THE ROYAL STATISTICAL SOCIETY

2008 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, 2008. Question 1

(i) (a) Let p be the true proportion of all small hotels in the county which use the local tourist information centre to obtain bookings. The sample estimate is $\hat{p} = 45/60 = 0.75$. The variance of $\hat{p}$ is estimated as

$$\frac{\hat{p}(1-\hat{p})}{n} = \frac{0.75 \times 0.25}{60} = 0.003125$$

so the standard deviation of $\hat{p}$ is estimated by $\sqrt{0.003125} = 0.0559$.

Hence an approximate 95% confidence interval for p is given by $0.75 \pm (1.96 \times 0.0559)$, i.e. it is (0.64, 0.86). With 95% probability of being right, we can say that the range 64% to 86% covers the true percentage of hotels which use the tourist information centre.

(b) The expression for the variance of $\hat{p}$ used above is appropriate for an infinite population. Theory shows that, when the population is finite, the variance as obtained for an infinite population should be multiplied by the factor $(1 - f)$, where $f = n/N$ where n is the size of the sample and N is the size of the population. f is called the sampling fraction and $1 - f$ the finite population correction factor. Here, the total number of hotels in the county is approximately $N = 6000$ which, though large, is finite. However, the sample size $n = 60$ is only 1% of N, so $f = 0.01$ and $1 - f = 0.99$. So, in this survey, the finite population correction would make very little difference to the estimated variance or the confidence interval and so it need not be used.

[Note: the finite population correction is ignored in part (ii).]

(ii) (a) The required interval has limits $\bar{x} \pm 1.96s/\sqrt{n}$ in the usual notation. Thus it is (in £) $50 \pm (1.96 \times 15/\sqrt{60})$, i.e. £$(50 \pm 3.80)$, or £46.20 to £53.80.

(b) For a 99% interval, we are more confident (99% instead of 95%) that the interval contains the true value of the mean. Therefore the interval must be wider. Mathematically, the factor 1.96 is replaced by 2.576.

(c) For the precision required, $1.96s/\sqrt{n}$ must be ≤ 3, i.e. $1.96 \times 15/\sqrt{n} \leq 3$. Hence $\frac{1.96 \times 15}{3} \leq \sqrt{n}$, or $n \geq \left(\frac{1.96 \times 15}{3}\right)^2 = 96.04$. The minimum size of sample is therefore 97.

Higher Certificate, Module 8, 2008. Question 2

(i) (a) The limits of the interval are "Sample mean ± (t × SE of mean)", where t is the two-tail 5% point for t_{19} which is 2.093 (19 df since the sample size is 20 for this division). Thus the interval is 46 ± (2.093 × 2.5), i.e. 46 ± 5.23 or 40.8 to 51.2 hours.

(b) The confidence interval lies entirely above the target of 40 hours. This suggests that the Internal Communications division is not meeting the target value, in that staff appear to be working longer hours on the whole.

(ii) (a) For all employees, the sample mean is

$$\bar{y} = \{(300 \times 41) + (400 \times 40) + (100 \times 46)\} / 800 = 41.125 \text{ hours}.$$

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

$$\left(\frac{300}{800}\right)^2 (1.5)^2 + \left(\frac{100}{200}\right)^2 (2.0)^2 + \left(\frac{100}{800}\right)^2 (2.5)^2 = 1.4141$$

and so $\text{SE}(\bar{y}) = 1.189$ (hours). Since this is based on a combined sample size of 120 observations, we may reasonably use N(0, 1) instead of t in calculating the interval, which is therefore obtained as 41.125 ± (1.96 × 1.189), i.e. 41.125 ± 2.33 or 38.8 to 43.5 hours.

(b) This interval contains the target of 40 hours, so there is no evidence to suggest that the target of 40 hours is not being met over all employees.

(iii) The small IC division seems very likely to be working above target, whereas the two other much larger divisions do not (40 would be comfortably inside any interval calculated for IM). Overall the target is apparently being met, but quoting the overall figure obscures differences between divisions.

Solution continued on next page

(iv) Stratification subdivides a population into groups; each group is reasonably homogeneous within itself, but systematic differences between groups may exist. With stratification, the precision of overall population estimates is increased and information on the separate groups can be obtained. In this survey, the divisions would form the groups and there are apparent differences between them, both in mean and in variability.

(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 120 with the population strata sizes in the ratio 3:4:1, the n_h will be 45, 60 and 15 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 $N_h s_h$. We need to find the value of s_h for each stratum. Each is simply given by $\sqrt{n_h}\, SE(\bar{y}_h)$, so they are respectively 10.61, 14.14 and 11.18. Thus the values of $N_h s_h$ are 3183, 5656 and 1118, with a total of 9957. Therefore, with n = 120, the values of n_h are 38.36, 68.17 and 13.47, so we take them as 38, 68 and 14.

Higher Certificate, Module 8, 2008. Question 3

We ignore finite population corrections in this question as the sampling fractions are very small (see solution to question 1(i)(b) above).

(i) (a) Let p_F be the true proportion of female residents who approve of the proposed supermarket scheme. The sample estimate is $\hat{p}_F = 32/100 = 0.32$. The variance of $\hat{p}_F$ is estimated as

$$\frac{\hat{p}_F(1-\hat{p}_F)}{n_F} = \frac{0.32 \times 0.68}{100} = 0.002176$$

so the standard deviation of $\hat{p}_F$ is estimated by $\sqrt{0.002176} = 0.0467$.

Hence an approximate 95% confidence interval for p_F is given by $0.32 \pm (1.96 \times 0.0467)$, i.e. it is (0.23, 0.41).

The analysis for p_M, the true proportion of male residents who approve of the proposed supermarket scheme, is similar. The sample estimate is $\hat{p}_M = 65/100 = 0.65$, and the variance of $\hat{p}_M$ is estimated as

$$\frac{\hat{p}_M(1-\hat{p}_M)}{n_M} = \frac{0.65 \times 0.35}{100} = 0.002275$$

so the standard deviation of $\hat{p}_M$ is estimated by $\sqrt{0.002275} = 0.0477$.

Hence an approximate 95% confidence interval for p_M is given by $0.65 \pm (1.96 \times 0.0477)$, i.e. it is (0.56, 0.74).

These confidence intervals do not overlap, by a large margin. They strongly suggest that there is a very large difference between the views of the sexes.

(b) For the whole town,

$$\hat{p} = \{(28000 \times 0.32) + (21000 \times 0.65)\} / 49000 = 0.46,$$

and the variance of $\hat{p}$ is estimated as

$$\left(\frac{28000}{49000}\right)^2 \times 0.002176 + \left(\frac{21000}{49000}\right)^2 \times 0.002275 = 0.0011284.$$

The standard deviation of $\hat{p}$ is estimated by $\sqrt{0.0011284} = 0.0336$. Thus the required 95% confidence interval is $0.46 \pm (1.96 \times 0.0336)$, i.e. it is (0.39, 0.53).

Solution continued on next page

(ii) (a) Bias arises when the sample actually taken differs substantially from the whole population in respect of the characteristic(s) being measured. This can arise from faulty technique, refusal to participate, poor (or no) sampling frame, etc.

(b) He appears to be thinking that the average ages of the two groups, male and female, questioned would be very different. Although females live longer than males on the whole, it is unlikely that the oldest females would often be out shopping on the main street, so this seems a weak argument. If he is supposing that there would be a substantial number of non-retired males who would probably not have been questioned, he may have a better argument. Without more information, he cannot reliably say anything.

(c) This was probably a quota sample of some sort. We need to know the following.

- What time of day, which day(s) of the week, the survey took place
- What instructions the interviewers were given about selection
- Exactly how the questions were worded and asked
- What response rate there was for requests to be interviewed
- What, if any, questions besides the main one were asked.

Additional questions on where people live, their age, whether or not they are retired, whether or not they are regular shoppers on the main street, where else they shop and how often, could be asked, in the hope of relating answers to some factor such as age. Views on supermarkets in general might be included, also whether the shoppers use a car or some other method of reaching shops.

Higher Certificate, Module 8, 2008. Question 4

(i) Advantages include: minimal administrative costs (postage, printing, paper); data could be collected quickly. Disadvantages include: no response from those who do not use the website regularly; some fields of study may be less suited to this form of learning, for example practical work may not be covered properly; a target population needs to be identified if results are to be useful.

(ii) The target population must be agreed with the development team: is it full-time staff and students, on which courses, are there part-time people who should be included? It must include everyone (who can be identified) that the materials are aimed at.

A sampling frame is required, e.g. student and staff lists, identified perhaps by student registration numbers and staff payroll lists.

Stratification into staff/student, individual courses and/or levels of study would be wise (post-hoc methods can sometimes achieve this). Some courses may be better served than others by using web material.

Paper questionnaires sent to a randomly chosen sample, or interviews if more in-depth answers are required, could be used (but are expensive). A pilot trial would be wise.

More than one method, e.g. questionnaire/interview and website, could be used if possible – but take care when combining results from different methods.

[In the examination, credit was given for all constructive ideas.]