University of Waterloo STAT 332 – W. H. Cherry

Assignment 1

A1 – 1. (a) An experimenter wishes to estimate the average annual hydro bill per family in the City of Waterloo. Suggest a possible *frame(s)* for a survey carried out to obtain the required estimate, and briefly discuss the strength(s) and weakness(es) of your suggestion.

(b) A University of Waterloo administrator wishes to investigate attitudes towards the University's Co-op. Program among University of Waterloo students. To conduct a sample survey, the administrator takes the class list for a large first-year accounting course at the university and obtains, by equiprobable selecting (EPS), 100 of the 340 students on the list. For this investigation, describe briefly:

(i) the *target population*; (ii) the *study population* and the *frame(s)*;
(iii) the *respondent population* and the *sample*; (iv) a *response* and a *population attribute*;
(v) two categories of error that will impose limitations on the attribute estimates obtained from the survey.

A1 – 2. A researcher wishes to estimate the proportion ($\mathbf{P}$) of people in a large population who have ever used narcotic drugs. She prepares a box containing 100 cards, 20 of which contain Question *B* and 80 Question *D*. Each person who is interviewed obtains a card by equiprobable selecting *with* replacement from the box and answers the question it contains. Since only the person being interviewed knows which question he or she is answering, confidentiality is assured and so the researcher hopes that the answers will be truthful. It is known that one-sixth of birthdays fall in July or August.

Question *B*: *Were you born in July or August?*
Question *D*: *Have you ever used narcotic drugs?*

(a) Find an expression for the probability that a person answers 'yes'.
(b) If y people of n obtained by equiprobable selecting answer 'yes', find an expression for estimating $\mathbf{P}$.
(c) Find the proportion of the people who answer 'yes' that are responding to Question *D*.
(d) Briefly compare and contrast this *randomized response* method for obtaining information on a sensitive issue with *direct* questioning of the people selected from the respondent population about narcotic drug use.

A1 – 3. The respondent population standard deviation $\mathbf{S}$ is defined as shown in the first expression at the right.

$$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2}$$

$$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\Big[\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - \mathbf{N}\overline{\mathbf{Y}}^2\Big]}$$

$$s = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^2}$$

(a) Show that this first expression is equivalent to the second expression at the right; the latter is usually more convenient for manual calculation of $\mathbf{S}$ (or $\mathbf{S}^2$).
(b) In nearly all sample surveys, $\mathbf{S}$ is unknown and is *estimated* by the *sample* standard deviation (s), defined in the third expression at the right. Show that the random variable S^2 is an *unbiased estimator* of $\mathbf{S}^2$ under EPS of the sample.

A1 – 4. A respondent population consists of $\mathbf{N} = 6$ units with the following values for a quantitative respone variate $\mathbf{Y}$: 9, 15, 8, 11, 21, 14.

(a) List all the possible samples of size 4, under equiprobable selecting (EPS), from this population.
(b) Tabulate the value ($\overline{y}$) of the *average* for each sample, together with the *probability of selecting* the sample.
(c) Use the probability function in (b) to find the *mean* and the *standard deviation* of the random variable ($\overline{Y}$) representing the sample average under EPS.
(d) Verify that the values obtained in (c) are, respectively, $\overline{\mathbf{Y}}$, the population average, and $\mathbf{S}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}$, the *standard deviation of the sample average* under equiprobable (or 'simple random') selecting, $s.d.(\overline{Y})$.
(e) For each sample, calculate the *sample standard deviation* (s), as defined in question **A1 – 3**(b). Using the probabilit function tabulated in (b), verify that the *mean* of the random variable S^2 is $\mathbf{S}^2$.
(f) For the response variate $\mathbf{Y}$, the population *average* ($\overline{\mathbf{Y}}$) is one attribute that describes the location (or 'centre') of the distribution of that response variate; an alternative is the population *median*. Find the median response (${}_{\mathrm{M}}\mathbf{Y}$ say) of the population of six units listed above.
(g) For each sample of four units in (a), find the sample *median* (${}_{\mathrm{M}}y$). Then repeat parts (b) and (c) to find the *mean* and the *standard deviation* of the random variable (${}_{\mathrm{M}}Y$) representing the sample median under EPS.
(h) Use your findings in (c) and (g) to compare and contrast the *bias* and the *variability* of the sample average and the sample median as estimators of $\overline{\mathbf{Y}}$ and ${}_{\mathrm{M}}\mathbf{Y}$ under EPS.

A1 – 5. The random variable $\overline{Y}$ representing the average of a set of observations obtained by equiprobable selecting, is to be used as an estimate of the mean μ of a normal distribution whose standard deviation is known to be 8 cm. What should be the size, n, of the sample so that, with a probability of 90%, the estimate ($\overline{y}$) will differ from the true value (μ) by at most one centimetre?

(*continued overleaf*)

A1 – 6. The attached page (#0.17-3) contains a (study or respondent) *population* of 80 circles, which are to be used for the sampling investigation described below. [If you wish, think of the circles as representing fish in a lake or tumours removed during surgery.] Give a clear, detailed presentation of your procedure and/or results at *each* stage of the investigation.

(a) Label the circles 00, 01, ..., 78, 79 in any order and use the table of equiprobable digits ('random numbers') provided (on page #0.17-4) to obtain *4* circles by equiprobable selecting (EPS). You could start in the table at row and column numbers provided by two separate readings of the second hand on your watch or wall clock.

(b) Measure the diameter of each circle chosen; for convenience, the diameters are all multiples of 3 mm, so you should record them as 3, 6, 9, etc., mm. Find the average diameter of the four circles in your sample (the *sample average*); for convenience, record this average in one-fourths of a millimetre.

(c) Select three more samples (to make *four* in all) as in (a), using different parts of the equiprobable digit table, and carry out the calculations described in (b) for each sample.

(i) Was any circle selected more than once in your four samples? Discuss your answer briefly.

(ii) Illustrate the differences in your four sample averages by showing them on an appropriate section of the real number line marked off in scale divisions of one-sixteenth of a millimetre.

(d) Using another part of the equiprobable digit table, select *16* circles by EPS and find the sample average diameter in sixteenths of a millimetre. Comment briefly on how this average compares with those found from the four smaller samples; add this average, with a suitable label to distinguish it, to your real number line in (c) (ii).

(e) Another selecting process from the population of circles is to close your eyes and drop your pen or pencil point 'at random' on page #0.17-3; obtain a sample of size $n = 4$ by this process (*haphazard* selecting).

(i) Explain briefly whether this process is equivalent to EPS this situation.

(ii) Comment briefly on how the sample average compares with those found in (a), (b) and (c). Add this average, with a suitable label, to your real number line in (c) (ii).

(f) Yet another selecting process is to choose what the investigator (*you*, in this instance) considers to be a 'representative' sample of the population (*judgement* selecting). Use this process to select four circles and comment on how this sample average compares with those found previously; add this average, with a suitable label, to your real number line in (c) (ii). In situations like the present one, the sample average obtained under judgement selecting usually *over*estimates the population average ($\bar{\mathbf{Y}} = 12$ mm in this question).

(g) Suppose that we think of the circles as representing the cross-sections of trees at a specified height (*e.g.*, chest height) above the ground. Foresters who wish to estimate the total amount of lumber in a woodlot select trees for measurement by a process equivalent to dropping a pen or pencil [as in (e) above]. Discuss briefly the advantages and disadvantages of such a selecting process in this context.

(h) Using each of the five samples obtained by EPS in (a) to (d) above, find an approximate 90% confidence interval for the true average diameter ($\bar{\mathbf{Y}}$) of the population of 80 circles. If $\bar{\mathbf{Y}}$ is *actually* 12 mm, comment briefly on your five intervals. Also, if appropriate, find an approximate 90% confidence interval for $\bar{\mathbf{Y}}$ from your sample in (e) and in (f), and comment briefly on the meaning of these two intervals.

A1 – 7. (a) For a population of $\mathbf{N}$ tickets numbered $\mathbf{Y} = 1, 2, \ldots., \mathbf{N}$, show that the average ($\bar{\mathbf{Y}}$) and the standard deviation ($\mathbf{S}$) are as given at the right.

$$\bar{\mathbf{Y}} = \frac{\mathbf{N}+1}{2}; \qquad \mathbf{S} = \sqrt{\frac{\mathbf{N}(\mathbf{N}+1)}{12}}$$

(b) If n tickets are selected equiprobably with*out* replacement (EPSWOR) from the $\mathbf{N}$ tickets, and the random variable T_n is the *total* of the numbers on these tickets, establish the middle two results at the right.

$$E(T_n) = n\bar{\mathbf{Y}}; \qquad s.d.(T_n) = n\mathbf{S}\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}}$$

(c) If the n tickets in (b) are selected *with* replacement (EPSWIR), establish the lower two results at the right.

$$E(T_n) = n\bar{\mathbf{Y}}; \qquad s.d.(T_n) = n\mathbf{S}\sqrt{\frac{1}{n} - \frac{1}{n\mathbf{N}}}$$

(d) In light of the different selecting processes in (b) and (c), compare and contrast the results for the mean and standard deviation of T_n. [This comparison is aided by *not* simplifying algebraically the expression for $s.d.(T_n)$ when selecting is *with* replacement.]