HONG KONG STATISTICAL SOCIETY

2015 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) Estimates of the total number of animals in the herd may be biased. The extent of bias will depend largely on (1) how well the survey takers were able to locate the sampling units from the air and define their borders without reference to caribou groups; [1 mark] (2) how adequately they were able to search the area, i.e., data might be lost due to poor sighting conditions, inadequate coverage, pilot and observer abilities, etc; [1 mark]; (3) how accurately they were able to count the animals in the photographs, i.e., locating animals on photographs is quite difficult, an exact count might be almost impossible due to poor sighting conditions or large concentrations of animals; [1 mark] and (4) errors caused by major movements of the herd taking place during the survey and the effects of local movements which might be initiated by the activity of the aircraft. [1 mark]

Note, marks will be awarded for other sensible comments.

(ii) The variance of the estimator of the population total is $N^2 V(\bar{y}_{st}) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h}$ [1 mark]

N_h = total number of units in stratum h

S_h = population standard deviation in stratum h

n_h = sample size (number of units) taken in stratum h

$W_h = \frac{N_h}{N}$ = weight for stratum h where N is the total number of units in population

[2 marks]

(iii) The estimated total is $\hat{Y}_{st} = N \bar{y}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h$ (where there are L strata), so

$\hat{Y}_{st}$ = (400×24.1) + (40×25.6) + (100×267.6) + (40×179.0) + (70×293.7) + (120×33.2)

= 69127 animals [2 marks: 1 formula 1 answer]

The estimated variance is $N^2 V(\bar{y}_{st}) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{s_h^2}{n_h}$

$V(\hat{Y}_{st}) = 400 \times (400\text{-}98) \times \frac{74.7^2}{98} + \ldots = 84123268.30,$

so the estimated standard error is 9171.9 animals [2 marks: 1 calculations 1 answer]

Assume $\hat{Y}$ is normally distributed around the population value. An approximate 95% confidence interval for the total number of animals in this region: $\hat{Y}_{st} \pm 1.96$ se $(\hat{Y}_{st})$. So, the interval is 69127 ± 17976.92, i.e., 51150.1 to 87103.9 animals. [2 marks: 1 formula 1 answer]

Note, the use of 2 instead of the nominal two-sided value of 1.96 will be allowed (applies here and throughout the document).

(iv) As is clear from the data, the six strata split into three with fairly low density of caribou and three with much higher density. [1 mark] There are also some variations in the values of s_h between the six strata. [1 mark] Stratified sampling ensures that all these six strata will be represented adequately [1 mark], and that an estimate of the total number of animals will have a smaller standard deviation than for simple random sampling. [1 mark]

Optimal allocation aims to minimise the variance Var ($\hat{Y}_{st}$) for fixed total sample size, n. [1 mark] As well as allocating more sample to strata with higher density (per proportional allocation) [1 mark], optimal allocation will allocate more to those with larger standard deviations, so the precision is comparable with those having lower variability. [1 mark] Proportional allocation is likely to lead to a considerably larger value of Var ($\hat{Y}_{st}$).

Question 2

(i) Advantages of collecting dietary information by a diary, rather than relying solely on a questionnaire include: [3 marks]

a. Does not usually rely on memory to any extent, i.e., by recording foods as they are consumed, the problem of omission may be lessened and the foods more fully described. More accurate record of what was eaten and when.

b. Can help to overcome the problems associated with collecting sensitive information by personal interview; does not usually involve any interviewing if a good layout is used.

c. Can provide more accurate information on a respondent's behaviour and experiences on a daily basis.

Disadvantages include [3 marks]

a. Generally more expensive than a personal interview alone. Interviewers usually make at least two visits and are often expected to spend time checking the diary with the respondent.

b. Fewer people may be willing to participate in the first place and some of those who do will not complete, i.e., higher dropout rate. Limited use in some population groups (e.g., low literacy, recent immigrants, and some elderly groups).

c. Respondents may change their behaviour as a result of keeping a diary.

d. Tedious for respondents to complete, no guarantee that entries are made at the time

Note, marks will be awarded for other sensible comments.

(ii) Direct questions on sensitive and highly personal matters such as types and quantities of food and drink consumed may lead to refusal to answer or to people giving incorrect answers. Bias can result if people who do not complete the questionnaire or the dietary diary have different characteristics from those who do. [1 mark] Non-responders often tend towards being overweight or obese rather than normal range. It is unwise to treat the replies of the responders as giving a fully accurate picture. [1 mark]

Demographic and social factors, and responses to other questions on lifestyle, may vary between the obese and non-obese. [1 mark] Missing responses can sometimes be satisfactorily imputed by matching non-response people, for these factors and lifestyles, with those who have replied. [1 mark]

The usual reasons for non-response in all surveys also apply, such as non-availability at the time an interviewer visits; and these might well not be the same for obese as for non-obese. [1 mark] Respondent conditioning, incomplete recording of information and under-reporting, inadequate recall, insufficient cooperation and sample selection are all common reasons for bias in a diary. [2 marks for covering a range of reasons].

Note, marks will be awarded for other sensible comments.

(iii) The words in italics (*never*, *occasionally*, *frequently*) are vague with no precise meaning that will be understood by everyone, particularly in different age groups. [1 mark] No time period is suggested: is it over a year, or a month, or in winter, summer, etc? [1 mark]

A possible series of questions might be:

1. How many days per week do you engage in moderate or vigorous activity (such as brisk walking, jogging, biking, aerobics, etc) in addition to your normal daily routine? [3 marks: 1 question, 1 frequency scale, 1 time-bound/definition of activity]

- ☐ 0 days
- ☐ 1 day
- ☐ 2 days
- ☐ 3 days
- ☐ 4 days
- ☐ 5 days
- ☐ 6 days
- ☐ 7 days
- ☐ Don't know/no response
- ☐ Refusal

2. How long do you usually engage in moderate or vigorous activities on these days? [2 marks: 1 question, 1 frequency scale]

- ☐ Less than 30 minutes
- ☐ 30 to 60 minutes
- ☐ More than 60 minutes
- ☐ Don't know/no response
- ☐ Refusal

Note, marks will be awarded for other sensible answers.

Alternative to question 1

- ☐ More than twice a week
- ☐ Twice a week
- ☐ Once a week
- ☐ Once a fortnight
- ☐ Once a month
- ☐ Never
- ☐ Don't know/no response
- ☐ Refusal

Question 3

(i) (a) Let y_i denote the number of persons in household.

$\sum y_i = (434\times1) + (525\times2) + (247\times3) + (200\times4) + (65\times5) + (20\times6) + (5\times7) + (4\times8) = 3537$

$\sum y_i^2 = (434\times1) + (525\times4) + (247\times9) + \ldots\ldots + (4\times64) = 10803$ [2 marks]

The simple random sample estimate is

$\bar{y} = \frac{3537}{1500} = 2.4$ persons per household. [1 mark]

To find the estimated variance underlying this, first calculate

$s_y^2 = \frac{1}{1499}\left(10803 - \frac{3537^2}{1500}\right) = \frac{2462.754}{1499} = 1.6429$ [1 mark]

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

$(1-f)\frac{s_y^2}{n} = \frac{0.935\times1.6429}{1500}$, where $(1-f) = \frac{N-n}{N}$ is the finite population correction.

Therefore, the standard error is $= \sqrt{0.00102} = 0.032$ [2 marks; 1 formula 1 answer]

Assume $\bar{y}$ is normally distributed around the population value. An approximate 95% confidence interval for $\bar{Y}$ is given by $2.4 \pm 1.96\ se(\bar{y})$. Limits are 2.4 ± 0.063, i.e., 2.337 to 2.463, i.e., 2.3 to 2.5. [2 marks: 1 method 1 answer]

Note, here and in part (b) accept answers ignoring finite population correction (has minimal effect on standard error/CI).

(b) Let p be the proportion of one-person households.

For simple random sampling, the sample estimate is $\hat{p} = \frac{434}{1500} = 0.29$ [1 mark] and the variance of $\hat{p}$, is estimated as $(1-f)\frac{\hat{p}(1-\hat{p})}{n-1} = \frac{0.935\times0.29\times0.71}{1499} = 0.00013$ [1 mark]

Assume $\hat{p}$ is approximately normally distributed about the population value p. Approximate 95% limits for p are $\hat{p} \pm 1.96\ se(\hat{p})$. Limits are 0.29 ± 0.022, i.e., 0.27 to 0.31. [2 marks: 1 method 1 answer]

Note, accept a divisor 'n' (=1500) in the formula for the estimated variance of $\hat{p}$.

(ii) Let x_i denote number of bedrooms per household. Since number of persons and number of bedrooms varies by household [1 mark], an estimate of the number of persons per bedroom is a ratio of two variables, i.e ., $\hat{P} = \frac{\Sigma y}{\Sigma x}$. [1 mark] $\hat{P}$ is a biased estimator. The bias is likely to be negligible in large samples. [1 mark]

(iii) What constitutes a "household"? Is it the same as a "dwelling"? [1 mark] If not, should a multi-household dwelling have only one household sampled from it? [1 mark]

Which measure of overcrowding, i.e., average number of persons per dwelling, persons per room, persons per bedroom, bedroom standard (based on consideration of age, sex, marital status and relationship of household members)? [1 mark] What counts as a room, definition of bedroom or acceptable sleeping arrangement, etc?

Criteria for defining crowding, i.e., is it 1.5, 2 or 3 persons per bedroom; 2 or more bedrooms below the bedroom standard? [1 mark]

Should single person households be excluded from the calculations (since they have the lowest possible occupancy rate of 1.0)? [1 mark]

Note, marks will be awarded for other sensible comments.

Question 4

(a) As the topic is a sensitive one, self-completion questionnaires (no interviewer present) are more likely to elicit honest responses than telephone interviews. The respondent would have time to consider their answers, refer to records or consult with others. [1 mark] The questionnaire could use closed questions and examples to prompt the respondent's memory and ensure that the respondent knew what kinds of incidents constituted crimes. [1 mark] In addition, not everyone will have a telephone; those without telephones are likely to qualitatively differ from those with telephones, particularly socioeconomically. This could have negative implications for surveys on crime victimization. [1 mark] In a telephone call it could be difficult for the interviewer to prove their identity, assure confidentiality and establish rapport. [1 mark]; more concentration is needed in a telephone interview and respondents are likely to become tired if the interview is long. [1 mark]

Advantages of telephone interviews compared with self-completion questionnaires is that the researcher will know quickly whether or not the selected sample member is willing to respond, so less time and money is spent on following up non-responders. [1 mark]

Note, marks will be awarded for other sensible comments.

(b) Convenience sampling would consist of taking crates on or very near the outside of the truckload (in practice perhaps simply from the top of the load), and picking oranges on or very close to the top layer of each create. [1 mark] There may well be trends in quality from top to bottom, and sides to middle, of the truck [1 mark] (and of course the buyer simply does not know at the outset whether there are any such trends). So there is considerable risk of a biased estimate of whatever is being measured. [1 mark]

Cluster sampling requires identification of 'clusters' that are expected to behave reasonably similarly to the entire population. This should avoid the problems outlines above. [1 mark]

One method could be to use the crates as clusters. [1 mark] A one-stage scheme would then consist of taking a simple random sample of the crates and then inspecting all the oranges in the chosen crate. [1 mark] This would be feasible if the crates are all going to be unloaded anyway. It could be extended to a two-stage cluster scheme by defining a further level of clusters within the crates – may be layers of oranges. [1 mark]

Note, marks will be awarded for other sensible comments.

[For information: An alternate two-stage scheme that could be envisaged consists of using the layers of crates (or perhaps vertical piles of crates) on the lorry as the first stage clusters, selecting a random sample of these, and then selecting crates from within each chosen cluster.]

(c) $N = 11500$, sample size is n, population $S = 35.38$, width of 95% CI is to be 2×3.0

Assume the sample mean, $\bar{y}$, is normally distributed about the population value $\bar{Y}$ and its standard error is $\sqrt{\{(1-\frac{n}{N})\frac{S^2}{n}\}}$. [1 mark] The 95% confidence interval for $\bar{Y}$ is given by $\bar{y} \pm 1.96\sqrt{\text{Var}(\bar{y})}$. Require $1.96\sqrt{\text{Var}(\bar{y})} \leq 3.0$. [1 mark]

That is, $1.96 \sqrt{\{(1-\frac{n}{11500})\frac{(35.38)^2}{n}\}} \leq 3.0$ or $(\frac{3.0}{1.96\times 35.38})^2 = \frac{1}{n} - \frac{1}{11500}$ [1 mark]

This gives $n = 510.58$. Thus, we need n of about 510. [1 mark]

Note, accept n=535, ignoring finite population correction.

Practical issues: how to define students' average weekly earnings, e.g., average of all term-time weeks? [1 mark]; how to locate the target population, there will be a full list of registered students, but not of those in paid work during term time, issues of inflating the sample size. [1 mark]

Note, marks will be awarded for other sensible comments.