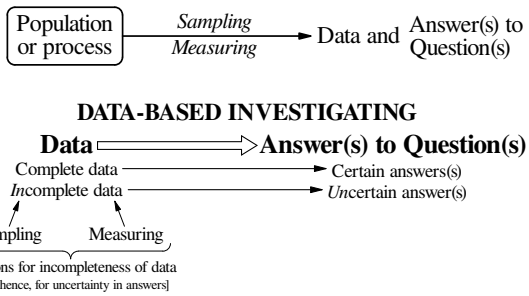
University of Waterloo W. H. Cherry

RESPONSE MODELS IN STAT 231: Modelling Sample Error and Measurement Error

As summarized in the two schemas at the right, statistics is concerned with *data-based investigating* of some population or process to *answer* one or more (statistical) *questions* of interest.



- If the investigating yields *complete* information, we can obtain a *certain* answer; that is, an answer we can *know* is correct.
- If the investigating yields *in*complete information, we can*not* know an answer is correct (an *un*certain answer) – in fact, it is *un*likely a *numerical* answer (an 'estimate') is (exactly) correct;
  - sampling and measuring yield data (and, hence, information) that are *inherently* incomplete.

In this Highlight #74, we develop a probability a model for the investigative processes of sampling (which involves (selecting and estimating) and measuring; this model allows us (*e.g.*, in Statistical Highlight #2) to quantify the *likely* size of sample error and measurement error – that is, to quantify uncertainty from these two sources – for numerical answers to some types of questions.

1. Ideas from Probability

To go from probability to statistics, ideas we need from probability modelling (*e.g.*, in STAT 230) are as follows.

- Using the normal distribution to model the shape of appropriate data distributions, as summarized by the probability statement (HL74.1) about the random variable Y at the right.

$$Y \sim N(\mu, \sigma) \quad \text{-----(HL74.1)}$$

- A random variable which is a sum or an average of (probabilistically independent) normal random variables also has a normal distribution and its mean and standard deviation can be expressed in terms of the mean(s) and standard deviation(s) of the random variables that make up the sum or average. For example, for n probabilistically independent $N(\mu, \sigma)$ random variables Y_j, $j = 1, 2, ..., n$, as given in equations (HL74.2) and (HL74.3) at the right below:
  - their sum (T) [or *total*] has a normal distribution with a *larger* mean and standard deviation than the individual random variables by respective factors of n and $\sqrt{n}$.

$$T \sim N(n\mu, \sigma\sqrt{n}) \quad \text{-----(HL74.2)}$$

  - their *average* ($\overline{Y}$) has a normal distribution with the *same* mean μ as the individual random variables but a standard deviation that is *smaller* by a factor of $1/\sqrt{n}$.

$$\overline{Y} \sim N(\mu, \sigma\sqrt{\frac{1}{n}}) \quad \text{-----(HL74.3)}$$

If the requirement for *normality* of the distribution of the individual random variables in a total or an average is relaxed, the Central Limit Theorem (CLT) approximation allows us to write equations (HL74.4) and (HL74.5); the *accuracy* of this approximate normality of a total or an average improves as:

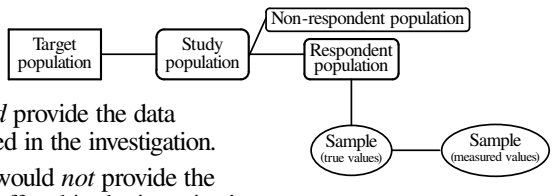
- the value of n gets *larger*;
- the distribution(s) of the Y_js get *more symmetrical*.

$$T \div N(n\mu, \sigma\sqrt{n}) \quad \text{-----(HL74.4)}$$

$$\overline{Y} \div N(\mu, \sigma\sqrt{\frac{1}{n}}) \quad \text{-----(HL74.5)}$$

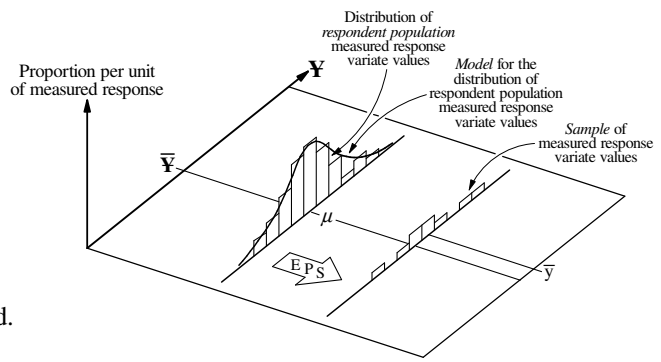
2. Ideas from Statistics

The schema at the right reminds us that data-based investigating in statistics is concerned with five groups of elements or units and their **attributes**: a quantity defined as a function of response (and, perhaps, explanatory) variates over the group (*e.g.*, an average).



- **Target population**: the group of elements to which the investigator(s) want Answer(s) to the Question(s) to apply.
- **Study population**: a group of elements *available* to an investigation.
- **Respondent population:** those elements of the study population that *would* provide the data requested under the incentives for response offered in the investigation.
- **Non-respondent population:** those elements of the study population that would *not* provide the data requested under the incentives for response offered in the investigation.
- **Sample**: the group of units selected from the respondent population *actually used* in an investigation – the sample is a *sub*set of the respondent population (as indicated by the *vertical* line in the schema above).

As indicated pictorially in the 3-dimensional diagram at the right, we think of a *respondent* population as having a *distribution* of values for some measured *response* variate ($\mathbf{Y}$) of interest; this distribution can be displayed as a *histogram*. Simple population *attributes* of interest are:



- Size: $\mathbf{N}$ [the number of *elements*];
- Location: $\overline{\mathbf{Y}}$ [the *average*];
- Variation: $\mathbf{S}$ [the (*data*) *standard deviation*];

Unless $\mathbf{N}$ is to be estimated, it is usually considered to be:

- known (or ignored), and
- large compared to the size (n) of any *sample* (to be) selected.

(*continued overleaf*)

University of Waterloo W. H. Cherry

The probability model (HL74.1) for the shape of the histogram is included in the diagram overleaf on page HL74.1 at the bottom right, which also reminds us that μ, the mean of the model, represents the respondent population average, $\overline{\mathbf{Y}}$, of the response variate.

The *sample* average is $\overline{y}$ and equation (HL74.3) is a model for the corresponding random variable $\overline{Y}$; the method of *selecting* the sample [equiprobable selecting (EPS)] is (part of) the reason we can treat the measured response variate values, y_j, j = 1, 2, ..., n, of the n units in the sample as values, y_j, j = 1, 2, ..., n, of random variables, Y_j, j = 1, 2, ..., n, for the model (HL74.3).

3. The Set of All Possible Samples from a Population

Our interpretation of model quantities involves the idea of repeating over and over processes like selecting and measuring; for example, for the random variable Y_j, representing the response of the jth unit selected equiprobably for the sample, we say:

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.

To pursue this idea of repeating over and over (or of *repetition*), we first discuss the set of all possible samples of size n that can be selected (without replacement) from a population of size $\mathbf{N}$. There are $\mathbf{N}^{(n)}$ such samples if the *order* of selection is taken into account; for example, there are $4^{(2)} = 12$ such samples of size 2 for a population of 4 elements.

Example HL74.1: A respondent population of $\mathbf{N} = 4$ elements has the following integer *true* $\mathbf{Y}$-values for its response variate: 1, 2, 4, 5 (so that: $\overline{\mathbf{Y}} = 3$, $\mathbf{S} \simeq 1.8257$).

Table HL74.1 at the right lists the 12 possible ordered samples of size n = 2 that can be selected; *equiprobable* selecting (EPS) is assumed. For discussion below, the values of the average ($\overline{y}$) and sample error for each sample are also given.

Table HL74.1
EPS of n = 2 units

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

Looking down the column of values for the *first* unit in the 12 samples, we see they are three copies of the values of the response variate of the elements of the respondent population; we see the same in the column of 12 values for the *second* unit. This result holds in general – in the set of $\mathbf{N}^{(n)}$ possible ordered samples of size n, selected from a population of size $\mathbf{N}$, the jth unit comprises $(\mathbf{N}-1)^{(n-1)}$ copies of the values of the response variate of the elements of the respondent population. Hence, *under EPS*, the values of the unit in *any* position j of the set of all possible ordered samples of size n can be modelled like the values of the response variate of the elements of the respondent population – see equation (HL74.6) at the right; in statistics, it is more convenient to write equation (HL74.6) as the *response model* (HL74.7).

$$ {}_{\mathrm{T}}Y_j \sim N(\mu, \sigma_{\mathrm{s}}), \quad j = 1, 2, ..., n, \quad \text{EPS} \qquad \text{-----(HL74.6)} $$

$$ {}_{\mathrm{T}}Y_j = \mu + {}_{\mathrm{T}}R_j, \quad j = 1, 2, ..., n, \quad {}_{\mathrm{T}}R_j \sim N(0, \sigma_{\mathrm{s}}), \quad \text{independent, EPS} \qquad \text{-----(HL74.7)} $$

NOTES: 1. The context of Example HL74.1 is *true* response variate values, indicated by the suffix T on the random variables ${}_{\mathrm{T}}Y_j$ and ${}_{\mathrm{T}}R_j$ (the **residuals**); conceptually, we are in the *first* sample ellipse in the schema at the lower right overleaf on page HL74.1 (see also Note 8 on page HL74.4).

- Another difference between equations (HL74.1) and (HL74.6) is the subscript s on σ_{s}, which indicates the variation quantified by this (probabilistic) standard deviation arises from the *selecting* (or 'sampling') process.

2. Because any sample position contains *multiple* copies (not just one copy) of the values of the respondent population responses, its (data) standard deviation is slightly *smaller* than the respondent population (data) standard deviation $\mathbf{S}$, which is represented by σ_{s} in the model. This discrepancy is of no consequence for the present discussion in the usual practical situation of $\mathbf{N}$ large and $\mathbf{N} \gg$ n (see also Note 4 on the facing page HL74.3).

3. The second and third columns of Table HL74.1 in Example HL74.1 above illustrate two other matters.

- Sample error – the difference between the value of the sample and respondent population attributes (here, averages) – has an *average* of *zero* over the set of all possible samples; stated another way, the average of the set of all possible sample averages is the respondent population average. This is the meaning of saying that, under EPS, the sample average is an *unbiased estimator* of the respondent population average. Symbolically, we write:

$$ E({}_{\mathrm{T}}\overline{Y}) = \overline{\mathbf{Y}} \text{ or } E({}_{\mathrm{T}}\overline{Y}) - \overline{\mathbf{Y}} = 0 \text{ [or, in the model: } E({}_{\mathrm{T}}\overline{Y}) = \mu \text{ or } E({}_{\mathrm{T}}\overline{Y}) - \mu = 0\text{]}. \qquad \text{-----(HL74.8)} $$

This matter is illustrated more broadly in Appendix 3, which starts on page HL74.5 of this Highlight #74.

  - EPS is needed for unbiasedness when estimating $\overline{\mathbf{Y}}$ to ensure that all samples are *equally likely*, as reflected in calculating the *average* sample error.

- For the four $\mathbf{Y}$-values of the response variate in this population, there *are* samples whose average is *equal* to the respondent population average, $\overline{\mathbf{Y}}$; however, there are many populations for which *no* sample of a given size (or, conceivably, of *any* size other than $\mathbf{N}$) has an average of $\overline{\mathbf{Y}}$; *i.e.*, *no* sample has *zero* sample error.
  - This is one reason a numerical Answer obtained by sampling is likely to be at least a little off the truth.

(*continued*)

University of Waterloo W. H. Cherry

RESPONSE MODELS IN STAT 231: Modelling Sample Error and Measurement Error (continued 1)

4. A Probability Model for Selecting Equiprobably

Example HL74.1 on the facing page HL74.2 illustrates properties of the set of all possible samples of a population under EPS; we now argue from this case to the process of selecting equiprobably (a sample) *over and over.* Provided 'over and over' means selecting a number of times that is *large* in relation to $(\mathbf{N}-1)^{(n-1)}$, because any sample is *equally* likely, the set of samples will be (roughly) many copies of the set of all possible samples. Subject to this new caveat of not having *exactly* equal proportions of each possible sample in the set, together with a slightly smaller standard deviation (see Note 2 on the facing page HL74.2), we can again use the model (HL74.7). Hence, for the behaviour of the sample *average* when selecting over and over, we use the model (HL74.3), rewritten above as equation (HL74.9).

$$ {}_{\mathrm{T}}\overline{Y} \sim N\left(\mu, \sigma_{\mathrm{s}}\sqrt{\frac{1}{\mathrm{n}}}\right), \text{ EPS} \quad \text{-----(HL74.9)} $$

* The idea of obtaining, when selecting over and over, roughly *equal* proportions of each member of the set of all possible samples is like tossing a fair coin over and over, where we expect to observe roughly *equal* proportions of heads and tails.

NOTES: 4. Note 2 and the paragraph above refer to the standard deviation of ${}_{\mathrm{T}}Y_j$ as being slightly *smaller* than σ_{s}; using a different approach in Statistical Highlight #77, the standard deviation of ${}_{\mathrm{T}}Y_j$ is actually as given in equation (HL74.10); the multiplier of $(1-1/\mathbf{N})$ under a square root is obviously close to 1 in value for most populations encountered in practice, which have *many* elements.

$$ s.d.({}_{\mathrm{T}}Y_j) = \mathbf{S}\sqrt{1-\frac{1}{\mathbf{N}}} \quad \text{-----(HL74.10)} $$

$$ s.d.({}_{\mathrm{T}}\overline{Y}) = \mathbf{S}\sqrt{\frac{1}{\mathrm{n}}-\frac{1}{\mathbf{N}}} \quad \text{-----(HL74.11)} $$

- The consequence of equation (HL74.10) is that the standard deviation of ${}_{\mathrm{T}}\overline{Y}$ is as given in equation (HL74.11) above; in the usual situation in practice, where $\mathbf{N} \gg \mathrm{n}$, this standard deviation becomes effectively that of equation (HL74.9), where σ_{s} is the model parameter representing the respondent population (data) standard deviation $\mathbf{S}$.

5. The assumption of a *normal* model for the shape of the histogram of the (true) response variate $\mathbf{Y}$-values in the respondent population can be relaxed if we can assume that the Central Limit Theorem will provide adequate *approximate* normality of the random variable ${}_{\mathrm{T}}\overline{Y}$ representing the sample average ${}_{\mathrm{T}}\overline{\mathrm{y}}$ under repetition.

5. An EPS-Based Probability Model for Measuring

The response model (HL74.7), with *normal* residuals, for equiprobable selecting over and over can be adapted to model the process of *measuring* (independently) over and over the value of a variate for a unit, for two reasons:

* We know from histograms like those for the paper thickness data in Statistical Highlight #39 (and for the coin weights in Figure 6.2 of the STAT 231 2004 Course Materials) that normal distributions are a reasonable model for values produced by (some) measuring processes.
* We can model measuring as a selecting process: possible values for the variate being measured could be written on slips of paper, the slips placed in a box and a slip selected by EPS; the number on the slip is then regarded as the 'measured' value; measuring over and over would be modelled as selecting slips equiprobably *with* replacement over and over.

We therefore write the reponse model for measuring as equation (HL74.12) at the right, where:

$$ {}_{\mathrm{M}}Y = \tau + \delta + {}_{\mathrm{M}}R, \quad {}_{\mathrm{M}}R \sim N(0, \sigma_{\mathrm{M}}), \quad \text{independent, EPS} \quad \text{-----(HL74.12)} $$

${}_{\mathrm{M}}Y$ is a random variable whose distribution represents the possible values of the measurement of the response variate of a unit, if the measuring process were to be repeated over and over on this unit.

τ is a model parameter which represents the *true value* of the response variate of the unit being measured.

δ is a model parameter (called the *bias*) which represents the ***in****accuracy* of the measuring process; the value of δ *quantifies* the inaccuracy of the measuring process – as inaccuracy *in*creases (*i.e.*, as accuracy *de*creases), δ *in*creases.

R is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the value of the measurement of the response variate of the unit being measured, if the measuring process were to be repeated over and over on this unit.

σ_{M} the *standard deviation* of the normal model for the distribution of the residual, is a model parameter (called the *variability*) which represents the ***im****precision* of the measuring process and describes measuring variation if the measuring process were to be repeated over and over on a unit; the value of σ_{M} *quantifies* the imprecision of the measuring process – as imprecision *in*creases (*i.e.*, as precision *de*creases), σ_{M} *in*creases.

For a *calibrated* measuring system, where the inaccuracy has been (essentially) removed using a defined standard, the parameter δ can be omitted and the model written as equation (HL74.13).

$$ {}_{\mathrm{M}}Y = \tau + {}_{\mathrm{M}}R, \quad {}_{\mathrm{M}}R \sim N(0, \sigma_{\mathrm{M}}), \quad \text{independent, EPS} \quad \text{-----(HL74.13)} $$

NOTES: 6. The absence of the subscript j on ${}_{\mathrm{M}}Y$ and ${}_{\mathrm{M}}R$ in the response models (HL74.12) and (HL74.13) is because we are modelling the process of measuring *once* the value of a variate for a unit.

7. The suffix M on ${}_{\mathrm{M}}Y$ (and ${}_{\mathrm{M}}R$) in the models (HL74.12) and (HL74.13) reminds us we are dealing with a *measured* value of a response variate, as distinct from ${}_{\mathrm{T}}Y_j$ in equation (HL74.7), the corresponding *true* value for the jth unit in the sample.

(*continued overleaf*)

University of Waterloo W. H. Cherry

6. A Probability Model for Selecting Equiprobably *and* Measuring

To obtain a response model for both selecting *and* measuring, two processes responsible for *in*complete data and, hence, for *un*certain Answers, we combine the response models (HL74.7) and (HL74.13); the resulting model [equation (HL74.16) below] describes the processes of EPS of n units from a respondent population (for which $\mathbf{N} \gg$ n) and measuring their response variate values once each with a *calibrated* measuring system.

Equation (HL74.7) is given again at the right, and equation (HL74.13) is rewritten as equation (HL74.14), where τ has been replaced by the *random variable* ${}_{\mathrm{T}}Y_j$ because we are measuring (once) the response of the jth unit in the sample *selected by EPS*; ${}_{\mathrm{M}}Y_j$ has a subscript j for the same reason. Also, the suffixes T and M on the residuals are to distinguish the residuals in the two models for selecting and measuring; these two sets of residuals become *one* set denoted R_j in the combined model (HL74.16) – see equation (HL74.17) below.

$${}_{\mathrm{T}}Y_j = \mu + {}_{\mathrm{T}}R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad {}_{\mathrm{T}}R_j \sim N(0, \sigma_{\mathrm{s}}), \quad \text{independent, EPS} \qquad \text{-----(HL74.7)}$$

$${}_{\mathrm{M}}Y_j = {}_{\mathrm{T}}Y_j + {}_{\mathrm{M}}R_j, \; j = 1, 2, \ldots, \mathrm{n}, \; {}_{\mathrm{M}}R_j \sim N(0, \sigma_{\mathrm{M}}), \quad \text{independent, EPS} \qquad \text{----(HL74.14)}$$

Then, writing ${}_{\mathrm{T}}Y_j$ in equation (HL74.14) as its expression in (HL74.7), we obtain equation (HL74.15), which we rewrite as equation (HL74.16), where σ quantifies the *overall* variation due to *both* selecting and measuring [equation (HL74.17)].

$${}_{\mathrm{M}}Y_j = \mu + {}_{\mathrm{T}}R_j + {}_{\mathrm{M}}R_j, \quad j = 1, 2, \ldots, \mathrm{n} \qquad \text{----(HL74.15)}$$

$${}_{\mathrm{M}}Y_j = \mu + R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad R_j \sim N(0, \sigma), \quad \text{independent, EPS} \qquad \text{----(HL74.16)}$$

$$R_j = {}_{\mathrm{T}}R_j + {}_{\mathrm{M}}R_j, \quad \sigma = \sqrt{\sigma_{\mathrm{s}}^2 + \sigma_{\mathrm{M}}^2} \qquad \text{----(HL74.17)}$$

NOTES: 8. In this Highlight #74, we have developed a model for selecting *and* measuring by combining separate models for the two processes; in doing so, we distinguish true and measured response variate values (${}_{\mathrm{T}}Y_j$ and ${}_{\mathrm{M}}Y_j$). After discussing this distinction here, the data we encounter elsewhere are essentially *always* obtained by measuring and explicit consideration of the true value ${}_{\mathrm{T}}Y_j$ of equation (HL74.7) seldom arises. For convenience, we will therefore usually write the model (HL74.16) as at the right with*out* the suffix M.

$$Y_j = \mu + R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad R_j \sim N(0, \sigma), \quad \text{independent, EPS} \qquad \text{----(HL74.16)}$$

- ${}_{\mathrm{M}}Y_j$ and Y_j in equation (HL74.16) is the quantity defined early in Section 3 on the second side (page HL74.2) of this Highlight #74 – this definition involves *both* selecting and measuring; conceptually, we have reached the *second* sample ellipse in the schema at the lower right of the first side (page HL74.1) of this Highlight #74.

9. Based on the model (HL74.16), equation (HL74.18) at the right describes the behaviour of the random variable representing the sample average, $\overline{\mathrm{y}}$, subject to variation from *both* equiprobable selecting and measuring.

$$\overline{Y} \sim N(\mu, \sigma\sqrt{\tfrac{1}{\mathrm{n}}}), \quad \text{EPS} \qquad \text{----(HL74.18)}$$

- Despite the apparent similarity of equations (HL74.3) and (HL74.18), their derivations and the interpretation of $\overline{Y}$, μ and σ differ in statistically vital ways.

7. Appendix 1: A Comparison of Approaches to Modelling the Behaviour of $\overline{Y}$

Three approaches to modelling the behaviour of the sample average, as a basis for estimating the respondent population average, are discussed in this Highlight #74; this Appendix 1 presents an overview of the strengths and weaknesses of each.

* *Probability modelling* starts with the normal model (HL74.1) and then uses probability theory for random variables to obtain the model (HL74.3); the weakness of this approach for *statistics* is that the ideas of Sections 2 to 6, and the assumptions they make *ex*plicit, are only *im*plicit. It is likely to be unclear to the beginning student why statements like equations (HL74.1) and (HL74.3) can be used to describe the real-world processes of selecting (and measuring, if it is even considered in this approach) and it is easy for statistical practice based on the probability approach to overlook essential components of the Plan.

$$Y \sim N(\mu, \sigma) \qquad \text{-----(HL74.1)}$$

$$\overline{Y} \sim N(\mu, \sigma\sqrt{\tfrac{1}{\mathrm{n}}}) \qquad \text{-----(HL74.3)}$$

* The *response model* (HL74.16), as developed on the basis of the *same* probability models (HL74.1) and (HL74.3), recognizes explicitly Question formulation (via consideration of the target, study and respondent populations) and properties of the selecting and measuring processes needed ed as a basis for the relevant probability models, written as the response models (HL74.7) and (HL74.14). This approach therefore overcomes the weaknesses of the probability modelling approach, albeit at the cost of more detail in presentation; the final probability expression for the behaviour of $\overline{Y}$, involving variation from *both* selecting and measuring, is equation (HL74.18). A weakness of this approach is that it does not recognize explicitly the finite size ($\mathbf{N}$ elements) of the respondent population.

$${}_{\mathrm{T}}Y_j = \mu + {}_{\mathrm{T}}R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad {}_{\mathrm{T}}R_j \sim N(0, \sigma_{\mathrm{s}}), \text{ indep., EPS} \qquad \text{-----(HL74.7)}$$

$${}_{\mathrm{M}}Y_j = {}_{\mathrm{T}}Y_j + {}_{\mathrm{M}}R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad {}_{\mathrm{M}}R_j \sim N(0, \sigma_{\mathrm{M}}), \text{ indep., EPS} \qquad \text{----(HL74.14)}$$

$$Y_j = \mu + R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad R_j \sim N(0, \sigma), \text{ indep., EPS} \qquad \text{----(HL74.16)}$$

$$\overline{Y} \sim N(\mu, \sigma\sqrt{\tfrac{1}{\mathrm{n}}}), \quad \text{EPS} \qquad \text{----(HL74.18)}$$

* The so-called *design-based* approach developed in Statistical Highlight #77, based on the equiprobable selecting of the sampling protocol, incorporates the finite size ($\mathbf{N}$ elements) of the respondent population, but it cannot easily take direct account of measurement error. Its expression for the behaviour of ${}_{\mathrm{T}}\overline{Y}$ is equation (HL74.19); the *approximate* normality comes from the Central Limit Theorem approximation.

$${}_{\mathrm{T}}\overline{Y} \div N(\overline{\mathbf{Y}}, \mathbf{S}\sqrt{\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}}), \quad \text{EPS} \qquad \text{----(HL74.19)}$$

Estimating $\overline{\mathbf{Y}}$ using the models (HL74.3), (HL74.18) and (HL74.19) requires an *estimate* of σ or $\mathbf{S}$, customarily taken as the *sample*

(*continued*)

University of Waterloo W. H. Cherry

RESPONSE MODELS IN STAT 231: Modelling Sample Error and Measurement Error (continued 2)

standard deviation, s. Because s is calculated from *measured* values of the response variate, its value *includes* variation due to measuring *as well as* selecting; this inclusion is modelled in (HL74.18) but is a *fortuitous* benefit for (HL74.3) and (HL74.19).

Thus, the model (HL74.18), derived from the response model (HL74.16), provides a basis for statistical practice that is missing from (HL74.3); the additional insight from (HL74.19) is that the multiplier of $\mathbf{S}$ in the standard deviation of ${}_{\mathrm{T}}\bar{Y}$ has weak dependence on $\mathbf{N}$; ignoring $\mathbf{N}$ in (HL74.18) is justified provided the sample size is a *small proportion* of the respondent population size ($n \ll \mathbf{N}$).

8. Appendix 2: Using the Response Model for Measuring

When investigating the inaccuracy and imprecision of a measuring system [often as part of broader data-based investigating to answer Question(s) of interest], a common approach is to make m independent measurements of the same value. The appropriate response models are shown at the right as equations (HL74.20) and (HL74.21); the former is applicable when measuring a *known* value to investigate both inaccuracy *and* imprecision, whereas the latter is used when investigating only imprecision; the two forms of (HL74.20) differ in whether the measured value or its difference from the true value is taken as the response variate – one or the other may be more convenient, depending on context.

$$ {}_{\mathrm{M}}Y_j = \tau + \delta + {}_{\mathrm{M}}R_j,\ \ j = 1, 2, \ldots, \mathrm{m},\ \ {}_{\mathrm{M}}R_j \sim N(0, \sigma_{\mathrm{M}}),\ \ \text{independent,}\ \ \text{EPS} \quad \text{----(HL74.20)} $$

$$ {}_{\mathrm{M}}Y_j - \tau = \delta + {}_{\mathrm{M}}R_j,\ \ j = 1, 2, \ldots, \mathrm{m},\ \ {}_{\mathrm{M}}R_j \sim N(0, \sigma_{\mathrm{M}}),\ \ \text{independent,}\ \ \text{EPS} \quad \text{----(HL74.20)} $$

$$ {}_{\mathrm{M}}Y_j = \tau + {}_{\mathrm{M}}R_j,\ \ j = 1, 2, \ldots, \mathrm{m},\ \ {}_{\mathrm{M}}R_j \sim N(0, \sigma_{\mathrm{M}}),\ \ \text{independent,}\ \ \text{EPS} \quad \text{----(HL74.21)} $$

Example HL74.2: A medical laboratory purchases a standard, certified to contain 200 mg/dL of cholesterol, to calibrate its process for measuring serum cholesterol levels; data from 15 measurements taken over an 8-hour day were:

Table HL74.2:

																Av.	**S.d.**
Measured value (${}_{\mathrm{M}}y_j$)	194	196	202	198	195	188	203	200	196	195	198	202	203	201	199	198	4.1231
Measured – true value (${}_{\mathrm{M}}y_j - \tau$)	−6	−4	2	−2	−5	−12	3	0	−4	−5	−2	2	3	1	−1	−2	4.1231

These data show that: the estimate of measuring *inaccuracy* (represented in the model by δ) is −2 mg/dL;
the estimate of measuring *imprecision* (represented in the model by σ_{M}) is 4.1 mg/dL.

With inaccuracy estimated to be about half the magnitude of imprecision, the former may have little practical importance here, but more detailed interpretation of these data requires extra-statistical knowledge.

NOTE: 10. Under the idealizations of the models (HL74.20) and (HL74.21), it can be shown that, for measuring processes:

- inaccuracy is *un*affected by averaging – looking ahead, one average is the estimate of the *intercept* of the centred form of the straight-line model in simple linear regression (see page HL73.2 in Statistical Highlight #73).
- inaccuracy is *removed* by differencing – differences occur when comparing and when calculating (data) standard deviations and the *slope* of the straight-line regression model (see page HL73.2 in Statistical Highlight #73).

9. Appendix 3: The Set of All Possible Samples from a Population

Section 3 introduced the idea of all possible samples that can be selected from a population. The sets of ten and nine histograms on pages HL74.6 and HL74.7 illustrate this idea for a population of $\mathbf{N} = 10$ elements with response variate values $\mathbf{Y} = 1, 2, \ldots, 10$. These histograms, for sample averages and sample (data) standard deviations, show important properties of the selecting process.

- The sample average takes a value which is determined by the particular sample selected; these values, over the set of all possible samples of a given size, form a *distribution*; under EPS, *this is the distribution of the random variable* $\bar{Y}$, a result that could be called **The Fundamental Theorem of Statistics** by analogy with The Fundamental Theorem of Calculus.
  - The mean of this distribution is the *population* average – this is the *unbiasedness* under EPS of the sample average as an estimator of the population average (see the second column of Table HL74.3 in Appendix 4 on page HL74.8).
  - The standard deviation of this distribution is given by equation (HL74.11) except that, because *s.d.*$(\bar{Y})$ is a *probabilistic* standard deviation and $s_{\bar{y}}$ (given in the third column of Table HL74.3 in Appendix 4) is a *data* standard deviation, the former is $\sqrt{k/(k-1)}$ times the latter, where k is $\binom{\mathbf{N}}{n}$, the number of possible (unordered) samples of size n.
- The distribution of sample averages reminds us that for the *particular* sample selected when executing the Plan for an investigation, its sample error is *un*known – its attribute may be close to the population attribute or (sometimes) it may not be; that is, an Answer has sampling *uncertainty*, meaning we cannot *know* how close an Answer obtained by sampling is to the truth.
- The *de*creasing *variation* of the values of sample averages with *in*creasing sample size, visible as the decreasing width down the page of the histograms of sample averages, illustrates sampling *imprecision* decreasing with increasing *sample size*; *i.e.*, under EPS, increasing sample size decreases the average magnitude of sample error and so decreases sampling uncertainty.
- The change in distribution *shape*, most noticable as the sample size increases from 1 to 2, illustrates the *idea* behind the Central Limit Tehoerem – a *non*-normal (uniform) distribution starting to show the central 'heaping' of the normal distribution.
  - The heaping, which first appears when $n = 2$, persists up to $n = 8$ and then disappears; this is likely the effect of *dependence* among samples, which consist of most of the elements in the population, becoming dominant over the heaping process reminiscent of the Central Limit Theorem.

(*continued overleaf*)

University of Waterloo W. H. Cherry

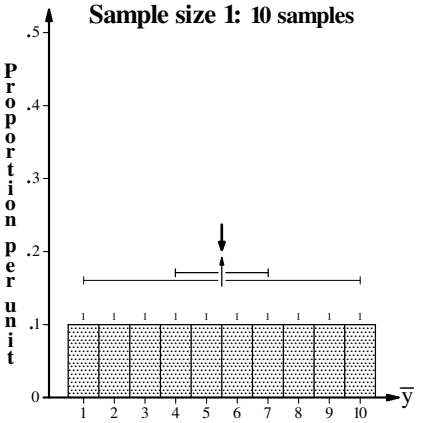
9. Appendix 3: The Set of All Possible Samples from a Population (continued)

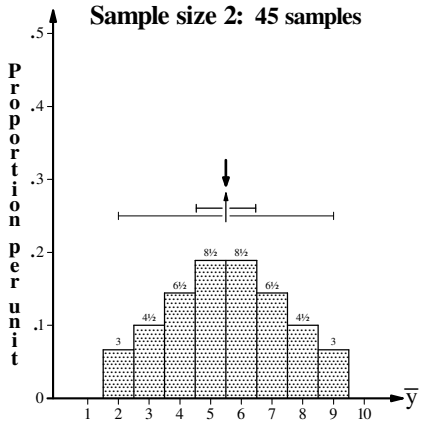
These ten histograms show the distribution of the *averages* of all possible samples of a given size that can be selected without replacement from a population of $N = 10$ elements, whose response variate values are $Y = 1, 2, ..., 10$. [A histogram of these ten population *responses* would be like that given at the right for the averages of the ten possible samples of size 1.]

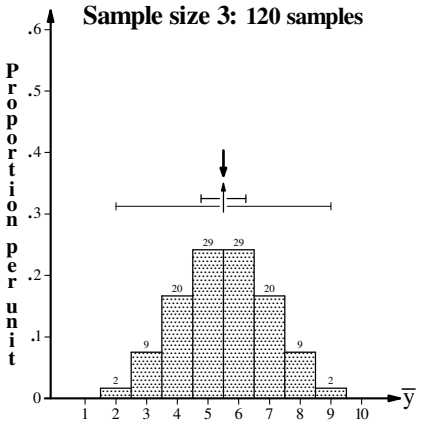
The decreasing vertical axis scale unit down the page reduces the visual impact of increasing histogram height; bar *frequencies* are given by the numbers at the tops of the bars.

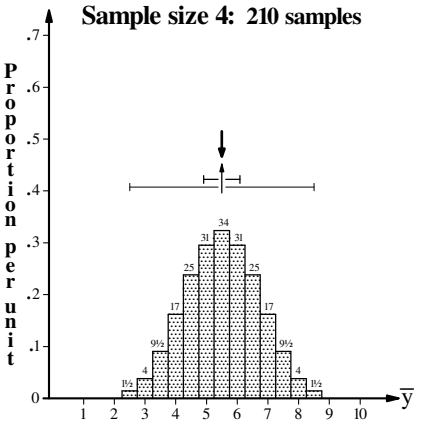
Above each histogram, except the one at the bottom right of the page:

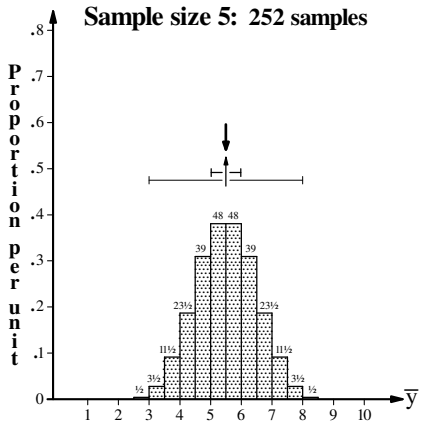
- the upper bold arrow (↓) indicates the value of the *population* average $\bar{Y} = 5.5$;
- the lower arrow (↑) indicates the *average* of the set of *sample* averages ($\bar{\bar{y}} = 5.5$);
- the upper bar (⟷) crossing the lower arrow has a length equal to the value of the (data) *standard deviation* of the set of sample averages;
- the lower such bar shows the *range* of the set of sample averages.

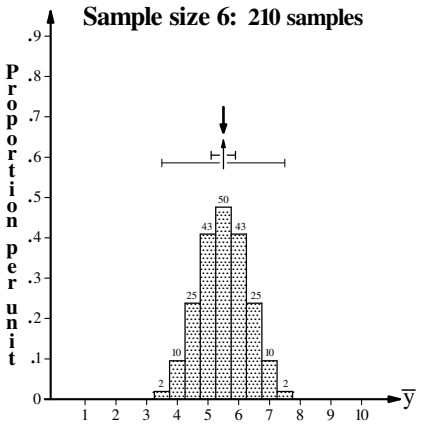


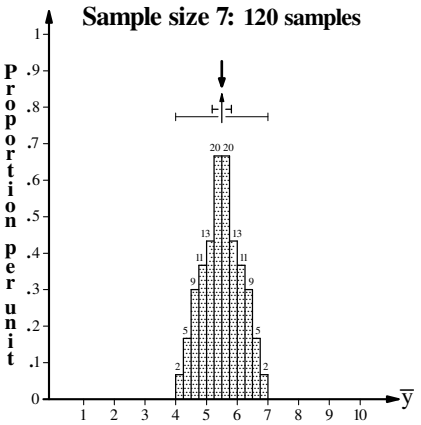


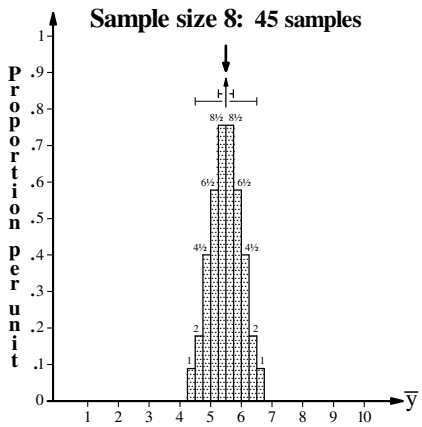


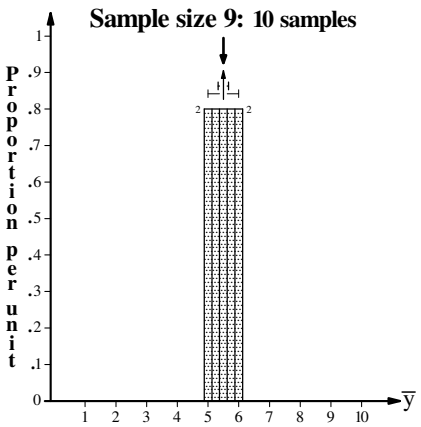


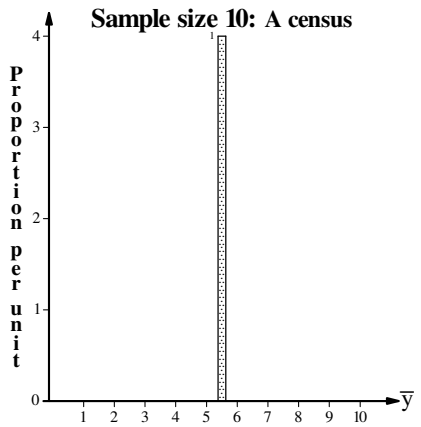












(*continued*)

University of Waterloo W. H. Cherry

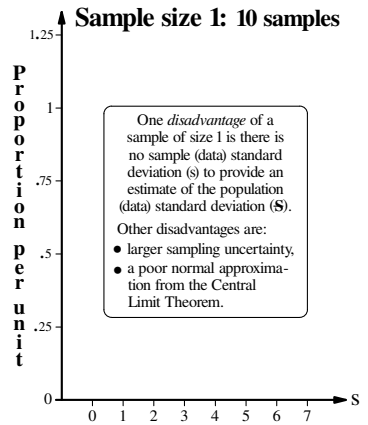
RESPONSE MODELS IN STAT 231: Modelling Sample Error and Measurement Error (continued 3)

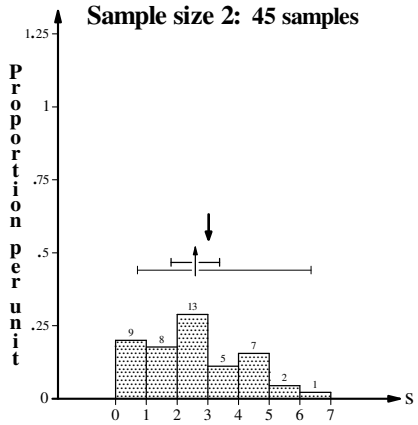
These nine histograms show the distribution of the (data) *standard deviations* of all possible samples of a given size that can be selected without replacement from a population of $N = 10$ elements, whose response variate values are $Y = 1, 2, ..., 10$.
[A histogram of these ten population *responses* would be like that given at the top right of the facing page HL74.6 for the averages of the ten possible samples of size 1.]

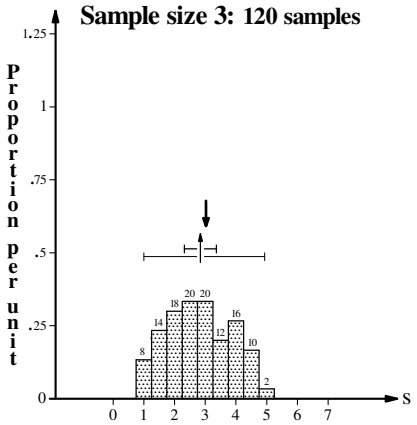
The decreasing vertical axis scale unit down the page reduces the visual impact of increasing histogram height; bar *frequencies* are given by the numbers at the tops of the bars.

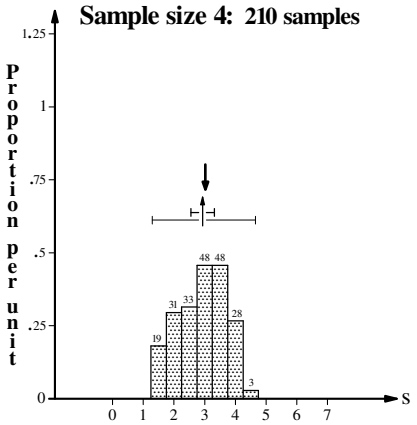
Above each histogram, except the one at the bottom right of the page:

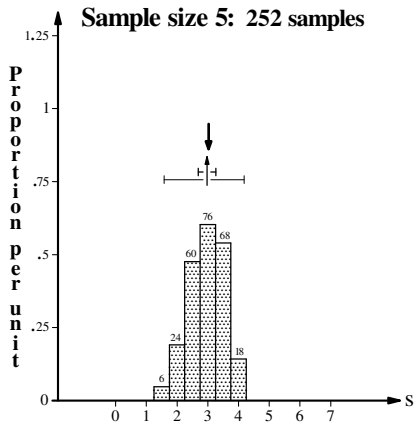
- the upper bold arrow (↓) indicates the value of the *population* standard deviation $S = 3.027\,650$;
- the lower arrow (↑) indicates the *average* ($\bar{s}$) of the set of *sample* (data) standard deviations;
- the upper bar (⊢——) crossing the lower arrow has a length equal to the value of the (data) *standard deviation* of the set of sample (data) standard deviations;
- the lower such bar shows the *range* of the set of sample (data) standard deviations.

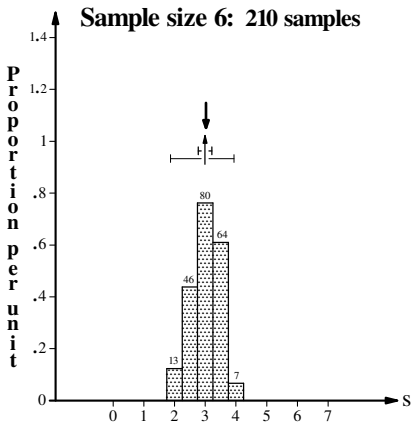


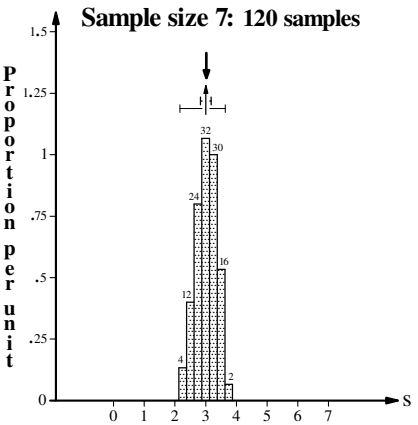


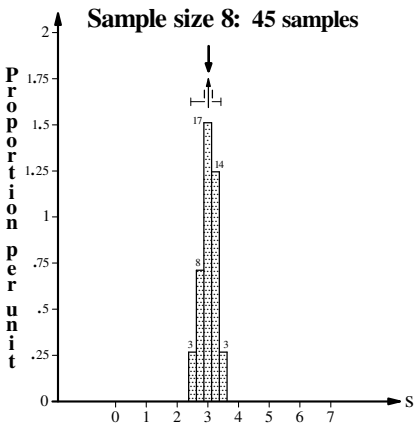


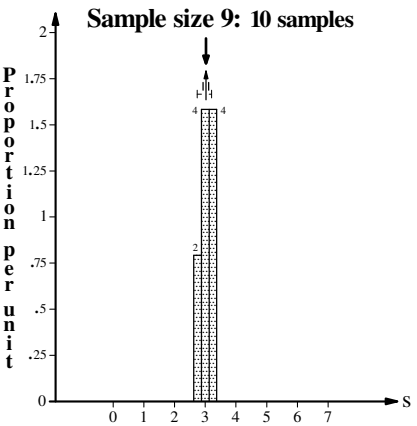


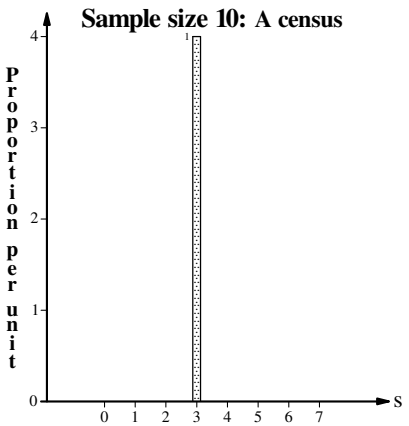












(*continued overleaf*)

University of Waterloo W. H. Cherry

9. Appendix 3: The Set of All Possible Samples from a Population (continued)

* The histograms of sample (data) standard deviations illustrate two differences from sample averages:
  - their distributions are narrower (even with a *smaller* scale unit on the horizontal axis) and are not symmetrical;
  - the sample (data) standard deviation is *not* an unbiased estimator of the population (data) standard deviation, but the estimating bias decreases as sample size increases [see Appendix 3 on pages HL77.9 and HL77.10 in Statistical Highlight #77.]
    - Estimating bias is *unlike* sampling and measuring inaccuracy which do *not* decrease with increasing sample size.

10. Appendix 4: Data Sets for Appendix 3

The histograms in Appendix 3 were constructed from the data tabulated below and on the facing page HL74.9. The average ($\overline{\overline{y}}$) and (data) standard deviation ($s_{\overline{y}}$) of the sample averages for each sample size (n) are given in the second and third columns of the table below, as are the average ($\overline{s}$) and (data) standard deviation (s_s) of the sample (data) standard deviations on the facing page HL74.9.

Table HL74.3

n	$\overline{\overline{y}}$	$s_{\overline{y}}$	Ordered Sample Averages ($\overline{y}$)
1	11/2	3.0277	1 2 3 4 5 6 7 8 9 10
2	11/2	1.9365	3/2 2 5/2 5/2 3 3 7/2 7/2 7/2 4 4 4 9/2 9/2 9/2 9/2 5 5 5 5 11/2 11/2 11/2 11/2 11/2 6 6 6 6 13/2 13/2 13/2 13/2 7 7 7 15/2 15/2 15/2 8 8 17/2 17/2 9 19/2
3	11/2	1.4686	2 7/3 8/3 8/3 3 3 3 10/3 10/3 10/3 10/3 11/3 11/3 11/3 11/3 11/3 4 4 4 4 4 4 4 13/3 13/3 13/3 13/3 13/3 13/3 13/3 13/3 14/3 14/3 14/3 14/3 14/3 14/3 14/3 14/3 14/3 5 5 5 5 5 5 5 5 5 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 6 6 6 6 6 6 6 6 6 6 19/3 19/3 19/3 19/3 19/3 19/3 19/3 19/3 19/3 20/3 20/3 20/3 20/3 20/3 20/3 20/3 20/3 7 7 7 7 7 7 7 22/3 22/3 22/3 22/3 22/3 23/3 23/3 23/3 23/3 8 8 8 25/3 25/3 26/3 9
4	11/2	1.1754	5/2 11/4 3 3 13/4 13/4 13/4 7/2 7/2 7/2 7/2 7/2 15/4 15/4 15/4 15/4 15/4 15/4 4 4 4 4 4 4 4 4 4 17/4 17/4 17/4 17/4 17/4 17/4 17/4 17/4 17/4 17/4 17/4 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 19/4 19/4 19/4 19/4 19/4 19/4 19/4 19/4 19/4 19/4 19/4 19/4 19/4 19/4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 21/4 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 23/4 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 25/4 25/4 25/4 25/4 25/4 25/4 25/4 25/4 25/4 25/4 25/4 25/4 25/4 25/4 13/2 13/2 13/2 13/2 13/2 13/2 13/2 13/2 13/2 13/2 13/2 13/2 13/2 27/4 27/4 27/4 27/4 27/4 27/4 27/4 27/4 27/4 27/4 7 7 7 7 7 7 7 7 29/4 29/4 29/4 29/4 29/4 29/4 15/2 15/2 15/2 15/2 15/2 31/4 31/4 31/4 8 8 33/4 17/2
5	11/2	0.9593	3 16/5 17/5 17/5 18/5 18/5 18/5 19/5 19/5 19/5 19/5 19/5 4 4 4 4 4 4 4 21/5 21/5 21/5 21/5 21/5 21/5 21/5 21/5 22/5 22/5 22/5 22/5 22/5 22/5 22/5 22/5 22/5 22/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 23/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 24/5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 26/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 27/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 28/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 29/5 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 31/5 32/5 32/5 32/5 32/5 32/5 32/5 32/5 32/5 32/5 32/5 32/5 32/5 32/5 32/5 33/5 33/5 33/5 33/5 33/5 33/5 33/5 33/5 33/5 33/5 34/5 34/5 34/5 34/5 34/5 34/5 34/5 34/5 34/5 7 7 7 7 7 7 7 36/5 36/5 36/5 36/5 36/5 37/5 37/5 37/5 38/5 38/5 39/5 8
6	11/2	0.7836	7/2 11/3 23/6 23/6 4 4 4 25/6 25/6 25/6 25/6 25/6 13/3 13/3 13/3 13/3 13/3 13/3 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 9/2 14/3 14/3 14/3 14/3 14/3 14/3 14/3 14/3 14/3 14/3 14/3 29/6 29/6 29/6 29/6 29/6 29/6 29/6 29/6 29/6 29/6 29/6 29/6 5 5 5 5 5 5 5 5 5 5 5 5 5 5 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 31/6 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 16/3 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 11/2 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 17/3 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 35/6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 37/6 37/6 37/6 37/6 37/6 37/6 37/6 37/6 37/6 37/6 37/6 37/6 37/6 19/3 19/3 19/3 19/3 19/3 19/3 19/3 19/3 19/3 19/3 13/2 13/2 13/2 13/2 13/2 13/2 13/2 13/2 13/2 20/3 20/3 20/3 20/3 20/3 20/3 20/3 41/6 41/6 41/6 41/6 7 7 7 43/6 43/6 22/3 15/2
7	11/2	0.6294	4 29/7 30/7 30/7 31/7 31/7 31/7 32/7 32/7 32/7 32/7 33/7 33/7 33/7 33/7 33/7 34/7 34/7 34/7 34/7 34/7 34/7 34/7 5 5 5 5 5 5 5 5 36/7 36/7 36/7 36/7 36/7 36/7 36/7 36/7 36/7 36/7 37/7 37/7 37/7 37/7 37/7 37/7 37/7 37/7 37/7 37/7 37/7 38/7 38/7 38/7 38/7 38/7 38/7 38/7 38/7 38/7 38/7 39/7 39/7 39/7 39/7 39/7 39/7 39/7 39/7 39/7 39/7 40/7 40/7 40/7 40/7 40/7 40/7 40/7 40/7 40/7 40/7 41/7 41/7 41/7 41/7 41/7 41/7 41/7 41/7 41/7 6 6 6 6 6 6 6 6 43/7 43/7 43/7 43/7 43/7 43/7 43/7 44/7 44/7 44/7 44/7 44/7 45/7 45/7 45/7 45/7 46/7 46/7 46/7 47/7 47/7 48/7 7
8	11/2	0.4841	9/2 37/8 19/4 19/4 39/8 39/8 5 5 5 41/8 41/8 41/8 21/4 21/4 21/4 21/4 43/8 43/8 43/8 43/8 11/2 11/2 11/2 11/2 11/2 45/8 45/8 45/8 45/8 23/4 23/4 23/4 23/4 47/8 47/8 47/8 6 6 6 49/8 49/8 25/4 25/4 51/8 13/2
9	11/2	0.3364	5 46/9 47/9 16/3 49/9 50/9 17/3 52/9 53/9 6

ACKNOWLEDGEMENTS: The numbers in the tables in Appendix 4 were kindly calculated by Professor R.W. Oldford using Quail. The development of the model for sample and measurement error in Section 6 on page HL74.4 is based on ideas in Chapter 5 of the 2004 Course Notes for Statistics 231, by Professors R.W. Oldford and R.J. MacKay.

(*continued*)

RESPONSE MODELS IN STAT 231: Modelling Sample Error and Measurement Error (continued 4)

Table HL74.4

n	$\bar{s}$	s_s	Ordered Sample (Data) Standard Deviations (s)																							
2	2.5927	1.5811	0.7071	0.7071	0.7071	0.7071	0.7071	0.7071	0.7071	0.7071	0.7071	1.4142	1.4142	1.4142	1.4142	1.4142	1.4142	1.4142	1.4142	2.1213	2.1213	2.1213	2.1213	2.1213	2.1213	2.8284
			2.8284	2.8284	2.8284	2.8284	2.8284	3.5355	3.5355	3.5355	3.5355	3.5355	4.2426	4.2426	4.2426	4.2426	4.9497	4.9497	4.9497	5.6568	5.6568	6.3639					
3	2.8443	1.0418	1	1	1	1	1	1	1	1	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	1.5275	2	2	2
			2	2	2	2.0817	2.0817	2.0817	2.0817	2.0817	2.0817	2.0817	2.0817	2.0817	2.0817	2.0817	2.0817	2.5166	2.5166	2.5166	2.5166	2.5166	2.5166	2.5166	2.5166	2.5166	2.5166	2.5166
			2.6458	2.6458	2.6458	2.6458	2.6458	2.6458	2.6458	2.6458	2.6458	2.6458	3	3	3	3	3.0551	3.0551	3.0551	3.0551	3.0551	3.0551	3.0551	3.0551	3.0551	3.2146	3.2146	3.2146
			3.2146	3.2146	3.2146	3.2146	3.2146	3.5119	3.5119	3.5119	3.5119	3.5119	3.5119	3.6056	3.6056	3.6056	3.6056	3.6056	3.6056	3.7859	3.7859	3.7859	3.7859	3.7859	3.7859	3.7859	4	4
			4.0415	4.0415	4.0415	4.0415	4.1633	4.1633	4.1633	4.1633	4.3589	4.3589	4.3589	4.3589	4.5092	4.5092	4.5826	4.5826	4.7258	4.7258	4.9329	4.9329					
4	2.9317	0.7580	1.2910	1.2910	1.2910	1.2910	1.2910	1.2910	1.2910	1.7078	1.7078	1.7078	1.7078	1.7078	1.7078	1.7078	1.7078	1.7078	1.7078	1.7078	1.7078	1.8257	1.8257	1.8257	1.8257	1.8257	1.8257
			2.0817	2.0817	2.0817	2.0817	2.0817	2.1602	2.1602	2.1602	2.1602	2.1602	2.1602	2.1602	2.1602	2.1602	2.1602	2.2174	2.2174	2.2174	2.2174	2.2174	2.2174	2.2174	2.2174	2.2174	2.2174
			2.3805	2.3805	2.3805	2.3805	2.3805	5/2	5/2	5/2	5/2	5/2	5/2	5/2	5/2	2.5820	2.5820	2.5820	2.5820	2.6300	2.6300	2.6300	2.6300	2.6300	2.6300	2.6300	2.6300	2.6300
			2.6458	2.6458	2.6458	2.6458	2.6458	2.6458	2.6458	2.6458	2.7538	2.7538	2.7538	2.7538	2.7538	2.7538	2.7538	2.7538	2.8868	2.8868	2.8868	2.9439	2.9439	2.9439	2.9439	2.9439	2.9439
			2.9439	2.9439	2.9439	2.9439	2.9861	2.9861	2.9861	2.9861	2.9861	2.9861	3.0957	3.0957	3.0957	3.0957	3.0957	3.0957	3.1091	3.1091	3.1091	3.1091	3.1091	3.1091	3.1091	3.1091	3.1091
			3.1623	3.1623	3.1623	3.1623	3.1623	3.1623	3.3040	3.3040	3.3040	3.3040	3.3040	3.3040	3.3040	3.3040	3.3040	3.3040	3.3665	3.3665	3.4034	3.4034	3.4034	3.4034	3.4157	3.4157	3.4157
			3.4157	7/2	7/2	7/2	7/2	3.5119	3.5119	3.5119	3.5590	3.5590	3.5590	3.5590	3.5940	3.5940	3.5940	3.5940	3.5940	3.5940	3.5940	3.5940	3.6515	3.6515	3.6968	3.6968	3.6968
			3.6968	3.6968	3.7417	3.7417	3.7749	3.7749	3.8622	3.8622	3.8622	3.8622	3.8622	3.8622	3.8730	3.8730	3.8730	3.9158	3.9158	4.0311	4.0311	4.0311	4.0311	4.0415	4.0415	4.0825	4.0825
			4.0825	4.0825	4.1130	4.1130	4.2032	4.2426	4.2426	4.4253	4.4253	4.6547															
5	2.9719	0.5793	1.5811	1.5811	1.5811	1.5811	1.5811	1.5811	1.9235	1.9235	1.9235	1.9235	1.9235	1.9235	1.9235	1.9235	1.9235	1.9235	2.0736	2.0736	2.0736	2.0736	2.0736	2.0736	2.0736	2.0736	2.0736
			2.0736	2.2361	2.2361	2.2361	2.2361	2.3022	2.3022	2.3022	2.3022	2.3022	2.3022	2.3022	2.3022	2.3875	2.3875	2.3875	2.3875	2.3875	2.3875	2.3875	2.3875	2.4083	2.4083	2.4083	2.4083
			2.4083	2.4083	2.4083	2.4083	2.5495	2.5495	2.5495	2.5495	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.5884	2.7019	2.7019	2.7019
			2.7019	2.7019	2.7019	2.7019	2.7019	2.7019	2.7019	2.7019	2.7019	2.7386	2.7386	2.7386	2.7386	2.7386	2.7386	2.7749	2.7749	2.7749	2.7749	2.7749	2.7749	2.8636	2.8636	2.8636	2.8636
			2.8636	2.8636	2.8810	2.8810	2.8810	2.8810	2.8810	2.8810	2.9155	2.9155	2.9155	2.9155	2.9155	2.9155	2.9155	2.9155	2.9665	2.9665	2.9665	2.9665	3.0332	3.0332	3.0332	3.0332	3.0496
			3.0496	3.0496	3.0496	3.0496	3.0496	3.0496	3.0496	3.0496	3.0496	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1145	3.1623	3.1623
			3.1623	3.1623	3.1623	3.1623	3.1937	3.1937	3.1937	3.1937	3.2094	3.2094	3.2094	3.2094	3.2094	3.2094	3.2094	3.2094	3.2711	3.2711	3.2711	3.2711	3.2711	3.2711	3.3466	3.3466	3.3466
			3.3466	3.3615	3.3615	3.3615	3.3615	3.3615	3.3615	3.3615	3.3615	3.3912	3.3912	3.3912	3.3912	3.3912	3.3912	3.4205	3.4205	3.4351	3.4351	3.4351	3.4351	3.4928	3.4928	3.5071	3.5071
			3.5071	3.5071	3.5355	3.5355	3.5355	3.5355	3.5355	3.5355	3.5637	3.5637	3.5637	3.5637	3.5637	3.5637	3.5637	3.5637	3.5777	3.5777	3.6469	3.6469	3.6469	3.6469	3.6469	3.6469	3.6469
			3.6469	3.6742	3.6742	3.7014	3.7014	3.7014	3.7014	3.7014	3.7014	3.7815	3.7815	3.8341	3.8341	3.8471	3.8471	3.8730	3.8730	3.9115	3.9115	3.9623	3.9623	4.0373	4.0373	4.0866	4.0866
			4.1833	4.1833																							
6	2.9939	0.4516	1.8708	1.8708	1.8708	1.8708	1.8708	2.1602	2.1602	2.1602	2.1602	2.1602	2.1602	2.1602	2.1602	2.3166	2.3166	2.3166	2.3166	2.3166	2.3166	2.3166	2.3166	2.3664	2.3664	2.3664	2.3664
			2.4290	2.4290	2.4290	2.4833	2.4833	2.4833	2.4833	2.4833	2.4833	2.5820	2.5820	2.5820	2.5820	2.5820	2.5820	2.6077	2.6077	2.6077	2.6077	2.6077	2.6077	2.6394	2.6394	2.6394	2.6394
			2.6394	2.6394	2.7325	2.7325	2.7325	2.7325	2.7386	2.7386	2.7386	2.7869	2.7869	2.7869	2.7869	2.7869	2.7869	2.8048	2.8048	2.8048	2.8048	2.8048	2.8048	2.8284	2.8284	2.8284	2.8284
			2.8577	2.8577	2.8577	2.8577	2.8810	2.8810	2.8810	2.8810	2.8810	2.8810	2.8810	2.8983	2.8983	2.9269	2.9269	2.9269	2.92689	2.9439	2.9439	2.9439	2.9439	3.0111	3.0111	3.0111	3.0111
			3.0166	3.0332	3.0332	3.0332	3.0332	3.0605	3.0605	3.0605	3.0605	3.0605	3.0605	3.0768	3.0768	3.0768	3.0768	3.0822	3.0822	3.0822	3.0822	3.1411	3.1411	3.1411	3.1411	3.1411	3.1411
			3.1623	3.1623	3.1885	3.1885	3.1885	3.1885	3.1885	3.1885	3.1885	3.1885	3.2042	3.2042	3.2249	3.2249	3.2660	3.2660	3.2660	3.2660	3.2660	3.2660	3.2711	3.2711	3.2711	3.2711	3.2711
			3.2711	3.2711	3.3116	3.3116	3.3116	3.3116	3.3116	3.3116	3.3116	3.3116	3.3116	3.3116	3.3267	3.3267	3.3466	3.3466	3.3862	3.3862	3.3912	3.3912	3.3912	3.4059	3.4059	3.4303	3.4303
			3.4448	3.4448	3.4641	3.4641	3.4881	3.4881	3.5024	3.5024	3.5071	3.5071	3.5449	3.5449	3.5590	3.5590	3.5590	3.5590	3.5777	3.5777	3.6194	3.6560	3.6560	3.6560	3.6560	3.6697	3.6697
			3.7283	3.7417	3.7417	3.7639	3.7639	3.7639	3.7639	3.8297	3.8297	3.9370															
7	3.0075	0.3505	2.1602	2.1602	2.1602	2.1602	2.4103	2.4103	2.4103	2.4103	2.4103	2.4103	2.5635	2.5635	2.5635	2.5635	2.5635	2.5635	2.6367	2.6367	2.6367	2.6367	2.6367	2.6367	2.6458	2.6458	2.6904
			2.6904	2.6904	2.6904	2.7946	2.7946	2.7946	2.7946	2.8200	2.8200	2.8200	2.8200	2.8702	2.8702	2.8702	2.8702	2.8785	2.8785	2.8785	2.8785	2.9114	2.9114	2.9439	2.9439	2.9921	2.9921
			2.9921	2.9921	2.9921	2.9921	3.0237	3.0237	3.0237	3.0237	3.0394	3.0394	3.0394	3.0394	3.0394	3.0394	3.0551	3.0551	3.1015	3.1015	3.1015	3.1015	3.1091	3.1091	3.1320	3.1320	3.1320
			3.1320	3.1320	3.1320	3.1472	3.1472	3.1848	3.1848	3.2071	3.2071	3.2514	3.2514	3.2514	3.2514	3.2587	3.2587	3.2660	3.2660	3.2878	3.2878	3.3381	3.3381	3.3523	3.3523	3.3523	3.3523
			3.3665	3.3665	3.4087	3.4087	3.4365	3.4365	3.4365	3.4365	3.4641	3.4641	3.4983	3.4983	3.5051	3.5051	3.5322	3.5322	3.5989	3.5989	3.6384	3.6384					
8	3.0165	0.2627	2.4495	2.4495	2.4495	2.6693	2.6693	2.6693	2.6693	2.8158	2.8158	2.8158	2.8158	2.8785	2.9001	2.9001	2.9001	2.9001	2.9155	2.9155	2.9277	2.9277	3.0208	3.0208	3.0443	3.0443	3.1053
			3.1053	3.1168	3.1168	3.1368	3.1368	3.1623	3.2071	3.2071	3.2486	3.2486	3.2705	3.2705	3.2842	3.2842	3.3381	3.3700	3.3700	3.3780	3.3780	3.4226					
9	3.0229	0.1789	2.7386	2.7386	2.9345	2.9345	3.0732	3.0732	3.1623	3.1623	3.2059	3.2059															
10	3.0277	---	3.0277																								

(*continued overleaf*)

University of Waterloo W. H. Cherry

11. Appendix 5: Sampling Protocols Beyond EPS [The title matter of Statistical Highlight #85.]

Equiprobable (or simple random) selecting of units consisting of *individual elements* from an *un*stratified (respondent) population is useful for modelling the selecting process but, in practice, more complex sampling protocols are used. Two such protocols are:

- **cluster selecting:** selecting equiprobably units from the (respondent) population that are *groups* of elements – clusters may be of *equal* size (*e.g.*, cardboard boxes of 24 cans of soup) or *un*equal size (*e.g.*, households);
- **stratified selecting:** subdividing the (respondent) population into groups (called **strata**) so that elements with*in* a stratum have *similar* response variate values ('homogenity of strata') and elements in different strata *differ* as much as practicable from each other; the sample is obtained by equiprobable selecting of units consisting of individual elements from *each* stratum.

[The element-unit distinction is discussed in Appendix 1 on pages HL77.8 and HL77.9 in Statistical Highlight #77.]

Example HL74.3 below illustrates the effects of *clustering* and *stratifying* on *sampling imprecision*, also bearing in mind that:

- *clustering* is commonly used because of the availability of a clustered *frame*, thereby avoiding the cost of generating a frame as part of the investigating – a **frame** can be thought of as a *list* of the (respondent) population sampling units (here, clusters);
- *stratifying* is commonly used because it provides (the often useful additional) *subdivision* of Answers by stratum.

Example HL74.3: A respondent population of $\mathbf{N} = 4$ elements (or units) has the following integer $\mathbf{Y}$-values for its response variate:

1, 2, 4, 5 [so that the population average and (data) standard deviation are: $\overline{\mathbf{Y}} = 3$, $\mathbf{S} \simeq 1.8257$];

we examine the *sampling imprecision*, under equiprobable selecting (EPS) with a sample size of $n = 2$, of the random variable $\overline{Y}$, whose values are the sample average $\overline{y}$, as an estimator of $\overline{\mathbf{Y}}$, using three sampling protocols:

- EPS of two units, each consisting of one element, from the *un*stratified population;
- EPS of one *cluster*, of size $L = 2$ elements, from the *un*stratified population;
- EPS of one unit, consisting of one element, from each of two *strata* of size $\mathbf{N}_1 = \mathbf{N}_2 = 2$.

Note that *each* estimator is *un*biased, because $E(\overline{Y}) = \overline{\mathbf{Y}}$ or $E(\overline{Y}) - \overline{\mathbf{Y}} = 0$ [the *average* sample error is *zero*].

Table HL74.5
Unstratified population
EPS of two *elements*

Sample	$\overline{y}$	Error	
(1, 2)	1½	−1½	large
(1, 4)	2½	−½	medium
(1, 5)	3	0	small
(2, 4)	3	0	small
(2, 5)	3½	½	medium
(4, 5)	4½	1½	large

Designation of sample error as *large*, *medium* or *small* is *only* for convenience in the context of Example HL74.3.

Table HL74.6a
Unstratified population
Clusters: [1, 2], [4, 5]
EPS of one *cluster* ($L = 2$)

Sample	$\overline{y}$	Error	
(1, 2)	1½	−1½	large
(4, 5)	4½	1½	large

Table HL74.6b
Unstratified population
Clusters: [1, 4], [2, 5]
EPS of one *cluster* ($L = 2$)

Sample	$\overline{y}$	Error	
(1, 4)	2½	–½	medium
(2, 5)	3½	½	medium

Table HL74.6c
Unstratified population
Clusters: [1, 5], [2, 4]
EPS of one *cluster* ($L = 2$)

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

Table HL74.7a
Stratified population
Strata: [1, 2], [4, 5]
EPS of one *element* per stratum

Sample	$\overline{y}$	Error	
(1, 4)	2½	–½	medium
(1, 5)	3	0	small
(2, 4)	3	0	small
(2, 5)	3½	½	medium

Table HL74.7b
Stratified population
Strata: [1, 4], [2, 5]
EPS of one *element* per stratum

Sample	$\overline{y}$	Error	
(1, 2)	1½	−1½	large
(1, 5)	3	0	small
(2, 4)	3	0	small
(4, 5)	4½	1½	large

Table HL74.7c
Stratified population
Strata: [1, 5], [2, 4]
EPS of one *element* per stratum

Sample	$\overline{y}$	Error	
(1, 2)	1½	−1½	large
(1, 4)	2½	–½	medium
(2, 5)	3½	½	medium
(4, 5)	4½	1½	large

Example HL74.3 illustrates that:

- The effect on (sampling) imprecision of clustering and of stratifying depends on how each is *implemented* in the Plan – that is, it depends on this component of the *sampling protocol*.
- Clustering and stratifying affect imprecision by determining which of the *possible* samples of size n have non-zero selecting probabilities.
- Decreased imprecision is favoured by *heterogeneity of clusters* but by *homogeneity of strata* with respect to the response(s) of interest.
  - In the middle and right-hand columns of Example HL74.3, heterogeneity increases *down* the three clustered sampling protocols, homogeneity increases *up* the three stratified protocols.

[**As an exercise**, quantify the sample error *variation* by calculating the relevant (data) *standard deviation* for each of the seven sampling protocols; comment on what is illustrated by the values obtained.]

- There is a sense in which clustering is *passively* accepted in the interests of reducing investigation cost, whereas stratifying may be *actively* imposed by the investigator(s) on [or may be a natural feature of] the study (or respondent) population.
- While EPS from an *un*stratified population implies *equal* inclusion probabilities for all population *elements*, the converse does *not* hold – in the three clustered and three stratified sampling protocols, all *elements* have equal inclusion probabilities but all six *samples* of size 2 are *not* equally probable [four samples and two samples (respectively) have *zero* selecting probability].