University of Waterloo W. H. Cherry

SAMPLING and SURVEY SAMPLING: Estimating a Population Average or Total

1. Background Information

The mathematics on pages HL77.4 and HL77.5 in Section 5 of this Highlight #77 is a centrepiece of the theory of survey sampling ('design-based' inference – recall Section 7 in Statistical Highlight #74); we can appreciate it more fully with the following background information.

∗ In introductory probability, we use *upper* case *italic* letters (usually near the end of the alphabet, like Y, R and S) for random variables and the corresponding *lower* cases letters (*e.g.*, y, r and s) for their values. We now further distinguish *upper*-case **bold** letters for *population* quantities; for example, $\mathbf{Y}_\mathrm{i}$ is the response variate for the i*th* element in the respondent population.
- – The line through population symbols is to make distinguishable *italic* and **bold** upper-case *hand*written letters and we *say*, for instance, '$\mathbf{Y}_\mathrm{i}$ cross' for the response variate of the i*th* respondent population element.

Table HL77.1:

SYMBOL: Random variable	Value	Respondent population	DESCRIPTION
Y	y	$\mathbf{Y}$	Response variate
–	j	i	Summation index
–	x	$\mathbf{X}$	Focal explanatory variate
–	z	$\mathbf{Z}$	Explanatory variate
–	n	$\mathbf{N}$	Number of units/elements
$\overline{Y}$	$\overline{y}$	$\overline{\mathbf{Y}}$	Average (sum ÷ number)
R	r	$\mathbf{R}$	Residual [or Ratio]
S	s	$\mathbf{S}$	Standard deviation

∗ In introductory statistics courses, the number of elements in the population (also called the population *size*) is seldom considered explicitly; in this Highlight #77, this attribute is denoted $\mathbf{N}$ ('$\mathbf{N}$ cross').
- – Including the population size in survey sampling theory is sometimes called dealing with a *finite* population, but this is *unhelpful* terminology (perhaps carrying over from mathematics where infinity often arises). A population in statistics is a real-world entity and so, by its nature, has a finite number of elements – see also Note 2 on page HL77.5 and Appendix 4 on page HL77.9.

∗ When investigating a Question with a *descriptive* aspect [one whose Answer will involve primarily values for population/process attributes (past, present, future)], a useful way to think of (or to 'model') the response variate of a respondent population element is as shown in equation (HL77.1) at the right;

$$\mathbf{Y}_\mathrm{i} = \overline{\mathbf{Y}} + \mathbf{R}_\mathrm{i} \qquad \text{-----(HL77.1)}$$

this model is useful because it expresses a quantity we can observe ($\mathbf{Y}_\mathrm{i}$) in terms of an attribute of interest ($\overline{\mathbf{Y}}$, the population *average*) and a quantity ($\mathbf{R}_\mathrm{i}$, the *population residual* for element i) whose behaviour is amenable to probability modelling which, ultimately, enables us to quantify imprecision due to sample error and (in some contexts) measurement error.
- – The **response model** for a *non*-comparative Plan – the type of Plan usually appropriate to answer a Question with a descriptive aspect – involving equiprobable (simple random) selecting (EPS) of n units from an *un*stratified population is:

$$Y_j = \mu + R_j, \quad j = 1, 2, ..., \mathrm{n}, \quad R_j \sim N(0, \sigma), \quad \text{prob. independent}, \quad \text{EPS}, \qquad \text{-----(HL77.2)}$$

where Y_j is a random variable whose distribution represents the possible values of the measured response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over;

R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over.

The symbol μ in the model (HL77.2), and two related entities $\hat{\mu}$ and $\tilde{\mu}$, have the following meanings:

μ: a parameter representing the *average* ($\overline{\mathbf{Y}}$) of the measured response variate of the elements of the respondent population.

$\hat{\mu}$: the least squares *estimate* of μ – a *number* whose value is derived from an appropriate set of *data*;
- ○ for the model (HL77.2), $\hat{\mu} = \overline{y}$, reflecting the intuitive idea that, under EPS, the measured sample average estimates the measured respondent population average;

$\tilde{\mu}$: the least squares *estimator* of μ – a *random variable* whose distribution represents the possible values of the estimate $\hat{\mu}$ if the selecting, measuring and estimating processes were to be repeated over and over;
- ○ here, $\tilde{\mu} = \overline{Y}$, the random variable representing the sample average under EPS.

∗ In Statistical Highlight #72, we met σ, and two related entities $\hat{\sigma}$ and $\tilde{\sigma}$, in the context of response models like (HL77.2) above;

σ: the (probabilistic) *standard deviation* of the normal model for the distribution of the residual, is a model parameter representing the (data) *standard deviation* ($\mathbf{S}$) of the measured response variate of the respondent population elements; this (data) standard deviation (and, hence, σ) *quantifies* the *variation* of the measured response variate over the elements of the respondent population – as this variation increases, so does $\mathbf{S}$ (and, hence, so does σ).
- ○ two *other* characteristics of variation are its *location* and its *shape* – these are, respectively, 0 and normal for (HL77.2);

$\hat{\sigma}$: the least squares *estimate* of σ – a *number* whose value is derived from an appropriate set of *data*;

$\tilde{\sigma}$: the least squares *estimator* of σ – a *random variable* whose distribution represents the possible values of the estimate $\hat{\sigma}$ if the selecting, measuring and estimating processes were to be repeated over and over.

How σ is *estimated*, reflected by differing expressions for $\hat{\sigma}$, depends on the sampling protocol in the Plan for the investigation and the response model apporiate for this Plan; for the model (HL77.2), $\hat{\sigma}$ is given by equation (HL77.3).

$$\hat{\sigma} = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^2} \qquad \text{-----(HL77.3)}$$

∗ In *this* Highlight #77, we meet overleaf (page HL77.2) *four* quantities called '*standard deviation*'; the first two and an associ-

(*continued overleaf*)

ated quantity S correspond to the three σs given overleaf on page HL77.1 but, *un*like $\hat{\sigma}$, s is called a standard deviation.

- $\mathbf{S}$: the respondent *population* (data) standard deviation – it is defined in Section 4 in Table HL77.5 on page HL77.4 and is a *number* which quantifies the variation over the respondent population of the response variate $\mathbf{Y}$ about its average $\overline{\mathbf{Y}}$;
  - ○ like most population attributes except $\mathbf{N}$, usually the value of $\mathbf{S}$ is *un*known;
- s: the *sample* (data) standard deviation – it is defined in Table HL77.5 on page HL77.4 and is a *number* which quantifies the variation over the sample of the response y about its average $\overline{\mathrm{y}}$; the expression for s in Table HL77.5 is like (HL77.3) at the bottom right overleaf on page HL77.1 for $\hat{\sigma}$ in the model (HL77.2).
  - ○ under EPS (and the Plan), we take s = s to estimate $\mathbf{S}$ – that is, to provide a value we can use for (unknown) $\mathbf{S}$;
  - ○ this Highlight #77 is concerned with only *one* sampling protocol – EPS from an unstratified population to estimate an average or total – and so there is only *one* expression for the estimate (s = s) of $\mathbf{S}$;
- S: the *estimator* corresponding to s – under EPS (and the Plan), it is a *random variable*, of which s is one (realized) *value*;

Table HL77.2: SUMMARY OF STANDARD DEVIATIONS

Response Models	Survey Sampling
σ Model parameter	$\mathbf{S}$ Respondent population standard deviation – an attribute
$\hat{\sigma}$ Estimate of σ	s Sample standard deviation – we take s = s to estimate $\mathbf{S}$
$\tilde{\sigma}$ Estimator of σ	S Estimator corresponding to s – a random variable

+ $s.d.(\overline{Y})$: the standard deviation of the sample average – under EPS (and the Plan), it provides a theoretical basis for quantifying uncertainty due to sample error in estimates of respondent population attributes like an average or total;
+ $\widehat{s.d.}(\overline{Y})$: the *estimated* standard deviation of the sample average which, under EPS (and the Plan), is the basis for calculating *values* for the end points of confidence intervals for respondent population attributes like an average or total.

The expressions for $s.d.(\overline{Y})$ and $\widehat{s.d.}(\overline{Y})$ differ in that $\mathbf{S}$ is replaced by its estimate s – see equation (HL77.9) and (HL77.17) on page HL77.5. [In a non-probability sampling context concerned only with *data*, s would usually be denoted s.]

∗ We capitalize on having *two* words – average and mean – in English to make a useful distinction for a measure of *location*:
- the *average* is a measure of location for a set of *data*;
- the *mean* is a measure of location for (the distribution of) a *random variable.*

However, for the magnitude of *variation* there is only *one* term – standard deviation – for the commonly-used measure, and this can be a source of confusion. Ideally, we would like:
- the *(new word)* as a measure of variation for a set of *data*,
- the *standard deviation* as a measure of variation for a *random variable*,

but the use of 'standard deviation', regardless of context, is too well-established in statistics for this ideal to be attainable. A compromise, to assist beginning students, is to distinguish a *data* standard deviation from a *probabilistic* standard deviation – see Table HL77.3 at the right. Note that *we* use one symbol (*e.g.*, $\mathbf{S}$, s) for a data standard deviation and the abbreviation *s.d.* for a probabilistic standard deviation.

Table HL77.3

Respondent population standard deviation	$\mathbf{S}$	*data* standard deviation
Sample standard deviation	s	
Standard deviation of the sample average	$s.d.(\overline{Y})$	*probabilistic* standard deviation
Estimated standard deviation of the sample average	$\widehat{s.d.}(\overline{Y})$	

The previous Statistical Highlight #76 portrays visually the distinction between the sample standard deviation (s; represented by the 16 'hooked' horizontal lines in each diagram) and the standard deviation of the sample average [$s.d.(\overline{Y})$; as estimated from the 16 sample averages and denoted $s_{\overline{y}}$ near the lower right-hand corner of each diagram].

∗ The standard deviation of $\overline{Y}$ is sometimes referred to as the *standard error* of $\overline{Y}$ (*e.g.*, Barnett, pp. 26, 45) but this term is avoided in these Statistical Highlights because it is used by different authors for *both* $s.d.(\overline{Y})$ and $\widehat{s.d.}(\overline{Y})$ (see also Cochran, pages 24, 25-27 and 53), potentially confusing a quantity and its estimate. [References are given on page HL77.8 in Section 7.]

∗ The following suggestions may help avoid confusion arising from (careless use of) the terminology discussed above.
- ● when you encounter the word *mean*, be sure you understand whether it refers to:
  - an average of data (and whether the data are from a *sample* or a *census*), **OR**
  - a random variable [and whether it is an individual random variable or a (linear) combination (*e.g.*, an average, sum or difference)], **OR**
  - a parameter of a response model or probability model.
- ● when you encounter the term *standard deviation* or *standard error*, be sure you understand whether it refers to:
  - the variation of data (and whether the data are from a *sample* or a *census*), **OR**
  - a random variable [and whether it is an individual random variable or a (linear) combination (*e.g.*, an average, sum or difference)], **OR**
  - a parameter of a response model or probability model.
- ● when you encounter the word *inaccuracy*, remember that it is a real-world quantity and is defined only in the context of *repetition* of a *process* – like selecting or measuring.
  - *Estimating* bias (a model quantity) *differs* from inaccuracy in that it *de*creases with *in*creasing sample size and so may not be of much practical concern in actual sample surveys.

  [There is further discussion of bias in Appendix 5, Appendix 6 and Appendix 7 on pages HL77.10 to HL77.13.]

(*continued*)

University of Waterloo W. H. Cherry

SAMPLING and SURVEY SAMPLING: Estimating a Population Average or Total (continued 1)

2. Equiprobable (Simple random) Selecting

The *definition* of equiprobable selecting (EPS) is: If a sample of n units is obtained from a respondent population of $\cancel{\mathbf{N}}$ units in such a way that every *sample* of size n has an *equal* probability of being selected, the selecting process is called **equiprobable selecting**. [Elsewhere, you may see it called **simple random selecting** (SRS).]
In practice, we often think of EPS as being *implemented* by selecting each *unit* of the sample equiprobably ('at random') and with*out* replacement ('EPSWOR') from the (unstratified) respondent population. [The element-unit distinction is discussed in Appendix 2 on page HL77.9.]

Because we commonly think of EPS in terms of how we select the *units*, we may overlook the fact that the definition is in terms of *sample* probabilities. In particular, we need to recognize that, while the definition implies that each *unit* has the same inclusion probability of n/$\cancel{\mathbf{N}}$, there are selecting processes with equal unit inclusion probabilities that are *not* EPS. An illustration is given at the right below; for this respondent population of $\cancel{\mathbf{N}}$ = 4 units (or elements), six samples of size n = 2 can be obtained by EPS but only *two* such samples are obtained by *systematic* selecting; however, provided the starting point of the systematic selecting process is chosen equiprobably, any unit has an inclusion probability of ½ under *either* process.
The same point is made by saying that, under systematic selecting, two of the six possible samples of size 2 have probability ½ and four have *zero* probability (see also Appendix 5 on page HL21.6 in Statistical Highlight #21).

$\cancel{\mathbf{N}}$ = 4 **Population units** (or **elements**)
1, 2, 3, 4;
the samples of size 2 are:
EPS: (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4);
systematic selecting: (1, 3), (2, 4).

The emphasis in statistics on EPS (or its equivalent) is because it is the basis of theory which provides:

- *unbiased* estimating of a population average (an attribute commonly of interest);
- a connection between sampling imprecision and *sample size* (or level of *replicating*);
- an expression for a *confidence interval* for a population average – such an interval, under suitable modelling assumptions, *quantifies* sampling and measuring imprecision (as demonstrated in Statistial Highlight #74).

[These three provisions of statistical theory refer to behaviour under *repetition* – Answer(s) obtained in a *particular* investigation *remain* uncertain, as reflected by their limitations.]
EPS does *not*, of itself, *reduce* sample error or sampling imprecision, as implied in (wrong) statements such as:

- ○ EPS generates a *representative* sample – see also Appendix 3 on page HL77.9;
- ○ EPS generates a sample which provides a proper basis for *generalization*;

!

as well as misrepresenting the statistical benefits of using EPS, such statements confuse repetition (the *process* of EPS) with a component of a *particular* investigation (the actual sample). A *correct* statement is:
EPS, *in conjunction with adequate replicating* (or an *adequate sample size*), provides for quantifying sampling imprecision and so allows a particular investigation to obtain an Answer with acceptable limitation (in the Question context) due to sample error.

- What constitutes *acceptable* limitation depends on the investigation requirements for its Answer(s); for instance, in a poll to estimate one or more proportions, an acceptable limitation may be quantified as the proportion(s) estimated to within 2 percentage points 19 times out of 20. [Limitations may also be imposed by the *resources* available for the investigating].

3. Sample Size and Sample Error under EPS

Example HL77.1: A respondent population of $\cancel{\mathbf{N}}$ = 4 units has the following integer $\cancel{\mathbf{Y}}$-values for its response variate:

1, 2, 4, 5 (so that: $\bar{\cancel{\mathbf{Y}}} = 3$, $\cancel{\mathbf{S}} \simeq 1.8257$);

we examine the behaviour of *sample error* under EPS as the sample size *in*creases from 1 to 2 to 3 to 4.
The number at the bottom of the four error columns of Tables HL77.4 below is the *average magnitude* of the sample error for that sample size.

Table HL77.4a
EPS of n = 1 unit

Sample	$\bar{y}$	Error
(1)	1	−2
(2)	2	−1
(4)	4	1
(5)	5	2
		1½

Table HL77.4b
EPS of n = 2 units

Sample	$\bar{y}$	Error
(1, 2)	1½	−1½
(1, 4)	2½	−½
(1, 5)	3	0
(2, 4)	3	0
(2, 5)	3½	½
(4, 5)	4½	1½
		⅔

Table HL77.4c
EPS of n = 3 units

Sample	$\bar{y}$	Error
(1, 2, 4)	2⅓	−⅔
(1, 2, 5)	2⅔	−⅓
(1, 4, 5)	3⅓	⅓
(2, 4, 5)	3⅔	⅔
		½

Table HL77.4d
EPS of n = 4 units

Sample	$\bar{y}$	Error
(1, 2, 4, 5)	3	0
		0

Example HL77.1 reminds us of general results under EPS that follow from the theory in Section 5 on pages HL77.4 and HL77.5:

- as the sample size *in*creases, the average magnitude (and, hence, the standard deviation) of sample error *de*creases – this is what we mean when we say that increasing sample size *de*creases sampling *imprecision* under EPS;

(*continued overleaf*)

- taking the *sign* of sample error into account, the average error is *zero* for each n – this is what we mean by saying that, under EPS, (the random variable representing) the sample average is an *unbiased* estimator of the respondent population average;
  - note that *both* the selecting method *and* the population attribute and its estimator are involved in this statement;
  - another statement with these components, which contrasts with the statement above about $\bar{Y}$, is that, for the population attribute which is the *ratio* of the average of two response variates ($\mathbf{R} = \bar{\mathbf{Y}}/\bar{\mathbf{X}}$), the sample ratio $r = \bar{y}/\bar{x}$ is *biased* [$E(R) \neq \mathbf{R}$] under EPS but *un*biased if the first sample unit is selected with probability proportional to its $\mathbf{X}$-value and the remainder selected equiprobably (see Cochran, page 175 and see also Note 23 on pages HL77.11 and HL77.12).
- there is no *sample* error when a *census* is taken – when *all* units of the respondent population are selected.

We also see that there can be a sample size(s) for which *none* of its $\binom{\mathbf{N}}{\mathrm{n}}$ samples has *zero* sample error – *no* sample has $\bar{y} = \bar{\mathbf{Y}}$.

4. Notation

Table HL77.5 below extends Table HL77.1 on page HL77.1 to define the notation used in the theory developed in this Highlight #77; the last column of Table HL77.5 includes the 'model'. It is a model in the sense of being an *idealization* or *mathematical abstraction* involving the equal (sample) probabilities attained under EPS (and the Plan); this sense of 'model' differs from that of a symbolic expression like a response model [such as equation (HL77.2) on the first side (page HL77.1) of this Highlight #77].

Table HL77.5:

....QUANTITY......	RESPONDENT POPULATION	.........SAMPLE [MODEL].....................
Size (elements/units)	$\mathbf{N}$	n
Response	$\mathbf{Y}_{\mathrm{i}}$ (i = 1, 2, ..., $\mathbf{N}$)	y_j (j = 1, 2, ..., n) [r.v.s are Y_j with value y_j]
Average	$\bar{\mathbf{Y}} = \frac{1}{\mathbf{N}}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}} = \frac{1}{\mathbf{N}}\,{}_{T}\mathbf{Y}$	$\bar{\mathrm{y}} = \frac{1}{\mathrm{n}}\sum_{j=1}^{\mathrm{n}}\mathrm{y}_j$ [r.v. is $\bar{Y}$ with value $\bar{y}$]
Total	${}_{T}\mathbf{Y} = \mathbf{N}\bar{\mathbf{Y}} = \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}$	$\bar{\mathrm{y}} = {}_{T}\mathrm{y} = \mathbf{N}\bar{\mathrm{y}}$ [r.v. is ${}_{T}Y$ with value
Standard deviation	$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\mathrm{SS}_{\mathbf{Y}}} \equiv \sqrt{\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\mathbf{Y}_{\mathrm{i}} - \bar{\mathbf{Y}})^2}$	$\mathrm{s} = \sqrt{\frac{1}{\mathrm{n}-1}\mathrm{SS}_{y}} \equiv \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(\mathrm{y}_j - \bar{\mathrm{y}})^2}$ [r.v. is S with value s]

${}_{T}y$, the *estimate* of the population total ${}_{T}\mathbf{Y}$, is *not* the *sample* total, which is $\mathrm{n}\bar{y} = \sum_{j=1}^{\mathrm{n}} y_j$ and is usually *not* a sample attribute of interest.

5. Estimating $\bar{\mathbf{Y}}$, the Respondent Population Average

We want both a value (or *point estimate*) for this respondent population attribute *and* a measure of the (sampling) uncertainty of the estimate, for which we use a confidence interval.

To develop the relevant theory, we first establish results for $E(Y_j)$, $s.d.(Y_j)$ and $cov(Y_j, Y_l)$:

(i) $E(Y_j) = \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}\cdot\Pr(Y_j = \mathbf{Y}_{\mathrm{i}})$ (the mean of a discrete random variable).

We can find $\Pr(Y_j = \mathbf{Y}_{\mathrm{i}})$ in any of three ways:

(a) because every possible ordered sample is equally probable under equiprobable selecting, any population unit is equally probable at any position in the sample and, because there are $\mathbf{N}$ units in the population, this probability is $1/\mathbf{N}$;

(b) ordered counting: $$\frac{\text{number of ordered samples with } Y_j = \mathbf{Y}_{\mathrm{i}}}{\text{total number of samples of size n}} = \frac{(\mathbf{N}-1)^{(\mathrm{n}-1)}}{\mathbf{N}^{(\mathrm{n})}} = \frac{1}{\mathbf{N}}; \qquad \text{-----(HL77.4)}$$

(c) *un*ordered counting: $$\frac{\text{number of } un\text{ordered samples with } \mathbf{Y}_{\mathrm{i}} \text{ at any position}}{\text{total number of samples of sze n}} \cdot \frac{1}{\text{number of sample positions}} = [\tbinom{\mathbf{N}-1}{\mathrm{n}-1}/\tbinom{\mathbf{N}}{\mathrm{n}}]\cdot[1/\mathrm{n}] = \frac{1}{\mathbf{N}};$$

$$\therefore \quad E(Y_j) = \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}\cdot\frac{1}{\mathbf{N}} = \frac{1}{\mathbf{N}}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}} = \bar{\mathbf{Y}}. \qquad \text{-----(HL77.5)}$$

(ii) $E(Y_j^2) = \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2\cdot\Pr(Y_j = \mathbf{Y}_{\mathrm{i}}) = \frac{1}{\mathbf{N}}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2$ [using the result for $\Pr(Y_j = \mathbf{Y}_{\mathrm{i}})$ from (**i**)];

$$\therefore \quad s.d.(Y_j) = \sqrt{E(Y_j^2) - [E(Y_j)]^2} = \sqrt{\frac{1}{\mathbf{N}}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2 - \bar{\mathbf{Y}}^2} = \sqrt{\frac{1}{\mathbf{N}}\Big[\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2 - \mathbf{N}\bar{\mathbf{Y}}^2\Big]} = \sqrt{\frac{\mathbf{N}-1}{\mathbf{N}}}\,\mathbf{S}. \qquad \text{-----(HL77.6)}$$

(iii) $E(Y_jY_l) = \sum_{\mathrm{i}=1}^{\mathbf{N}}\sum_{\mathrm{k}\neq\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}\mathbf{Y}_{\mathrm{k}}\cdot\Pr(Y_j = \mathbf{Y}_{\mathrm{i}}, Y_l = \mathbf{Y}_{\mathrm{k}})$ (the mean of a product of discrete random variables).

But from (**i**), because $\Pr(A \cap B) = \Pr(A)\cdot\Pr(B|A)$, $\Pr(Y_j = \mathbf{Y}_{\mathrm{i}}, Y_l = \mathbf{Y}_{\mathrm{k}}) = \Pr(Y_j = \mathbf{Y}_{\mathrm{i}})\cdot\Pr(Y_l = \mathbf{Y}_{\mathrm{k}}|Y_j = \mathbf{Y}_{\mathrm{i}}) = \frac{1}{\mathbf{N}}\cdot\frac{1}{\mathbf{N}-1}$;

$$\therefore \quad E(Y_jY_l) = \frac{1}{\mathbf{N}}\cdot\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}\sum_{\mathrm{k}\neq\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{k}} = \frac{1}{\mathbf{N}}\cdot\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}(\mathbf{N}\bar{\mathbf{Y}} - \mathbf{Y}_{\mathrm{i}}) = \frac{1}{\mathbf{N}}\cdot\frac{1}{\mathbf{N}-1}\Big[\mathbf{N}\bar{\mathbf{Y}}^2 - \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2\Big] \quad (\text{because } \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}} = \mathbf{N}\bar{\mathbf{Y}});$$

$$\therefore \quad cov(Y_j, Y_l) = E(Y_jY_l) - E(Y_j)\cdot E(Y_l)$$
$$= \frac{1}{\mathbf{N}(\mathbf{N}-1)}\Big[\mathbf{N}^2\bar{\mathbf{Y}}^2 - \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2\Big] - \bar{\mathbf{Y}}\cdot\bar{\mathbf{Y}} = \frac{\mathbf{N}^2\bar{\mathbf{Y}}^2 - \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2 - \mathbf{N}(\mathbf{N}-1)\bar{\mathbf{Y}}^2}{\mathbf{N}(\mathbf{N}-1)} = \frac{-\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2 + \mathbf{N}\bar{\mathbf{Y}}^2}{\mathbf{N}(\mathbf{N}-1)} = -\frac{\mathbf{S}^2}{\mathbf{N}}. \qquad \text{-----(HL77.7)}$$

(*continued*)

University of Waterloo W. H. Cherry

SAMPLING and SURVEY SAMPLING: Estimating a Population Average or Total (continued 2)

Hence, when we use $\overline{Y}$ (the random variable representing the sample average) as an estimator of $\overline{\mathbf{Y}}$ (the respondent population average), under EPS we have:

$$E(\overline{Y}) = E[\tfrac{1}{n}\sum_{j=1}^{n} Y_j] = \tfrac{1}{n}E[\sum_{j=1}^{n} Y_j] = \tfrac{1}{n}\sum_{j=1}^{n} E(Y_j) = \tfrac{1}{n}[\underbrace{\overline{\mathbf{Y}}+\overline{\mathbf{Y}}+ \ldots\ldots +\overline{\mathbf{Y}}}_{\text{n terms from } (\mathbf{i})}] = \overline{\mathbf{Y}};$$ *i.e.*, $\overline{Y}$ is an *unbiased* estimator of $\overline{\mathbf{Y}}$ under EPS; -----(HL77.8)

$$s.d.(\overline{Y}) = s.d.[\tfrac{1}{n}\sum_{j=1}^{n} Y_j] = \tfrac{1}{n}\, s.d.[\sum_{j=1}^{n} Y_j]$$

$$= \tfrac{1}{n}\sqrt{\sum_{j=1}^{n}[s.d.(Y_j)]^2 + \sum_{j=1}^{n}\sum_{l\neq j=1}^{n} cov(Y_j, Y_l)} = \tfrac{1}{n}\sqrt{\underbrace{n(\mathbf{N}-1)\mathbf{S}^2/\mathbf{N}}_{\substack{\text{n equal terms}\\ \text{from } (\mathbf{ii})}} + \underbrace{n(n-1)[-\mathbf{S}^2/\mathbf{N}]}_{\substack{n^2-n \text{ equal}\\ \text{terms from } (\mathbf{iii})}}} = \mathbf{S}\sqrt{\tfrac{1}{n} - \tfrac{1}{\mathbf{N}}}$$ -----(HL77.9)

[the standard deviation of the sample average under EPS (from an *un*stratified respondent population)].

Thus, the distribution of (the estimator of) the sample average under EPS is: $\overline{Y} \dot{\sim} N(\overline{\mathbf{Y}}, \mathbf{S}\sqrt{\tfrac{1}{n} - \tfrac{1}{\mathbf{N}}})$ -----(HL77.10)

NOTES: 1. Equation (HL77.9) for $s.d.(\overline{Y})$ shows that the *finite* size of the respondent population modifies the random variable model expression by including a second term, $1/\mathbf{N}$, under the square root multiplying $\mathbf{S}$. Other matters of interest are:

- when $n = \mathbf{N}$, $s.d.(\overline{Y}) = 0$, reminding use that, at least in principle, sample error is *zero* in a census.
- when $n \ll \mathbf{N}$ (*i.e.*, when the sample size is a small proportion of the respondent population size, say 5% or less), the expression for $s.d.(\overline{Y})$ becomes essentially the more familiar form $\mathbf{S}\sqrt{\tfrac{1}{n}}$. -----(HL77.11)
- the form of the square root multiplying $\mathbf{S}$ means that the precision of estimating $\overline{\mathbf{Y}}$ by $\overline{Y}$ under EPS is determined primarily by the *sample* size and only weakly by the *population* size.
  - This insight of statistical theory is counter-intuitive – there is essentially the *same* sampling imprecision in a national poll of 1,500 people selected from a population of 30 million Canadians or 300 million Americans.

2. The expression (HL77.9) may be written as shown in equations (HL77.12) and (HL77.13) at the right. The former, where $\mathbf{S}$ has been replaced by its expression in terms of $\mathbf{Y}$, is of interest to compare with equation (HL77.14) below; equation (HL77.13) gives the standard deviation of $\overline{Y}$ as the familiar form (HL77.11) multiplied by the square root of a *finite population correction* $1-f$, where $f = n/\mathbf{N}$ is the *sampling fraction*. BUT:

$$s.d.(\overline{Y}) = \sqrt{\frac{1}{\mathbf{N}-1}(\frac{1}{n} - \frac{1}{\mathbf{N}})\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2}$$ -----(HL77.12)

$$s.d.(\overline{Y}) = \sqrt{(1-f)}\,\mathbf{S}\sqrt{\frac{1}{n}}$$ -----(HL77.13)

- Thinking of $s.d.(\overline{Y})$ as an 'infinite population' result times a 'correction factor' unhelpfully encourages confusing a *model* with the real world – recall the comment below Table HL77.1 at the end of the second asterisk (∗) on page HL77.1.

3. The *coefficient of variation* (*c.v.*) of $\overline{Y}$ [a measure of *relative* imprecision] is given in equation (HL77.14) at the right. Relative imprecision *de*creases [*i.e.*, $s.d.(\overline{Y})$ becomes *smaller relative to* $\overline{\mathbf{Y}}$] as n becomes larger and when the $\mathbf{Y}_i$ have less variation about their average $\overline{\mathbf{Y}}$.

$$c.v.(\overline{Y}) \equiv \frac{s.d.(\overline{Y})}{\overline{\mathbf{Y}}} = \sqrt{\frac{1}{\mathbf{N}-1}(\frac{1}{n} - \frac{1}{\mathbf{N}})\sum_{i=1}^{\mathbf{N}}(\frac{\mathbf{Y}_i}{\overline{\mathbf{Y}}} - 1)^2}$$ -----(HL77.14)

4. $\overline{Y}$ is the linear unbiased estimator of $\overline{\mathbf{Y}}$ with smallest standard deviation based on a sample of size n units selected by EPS (*e.g.*, see Barnett, pp. 26-27 and see also Appendix 2 in Figure 2.17 of the STAT 332 Course Materials).

5. The expression (HL77.9) for $s.d.(\overline{Y})$ under EPS is useful in three ways:
- it gives the imprecision of the estimator $\overline{Y}$;
- it allows us to calculate the approximate sample size needed to attain a specified imprecision for estimating $\overline{\mathbf{Y}}$ – see Figure 2.13 of the STAT 332 Course Materials.
- it allows us to compare the efficiency of $\overline{Y}$ with that of *other* estimators of $\overline{\mathbf{Y}}$.

A practical difficulty in using the expression (HL77.9) above for $s.d.(\overline{Y})$ is that $\mathbf{S}$ is usually *un*known; a way around this difficulty is to take the *sample* standard deviation s as s, an estimate of $\mathbf{S}$, ostensibly because of the following:

$$s^2 = \frac{1}{n-1}[\sum_{j=1}^{n} y_j^2 - n\overline{y}^2] = \frac{1}{n-1}\sum_{j=1}^{n} y_j^2 - \frac{n}{n-1}\overline{y}^2;$$ -----(HL77.15)

$$\therefore\ E(S^2) = \frac{1}{n-1}E[\sum_{j=1}^{n} Y_j^2] - \frac{n}{n-1}E(\overline{Y}^2) = \frac{1}{n-1}\underbrace{n\frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2}_{\text{n equal terms from } (\mathbf{ii})} - \frac{n}{n-1}\{\overline{\mathbf{Y}}^2 + [s.d.(\overline{Y})]^2\}$$

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

$$= \frac{n}{n-1}\{\frac{\mathbf{N}-1}{\mathbf{N}}\mathbf{S}^2 - (\frac{1}{n} - \frac{1}{\mathbf{N}})\mathbf{S}^2\} = \mathbf{S}^2;$$ -----(HL77.16)

$\{[s.d.(\overline{Y})]^2 = E(\overline{Y}^2) - [E(\overline{Y})]^2$, and: $E(\overline{Y}) = \overline{\mathbf{Y}}$ so that $E(\overline{Y}^2) = \overline{\mathbf{Y}}^2 + [s.d.(\overline{Y})]^2\}$

i.e., the random variable S^2 [with value s^2 taken as (the value of) the *square* of the sample (data) standard deviation under EPS] is an *un*biased estimator of $\mathbf{S}^2$, the *square* of the respondent population (data) standard deviation [but see Appendix 5 on page HL77.10].

Thus, the *estimated* standard deviation of $\overline{Y}$ under EPS is given by: $\hat{s.d.}(\overline{Y}) = s\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}}$. -----(HL77.17)

On the basis of the approximate normality of the distribution of $\overline{Y}$ as a consequence of the Central Limit Theorem, and arguing in a general way from the use of the t distribution in normal theory when the population standard deviation is estimated by the sample standard deviation, the theory developed above leads to an interval involving random variables:

$$I = [\overline{Y} - {}_{\alpha}t^{*}_{\mathrm{n}-1}S\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}},\ \overline{Y} + {}_{\alpha}t^{*}_{\mathrm{n}-1}S\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}] \qquad \text{-----(HL77.18)}$$

such that $\Pr(I \ni \overline{\mathbf{Y}}) \simeq 100(1-\alpha)\%$, where ${}_{\alpha}t^{*}_{\mathrm{n}-1}$ is the $100(1-\alpha/2)th$ percentile of the $t_{\mathrm{n}-1}$ distribution. For calculating an approximate $100(1-\alpha)\%$ *confidence interval* for $\overline{\mathbf{Y}}$, we use:

$$\overline{y} \pm {}_{\alpha}t^{*}_{\mathrm{n}-1}s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = [\overline{y} - {}_{\alpha}t^{*}_{\mathrm{n}-1}s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}},\ \overline{y} + {}_{\alpha}t^{*}_{\mathrm{n}-1}s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}]. \qquad \text{-----(HL77.19)}$$

NOTES: 6. The *first* expression in (HL77.19) is convenient when assessing sampling imprecision; the *second* is a more *direct* Answer.

7. The values of sample attributes determine two characteristics of the approximate confidence interval for $\overline{\mathbf{Y}}$ – the sample average defines its *centre*, the sample standard deviation determines its *width*; *both* characteristics (centre *and* width) of a confidence interval may be adversely affected by compromised selecting or measuring processes.

8. Using the t distribution requires that the population element (or unit) responses be *probabilistically independent* and *normally* distributed; in equiprobable selecting (with*out* replacement), successive observations are (weakly) dependent and not (necessarily) normally distributed, so this use of the t distribution has a weakened theoretical basis.
   - This weaker theoretical basis is one reason why the confidence interval expressions (HL77.19) are approximate.

9. An important consideration in assessing the agreement of the *stated* and *actual* levels of a confidence interval is how large the sample size needs to be for reasonable normality of $\overline{Y}$ as a consequence of the Central Limit Theorem; unfortunately, there is no reliable general rule but, when the deviation from normality is mainly a positive skewness, a crude rule (see Cochran, page 42) which is occasionally useful is that n should be greater than $25G_1^2$, where:

$$G_1 = \frac{1}{\mathbf{N}\mathbf{S}^3}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^3, \qquad \text{[estimated from the sample as: } g_1 = \frac{1}{\mathrm{n}s^3}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^3]. \qquad \text{-----(HL77.20)}$$

- The approximate normality of the distribution of $\overline{Y}$ is a second (related) reason why the confidence interval expressions (HL77.19) are approximate.

10. The results derived in (**i**), (**ii**) and (**iii**) on the fourth side (page HL77.4) of this Highlight, which provide the theoretical basis for the confidence interval expressions, all involve the equal *unit* selection probabilities that are a consequence of EPS and, in (**iii**), the *joint* probability $1/\mathbf{N}(\mathbf{N}-1)$, which comes from the formal requirement for equiprobable *samples* under EPS. There is thus ***no*** basis for using these expressions to calculate a confidence interval from a sample obtained by *other* selecting methods (accessibility, haphazard, judgement, quota, systematic, volunteer, etc.).
    - Likewise, throughout this Highlight #77, lower-case *italic* ys (values of random variables) are taken as the measured sample response variate *data* values (roman ys) on the basis of EPS as the sample selecting process. For *other* (*non*-probability) selecting processes, there is no reasonable basis for treating the model ys of the foregoing theory as the sample data ys (see Appendix 1 on pages HL77.8 and HL77.9 and also Note 11 at the top of page HL2.14 of Statistical Highlight #2).

11. The theory leading to equation (HL77.9) overleaf on page HL77.5 considers *only* sample error but using equation (HL77.17) when calculating a confidence interval for $\overline{\mathbf{Y}}$ or ${}_{\mathrm{T}}\mathbf{Y}$ involves using the *measured* sample y_js to calculate s (and s). As a consequence, the confidence expressions (HL77.19) above for $\overline{\mathbf{Y}}$ and (HL77.23) near the bottom of the facing page HL77.7 for ${}_{\mathrm{T}}\mathbf{Y}$ quantify *both* sample error *and* (fortuitously) measurement error.
    - We see that this is so by considering a respondent population whose elements all have the *same* $\mathbf{Y}$-value; variation in the y_js would then reflect *only* measurement error. Hence, in the usual case of *varying* $\mathbf{Y}_i$s, the measured y_j values reflect both sample and measurement error.
    - The confidence interval expressions (HL77.19) above and (HL77.23) on page HL77.7 quantify the *combined* uncertainty due to sample error and measurement error – their effects could be estimated *individually* if replicate measurements were to be made on the sample units, but this is rare in sample surveys because there would be little benefit, extra cost and the difficulty of maintaining (real-world) *independence* of replicate measurements, especially when the population elements are humans and the measuring instrument is a questionnaire.

    This matter is the survey sampling analogue of the theory developed in Statistical Highlight #74 for the model (HL77.2) on page HL77.1 – for example, recall equation (HL74.18) on page HL74.4.
    - It would be useful if the ('finite population') theory in *this* Highlight #77 could inform that of Statistical Highlight #74 so that it would be correct, when the population size is $\mathbf{N}$, to write equation (HL74.18) on page HL74.4 as equation (HL77.21).

$$\overline{Y} \sim N(\mu, \sigma\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}) \qquad \text{----(HL77.21)}$$

Example HL77.2: In an audit of hospital accounts, 200 accounts were obtained by equiprobable selecting from a total of 1,000 accounts; for all 200 accounts, the sample average was $\overline{\mathrm{y}}$ = \$394.22 and the sample standard deviation was s = \$21.10. Find an approximate 90% confidence interval for the *average* amount per account ($\overline{\mathbf{Y}}$) at the hospital.

University of Waterloo W. H. Cherry

SAMPLING and SURVEY SAMPLING: Estimating a Population Average or Total (continued 3)

Solution: We have: $\mathbf{N} = 1{,}000$, $\mathrm{n} = 200$, $\bar{\mathrm{y}} = \$394.22\ (= \bar{y})$, $\mathrm{s} = \$21.10\ (= s)$, ${}_{.1}t^*_{199} = 1.65255$ for 90% confidence.

Then: $s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 21.10\sqrt{\frac{1}{200} - \frac{1}{1{,}000}} = \$1.334\,481\,173.$

Hence, an approximate 90% confidence interval for $\overline{\mathbf{Y}}$, the average amount per account at the hospital, is:

$\bar{y} \pm 1.65255 \times \hat{s.d.}(\bar{Y}) = 394.22 \pm 1.65255 \times 1.334\,482 \implies (392.01,\ 396.43)$ or about (\$392, \$397).

NOTES: 12. In calculating the 90% confidence interval for $\overline{\mathbf{Y}}$, the value of ${}_{.1}t^*_{199} = 1.65255$ has been obtained by *linear interpolation* from the relevant entries (*viz*., 1.65291 and 1.65251) from a *t*-table like Figures 6.6 and 2.4 on pages 6.38 and 2.52 in the STAT 220 and STAT 332 Course Materials for 190 and 200 degrees of freedom.

13. In calculations like those in the solution of Example HL77.2, we must use and show enough significant figures to avoid rounding inaccuracy; however, it is an *essential* part of a proper solution to give a *final* answer rounded to a number of figures appropriate to the Question context.

Example HL77.3: In the same hospital as in Example HL77.2, n = 9 accounts were obtained by equiprobable selecting from the total of 484 open accounts; the data, and their numerical summaries, were as follows:

\$333.50 332.00 352.00 343.00 340.00 341.00
\$345.00 342.50 339.00

$\sum_{j=1}^{9} \mathrm{y}_j = 3{,}068.00$, $\sum_{j=1}^{9} \mathrm{y}_j^2 = 1{,}046{,}132.50.$

Find an approximate 95% confidence interval for the average amount per open account at the hospital.

Solution: The solution of this Example HL77.3 is like that of Example HL77.2 except *we* must calculate the values of $\bar{\mathrm{y}}$ and s from the numerical summaries of the sample data.

We have: $\mathbf{N} = 484$, $\mathrm{n} = 9$, $\bar{\mathrm{y}} = \frac{3{,}068.00}{9} = \$340.\dot{8}\ (= \bar{y})$, ${}_{.05}t^*_8 = 2.30600$ for 95% confidence,

$$\mathrm{s} = \sqrt{\frac{1{,}046{,}132.50 - 3{,}068.00^2/9}{8}} = \$5.972\,739\,112\ (= s).$$

Then: $s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 5.97274\sqrt{\frac{1}{9} - \frac{1}{484}} = \$1.972\,315\,621.$

Hence, an approximate 95% confidence interval for $\overline{\mathbf{Y}}$, the average amount per open account at the hospital, is:

$\bar{y} \pm 2.30600 \times \hat{s.d.}(\bar{Y}) = 340.\dot{8} \pm 2.30600 \times 1.972\,316 \implies (336.34,\ 345.44)$ or about (\$336, \$346).

NOTES: 14. Despite the small sample size of 9, the confidence interval is, as in Example HL77..2, relatively *narrow* (*i.e*., the Answer shows relatively *low* imprecision for estimating $\overline{\mathbf{Y}}$) because the value of the population standard deviation $\mathbf{S}$ is *small* in relation to the value of $\overline{\mathbf{Y}}$, as indicated by their estimates from the sample of about \$6 for s and about \$340 for $\bar{y}$.

- The small sample size in Example HL77.3, where *all* the sample data are given, is only for classroom convenience; a real sample survey like this would usually have a *much* larger sample size.

15. For interest, we can carry out the check discussed in Note 9 on the facing page HL77.6 for the adequacy of the sample size with respect to the assumed normality of $\bar{Y}$; for convenience, estimating the sum of cubes from the sample data is set out in Table HL77.6 at the right; dividing the sum of cubes by $\mathrm{n}s^3 = 1{,}917.622\,626$, we find $25g_1^2 = 4.453\,978$, which *is* less than n = 9 as the check requires.

Table HL77.6:

j	y_j	$\bar{y}$	$y_j - \bar{y}$	$(y_j - \bar{y})^3$
1	333.50	$340.\dot{8}$	$-7.3\dot{8}$	−403.401 405
2	332.00	$340.\dot{8}$	$-8.\dot{8}\dot{8}$	−702.331 959
3	352.00	$340.\dot{8}$	$11.\dot{1}$	1,371.742 116
4	343.00	$340.\dot{8}$	$2.\dot{1}$	9.408 779
5	340.00	$340.\dot{8}$	$-0.\dot{8}$	−0.702 332
6	341.00	$340.\dot{8}$	$0.\dot{1}$	0.001 372
7	345.00	$340.\dot{8}$	$4.\dot{1}$	69.482 854
8	342.50	$340.\dot{8}$	$1.6\dot{1}$	4.181 927
9	339.00	$340.\dot{8}$	$-1.\dot{8}$	−6.739 369
				341.641 983

6. Estimating ${}_{\mathrm{T}}\mathbf{Y}$, the Respondent Population Total

Under the assumption that the population size, $\mathbf{N}$, is a *known constant*, the theory of equiprobable selecting for estimating ${}_{\mathrm{T}}\mathbf{Y}$ is a straight-forward extension of the results for $\overline{\mathbf{Y}}$. Because the population total is ${}_{\mathrm{T}}\mathbf{Y} = \mathbf{N}\overline{\mathbf{Y}}$, its estimator is $\mathbf{N}\bar{Y}$; the standard deviation of this estimator is then $\mathbf{N} \times s.d.(\bar{Y})$. Hence, we obtain an interval involving random variables:

$$I = [\mathbf{N}\bar{Y} - {}_{\alpha}t^*_{\mathrm{n}-1}\mathbf{N}S\sqrt{\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}},\ \mathbf{N}\bar{Y} + {}_{\alpha}t^*_{\mathrm{n}-1}\mathbf{N}S\sqrt{\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}}] \qquad \text{-----(HL77.22)}$$

such that $\Pr(I \ni {}_{\mathrm{T}}\mathbf{Y}) \simeq 100(1-\alpha)\%$, where ${}_{\alpha}t^*_{\mathrm{n}-1}$ is the $100(1-\alpha/2)th$ percentile of the $t_{\mathrm{n}-1}$ distribution. For calculating an approximate $100(1-\alpha)\%$ *confidence interval* for ${}_{\mathrm{T}}\mathbf{Y}$, we use:

$$\mathbf{N}\bar{\mathrm{y}} \pm {}_{\alpha}t^*_{\mathrm{n}-1}\mathbf{N}s\sqrt{\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}} = [\mathbf{N}\bar{\mathrm{y}} - {}_{\alpha}t^*_{\mathrm{n}-1}\mathbf{N}s\sqrt{\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}},\ \mathbf{N}\bar{\mathrm{y}} + {}_{\alpha}t^*_{\mathrm{n}-1}\mathbf{N}s\sqrt{\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}}]. \qquad \text{-----(HL77.23)}$$

(*continued overleaf*)

Example HL77.4: A company was concerned about the time per week its 750 managers spent on unimportant tasks. For 50 managers obtained by equiprobable selecting, it was found that the average time spent on such tasks was 10.31 hours and the standard deviation was 1.5 hours. Find an interval estimate for the total person-hours spent per week by the 750 managers on these unimportant tasks.

Solution: Estimating $_{\mathrm{T}}\mathbf{Y}$ is like estimating $\overline{\mathbf{Y}}$ except that we must multiply by $\mathbf{N}$ in appropriate places; the solution of this Example HL77.4 therefore follows the pattern of Examples HL77.2 and HL77.3.

We have: $\mathbf{N} = 750$, $\mathrm{n} = 50$, $\overline{\mathrm{y}} = 10.31$ hours ($= \overline{y}$), $\mathrm{s} = 1.5$ hours ($= s$), $_{.05}t^{*}_{49} = 2.00958$ for 95% confidence.

Then: $s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 1.5\sqrt{\frac{1}{50} - \frac{1}{750}} = 0.204\,939\,015$ hours.

Hence, an approximate 95% confidence interval for $_{\mathrm{T}}\mathbf{Y}$, the total number of hours spent by the 750 managers on unimportant tasks, is:

$$\mathbf{N}\overline{y} \pm 2.00958 \times \mathbf{N} \times \widehat{s.d.}(\overline{Y}) = 750 \times 10.31 \pm 2.00958 \times 750 \times 0.204\,939 \implies (7{,}424,\ 8{,}041)$$

or about (7,400, 8,100) hours per week.

NOTES: 16. When the confidence *level* is not specifed in the Question, we take the default as 95%.

17. Population *totals* are often *large* numbers and so the widths of confidence intervals for $_{\mathrm{T}}\mathbf{Y}$ may be large in *absolute* terms but not necessarily large *relative* to the magnitude of $_{\mathrm{T}}\mathbf{Y}$.

18. It would be difficult to implement an accurate and precise measuring system for quantifying personal time useage for activities like those in Example HL77.4; this is why the end points of the final confidence interval have been rounded to only *two* significant digits.

Example HL77.5: One hundred water meters, obtained by equiprobable selecting from a community of 10,000 households, are monitored over a particular dry spell of weather. For all 100 meters, the sample average and standard deviation (in suitable units) are found to be 12.5 and 35.4 respectively. Find an approximate 99% confidence interval for the *total* water consumption in the community during the dry spell.

Solution: We have: $\mathbf{N} = 10{,}000$, $\mathrm{n} = 100$, $\overline{\mathrm{y}} = 12.5$ units ($= \overline{y}$), $\mathrm{s} = 35.4$ units ($= s$), $_{.01}t^{*}_{99} = 2.62641$ for 99% confidence.

Then: $s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 35.4\sqrt{\frac{1}{100} - \frac{1}{10{,}000}} = 3.522\,255\,527$ units.

Hence, an approximate 99% confidence interval for $_{\mathrm{T}}\mathbf{Y}$, the total water consumption of the 10,000 households over the dry spell, is:

$$\mathbf{N}\overline{y} \pm 2.62641 \times \mathbf{N} \times \widehat{s.d.}(\overline{Y}) = 10{,}000 \times 12.5 \pm 2.62641 \times 10{,}000 \times 3.522\,256 \implies (32{,}491,\ 217{,}509)$$

or about (32,000, 220,000) units.

NOTE: 19. The *wide* confidence interval (*i.e.*, the high *imprecision*) for estimating $_{\mathrm{T}}\mathbf{Y}$ in Example HL77.5 is mainly a consequence of a very *variable* population of household water consumptions: $\mathrm{s}/\overline{\mathrm{y}} = 283\%$. Because water consumption is an inherently *non*-negative quantity, these sample attribute values suggest a highly (positively) skewed population distribution of water consumptions which, in turn, raises concerns about the accuracy of the stated confidence *level* of an interval based on the t distribution. In a real sample survey, this matter would need to be followed up.

- The high imprecision for estimating $_{\mathrm{T}}\mathbf{Y}$ in Example HL77.5 could be managed by *stratifying* the population into groups of households more *homogeneous* with respect to their water consumptions; stratifying is discussed briefly in Statistical Highlight #63 on page HL63.4 in Note 6, in Highlight #74 on page HL74.10, in Highlight #88 on page HL88.12 in Note 18 and, in more detail, in Part 4 of the STAT 332 Course Materials.

7. REFERENCES:

1. Barnett, V. *Sample Survey Principles and Methods.* Second edition, Edward Arnold, London, 1991; (*First* edition: *Elements of Sampling Theory.* The English Universities Press Ltd., London, 1974).
2. Cochran, W. G. *Sampling Techniques.* John Wiley & Sons, Inc., New York, 3rd Edition, 1977.

8. Appendix 1: Notation – Managing Falsehood

To help maintain the key distinction between the real world and the model, *we* distinguish an *observed* (Roman) y from a *model* (italic) y – recall Table HL77.5 on the upper half of page HL77.4; this leads us to distinguishing the observed sample average and standard deviation, $\overline{\mathrm{y}}$ and s, from the model quantities $\overline{y}$ and s. This distinction in turn reminds us of the importance in the Design stage of the FDEAC cycle of developing a Plan that makes it reasonable to treat $\overline{\mathrm{y}}$ as $\overline{y}$ and s as s.

Failure to observe the foregoing distinctions makes it easier to overlook the question to be asked about *all* data: *How were they*

SAMPLING and SURVEY SAMPLING: Estimating a Population Average or Total (continued 4)

generated? If this question is *not* asked, it becomes more likely that statistical methods will be applied in situations where their underlying assumptions (like probability selecting, accurate measuring processes, independent measuring, distributional assumptions) are violated, leading to answers with such severe limitations that they are likely to embody falsehoods.

- Aside from notation, the process for calculating $\bar{y}$ and $\bar{y}$ is the *same*, and likewise for calculating s and s – this latter process is also that for $\hat{\sigma}$ in equation (HL77.3) at the bottom right of page HL77.1.
- Our error categories (more specifically study error, non-response error, sample error, measurement error and model error in the case of sample surveys) provide a framework for assessing the severity of the limitations imposed by error on answers.

9. Appendix 2: Population Elements and Population Units

As discussed in Section 1 in the middle of page HL75.1 in Statistical Highlight #75, we distinguish:

- **Elements:** the entities that make up a population; for example, a person is an element of the population of Canadians, but we recognize that many populations in data-based investigating have non-human or *in*animate elements.
- **Units:** the entities *selected* for the sample; a unit may be one element (*e.g.*, a person) or *more than* one (*e.g.*, a household).

This Highlight #77 is concerned with (survey) *sampling* and so refers in most places to units, but population attributes of interest (like $\mathbf{N}$, $\bar{\mathbf{Y}}$, ${}_T\mathbf{Y}$ and $\mathbf{S}$) refer to *elements*. In introductory courses like STAT 220 and STAT 231, we restrict attention to units which are elements so the distinction is of no consequence but in Figures 2.14 and 2.16 in STAT 332 for example, when units are *groups* of elements (as in *cluster* sampling), some expressions in the theory must be modified. This is illustrated in Table HL77.8 at the right below by comparing expressions in this Highlight #77 with those for selecting *equal*-sized clusters, like cardboard boxes in a supermarket that each contain, say, 24 cans of soup, or cartons from a component manufacturing process that each contain a set number (say, 10) of the component.

Table HL77.7 at the right gives additional notation we need, where the repondent population is considered as $\mathbf{N}$ elements of which n are selected (by EPS) for the sample, or as $\mathbf{M}$ clusters each of L elements, of which m are selected (by EPS) to yield a sample of mL elements.

Table HL77.7:

	Elements	Clusters	Relationships
Respondent population	$\mathbf{N}$	$\mathbf{M}$	$\mathbf{N} = \mathbf{M}\mathrm{L}$, $\mathrm{L} = \mathbf{N}/\mathbf{M}$
Sample	n	m	$\mathrm{n} = \mathrm{mL}$, $\mathrm{L} = \mathrm{n}/\mathrm{m}$

Also: $\bar{\mathbf{Y}}_{\mathrm{i}} = \frac{1}{\mathrm{L}}\sum_{k=1}^{\mathrm{L}}\mathbf{Y}_{\mathrm{i}k}$ is the average response of the ith population cluster,

$\bar{y}_j = \frac{1}{\mathrm{L}}\sum_{k=1}^{\mathrm{L}} y_{jk}$ is the average response of the jth sampled cluster,

the subscript *ec* in Table HL77.8 below denotes 'equal-sized clusters'.

Given the respondent population 'models' of $\mathbf{N}$ elements or $\mathbf{M}$ clusters, the structural similarity of corresponding expressions in the two columns of Table HL77.8 are clear; noteworthy points are:

- when estimating $\bar{\mathbf{Y}}$, $\bar{y}$ involves *element* responses y_j but $\bar{y}_{ec}$ involves cluster *average* responses $\bar{y}_j$;
- $\mathbf{S}_{ec}$ (estimated by s_{ec}), which quantifies variation of *cluster averages* in the respondent population, is to be distinguished from the variation of *element* responses quantified by $\mathbf{S}$ (estimated by s).

The cluster sampling expressions in the right-hand column of Table HL77.8 are taken from Figure 2.14 of the STAT 332 Course Materials. The theory for *un*equal-sized clusters is more complicated – see Figure 2.16 of those Course Materials.

Table HL77.8

EPS of elements	Page	EPS of clusters
$\bar{y} = \frac{1}{\mathrm{n}}\sum_{j=1}^{\mathrm{n}} y_j$	HL77.4	$\bar{y}_{ec} = \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}} \bar{y}_j$
$s = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \bar{y})^2}$	HL77.4	$s_{ec} = \sqrt{\frac{1}{\mathrm{m}-1}\sum_{j=1}^{\mathrm{m}}(\bar{y}_j - \bar{y}_{ec})^2}$
$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\mathbf{Y}_{\mathrm{i}} - \bar{\mathbf{Y}})^2}$	HL77.4	$\mathbf{S}_{ec} = \sqrt{\frac{1}{\mathbf{M}-1}\sum_{\mathrm{i}=1}^{\mathbf{M}}(\bar{\mathbf{Y}}_{\mathrm{i}} - \bar{\mathbf{Y}})^2}$
$s.d.(\bar{Y}) = \mathbf{S}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}$	HL77.5	$s.d.(\bar{Y}_{ec}) = \mathbf{S}_{ec}\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}}$
$\bar{y} \pm {}_{\alpha}t^*_{\mathrm{n}-1}s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}$	HL77.6	$\bar{y}_{ec} \pm {}_{\alpha}t^*_{\mathrm{m}-1}s_{ec}\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}}$

10. Appendix 3: Representative Sampling

The appealing intuitive idea of a 'representative sample' – one that 'looks like' the (respondent) population with respect to the attribute(s) of interest – is equivocal statistically for four reasons:

- a sample selected by EPS is unlikely to be 'representative' in the sense just given for *all* attributes of potential interest – for instance, a sample may have small [possibly (close to) zero] sample error for estimating $\bar{\mathbf{Y}}$ but large sample error for estimating $\mathbf{S}$;
- the sample, of itself, provides no information about its 'representativeness';
- there is no selecting process known to yield a 'representative' sample, except taking a census;
- the terminology tends to obscure the distinction between the individual case (the *particular* sample) and behaviour under repetition (the properties of the selecting *process*).

Less equivocal terminology is *representative sampling*, with its implication of a selecting *process* (like EPS) which, in conjunction with adequate replicating, provides for quantifying sampling imprecision and so allows a particular investigation to obtain an Answer with acceptable limitation in the Question context due to sample error. However, the writer's preference is to *avoid* in statistics the terms 'representative' and 'representativeness' when referring to a sample (or a sampling protocol).

- Kruskal and Mosteller devote 50 pages to discussing the (sometimes ill-defined) meanings in statistical contexts of **representative sampling** in three articles in the *International Statistical Review*, **47**, 13-24, 111-127, 245-265 (1979). [UW Library call number HA 11.I505]

NOTE: 20. An illustration, involving bivariate data, of another instance of sample-attribute dependence is:

(*continued overleaf*)

NOTE: (cont.) 20. ● when estimating the least squares *slope* of a straight-line relationship, sample points concentrated more near the *ends* of the interval of observation will reduce sampling imprecision (although this will *in*crease imprecision of any inference needed to show that the relationship *is* a straight line);

● similar considerations apply when estimating *correlation*, although estimating this attribute is rarely discussed.

11. Appendix 4: Terminology – Finite Populations

The order of topics in most current introductory statistics texts means that, before students can encounter ideas like $s.d.(\overline{Y})$ in equation (HL77.9) near the top of page HL77.5 or $s.d.(P)$ in equation (2.10.4) on page 2.81 in Figure 2.10, they are presented with two (likely new) ideas.

- For (much beloved of statisticians) n probabilistically independent identically distributed (abbreviated iid) random variables (Y_j, say) each with mean μ and standard deviation σ, the random variable $\overline{Y}$ that is their *average* has the *same* mean μ but a (smaller) standard deviation of $\sigma/\sqrt{\mathrm{n}}$ [equation (HL77.24)].

$$s.d.(\overline{Y}) = \sigma/\sqrt{\mathrm{n}} \equiv \sigma\sqrt{\frac{1}{\mathrm{n}}} \qquad \text{-----(HL77.24)}$$

- The standard deviation for a binomial *proportion* also has $\sqrt{\mathrm{n}}$ in its denominator, as discussed in Figure 2.10 near the bottom of page 2.82 in the second bullet (●) of Note 7.

However, with the accolades it tends to receive from statisticians, this useful insight that averages are *less* variable than their components by a factor of $\sqrt{\mathrm{n}}$ can become an ideal that other relevant statistical theory should try to emulate; in the case of survey sampling theory, pursuit of this 'ideal' is detrimental to statistics and its teaching in three ways.

∗ Equations (HL77.9) and (HL77.17) near the top and bottom of page HL77.5 are written in their second (equivalent) forms shown as equations (HL77.25) and (HL77.26) at the right, as are equations (2.10.4) and (2.10.6) from page 2.81 of Figure 2.10 of the STAT 332 Course Materials, where $f = \mathrm{n}/\mathbf{N}$; equation (2.10.6) is even rewritten as (the inelegant) equation (HL77.27) to yield (the mandatory?) 1/n under the square root. [$\mathbf{P}$ is the respondent population *proportion* of elements with some characteristic C and $\mathbf{Q} = 1-\mathbf{P}$ is the proportion with*out* this characteristic; p and q are the corresponding *sample* attribute values.]

$$s.d.(\overline{Y}) = \mathbf{S}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \equiv \sqrt{(1-f)}\,\mathbf{S}\sqrt{\frac{1}{\mathrm{n}}} \qquad \text{-----(HL77.25)}$$

$$\hat{s.d.}(\overline{Y}) = s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \equiv \sqrt{(1-f)}\,s\sqrt{\frac{1}{\mathrm{n}}} \qquad \text{-----(HL77.26)}$$

$$s.d.(P) = \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)} \equiv \sqrt{(1-f)\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}\frac{1}{\mathrm{n}}} \qquad \text{-----(2.10.4)}$$

$$\hat{s.d.}(P) = \sqrt{\frac{\mathrm{n}}{\mathrm{n}-1}pq\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)} \equiv \sqrt{(1-f)pq\frac{1}{\mathrm{n}-1}} \qquad \text{-----(2.10.6)}$$

$$= \sqrt{(1-f)\frac{\mathrm{n}}{\mathrm{n}-1}pq\frac{1}{\mathrm{n}}} \qquad \text{-----(HL77.27)}$$

$$s.d.(\overline{Y})_{\mathrm{EPSWIR}} = \sqrt{\frac{\mathbf{N}-1}{\mathbf{N}}}\,\mathbf{S}\sqrt{\frac{1}{\mathrm{n}}} \qquad \text{-----(HL77.28)}$$

- $(1-f)$ in the five second forms makes *less* clear the relative roles of n and $\mathbf{N}$ [see the third bullet (●) of Note 1 on the upper half of page HL77.5] and in equations (HL77.25) [and (HL77.26)] makes it appear (wrongly) that the (data) s.d.s $\mathbf{S}$ [and s] are equivalent to the (probabilistic) s.d. σ – comparing equations (HL77.24) and (HL77.28) at the right above shows that *actual* equivalence involves a premultiplier (although its value is typically *very* close to 1) for $\mathbf{S}$ – see also equation (HL77.6) on page HL77.4.
  - ○ Equation (HL77.28) is derived from the *first* term of equation (HL77.9) near the top of page HL77.5 when the *second* covariance term is zero due to probabilistic independence under equiprobable selecting *with* replacement (EPSWIR).
  - ○ We note that the *reciprocal* of the premultiplier of $\mathbf{S}$ in equation (HL77.28) occurs in equation (2.10.4) to convert the probabilistic (binomial proportion) s.d. $\mathbf{P}\mathbf{Q}$ to the *data* s.d. needed when selecting is with*out* replacement.
  - ○ When $\mathrm{n} = 1$, the RHS of equation (HL77.28) is that of equation (HL77.6) on page HL77.4.
  - ○ When $\mathrm{n} = 1$, equations (HL77.25) and (HL77.28) become the *same*, reminding us that selecting without and with replacement no longer differ for a sample of size one. Regrettably, this common special case of equations (HL77.25) and (HL77.28) has been made part of the (unprofitable) $\mathrm{n}-1$ *vs* n saga – see Statistical Highlights #94 and #100 (the lower half of page HL94.9, the middle of page HL100.8).

∗ $1-f$ is called the *finite population correction* (abbreviated f.p.c.); such a name conveys the bizarre misconception that we should think of the real world as a 'correction' to a model.

- As these Materials maintain throughout, populations in statistics are real-world entities with *finite* numbers of elements and they should *not* be confused with (sometimes useful) infinite *models*.

∗ The idea of an infinite 'population' is, in part, a confusion of selecting with*out* replacement, which ends when all elements have been selected, and selecting *with* replacement which can, in principle, continue indefinitely. However, as the discussion in Note 7 on page 2.82 in Figure 2.10 illustrates, an urn model has a *finite* number of elements even when selecting is with replacement (and so could continue indefinitely) under a binomial model.

12. Appendix 5: The Mean of S, $E(S)$

The justification in equation (HL77.17), at the bottom of page HL77.5, for using s to estimate $\mathbf{S}$ is compromised by the fact that S is *not* an unbiased estimator of $\mathbf{S}$. Because of the square root in the expression for s in Table HL77.5 on page HL77.4, there is no simple expression for the estimating bias of the corresponding random variable S under EPS, but we know that bias exists from the following argument, which is an illustration of Jensen's Inequality and uses the fact that the variance of any

SAMPLING and SURVEY SAMPLING: Estimating a Population Average or Total (continued 5)

(non-constant) random variable is positive. We have:

$$0 < var(S) = E(S^2) - [E(S)]^2 = \mathbf{S}^2 - [E(S)]^2 \quad \text{so that, taking square roots:} \quad E(S) - \mathbf{S} < 0. \qquad \text{-----(HL77.29)}$$

NOTE: 21. For the model (HL77.2) on page HL77.1, the mathematics is more tractable and leads to equation (HL77.30) at the right so that, because of equation (HL2.38) on page HL2.15 of Statistical Highlight #2, rewritten at the right as equation (HL77.31), the bias term multiplying σ on the RHS of equation (HL77.30) is the *mean* of a K_{n-1} distribution. Table HL2.10 of its values for n = 2 to 51 (*i.e.*, for 1 to 50 degrees of freedom) on page HL2.17 of Statistical Highlight #2 reminds us that estimating bias:

$$E(\tilde{\sigma}) = \sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\sigma \qquad \text{-----(HL77.30)}$$

$$\frac{\tilde{\sigma}}{\sigma} \sim K_{n-1} \quad \text{or:} \quad \tilde{\sigma} \sim \sigma K_{n-1} \qquad \text{-----(HL77.31)}$$

- decreases in magnitude with increasing sample size, *un*like (real-world) inaccuracy – see Example HL77.6 on page HL77.13;
- of S as an estimator of $\mathbf{S}$ is likely to be unimportant practically for the sample sizes used in most real sample surveys.

13. Appendix 6: Bias and Rms Error

For a random variable Y and some constant c, we have:

$$\begin{aligned} E\{[Y-c]^2\} &= E\{[E(Y) - c + Y - E(Y)]^2\} = E\{[E(Y)-c]^2 + [Y-E(Y)]^2 + 2[E(Y)-c][Y-E(Y)]\} \\ &= E\{[E(Y)-c]^2\} + E\{[Y-E(Y)]^2\} + 2E\{[E(Y)-c][Y-E(Y)]\} \\ &= [E(Y-c)]^2 + E\{[Y-E(Y)]^2\} + 2[E(Y)-c]E[Y-E(Y)] \end{aligned}$$

$$i.e., \quad E\{[Y-c]^2\} = [E(Y-c)]^2 + [s.d.(Y)]^2 \qquad \text{because } E[Y-E(Y)] \equiv 0. \qquad \text{-----(HL77.32)}$$

If we now think of Y as a random variable whose distribution represents the possible values of a *response variate* $\mathbf{Y}$ and c as a *true* value, the left-hand side of equation (HL77.32) is a *mean squared error* and $E(Y-c)$ in the first term on the right-hand side is a *bias*; we can therefore interpret equation (HL77.32) as:

$$\text{mean squared error} = \text{bias}^2 + \text{standard deviation}^2. \qquad \text{-----(HL77.33)}$$

Taking the square root so we are working on the *same* scale as the variate represented by Y, the **root mean squared error** is:

$$\text{rms error} = \sqrt{\text{bias}^2 + \text{standard deviation}^2}. \qquad \text{-----(HL77.34)}$$

Thus, the rms error is *one* concept that *combines* the two model quantities of bias and (probabilistic) standard deviation, or it can be used as a model for the two corresponding real-world entities of inaccuracy and imprecision.

○ The word 'mean' on the LHS of equation (HL77.34) invokes *behaviour under repetition*, consistent both with the RHS of this equation involving *model* quantities and with the two corresponding real-world entities being inaccuracy and imprecision.

Equation (HL77.34) provides useful insights about bias and variability in the context of survey sampling; different cases depend on how broad our focus is in terms of *which* true value c represents – see also the discussion and diagram showing four components of overall *error* on the upper half of page HL6.6 in Statistical Highlight #6.

* The narrowest focus is *measuring* when c is the true value of the response variate $\mathbf{Y}$; equation (2.3.34) is then:

$$\text{measuring rms error} = \sqrt{\text{measuring bias}^2 + \text{measuring standard deviation}^2}. \qquad \text{-----(HL77.35)}$$

* For measuring *and* sampling, c is the true value of the *respondent* population attribute of $\mathbf{Y}$ and then:

$$\begin{matrix}\text{measuring } and \text{ sampling} \\ \text{rms error}\end{matrix} = \sqrt{\text{measuring + sampling bias}^2 + \text{measuring and sampling standard deviation}^2}; \qquad \text{-----(HL77.36)}$$

NOTE: 22.
$$\begin{matrix}\text{Measuring and sampling} \\ \text{standard deviation}\end{matrix} = \sqrt{\text{measuring standard deviation}^2 + \text{sampling standard deviation}^2}. \qquad \text{-----(HL77.37)}$$

* For measuring *and* sampling *and* non-responding, c is the true value of the *study* population attribute of $\mathbf{Y}$ and then, under our assumption that non-response is *deterministic* (*not* stochastic):

$$\begin{matrix}\text{measuring and sampling and} \\ \text{non-responding rms error}\end{matrix} = \sqrt{\begin{matrix}\text{measuring + sampling} \\ \text{+ non-responding bias}^2\end{matrix} + \text{measuring and sampling standard deviation}^2}. \qquad \text{-----(HL77.38)}$$

* For measuring *and* sampling *and* non-responding *and* specifying, c is the true value of the *target* population attribute of $\mathbf{Y}$ and then, under our assumption that specifying the study population also is *deterministic*:

$$\begin{matrix}\text{measuring and sampling} \\ \text{and non-responding and} \\ \text{studying rms error}\end{matrix} = \sqrt{\begin{matrix}\text{measuring + sampling} \\ \text{+ non-responding} \\ \text{+ studying bias}^2\end{matrix} + \text{measuring and sampling standard deviation}^2}. \qquad \text{-----(HL77.39)}$$

NOTE: 23. In printed materials other than these Course Materials (*e.g.*, see Cochran, p. 15), equation (HL77.32) [or (HL77.33)] is usually discussed only with respect to *estimating* bias. Although estimating bias is a relatively minor topic in STAT 332, it is useful to recognize the following [recall also Example HL77.1 on pages HL77.3 and HL77.4]:

NOTE: (cont.) 23. ● **Estimating bias** (a *model* quantity) is the difference between the mean of an estimator and the value of the corresponding population attribute (or model parameter); for example, under EPS:

– the random variable $\overline{Y}$ representing the sample *average* $\overline{y}$ is an *un*biased estimator of the respondent population average $\overline{\mathbf{Y}}$ because, as shown in equation (HL77.8) near the top of page HL77.5 of this Highlight #77, $E(\overline{Y}) = \overline{\mathbf{Y}}$ or $E(\overline{Y}) - \overline{\mathbf{Y}} = 0$; BUT:

– the random variable R representing the sample ratio $r = \overline{y}/\overline{x}$ is a *biased* estimator of the respondent population ratio $\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}}$ because $E(R) \neq \mathbf{R}$ or $E(R) - \mathbf{R} \neq 0$, and likewise for S as an estimator of $\mathbf{S}$ as discussed on pages HL77.10 and HL77.11 in Appendix 5.

● Rms error is of interest because, while we prefer the likely smaller sample error of an *un*biased estimator of a population attribute, a *biased* estimator may occasionally have only *small* bias and appreciably *smaller* standard deviation than an available *un*biased estimator; we *may* then prefer the biased estimator with *smaller* rms error.

● *Un*like (real-world) inaccuracy, estimating bias *de*creases in magnitude with increasing sample size (as discussed in Appendix 5 on pages HL77.10 and HL77.11 in Note 21).

14. Appendix 7: Sampling Bias/Inaccuracy

∗ **Sample error:** the difference between [the (true) values of] the sample and respondent population attributes.
∗ **Sampling inaccuracy:** average sample error under repetition of selecting and estimating.
∗ **Sampling bias:** the model quantity representing sampling inaccuracy.
∗ **Sampling** involves *both* selecting and estimating.
∗ **Probability selecting:** a process for which every study population element has a known non-zero selection probability.
∗ **Estimating:** a process which uses statistical theory to derive the distribution of an *estimator* and data to calculate an (interval) *estimate.*

Sample error arises in the context of a *particular* investigation; inaccuracy (and bias) arise only under *repetition.*

Quantifying sample error is (of course) the reason for the theory presented in Section 5 on pages HL77.4 to HL77.6; incomplete knowledge that is the unavoidable consequence of sampling (and measuring) means that:

● the theory quantifies the behaviour of sample error *only* under repetition, **AND:**
● modelling *assumptions* (including the special case of *equi*probable selecting in this Highlight #77) must be met under the Plan for an investigation for the theory to be applicable – recall Note 10 near the middle of page HL77.6.

Because sampling involves selecting and estimating, sampling *bias* in equations (HL77.36), (HL77.38) and (HL77.39) overleaf on page HL77.11 [which *models* real-world sampling inaccuracy] involves two components:

○ **Selecting bias:** the model quantity representing selecting inaccuracy, which is average selection error under repetition of selecting; its context is *non*-probability selecting methods, like accessibility, convenience, haphazard, quota, volunteer or judgement selecting, although this introduces the substantial difficulty of modelling bias under *non*-probability selecting processes.

Another difficulty is that we are *rarely* in a position to identify selecting bias because of an absence of data from many repetitions of one of these (non-probability) selecting methods exectuted in exactly the same way each time in the same context. However, a fortuitous instance may be a sampling exercise used over more than a decade in teaching introductory statistics in the 4-year Bachelor of Mathematics program at the University of Waterloo.

A population of 100 'blocks' (irregular polygons cut from 6-mm grey plastic sheet, numbered from 1 to 100) was laid out on a table in the classroom and each of the (50 to 80) students selected a sample of 10 blocks by EPS (using a table of equiprobable digits) and by judgement selecting. From a list of the 100 block weights, each student calculated their two sample averages, which were then used by the instructor to construct, on an overhead projector at the front of the classroom, a bar-graph (in 2-gram intervals) of the averages from each selecting method.

– Under EPS, the bar-graph was usually centred close to the population average block weight (32.4 grams), was roughly normal (or at least roughly symmetrical), and had most of its values within about 10 grams of its centre.
– Under judgement selecting, the centre of the bar-graph was typically at least 40 grams (more than 20% too *high*), the shape was more 'ragged' and the width was appreciably greater than for EPS.

Although this is a *restricted* sampling context, the persistence of sampling *inaccuracy* under judgement selecting is noteworthy – no substantial exception to the characteristics noted above for the judgement-selecting bar-graph was observed in around one hundred classroom uses of the exercise. Regrettably, we cannot identify this phenomenon as *solely* selecting bias because we have no model for possible *estimating* bias under the (*non*-probability) judgement selecting process.

○ **Estimating bias:** (a *model* quantity that is) the difference between the mean of an estimator and the value of the corresponding population attribute (or model parameter) [as discussed above in Note 23].

– As demonstrated in equation (HL77.8) near the top of page HL77.5, under EPS when estimating $\overline{\mathbf{Y}}$, there is *zero* estimating bias and so it the does *not* contribute to sample error; this is *not* the case for estimating $\mathbf{R}$ under EPS – see the first bullet (●) of Note 23 at the top of this page HL77.12.

These matters are illustrated by the hypothetical data in Example HL77.6 on the facing page HL77.13.

(*continued*)

SAMPLING and SURVEY SAMPLING: Estimating a Population Average or Total (continued 6)

Example HL77.6: A respondent population of $\mathbf{N} = 5$ elements (or units) has the following integer values for two of their variates:

$\mathbf{Y}$: 4, 5, 6, 7, 8 and (correspondingly) $\mathbf{X}$: 2, 3, 4, 5, 6,

so that $\bar{\mathbf{Y}} = 6$, $\mathbf{S}_Y \simeq 1.5811$; $\bar{\mathbf{X}} = 4$, $\mathbf{S}_X \simeq 1.5811$; $\mathbf{R} = \bar{\mathbf{Y}}/\bar{\mathbf{X}} = 6/4 = 1\frac{1}{2}$, $\mathbf{S}_R \simeq 0.2662$;

we examine the behaviour of *sample error* under EPS as the sample size *in*creases from 1 to 2 to 3 to 4 to 5.

In the five Tables HL77.9 below, samples are given as (y; x), $r = \bar{y}/\bar{x} \equiv {}_Ty/{}_Tx$ and numbers at the bottom of a column are the *average* and the *average magnitude* of its entries; the behaviour of the sample error of $\bar{x}$ would mirror that of $\bar{y}$ and so is *not* included in the Tables. Key values from Tables HL77.9 are compared in Table HL77.10 at the right below them.

When estimating $\bar{\mathbf{Y}}$ under EPS, as in Example HL77.1 on pages HL77.3 and HL77.4, we see that:

- the random variable $\bar{Y}$ representing the sample *average* $\bar{y}$ is an *un*biased estimator of the respondent population average $\bar{\mathbf{Y}}$;
- average sample error *de*creases in magnitude as sample size *in*creases, meaning that sampling *im*precision *de*creases with increasing sample size;
- sample error is (of course) zero in a census (when $n = \mathbf{N}$).

When estimating $\mathbf{R}$ under EPS, we see that:

- the random variable R representing the sample ratio $r = \bar{y}/\bar{x}$ is a *biased* estimator of the respondent population ratio $\mathbf{R} = \bar{\mathbf{Y}}/\bar{\mathbf{X}}$ – when $n < \mathbf{N}$, the average of r is *not* (exactly) $1\frac{1}{2}$ (or the average sample error is *not* zero).
- average estimating bias *de*creases in magnitude as sample size *in*creases, meaning that sampling *in*accuracy *de*creases with increasing sample size;
  - it seems from Example HL77.6 that the inaccuracy of the two components of sampling – selecting and estimating – *each* has this relationship to sample size;
- estimating bias disappears in a census (when $n = \mathbf{N}$).

Table HL77.9a
EPS of n = 1 unit

Sample	$\bar{y}$	Error	r	Error
(4; 2)	4	−2	2	$\frac{1}{2}$
(5; 3)	5	−1	$1\frac{2}{3}$	$\frac{1}{6}$
(6; 4)	6	0	$1\frac{1}{2}$	0
(7; 5)	7	1	$1\frac{2}{5}$	$-\frac{1}{10}$
(8; 6)	8	2	$1\frac{1}{3}$	$-\frac{1}{6}$
Av.	6	0	1.58	0.08
|Av.|		$1\frac{1}{2}$		0.187

Table HL77.9b
EPS of n = 2 units

Sample	$\bar{y}$	Error	r	Error
(4,5; 2,3)	$4\frac{1}{2}$	$-1\frac{1}{2}$	$1\frac{4}{5}$	$\frac{3}{10}$
(4,6; 2,4)	5	−1	$1\frac{2}{3}$	$\frac{1}{6}$
(4,7; 2,5)	$5\frac{1}{2}$	$-\frac{1}{2}$	$1\frac{4}{7}$	$\frac{1}{14}$
(4,8; 2,6)	6	0	$1\frac{1}{2}$	0
(5,6; 3,4)	$5\frac{1}{2}$	$-\frac{1}{2}$	$1\frac{4}{7}$	$\frac{1}{14}$
(5,7; 3,5)	6	0	$1\frac{1}{2}$	0
(5,8; 3,6)	$6\frac{1}{2}$	$\frac{1}{2}$	$1\frac{4}{9}$	$-\frac{1}{18}$
(6,7; 4,5)	$6\frac{1}{2}$	$\frac{1}{2}$	$1\frac{4}{9}$	$-\frac{1}{18}$
(6,8; 4,6)	7	1	$1\frac{2}{5}$	$-\frac{1}{10}$
(7,8; 5,6)	$7\frac{1}{2}$	$1\frac{1}{2}$	$1\frac{4}{11}$	$-\frac{3}{22}$
	6	0	1.526	0.026
		$\frac{7}{10}$		0.096

Table HL77.9c
EPS of n = 3 units

Sample	$\bar{y}$	Error	r	Error
(4,5,6; 2,3,4)	5	−1	$1\frac{2}{3}$	$\frac{1}{6}$
(4,5,7; 2,3,5)	$5\frac{1}{3}$	$-\frac{2}{3}$	$1\frac{3}{5}$	$\frac{1}{10}$
(4,5,8; 2,3,6)	$5\frac{2}{3}$	$-\frac{1}{3}$	$1\frac{6}{11}$	$\frac{1}{22}$
(4,6,7; 2,4,5)	$5\frac{2}{3}$	$-\frac{1}{3}$	$1\frac{6}{11}$	$\frac{1}{22}$
(4,6,8; 2,4,6)	6	0	$1\frac{1}{2}$	0
(4,7,8; 2,5,6)	$6\frac{1}{3}$	$\frac{1}{3}$	$1\frac{6}{13}$	$-\frac{1}{26}$
(5,6,7; 3,4,5)	6	0	$1\frac{1}{2}$	0
(5,6,8; 3,4,6)	$6\frac{1}{3}$	$\frac{1}{3}$	$1\frac{6}{13}$	$-\frac{1}{26}$
(5,7,8; 3,5,6)	$6\frac{2}{3}$	$\frac{2}{3}$	$1\frac{3}{7}$	$-\frac{1}{14}$
(6,7,8; 4,5,6)	7	1	$1\frac{2}{5}$	$-\frac{1}{10}$
	6	0	1.511	0.011
		$\frac{7}{15}$		0.061

Table HL77.9d
EPS of n = 4 units

Sample	$\bar{y}$	Error	r	Error
(4,5,6,7; 2,3,4,5)	$5\frac{1}{2}$	$-\frac{1}{2}$	$1\frac{4}{7}$	$\frac{1}{14}$
(4,5,6,8; 2,3,4,6)	$5\frac{3}{4}$	$-\frac{1}{4}$	$1\frac{8}{15}$	$\frac{1}{30}$
(4,5,7,8; 2,3,5,6)	6	0	$1\frac{1}{2}$	0
(4,6,7,8; 2,4,5,6)	$6\frac{1}{4}$	$\frac{1}{4}$	$1\frac{8}{17}$	$-\frac{1}{34}$
(5,6,7,8; 3,4,5,6)	$6\frac{1}{2}$	$\frac{1}{2}$	$1\frac{4}{9}$	$-\frac{1}{18}$
	6	0	1.504	0.004
		$\frac{3}{10}$		0.038

Table HL77.9e
n = $\mathbf{N}$ = 5 units

Sample	$\bar{y}$	Error	r	Error
(4,5,6,7,8; 2,3,4,5,6)	6	0	$1\frac{1}{2}$	0

Table HL77.10	Estimating $\bar{\mathbf{Y}} = 6$			Estimating $\mathbf{R} = 1\frac{1}{2}$		
Sample size	Av. $\bar{y}$	Average error	Average |error|	Av. r	Average error	Average |error|
n = 1	6	0	$1\frac{1}{2}$	1.58	0.08	0.187
n = 2	6	0	$\frac{7}{10}$	1.526	0.026	0.096
n = 3	6	0	$\frac{7}{15}$	1.511	0.011	0.061
n = 4	6	0	$\frac{3}{10}$	1.504	0.004	0.038
n = $\mathbf{N}$	6	0	0	1.5	0	0

15. Appendix 8: A Survey Sampling Cautionary Tale

Although of general interest, the focus of the article reprinted below and overleaf on page HL77.14 is estimating a population *frequency*, the topic of Section 4 starting on page 2.86 in Figure 2.10.

- In a few places, wording in the article has been changed to align with terminology in these Course Materials.
- Equation (HL77.40) at the lower right is equation (2.10.9) from the bottom of page 2.81 in Figure 2.10 of the STAT 322 Course Materials.

BY WILLIAM GLENN, Ph.D.
Senior Statistician, Research Triangle Institute, Research Triangle Park, North Carolina.

My concern is with the quality of the kinds of estimates of the extent of drug abuse that you can get with a sample survey. I want to say, first of all, *that in order to use the sample survey properly, you must have a probability sample.* This is a sample in which every member of the population has a known nonzero probability of being selected. *Quota sampling*, frequently used in statewide drug-abuse surveys, is not probability sampling and therefore cannot yield estimates which are "good" from a statistical point of view. Reference to this point has been made by Dr. Cisin, and I am sure it is recognized by many others.

The quality of a population attribute estimate (such as the percentage of heroin users in a stated population) may be measured by the coefficient of variation (*c.v.*), estimated as in equation (HL77.40) below, where p is the estimate of $\mathbf{P}$, the proportion of the population having the characteristic of interest and n is the sample size. *c.v.* is simply the stan-

$$\widehat{c.v.}(P) \simeq \sqrt{\frac{1}{n}\left(\frac{1}{p} - 1\right)} \qquad \text{-----(HL77.40)}$$

(*continued overleaf*)

dard deviation of the estimator expressed as a fraction of the population attribute. The criterion for a good estimate, according to my colleagues who do a lot of survey sample work, is equation (HL77.41). This criterion is

$$c.v.(P) \leq 0.30 \qquad \text{----(HL77.41)}$$

saying that for an estimate of reasonable statistical quality, the standard error should not exceed 30 percent of the estimate. The particular choice of 30 percent is arbitrary, of course, but it is a value commonly cited by statisticians specializing in sample surveys.

We may use the criterion cited above to calculate the sample size needed as a function of the proportion **P**. Combining (HL77.40) and (HL77.41) and solving for n, we obtain:

$$n \geq (1 - \mathbf{P})/.09\mathbf{P}. \qquad \text{----(HL77.42)}$$

Suppose, for example, that **P** is of the order of 0.01 (about 1 percent of the population has the the characteristic of interest). The relationship (HL77.42) tells us that for a "good" estimate of **P**, our sample size should be at least 1,100.

We have seen reports on statewide surveys that give the percentage of regular heroin users as something less than one tenth of 1 percent. The report on a recent survey in North Caroline cites an estimated 2,760 regular users of heroin in a population of 3,731,520. This yields p = 0.00074 (*i.e.*, we are being told that about 7 persons in 10,000 in North Carolina are regular users of heroin). This small proportion could be encouraging to those who are concerned about heroin use in North Carolina, if they choose to believe it. But should anyone believe it? How good is the estimate? It was based on a sample size of 2,007 so, according to (HL77.40), has c.v. = 82 percent, which far exceeds the maximum criterion value. It is, therefore, a very poor estimate. Add to this the fact that it was obtained with a quota sample, and you must conclude that you have nothing with which statistical credibility can be associated.

This estimate of 2,760 regular users of heroin in North Carolina proves, upon close examination, to be totally incredible from any point of view – see the ABSTRACT below.

Suppose we had wished to estimate the percentage of regular heroin users in North Carolina by means of a sample survey. Suppose further that the true percentasge of users in the population is about one tenth of 1 percent. Our criterion tells us that we would need a sample size of at least 11,000. *Such a survey based on a probability sample at today's prices could cost half a million dollars.* Any reasonable person must conclude that spending that amount of money for that purpose makes no sense. There are, of course, those who would argue that if we can't afford a sample of 11,000, let's do the best we can with 2,000. This is fallacious reasoning because the low-quality estimate (based on 2,000 observations in this case) can give totally misleading information. The analogy that "we can't afford a Cadillac so we'll settle for a Chevrolet" can be pushed too far. The estimate with c.v. = 30 percent is a Chevrolet-class estimate. If I order something at a lower price, I may receive a beaten-up Volkswagen painted to look like a Chevrolet, and not know the difference until I look under the hood.

If I am the type who just rides off down the road before looking under the hood, I will never know the difference. That is how the analogy can be pushed too far; believe me, this is not an unfair description of what is happening.

In conclusion, I would like to emphasize the following four points.

- Sample surveys aimed at estimating the level of rare characteristics can be very expensive if conducted properly. They can be so expensive that their cost/benefit ratio is too high for serious consideration.
- Sample surveys not conducted properly can be even more expensive to society through the imputed cost of wrong (and therefore misleading) information and the misuse of available resources.
- The presentation of any estimate based on a sample survey should give the reader/user a frank statement about the likely accuracy of the estimate.
- We may not know *a priori* the ballpark in which an estimate may fall. If we elect to do the best we can with the resources available, our report should include a fair statement of what the data tell us. If that turns out to be less than our client wants, we must live with the facts. To do otherwise with drug-abuse data is to do a potential disservice to society.

ABSTRACT

Regular Users of Heroin Within the General Population of North Carolina.

A report prepared for the North Carolina Drug Authority in February, 1974, cites an estimate of 2,760 regular users of heroin within the general population of North Carolina (Table 100, p. 211). The report also provides the following information on the 2,760 regular users (Tables 101-105, pp. 212-216).

All 2,760 are in the age group 18-24
1,320 are male 1,440 are female
1,320 are black 1,440 are white
All are in the "Middle" socioeconomic status
1,320 are in "Lower/Middle" social position
1,440 are in "Upper/Lower Middle" social position
1,320 are employed in the category "Skilled/Semi-Skilled
1,440 are employed in the category "White Collar/Other Clerical
1,320 use heroin "at home"
1,440 use heroin "at a social gathering'"
1,320 also use non-controlled narcotics/prescription non-narcotic analgesics
1,440 also use barbiturates, non-barbiturate sedative-hypnotics, controlled narcotics (non-heroin), LSD, and cocaine
All 2,760 also use marihuana/hashish, psychotogens other than LSD, and methedrine/-methamphetamine
1,320 are moderate drinkers of alcohol
1,440 are heavy drinkers of alcohol
All 2,760 are "moderate or light" smokers of tobacco,
All 2,760 are regular users of prescription drugs.

The report also breaks this information down by geographic regions within the state, as follows:

Region 1: Western (32 counties)
Regular users of heroin: none (Table 198, p. 353).

Region 2: North Central (16 counties, including Durham and Orange)
Regular users of heroin: 1,440 (Table 268, p. 424)
All 1,440 are white, female, and in the age group 18-24 (Table 269, p. 425)
All 1,440 use heroin "at a special gathering" (Table 270, p. 426).

Region 3: South Central (20 counties, including Wake)
Regular users of heroin: 1,320 (Table 339, p. 495)
All 1,320 are black, male, and in the age group 18-24 (Table 339, p. 496)
All 1,320 use heroin "at home" (Table 340, p. 497).

Region 4: Eastern (32 counties)
Regular users of heroin: none (Table 406, p. 564).

Thus, the 2,760 regular users of heroin within the general population of North Carolina consist of the 1,440 white females in the north central region who use heroin at social gatherings and the 1,320 black males in the south central region who use heroin at home.

In a report dated March, 1974, were present "implications and recommendations" pertaining to the findings in the February, 1974, report.

Those implications which pertain to the regular users of heroin in the general population are the following:

100.0% are in the age group 18-24 (Table 19, p. 95)
47.8% are males; 52.7% are females (Table 29, p. 105)'
100.0% of those in Region 2 are females (Table 33, p. 109)
100.0% of those in Region 3 are males (Table 35, p. 111)
47.8% are black; 52.7% are white (Table 39, p. 115)
100.0% of those in Region 2 are white (Table 43, p. 119)
100.0% of those in Region 3 are black (Table 45, p. 121)
!00.0% are employed persons (Table 49, p. 125)
47.8% are employed males; 52.7% are employed females (Table 51, p. 127)
!00.0% are non-students (Table 53, p. 129)

These implications follow, of course, from the data cited in the February, 1974, report If we are to accept them as credible, we must agree, amoung other things, that:

1. Regular users of heroin in the general population of North Carolina are found only in the two central regions of the state (in 36 of the 100 counties).
2. In one of these regions, all of the users are white and female; in the other, all of them are black and male.
3. There are no white makes or black females who are regular users of heroin in the general population of North Carolina.
4. No students in the general population of North Carolina are regular users of heroin.
5. No persons over the age of 24 in the general population of North Carolina are regular users of heroin.

(*continued at the bottom of page HL79.2 in Statistical Highlight #79*)