[Note that use of N(0, 1) rather than t_{24}, i.e. 1.96 instead of 2.064, would also make very little difference.]

(ii) Here the researcher requires $ts/\sqrt{n}$ to be ≤ 1000. Inserting s = 3000 and t = 2.064 gives $\sqrt{n} \geq 6.192$, so that $n \geq 38.34$. So the smallest achieved sample size is 39.

[Using the finite population correction, and/or using 1.96 instead of 2.064 (or even using 2 as a working approximation), will make only a slight difference.]

Part (b)

Here we have a sample of size n = 40 and the estimated proportion $\hat{p}$ = 25/40 = 0.625. The underlying variance of $\hat{p}$ is estimated by (0.625)(0.375)/40 = 0.005859.

So an approximate 95% confidence interval for the true population proportion is given by 0.625 ± (1.96 × √0.005859), i.e. 0.625 ± 0.150, i.e. (0.475, 0.775) [or, in percentage terms, 47.5% to 77.5%].

Part (c)

In (a), the response rate is only 50%, so some follow-up reminder asking for a reply would be advisable. Some refusals are to be expected as the information is – or should be – confidential. Perhaps a larger sample should be selected to start with.

In (b), the incentive might have biased the results by encouraging replies from those who think they are "hard up" (though this may depend on which store the vouchers were for), so those who do not have loans could be under-represented in the sample.

Higher Certificate, Module 8, 2007. Question 2

[As in all questions of this nature, credit is given for all relevant comments and ideas.]

Part (i)

The website of a small provincial airport is most unlikely to attract viewers who represent the population of all who would be affected by such a change. Unless they are very keen on air travel in general, viewers of the website are likely to already be travellers and, worse still from the point of view of potential bias, travellers or potential travellers already likely to be using this airport. Viewers are very unlikely to have simply visited the website on a casual basis or for pure curiosity. More likely, viewers will be wanting to know if there is a flight they could use for their journeys; but they may not live near enough to the airport, or in the right (or "wrong"!) direction, to be affected in their day-to-day lives by the proposed major changes.

The local council, on the other hand, will want the views of all in their area who may be affected by factors that would be involved in an expansion, such as increase in road traffic and increase in noise. The council will want to publicise full details of matters such as flight paths, likely changes in local transport needs and any other items within their planning control. A simple poll on a relatively obscure website cannot achieve this.

Part (ii)

The whole council area should be covered. If possible, areas in neighbouring councils that are likely to be affected should be covered as well. Voters' lists could be used as a basis for a sampling frame of households. Stratification will be necessary, by factors such as proximity to the airport and relation to flight paths.

If there are businesses, industrial estates or shopping areas near to the site, they should be covered as well, treated as a separate stratum (or strata). Even if they are not run by residents in the area, they will be affected.

All strata should of course be covered, as is ensured by stratified sampling (this is stratified sampling, not cluster sampling).

Topics mentioned in part (i) should be covered, with proper explanations of likely new buildings, traffic expansion, etc. Opportunity should be given for opinions to be expressed as well as simple yes/no answers. Information on each sampled household (or other unit) should also be collected in the questionnaire – it is likely that some people will see local job opportunities while others will expect disruption to their lives, and this is likely to depend on household structure, age and present occupations. Differences between strata should be studied in the analysis.

Higher Certificate, Module 8, 2007. Question 3

Part (a)

(i) The overall mean is $\frac{1}{10000}\left\{(1200\times4)+(5000\times6.5)+(3800\times8)\right\}$ = 6.77 days.

(ii) The estimated standard errors of the mean in each pay grade are

for A: $\sqrt{1.5/40}$ = 0.194 days

for B: $\sqrt{2.5/40}$ = 0.25 days

for C: $\sqrt{3.0/40}$ = 0.274 days.

(iii) For the overall mean, the estimated variance is

$$\sum_h \left(\frac{N_h}{N}\right)^2\left(\frac{s_h^{\,2}}{n_h}\right) = \left\{\left(\frac{1200}{10000}\right)^2\left(\frac{1.5}{40}\right)+\left(\frac{5000}{10000}\right)^2\left(\frac{2.5}{40}\right)+\left(\frac{3800}{10000}\right)^2\left(\frac{3.0}{40}\right)\right\}$$

$$= 0.026995,$$

so the estimated standard error is √0.026995 = 0.1643 days. So the required approximate 95% confidence interval is given by 6.77 ± (1.96×0.1643), i.e. 6.77 ± 0.322, i.e. it is (6.45, 7.09) days.

Part (b)

(i)

Stratum	$N_h s_h$	$120N_h s_h/\Sigma N_h s_h$
A	$1200\sqrt{1.5} = 1469.7$	11.05
B	$5000\sqrt{2.5} = 7905.7$	59.45
C	$3800\sqrt{3.0} = 6581.8$	49.50
	$\Sigma N_h s_h = 15957.2$	

The sample sizes n_h should therefore be taken as 11, 59 and 50 respectively.

Optimum allocation reduces the standard errors of estimates for the whole population by sampling more intensively in the more variable strata. As we go down the pay grades the variability does increase, so optimal allocation should be beneficial.

(ii) Proportional allocation would set the n_h in the ratio 1200 : 5000 : 3800, so a sample of total size 120 would use values 14, 60, 46 respectively. This is very similar to the optimal allocation so precision is likely to be very similar either way.

Higher Certificate, Module 8, 2007. Question 4

(i) Stratification is useful when a population can be split into several distinct groups which it is thought may be different from each other in terms of the characteristic that is the subject of the survey and/or when information is required about each group as well as about the population as a whole. Estimates for the population as a whole should be more precise than if the sampling were at random over the whole population.

In this case the Finance Division is of particular interest and is thought to be different from the others, so it is useful to have this as one stratum. If the other divisions are not thought to be different from one another, they could all form one other stratum and no further stratification would be needed.

(ii) For the Finance Division, we have a sample of size $n = 100$ and the sample proportion is 0.4. The underlying variance is estimated by $(0.4)(0.6)/100 = 0.0024$.

So an approximate 95% confidence interval for the true population proportion is given by $0.4 \pm (1.96 \times \sqrt{0.0024})$, i.e. 0.4 ± 0.096, i.e. (0.304, 0.496) [or, in percentage terms, approximately 30% to 50%].

For the other divisions, we have a sample of size $n = 100$ and the sample proportion is 0.2. The underlying variance is estimated by $(0.2)(0.8)/100 = 0.0016$.

So an approximate 95% confidence interval for the true population proportion is given by $0.2 \pm (1.96 \times \sqrt{0.0016})$, i.e. 0.2 ± 0.078, i.e. (0.122, 0.278) [or, in percentage terms, approximately 12% to 28%].

These are the confidence intervals for the proportions of staff actively seeking work outside. They do not overlap, suggesting that there is a large difference between the two proportions; in part (iv), we proceed to a formal test.

(iii) The estimated total is $0.4 \times 1000 = 400$, and using the calculations in part (ii) an approximate 95% confidence interval for the total actively seeking work outside is (304, 496).

(iv) The null hypothesis to be tested is that there is no difference between the proportions p_F (Finance) and p_O (other). The estimated difference is $0.4 - 0.2 = 0.2$, and the underlying variance is estimated by

$$\frac{0.4\times0.6}{100}+\frac{0.2\times0.8}{100} = 0.004\,.$$

Thus the value of the test statistic is $0.2/\sqrt{0.004} = 3.162$, which we refer to N(0, 1). This is very highly significant and we have very strong evidence against the null hypothesis, so we may formally reject it.

Solution continued on next page

(v) In general it is not good policy to have names on questionnaires because it can cause bias in some people's replies. Also, some people may refuse to reply at all.

Perhaps the administrators hope to use the survey to discover actual problems, either for individuals or for certain groups of employees, and provide help in overcoming these, perhaps using other information that is already available on an employee database.