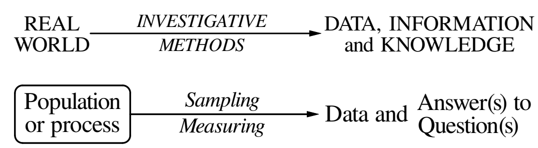
University of Waterloo W. H. Cherry

ERROR: Sample Error – Why Does Introductory Statistics Teaching Emphasize 'Random' Selecting?

1. Background I – What is Statistics About? [optional reading]

As summarized in the two schemas at the right, statistics is concerned with *data-based investigating* (or empirical problem solving) of the real world, which means investigating some population or process on the basis of *data* to *answer* one or more *questions*. For this introductory discussion, only the investigative methods of *sampling* and *measuring* are shown in the lower schema.



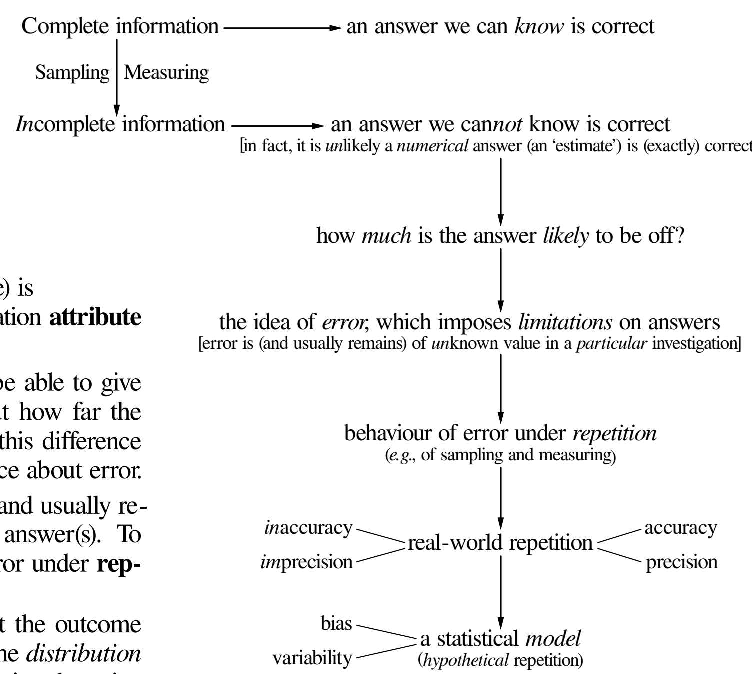
This introduction presupposes initially that data-based investigating is concerned with answer(s) to question(s) about a collection of **elements** which comprise a **population**. For example:

- ○ a computer manufacturer may want to assess warranty claims for one of its products – an *element* would be one item of the product and the *population* is all such items sold over some specified period;
- ○ a government department may want to know the proportion of Canadian adults who have concerns about the level of funding of the health care system – an *element* would be a Canadian adult and the *population* is all such Canadians;
- ○ a financial institution may wish to find people whose profile makes them more likely to accept an unsolicited offer of a credit card – an *element* would again be a person but the *population* is harder to specify; it could be people whose profile puts them at or above an acceptance level deemed likely to make the credit card offering profitable for the financial institution.

The question(s) to be answered may *sometimes* be concerned with an *individual* to be selected in future from the population.

2. Background II – Key Ideas: Uncertainty, Error and Repetition [optional reading]

A central issue in data-based investigating is that external constraints and the type of information we require *impose* sampling and measuring processes on most investigating; we then need to assess the likely 'correctness' of the answer from the investigating in light of the *uncertainty* introduced by these two processes. In essence, this is the problem of **induction**. An overview of this matter is given in the schema at the right below; the main ideas are summarized at the left.



- ∗ If we have *complete* information, we can obtain a *certain* answer; that is, an answer we can *know* is correct.
- ∗ If we have *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.
- ∗ An answer which is a *number* (like a sample average) is called a **point estimate** of the corresponding population **attribute** (the population average).
- ∗ To make such an answer more useful, we want to be able to give an **interval estimate**, a quantitative statement about how far the point estimate is *likely* to be from the true value – this difference is the **error** of the estimate. **Uncertainty** is ignorance about error.
- ∗ In a *particular* investigation, the size of the error is (and usually remains) *un*known; this is reflected in **limitations** on answer(s). To quantify uncertainty, we turn to the behaviour of error under **repetition** of the sampling and measuring processes;
  - – an analogy is tossing a coin – we cannot predict the outcome of a *particular* toss but *probability* can describe the *distribution* of outcomes under *repetition* of the process of tossing the coin.
- ∗ In the classroom, we can do *actual* repetition (repeating over and over) to demonstrate, for example, the statistical behaviour of the values of sample attributes (*e.g.*, averages) and of the values which arise from measuring processes;
  - – the two characteristics of *sign* and *magnitude* of (numerical) error lead, under repetition, to what we call **inaccuracy** and **imprecision**; the *inverses* of these two ideas – **accuracy** and **precision** – provide more familiar terminology, but we must realize that statistical methods (try to) manage the (undesirable) *former* to achieve needed levels of their (desirable) inverses.
- ∗ *Out*side the classroom, we recognize that *actual* repetition is usually not a viable option – we use instead *hypothetical* repetition based on an appropriate statistical **model** (for the selecting and measuring processes, for example);
  - – *we* help maintain the distinction between the real world and the model by using **bias** and **variability** as the *model* quantities which represent inaccuracy and imprecision in the real world.
- ∗ A statistical model, including its *assumptions*, allows us to achieve our aim of quantifying (some of) the uncertainty in our

 (*continued overleaf*)

* estimate(s); part of identifying the limitations on an Answer is to assess model error arising from the difference between ('idealized') modelling assumptions and the real-world situation (*e.g.*, see page 18.6 in Statistical Highlight #18).

There is further illustration in Statistical Highlight #87 of the ideas presented overleaf on page HL21.1.

3. Why Does Introductory Statistics Teaching Emphasize 'Random' Selecting? [The title matter of this Highlight #21.]

Sampling (comprising the processes of selecting and estimating) is involved in most data-based investigations for two reasons:

* primarily because resources are seldom available to gather data from *all* the elements of the population of interest:
  - error associated with sample attributes as estimates of population attributes is usually an acceptable trade-off for the reduced cost of using a sample;
  - greater timeliness of data and Answer(s) from a (smaller) sample than from the (larger) population may also be advantageous;
* secondarily, sampling is the only option in the *un*common situation where the measuring process *destroys* the element being measured (*e.g.*, firing shot gun cartridges to assess the quality of their manufacturing process).

Three important matters about the sample used in any data-based investigation are:

○ how the sample was selected; ○ the sample size; ○ the non-response rate.

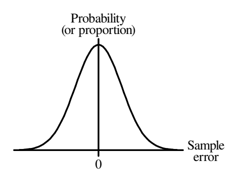
Our primary concern here is with the *first* of these matters; the usual terminology for the statistically desirable process is 'random selecting' but *we* use the more evocative **equiprobable selecting** (abbreviated '**EPS**'); we also discuss the *second* matter here.

Statistics emphasizes EPS (or its equivalent), particulary in introductory courses, because it is the basis of statistical theory which provides (under repetition):

- an (inverse) relationship between sampling imprecision and *sample size* (or degree of *replicating*);
- *unbiased* estimating (*i.e.*, *zero* sampling inaccuracy) of a population *average* (an attribute commonly of interest).
- an expression for a *confidence interval* (CI) for a population average – such an interval, under suitable modelling assumptions, *quantifies* sampling and measuring imprecision – see Statistical Highlight #2.

[Of course, an Answer obtained from a *particular* sample remains *uncertain*, as reflected by its limitations.]

- Also, a result from probability theory (the Central Limit Theorem) makes approximately *normal* (as illustrated at the right) the distribution of the averages of the set of all possible samples of a given size from the respondent population; as a consequence, *under EPS* there is a *higher* probability of selecting a sample with sample error of *smaller* magnitude, a *lower* probability of selecting one with *larger* sample error – see Statistical Highlight #74.
  - The *centre* (or 'average') of the (symmetrical) normal distribution in the diagram being at *zero* sample error is what is meant above by *unbiased* estimating – this matter is illustrated in Appendix 1 starting at the bottom of this page HL21.2, and also in Statistical Highlight #74.



EPS does *not*, of itself, *reduce* sample error or sampling imprecision, as implied in (wrong) statements such as:

○ EPS generates a *representative* sample – see Section 9 on page 77.9 in Statistical Highlight #77;

○ EPS generates a sample which provides a proper basis for *generalization*; **!**

as well as misrepresenting the statistical benefits from using EPS, such statements confuse repetition (the *process* of EPS) with a particular investigation (*a* sample). [Statements like these may arise from mistakenly interpreting language from statistical theory as referring to the sample obtained in an *individual* investigation when it actually refers to behaviour of the selecting *process* under *repetition*.] A *correct* statement is:

EPS provides for quantifying sampling imprecision and so, *in conjunction with adequate replicating* (or an *adequate sample size*), allows an Answer to be obtained with acceptable limitation imposed by sample error in the context of a particular investigation.

- What constitutes *acceptable* limitation imposed by sample error depends on the investigation requirements for Answer(s); such requirements are often quantified in terms of sampling imprecision.
  - An example is a proportion – like the percentage of working Canadians who do not contribute to their RRSP – to be estimated to within 2 percentage points with 95% probability or at a 95% level of confidence.
    + In this example, an Answer is to be obtained that is 'correct' (under repetition) about 95% of the time (*i.e.*, the CI *does* contain the population proportion) and 'wrong' about 5% of the time (the CI does *not* contain this proportion); such uncertainty, quantified (under repetition) in terms of probability or level of confidence, is *un*avoidable for an Answer from incomplete data (*i.e.*, an Answer obtained by *inductive* reasoning).

4. Appendix 1: Sample Size and Sample Error under EPS

This Appendix 1 illustrates (statistically useful) properties of EPS which are stated above in first two bullets (●) in the middle of this page HL21.2; their theoretical basis is given in Statistical Highlight #77 on pages HL77.4 and HL77.5.

Example HL21.1: 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 behaviour of *sample error* under EPS as the sample size *in*creases from 1 to 2 to 3 to 4.

University of Waterloo W. H. Cherry

ERROR: Sample Error – Why Does Introductory ... Emphasize ‘Random’ Selecting? (continued 1)

The number at the bottom of the four ‘error’ columns of Tables HL21.1 at the right is the *average magnitude* of the sample error for that sample size.

Tables HL21.1 illustrate results of theory based on EPS.

Table HL21.1a
EPS of n = 1 unit

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

Table HL21.1b
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 HL21.1c
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 HL21.1d
EPS of n = 4 units

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

- 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. [See Appendix 6 on the lower half of page HL21.8 for relevant definitions.]
- Taking the *sign* of sample error into account, the average error is *zero* in each case; *i.e.*, the random variable $\bar{Y}$, representing the sample average $\bar{y}$, is an *unbiased* estimator of the respondent population average $\mathbf{\bar{Y}}$ under EPS;
  - 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 the random variable $\bar{Y}$, is that for the population attribute which is the *ratio* of the average of two response variates ($\mathbf{R} = \mathbf{\bar{Y}}/\mathbf{\bar{X}}$), the random variable R, representing 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);
- There is no *sample* error when a **census** is taken – when *all* units/elements of the respondent population are selected.

REFERENCE: Cochran, W. G.: *Sampling Techniques.* Third edition, John Wiley & Sons, New York, 1977, ISBN 0-471-16240-X.

NOTE: 1. Although EPS provides unbiased estimating of an *average*, the same is *not* true for a standard deviation – for the random variable S representing the sample standard deviation s, under EPS: $E(S) - \mathbf{S} < 0$ -----(HL21.1)
– see Appendix 3 at the bottom of page HL77.9 in Statistical Highlight #77.

- However, it *is* true under EPS that: $E(S^2) = \mathbf{S}^2$ OR: $E(S^2) - \mathbf{S}^2 = 0,$ -----(HL21.2)
  but this is beside the point statistically because it is S (and s) [*not* S^2 (and s^2)] that occur in (useful) confidence interval expressions – see the bottom of page HL77.5 and the top of page HL77.6.

5. Appendix 2: Broader Perspectives on the Protocol for Selecting Units

The **protocol for selecting units**, sometimes called the **sampling protocol**, is (a description of) the process (to be) used to select, from the respondent population, the units that comprise the sample.

There are many processes used in practice to select samples; three of interest in these Statistical Highlights are:

○ **equiprobable** selecting, ○ **systematic** selecting, ○ **judgement** selecting.

∗ **Equiprobable (simple random) selecting [EPS (SRS)]:** all samples of size n units from a (respondent) population of size $\mathbf{N}$ units have probability $1/\binom{\mathbf{N}}{n}$ of being selected.

- What we call *equiprobable* selecting is likely to be called *simple random* (or *random*) selecting (or sampling) elsewhere.
- *Equiprobable* refers to a *process*; it is a contradiction in terms to refer to an equiprobable (or random) *sample.*
- The *definition* of EPS is in terms of *sample* selecting probabilities, not *unit* inclusion probabilities (a distinction pursued in Appendix 5 starting on page HL21.6 in this Highlight #21); consequences of this distinction for a sample of size n are:
  - + under EPS, the inclusion probability is $n/\mathbf{N}$ for each (element or) unit in the respondent population; BUT:
  - + even if the inclusion probability is $n/\mathbf{N}$ for each (element or) unit in the respondent population, the selecting process is *not necessarily* EPS – see Table HL21.6 on page HL21.6 of this Highlight #21;
  - + the sample selecting process is *not* EPS if, for each respondent population unit (or element), the inclusion probability is:
    ⊙ not equal to $n/\mathbf{N}$ OR: ⊙ not equal to that of all other unit(s) [or element(s)].

∗ **Systematic selecting:** one unit is selected by EPS from the first k units of the respondent (or study) population ($k \ll \mathbf{N}$) and then every k*th* unit is selected. [There is further discussion of systematic selecting in Note 8 on page HL21.8 in this Highlight #21.]

- Referring to the *first* k units of the respondent (or study) population implies an ordered (*e.g.*, alphabetic or numeric) list of these units; such a list (called a **frame**) may be real or conceptual (*e.g.*, a rule that would, if implemented, generate the list).
- For convenience, it is usually assumed that $\mathbf{N} = nk$ so *all* 1-in-k samples selected systematically are of the *same* size n.

∗ **Judgement selecting:** human judgement is used to select n units from the $\mathbf{N}$ units of the respondent population.

Experience shows that EPS (which includes systematic selecting from an equiprobably ordered frame) is *the* process for selecting the sample to answer a Question with a *descriptive* aspect, and that sample error under *judgement* selecting usually imposes an *un*acceptable limitation on an Answer, to the degree that such investigating is seldom a justifiable use of resources.

- Judgement selecting is included in this Highlight #21 and, for example, in Table HL6.1 at the bottom of page HL6.4 in Statistical Highlight #6 because, despite its lack of theoretical foundation, it is commonly used in investigations to answer a Question with a *causative* aspect, where EPS is often infeasible – see the discussion of experimental Plans in Section 4 on the lower half of page HL63.3 in Statistical Highlight #63 and see also Statistical Highlight #83.

Other named methods of selecting units for the sample, which are largely omitted from this discussion, include:

- **accessibility selecting:** selecting units (easily) *accessible* to the investigator(s) – for instance, the *top* layer in a basket of fruit or a truckload of potatoes or the *front* pallets or cartons in a large stack in a warehouse;
- **convenience selecting:** selecting units that are *conveniently* available to the investigator(s) – for instance, people with a medical condition of interest who are at a hospital or clinic nearby to the investigator(s);
- **haphazard selecting:** selecting units with*out* (conscious) preference by the investigator(s) – shoppers who pass the location of an interviewer in a mall or rats in a cage which are more easily caught for a laboratory test;
- **quota selecting:** selecting units according to values of specified explanatory variates (like sex, age, income for human units) so the sample distribution of each variate will (approximately) match that of the study population;
- **volunteer selecting:** asking for (human) volunteers, usually after a brief explanation of what the investigating will entail for units of the sample.

These names do not necessarily specify a *unique* selecting method – the first two methods overlap and all five involve some degree of 'accessibility' and/or 'convenience'.

Haphazard selecting is sometimes ***wrongly*** equated with 'random' selecting; *i.e.*, with our *equiprobable* selecting.

Quota selecting is a similar idea to **covering**: to try to manage sample error, the values of explanatory variates of the units of the sample are selected to cover the range of values of relevant response and/or explanatory variates that occur among (most of) the (elements or) units of the respondent (or study) population.

Covering is a guiding principle for *judgement* selecting; its chance of (partial) success is *in*creased by:

– greater replicating (*i.e.*, a larger sample size), AND:

– greater knowledge about the values of explanatory variates among the elements of the respondent (or study) population.

Volunteer selecting is *not* to be confused with **volunteer** (or **voluntary**) **response**, a phrase sometimes used to indicate that *human* units can (usually) *choose* whether to respond, *i.e.*, whether to provide the requested data; a separate (measuring) issue is whether these responses are correct or truthful (see also Note 4 on page HL38.7 in Statistical Highlight #38).

6. Appendix 3: Error Categories, Samples and Populations [optional reading]

Throughout these Statistical Highlights, we have recognized (as on page HL21.1, for instance) that statistics is concerned with data-based investigating of *populations* (or *processes*), but that resource constraints usually impose *sampling* with its components of *selecting* and *estimating*; thus, sampling is to be set against investigating *all* the (respondent) population elements (a census). At the right, two lists of investigative processes remind us that selecting and estimating are what distinguish investigating based on a sample and on a whole population – the processes of specifying the study population, obtaining responses, measuring variate values and comparing (for a Question with a *causative* aspect) are *common* to both types of Plan.

A Plan involving a .....

sample:	**population:**
Specifying	Specifying
Selecting	Responding
Responding	Measuring
Measuring	Comparing
Estimating	
Comparing	

Table HL21.2: ERROR CATEGORIES

Error category	**Arises with a .....** **Sample**	**Population**
Sample	Yes	No
Study	Yes	Yes
Non-response	Yes	Yes
Attribute measurement	Yes	Yes
Model	Yes	(Yes)
Comparison	Yes	Yes

We can classify *error categories* on the basis of whether they arise in the context of a Plan involving a sample and/or one involving a whole population as shown at the right in Table HL21.2; we see that:

- sample error arises *only* when the Plan involves a sample;
- study error, non-response error and attribute measurement error arise *regardless* of whether some of or all the (respondent) population elements (or units) are being investigated;
- because a (response) model is usually constructed to describe a *sampling* process, model error commonly arises in the context of a Plan involving a sample but models for *populations* are constructed in some investigating.
- comparison error arises only when the investigating involves a Question with a *causative* aspect and, like model error, is usually encountered in the context of a *sample* of units because most comparative Plans involve sampling, but comparison error in phenomena like Simpson's Paradox (Statistical Highlight #51) can arise when the Plan involves *either* a population *or* a sample.

NOTES: 2. *Specifying* the elements which comprise the study (or respondent) population is usually thought of in terms of a *frame*, defined overleaf near the bottom of page HL21.3.

3. Investigators may have the opportunity to trade study error and sample error; an illustration is an investigation with a target population of all Canadian adults and resources to select equiprobably 1,000 people for the sample.
   - A study population of all Canadians residing in Canada would have *smaller* study error but (likely) *larger* sample error because of greater variation among the elements of the study population; COMPERED WITH:

ERROR: Sample Error – Why Does Introductory ... Emphasize 'Random' Selecting? (continued 2)

NOTES: 3. ● a study population of all Canadian university students would have *larger* study error but (likely) *smaller* sample error because of smaller variation among the elements of the study population.
(cont.)

A Plan involving *smaller* study error and (likely) *larger* sample error is usually preferred because:

– study error requires *extra*-statistical knowledge to assess it and its behaviour can seldom be quantified; BUT:

+ under EPS, sampling theory describes the behaviour of sample (and, perhaps, measurement) error under repetition of the selecting and estimating processes, as previously discussed in this Highlight #21 and Statistical Highlight #74.

7. Appendix 4: Selecting 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 HL21.2 below illustrates the effects of *clustering* and *stratifying* on *sampling imprecision*, also bearing in mind that:

– *clustering* is commonly used because a clustered *frame* is already available, avoiding the cost of generating such a list;
– *stratifying* is commonly used because it provides (the often useful additional) *subdivision* of Answers by stratum.

Example HL21.2: A respondent population of $\mathbf{N} = 4$ elements 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 HL21.3
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 HL21.2.

Table HL21.4a
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 HL21.4b
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 HL21.4c
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 HL21.5a
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 HL21.5b
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 HL21.5c
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 HL21.2 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 HL21.2, heterogeneity increases *down* the three clustered sampling protocols, homogeneity increases *up* the three stratified protocols.

(*continued overleaf*)

[**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].

8. Appendix 5: Sample Selecting and Unit Inclusion Probabilities

To illustrate further the distinction between sample selecting and unit inclusion probabilities (from Section 5 on page 21.3) and ideas like clustering and stratifying, we consider a (respondent) population of $\mathbf{N} = 10{,}000$ elements and a sample of $n = 100$ elements, obtained using six protocols for selecting units (listed roughly in order of increasing complexity):

- ○ equiprobable selecting of 100 units (elements) from the *un*stratified population;
- ○ systematic selecting: selecting equiprobably 1 unit from the *first* 100 population units (elements) and then every 100*th* unit;
- ○ equiprobable selecting of 10 *clusters* of 10 elements from the population of 1,000 such clusters;
- ○ equiprobable selecting of 10 units (elements) from each of the 10 population *strata* each of $\mathbf{N}_h = 1{,}000$ elements $(h = 1, 2, \ldots, 10)$;
- ○ two-stage selecting: selecting 100 clusters equiprobably and then selecting 1 unit (element) equiprobably from each cluster;
- ○ two-stage selecting: selecting 2 strata equiprobably and then selecting 50 units (elements) equiprobably from each stratum.

Table HL21.6 below uses the symbol $\binom{\mathbf{N}}{n}$, the number of ways n items can be selected from $\mathbf{N}$ items if order of selecting is *un*important. [This symbol and its use are discussed in Figure 7.5 of the STAT 220 Course Materials].

Relevant calculations for this illustration are summarized in Table HL21.6 below, where the six protocols are now listed in order of *de*creasing number of possible samples. The *short* names for the protocols in the second column of Table HL21.6 should generally be avoided because their brevity can (temporarily) obscure the nature of, and differences among, the protocols.

Table HL21.6

Protocol for selecting units	**Short name**	**Number of samples**	**Ratio to EPS**	**. . . . . .Selecting or inclusion probability. . . . . Sample**	**. . . . . Unit . . . .**
EPS from an unstratified population	EPS	$\binom{10{,}000}{100} \simeq 6.5\times10^{241}$	1	1.5×10^{-242}	$\binom{1}{1}\binom{9{,}999}{99}/\binom{10{,}000}{100} = \frac{1}{100}$
2-stage EPS from a population in equal-sized clusters	2-stage cluster selecting	$\binom{1{,}000}{100}\binom{10}{1}^{100} \simeq 6.4\times10^{239}$	$\sim10^{-2}$	1.6×10^{-240}	$\binom{1}{1}\binom{999}{99}/\binom{1{,}000}{100}\bullet\binom{1}{1}\binom{9}{0}/\binom{10}{1} = \frac{1}{10}\bullet\frac{1}{10} = \frac{1}{100}$
EPS from a stratified population	Stratified selecting	$\binom{1{,}000}{10}^{10} \simeq 1.6\times10^{234}$	$\sim10^{-7}$	6.2×10^{-235}	$\binom{1}{1}\binom{999}{9}/\binom{1{,}000}{10} = \frac{1}{100}$
2-stage EPS from a stratified population	2-stage stratified slecting	$\binom{10}{2}\binom{1{,}000}{50}^{2} \simeq 4.0\times10^{171}$	$\sim10^{-70}$	2.5×10^{-172}	$\binom{1}{1}\binom{9}{1}/\binom{10}{2}\bullet\binom{1}{1}\binom{999}{49}/\binom{1{,}000}{50} = \frac{1}{5}\bullet\frac{1}{20} = \frac{1}{100}$
1-stage EPS from a population in equal-sized clusters	Cluster selecting	$\binom{1{,}000}{10} \simeq 2.6\times10^{23}$	$\sim10^{-218}$	3.8×10^{-24}	$\binom{1}{1}\binom{999}{9}/\binom{1{,}000}{10} = \frac{1}{100}$
1-in-100 systematic selecting from an unstratified population	Systematic selecting	$100 = 10^2$	$\sim10^{-240}$	$\frac{1}{100} = 10^{-2}$	$\frac{1}{100}$

The last four columns of (sometimes approximate) *numerical* table entries are, for each of the six protocols:

- ⊙ the number of samples that can be selected; *i.e.*, the size of the set of all possible samples;
- ⊙ the ratio of the number of samples a protocol can select to the number for EPS from an unstratified population;
- ⊙ the probability any *sample* is selected; here, the *reciprocal* of the number of samples [but see the first comment (+) below];
- ⊙ the probability any *unit* is included in the sample.
  - – In contrast to the *extreme* variation (over nearly 240 orders of magnitude) of the *sample* selecting probabilities among the protocols, the six *unit* inclusion probabilities are *all* 1 in 100 in this illustration (the *equality* of these six probabilities is a characteristic of this illustration, *not* a general result).
    - + Use of EPS at the one or both stages of each protocol means that, in *this* illustration, all *samples* the protocol can select are *equally* likely; as a consequence, the *sample* selecting probability for each protocol in the fifth ('Sample') column of Table HL21.6 above is the *reciprocal* of its number of samples [but see the first comment (○) in Note 7 at the bottom of the facing page HL21.7]. Thus, from the perspective of *samples*, the protocols are:
      ○ *alike* in having their possible samples *equi*probable; ○ *different* in their numbers of possible samples.
    - + Although *calculations* for *unit* inclusion probabilites for the one-stage and the two-stage clustered and stratified protocols have the *same* structure, they yield vastly different numbers of samples; there are also other important *statistical* distinctions between clustered and stratified protocols – see Example HL21.2 overleaf on page HL21.5.

NOTES: 4. EPS from an unstratified population yields the (exhaustive) set of all *possible* samples of a given size from a population of a given size; this set contains about 6.5×10^{241} samples when $\mathbf{N} = 10{,}000$ (elements or) units and $n = 100$ units.

- ● Each of the other five sampling protocols can select only a *sub*set of this (exhaustive) set of samples.

(*continued*)

ERROR: Sample Error – Why Does Introductory ... Emphasize 'Random' Selecting? (continued 3)

NOTES: (cont.) 4. ● – These five protocols are useful because EPS from an unstratified population can rarely meet Plan requirements.
– When these protocols are properly implemented, they *preferentially* exclude samples with an extreme value for an attribute like an average, thus *de*creasing sampling imprecision (as also in Example 21.2).

● As well as yielding all possible samples, EPS is emphasized in introductory discussions because it is:
– involved in more practically useful protocols like the last five in Table HL21.6 on the facing page HL21.6;
– the basis of sampling theory for quantifying the behaviour of sample error under repetition, *i.e*., for quantifying sampling imprecision, as previously discussed in Sections 3 and 4 in this Highlight #21.

Thus, we need to distinguish:

∗ **EPS from an unstratified population**: a protocol for selecting units which is seldom used in practice but which is the basis of sampling theory; FROM:

∗ **EPS** (unqualified): *part* of a protocol for selecting units which involves *other* statistical ideas like stratifying and/or clustering and/or systematic selecting – this is the more *common* usage of 'EPS'.

5. A simple image of how EPS is implemented is to have, in a box, a slip of paper labelled for each unit in the respondent (or study) population; the $\mathbf{N}$ slips are thoroughly mixed and then n are selected without replacement – the labels of these n slips specify the units that comprise the sample (or, more correctly, the selection).

● It is seldom recognized how much effort is needed to *really* mix ('randomize') a collection of items like tickets or slips of paper, whose 'rough' surfaces do not readily slide over each other; in contrast, there are striking images of *two* sets of 'slippery' plastic capsules in rotating drums used in the 1971 U.S. draft lottery in Program 8 entitled *Describing Relationships* of the video series *Against All Odds: Inside Statistics*.
– In a similar vein, the magician and statistician Persi Diaconis comments in Program 15 entitled *What is Probability?* of *Against All Odds: Inside Statistics* that most people do not realize it takes up to about seven *vigorous* shuffles to properly 'randomize' a deck of cards.

● In practice, EPS would usually be implemented with computer software that makes use of an *equiprobable digit* (or *random number*) *generator* – a source that is equally likely to generate any of the digits 0 to 9 at any position of a string of digits of specified length. Equiprobable digits are also available in printed tables – see the 'Probability Distribution Tables' earlier in these Materials. To use this approach, the (elements or) units of the respondent population (or frame) are usually thought of as being numbered (labelled) from 1 to $\mathbf{N}$.

6. In practice, selecting uses a **frame** – a real or conceptual *list* of the (respondent) population units; 'population' in the protocol descriptions in the first column of Table HL21.6 on the facing page HL21.6 could thus also be 'frame'.

● An advantage of two-stage selecting protocols is that, at the second stage, a frame is required *only* for those clusters or strata selected at the first stage (as in the following Note 7).
– A protocol for selecting units with three or more stages (see the following Note 7) enhances this advantage.

7. For telephone surveys – used for political polling and market research, for example – a **two-stage** selecting protocol for units is often employed:

● in the **first** stage, (listed) telephone numbers of a sample of households in the relevant geographic area(s) are generated equiprobably;

● in the **second** stage, the person who first answers the call to each household in the sample is asked to pass the call to the eligible household member (a Canadian citizen for a political poll, a homemaker for market research) who had the most recent birthday; this procedure implements (roughly) EPS of the eligible household members.

An advantage of this two-stage selecting process is that, when the units are *people* but there is a readily available (cheap) frame of *households* (*i.e*., **clusters**), the frame of household members need be generated *only* for those households in the sample and each such frame exists only in the mind of the person who first answers the telephone call [recall the first comment (–) immediately above Example HL21.2 on page HL21.5]. How accurately this first person follows the interviewer's instructions affects the degree to which EPS is achieved at the second stage.

○ Because households have differing numbers of members, unit inclusion probabilities are *un*equal at the second stage; these *two* stages of equiprobable selecting therefore do *not* achieve EPS overall (*un*like Table HL21.6).

○ Because of non-response, many more (typically about *four* times as many) households need to be selected at the first stage as are required for the final sample size; for example, a national poll of 1,500 people may require around 6,000 telephone numbers to be generated, and some of these may have to be called multiple times to reach the eligible household member – see the newspaper articles EM9342 [reprinted on the overleaf side (page HL78.2) of Statistical Highlight #78], EM9330 and EM9337 (reprinted in Statistical Highlight 16).

A multistage *stratified* selecting protocol for a sample survey of a large geographic area (like a Canadian province) may need to place each large *urban* area in its own stratum. For example, a sample survey in Ontario would rarely want to omit Toronto; its inclusion is *assured* by making Toronto a stratum with an inclusion probability of 1.

(*continued overleaf*)

NOTES: (cont.) 8. Systematic selecting is included in this Highlight #21 because it is commonly used in practice; however, *we* think of it as being *equivalent* to EPS by applying the restrictive assumption that the frame (from which every k*th* unit is selected for the sample) has the units arranged so any value of the response variate is equally likely to be anywhere on the list (an **equiprobably ordered frame** for a given response variate). Three illustrations are:

- If a list of UW Faculty of Mathematics students, arranged in alphabetical order by family name, is used as a frame for 1-in-8 systematic selecting, the sample of about 500 students would most likely be essentially equivalent to selecting the students equiprobably from the list if the Question(s) involve the level of student debt but *not* necessarily equivalent to EPS if the Question(s) involve country of birth.
- If a list of family physicians licensed in Ontario, arranged in alphabetic order by family name, is used as a frame for 1-in-100 systematic selecting, the sample of about 300 physicians would most likely be essentially equivalent to selecting the physicians equiprobably from the list if the Question(s) involve drug prescribing characteristics.
- If a list of all school teachers in Ontario, arranged in order by year of graduation, is used as a frame for 1-in-500 systematic selecting, the sample of about 300 teachers would most likely *not* be equivalent to selecting the teachers equiprobably from the list if the Question(s) involve remunerations levels [which tend to *in*crease with time since graduation (althought the roughly 'linear' form of this trend can be exploited statistically)].

Thus, in this Highlight #21, we think of two approaches to achieving equiprobability for the sample selecting process:

○ via an equiprobable **selecting process**, applied to a frame in *any* order;
○ via a *systematic* selecting process, applied to an equiprobably **ordered frame** (for a given response variate).

The second approach achieves (close to) equiprobability only under more restrictive conditions than the first approach.

9. In addition to our emphasis in this and other Statistical Highlights on (survey) sampling in the context of a *finite* respondent population of size **N** elements, the ideas of clustering and multistage selecting also require us to distinguish the **units** of the selecting process from the **elements** determined by the Question(s).

- For example, an investigation to answer Question(s) about people (*elements*) may use a frame of households (*units*).
  - *If* the Question(s) are about households, a unit and an element would be the *same* (a household).

Introductory statistics courses (like STAT 231, for example) tend, for simplicity, to largely ignore the element-unit distinction but, in doing so, it becomes easier to introduce the idea of an *infinite* 'population' with*out* emphasizing that this is *only* a (sometimes useful) *model* – real-world populations of interest in statistics are, by their nature, finite.

Elsewhere, elements may be called **elementary units** or **observation units**; units may be called **sampling units**.
Multistage sampling Plans have **primary sampling units**, **secondary sampling units**, etc., at their successive stages.

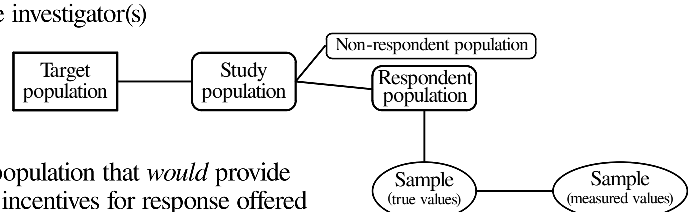
10. When answering a Question with a *descriptive* aspect, there is terminology – **sample survey** and **census** – to distinguish whether the investigating involves some or all of the elements (or units) of the respondent population.

- No such distinguishing terminology exists when answering a Question with a *causative* aspect, perhaps because essentially *all* such investigating involves *sampling.*

SOURCE: The discussion in Appendix 5 (which starts on page HL21.6) is based on material on page VII – 1 in MacKay, R.J. *Experimental Design and Sampling.* Course Notes for Statistics 332/362, University of Waterloo, Fall, 2005.

9. Appendix 6: Some Terminology [A full Glossary is given in Statistical Highlight #91.]

The schema at the right below shows five groups of elements which *we* distinguish for data-based investigating.



∗ **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 [such incentives arise predominantly when the elements are people, but *missing data* may also arise when elements are *in*animate].

∗ **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. [See also Note 1 on page HL6.3 in Statistical Highlight #6.]

∗ **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 the vertical line in the schema above reminds us).

∗ **Sample error:** the difference between [the (true) values of] the sample and respondent population attributes.

∗ **Sampling imprecision:** standard deviation of sample error (*i.e.*, its *haphazard* component exhibited as *variation*) under *repetition* of selecting and estimating.

∗ **Sampling inaccuracy:** average sample error (*i.e.*, its *systematic* component) under *repetition* of selecting and estimating.