University of Waterloo STAT 332 – W. H. Cherry

Assignment 1 Outline Solution

A1 – 1. **(a)** A possible *frame* for a sample survey carried out to estimate the average annual hydro bill per family in the City of Waterloo would be a list of all the addresses in the City prepared for property tax collection purposes.

Strengths of this list would be:

- it has already been drawn up so that no further expenditure of resources would be needed in this regard;
- it is likely to contain essentially *all* addresses, because the City requires such completeness as the basis of its property tax collection system.

Weaknesses of this list would be:

- it is not certain that the City would be prepared to make such a list available to a researcher – it might depend on how compelling a case the researcher could make for the release of the list;
- there would probably be difficulty identifying hydro consumption of *individual* units in multi-unit buildings, such as apartment blocks.

Because the response is a property of the *household*, there is no need for a *probabilistic* process for selecting a person within each household – the requirement is for the person to have access to the information about hydro consumption for that household.

(b) (i) The *target population*: all students (*currently* enrolled?) at the University of Waterloo;

(ii) the *study population*: students taking the first-year accounting course;
the sampling *frame*: the class list for the accounting course;

(iii) the *respondent population*: the students in the study population who *would* respond to the questionnaire under the incentives for response provided in the sampling protocol;
the *sample*: the students who respond from among the 100 selected;

(iv) a *response*: a student's degree of satisfaction with the Co-op. Program;
a *population attribute*: the *average* degree of satisfaction with the Co-op. Program for all students; **OR:**
the *proportion* of all students who are satisfied with the Co-op. Program;

(v) two categories of error that impose limitations on the sample survey answer are:

- *study* error – it seems *un*likely that the study population of students in the large first-year accounting course would adequately overlap the target population of *all* University of Waterloo students;
- *non-response* error – the sample attribute value obtained from the students who respond, from among the 100 students selected, may not be close enough to be useful to the attribute value of the respondent population (let alone to the attribute values of the study population or the target population).

A1 – 2. (a) $\frac{1}{30} + \frac{4}{5}Ᵽ$; (b) $p = \frac{5}{4}(\frac{y}{n} - \frac{1}{30})$; (c) $\frac{24Ᵽ}{1+24Ᵽ}$;

(d) Under the method of randomized response, *only* the respondent answering the question knows *which* of the two questions he or she answered; this confidentiality extending even to the investigator(s) is believed both to reduce non-response and to promote truthful answers among respondents, particularly to questions on sensitive issues.

Confidentiality with regard to the investigator(s) is lost in *direct* questioning, which may result in a lower response rate and/or a lower proportion of truthful answers to a question like D; this would *in*crease the limitation imposed on the survey Answer by this source of error.

A1 - 3. (a) The answer is given in the statement of the question; start by multiplying out the squared expression in brackets () inside the sigma sign.

(b) Show that $E(S^2) = Ꞩ^2$, as is demonstrated near the bottom of page 2.41 in Figure 2.3 of the Course Materials.

A1 - 4. (a) List the $\binom{6}{4} = 15$ possible samples of size 4, under equiprobable selecting, from this population.

(b) The probability function of a random variable consists of two components:

- the value(s) that the random variable can take on;
- the corresponding probabilities.

(c) Use the appropriate results from STAT 230 to find the mean and standard deviation of this discrete random variable; the answers are: $E(\bar{Y}) = 13$; $s.d.(\bar{Y}) = \sqrt{1.9} \simeq 1.378\,405$.

(d) If you do correctly the calculations indicated in the question, you will see a particular illustration of each of the two general results stated in the question.

(*continued overleaf*)

Assignment 1 Outline Solution (continued 1)

A1 – 4. (cont.) (e) This part of the question repeats for the sample *standard deviation* ideas like those for the sample *average* that are illustrated in parts (a), (b) and (c); if you do the calculations correctly, you will again see a particular illustration of the general result stated in the question.

(f) The population median is ${}_{M}\mathbf{Y} = 12\tfrac{1}{2}$.

(g) We find that: $E({}_{M}Y) = 12.3$; $s.d.({}_{M}Y) = \sqrt{2.19\dot{3}} \simeq 1.480\,991$.

(h) We see that under equiprobable selecting, *un*like the sample average, the sample median is a *biased* estimator of the population median, because $E({}_{M}Y) \neq {}_{M}\mathbf{Y}$;
in this somewhat special case of a *small* population and a *high* sampling fraction, the median is the *more* variable of the two estimators; more generally, the sample median is usually *less* variable than the sample average because it is less sensitive to extreme values of the response variate of the units in the sample.

NOTE: You might like to repeat parts (a) to (d) for samples of size 2 and compare them with those for samples of size 4 (no useful comparison of the average with the median can, of course, any longer be made).
The results are: $E(\overline{Y}) = 13$; $s.d.(\overline{Y}) = \sqrt{7.6} \simeq 2.756\,810$.

A1 – 5. Let the random variable Y_j represent the jth observation; we use the model: $Y_j \sim N(\mu, 8)$ and $\overline{Y} \sim N(\mu, 8/\sqrt{n})$;
hence, standardizing and using the $N(0, 1)$ table: $\Pr\left(-1.64485 \leq \dfrac{\overline{Y}-\mu}{8/\sqrt{n}} \leq 1.64485\right) = 0.90$,
so that: $\dfrac{1.64485 \times 8}{\sqrt{n}} = 1$ which gives: $n = 173.154 \simeq 174$;
i.e., a sample size of about 174 observations is needed to meet the conditions specified.

A1 – 6. This question provides an opportunity for each student to actually *perform* their own series of equiprobable and other selecting exercises; these exercises are concerned with illustrating the following matters:

- the use of a table of equiprobable digits to implement equiprobable selecting;
- different samples generally yield *different* estimates of the *same* population attribute;
- precision improves with increasing sample size: we expect in (h) to obtain a *narrower* confidence interval – *i.e.*, *reduced* limitation on the Answer – from the sample of size 16 in (d) than from the samples of size 4 in (c);
- *haphazard selecting* [in (e)] is biased towards the *larger* circles in this situation (because selection probability is proportional to area) and there is a tendency to choose circles nearer the *centre* of the page; it is easy to mistake haphazard selecting for equiprobable selecting;
- *judgement selecting* [in (f)] yields Answers with *severe* limitation – it often seems to *over*estimate the true value of the population attribute in situations like the present one;
- *probability selecting* is a necessary but not a sufficient condition for agreement of the *actual* and *stated* confidence *levels* of confidence intervals – confidence intervals calculated under the sampling protocols used in (e) and (f) have *no* basis in statistical theory, increasing the severity of the limitation imposd on Answers by this sourse.

Two further comments are:

- In (c)(i), there is quite a high probability [*viz.* $1 - 80^{(16)}/80^{16} = 0.7996 \simeq 0.80$ or about 80%] of choosing the same circle more than once.
- The method of selecting the sample in (g), as in (e), tends to choose the *larger* trees preferentially; this may be acceptable in the case of a woodlot, because it may be the larger trees that are of greater commercial interest (*e.g.*, because they usually contain a larger amount of useable lumber).

A1 – 7. (a) These results are obtained by straight-forward algebraic manipulation from the expressions for an average and a (data) standard deviation; remember the results for the sum, and the sum of the squares, of the first $\mathbf{N}$ integers (*e.g.*, see Question **A10 - 4).**

(b) This question allows you to practise deriving a specific case of the general results for the mean and (probabilistic) standard deviation of (n times) the sample average under equiprobable selecting. The reasoning can be based on that shown on pages 2.40 amd 2.41 in Figure 2.3 of the Course Materials; as well as the two results from (a), the result for each of the $n(n-1)$ *co*variance terms that you will need to establish is $-(\mathbf{N}+1)/12$.

(c) When the equiprobable selecting is *with* replacement, the difference from (b) is that the covariance terms are all *zero* (Why?); thus, (c) is a simpler version of (b).

Simplifying algegraically the second expression for $s.d.(T_n)$ leads (unhelpfully) in the direction of a spurious 'population standard deviation' with divisor $\mathbf{N}$, a matter discussed in Statistical Highlight #94 on page HL94.9 and in Statistical Highlight #100 on pages HL100.7 anf HL100.8.