11. Appendix 5: Sampling Protocols Beyond EPS

Equiprobable (or simple random) selecting of *individual* ('basic') *units* 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 ('basic') units – 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 units with*in* a stratum have *similar* response variate values ('homogenity of strata') and units in different strata *differ* as much as practicable from each other; the sample is obtained by equiprobable selecting of units from *each* stratum.

Example 5.8.3 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 5.8.3: A respondent population of $\mathbf{N} = 4$ 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 $\overline{Y}$, representing the sample average $\overline{y}$, as an estimator of $\overline{\mathbf{Y}}$, using three sampling protocols:

- EPS of individual *units* from the *un*stratified population;
- EPS of one *cluster* of size $L = 2$ from the *un*stratified population;
- EPS of one *unit* 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 5.8.5
Unstratified population
EPS of two *units*

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 5.8.3.

Table 5.8.6a
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 5.8.6b
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 5.8.6c
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 5.8.7a
Stratified population
Strata: [1, 2], [4, 5]
EPS of one *unit* 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 5.8.7b
Stratified population
Strata: [1, 4], [2, 5]
EPS of one *unit* 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 5.8.7c
Stratified population
Strata: [1, 5], [2, 4]
EPS of one *unit* 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 5.8.3 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 5.8.3, 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 is *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 *units*, the converse does *not* hold – in the three clustered and three stratified sampling protocols, all *units* 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].