University of Waterloo W. H. Cherry

SAMPLING and SURVEY SAMPLING: Selecting Protocols Beyond Equiprobable Selecting

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 HL85.1 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 HL85.1: 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 *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 HL85.1
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 HL85.1.

Table HL85.2a
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 HL85.3a
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 HL85.2b
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 HL85.3b
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 HL85.2c
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 HL85.3c
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 HL85.1 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 HL85.1, 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].

(*continued overleaf*)

Appendix: Equiprobable (Simple random) Selecting – 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; two protocols relevant here are as follows.

- ∗ **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; 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 HL85.4 below.
    - + 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.
  - – 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.

To illustrate further the distinction (introduced in the discussion above of EPS) between sample selecting and unit inclusion probabilities, and also 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 ($h = 1, 2, ..., 10$);
- ○ 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 HL85.4 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 HL85.4 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 HL85.4 should generally be avoided because their brevity can (temporarily) obscure the nature of, and differences among, the protocols.

Table HL85.4

				. . . . Selecting or inclusion probability.	
Protocol for selecting units	**Short name**	**Number of samples**	**Ratio to EPS**	**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 comment (+) on page HL85.3);
- ⊙ 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).

(*continued*)

University of Waterloo W. H. Cherry

SAMPLING and SURVEY SAMPLING: Selecting Protocols Beyond EPS (continued 1)

+ 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 HL85.4 on the facing page HL85.2 is the *reciprocal* of its number of samples [but see the first comment (○) overleaf in Note 5 in the middle of page HL85.4]. 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 HL85.1 on page HL85.1.

NOTES: 1. 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.
  - 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 (*e.g.*, see Statistical Highlight #21).
- 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 HL85.4 on the facing page HL85.2;
  - the basis of sampling theory for quantifying the behaviour of sample error under repetition, *i.e.*, for quantifying sampling imprecision – see Statistical 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'.

2. 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 population.

∗ **Judgement selecting:** human judgement is used to select n units from the $-\!\!\!\mathbf{N}$ (elements or) units of the respondent population. [Judgement selecting is discussed in Statistical Highlights #21 and #83.]

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).

3. 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 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

(*continued overleaf*)

NOTES: (cont.) 3.

- 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 **N**.

4. 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 HL85.4 on page HL85.2 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 5).
  - A protocol for selecting units with three or more stages (see the following Note 5) enhances this advantage.

5. 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 HL85.1 on page HL85.1]. 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 HL85.4).

○ 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.

6. Systematic selecting is included in this Highlight #85 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 #85, 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.

SOURCE: MacKay, R.J. *Experimental Design and Sampling.* Course Notes for Statistics 332/362, University of Waterloo, Fall, 2005, page VII – 1.