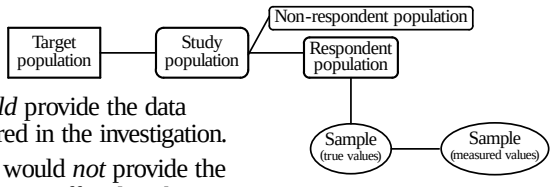
University of Waterloo W. H. Cherry

ERROR: Sample Error – Illustrations from the Set of All Possible Samples from a Population

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



- **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 elements selected from the respondent population *actually used* in an investigation – the sample is a *sub*set of the respondent population.

Associated with these groups of elements are:

- **attributes**: quantities defined as a function of response (and, perhaps, explanatory) variates over the group (*e.g.*, an average).

Example HL19.1: A respondent population of $N = 4$ elements has the following integer Y-values for its response variate:

1, 2, 4, 5 (so that: $\overline{Y} = 3$, $S \simeq 1.8257$).

Table HL19.1 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 ($\bar{y}$) and sample error for each sample are also given.

Table HL19.1
EPS of n = 2 units

Sample	$\bar{y}$	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

- 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 [modelled by the random variable $\overline{Y}$ (with values $\bar{y}$)] is an *unbiased estimator* of the respondent population average ($\overline{Y}$); symbolically, we write: $E(\overline{Y}) = \overline{Y}$ or $E(\overline{Y}) - \overline{Y} = 0$. -----(HL19.1)
  - EPS is needed for unbiasedness when estimating $\overline{Y}$ to ensure that all samples are *equally likely*, as reflected in calculating the *average* sample error.
- For the four Y-values of the response variate in this population, there *are* samples whose average is *equal* to the respondent population average, $\overline{Y}$; however, there are many populations for which *no* sample of a given size (or, conceivably, of *any* size other than N) has an average of $\overline{Y}$; *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.

To provide a broader illustration of properties of (the process of) EPS, using again the idea of the set of all possible samples, sets of ten and nine histograms [for (the attributes) sample average and sample (data) standard deviation] are given on the next two pages (HL19.2 and HL19.3) for a population of $N = 10$ elements with response variate values $Y = 1, 2, ..., 10$.

- 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* $\overline{Y}$, 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 (as also illustrated in Example HL19.1 above).
- 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 Theorem – 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; the latter 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.
- The histograms of sample (data) standard deviations (on page HL19.3) 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.
    - Estimating bias is *unlike* sampling and measuring inaccuracy which do *not* decrease with increasing sample size.

(*continued overleaf*)

University of Waterloo W. H. Cherry

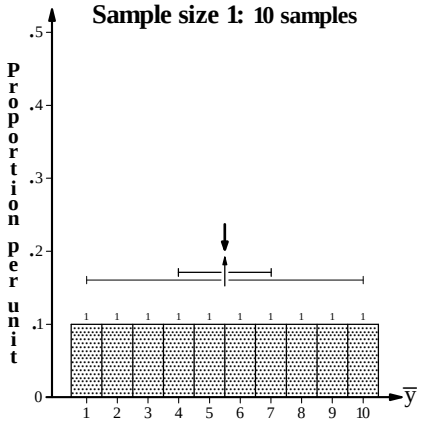
ACKNOWLEDGEMENT: The data for the histograms which follow were kindly calculated by Professor R.W. Oldford using Quail.

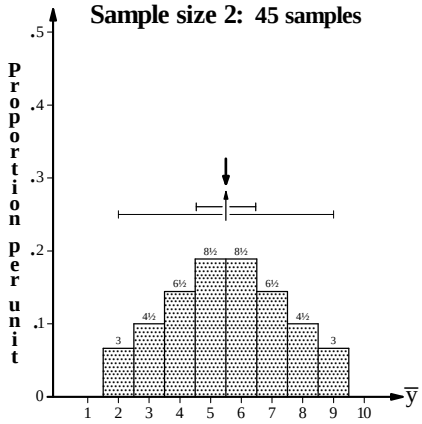
The ten histograms below show the distribution of the *averages* of all possible samples of a given size that can be selected without replacement from a population of $N = 10$ elements, whose response variate values are $Y = 1, 2, \ldots, 10$. [A histogram of these ten population *responses* would be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

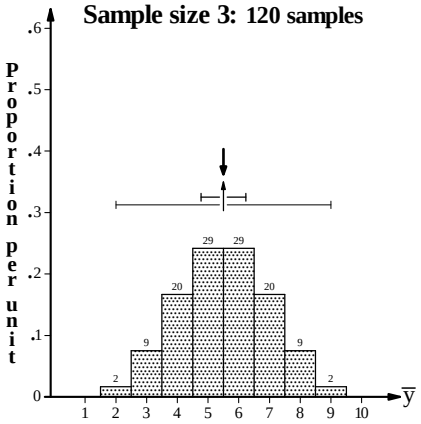
The decreasing vertical axis scale unit down the page reduces the visual impact of increasing histogram height; bar *frequencies* are given by the numbers at the tops of the bars.

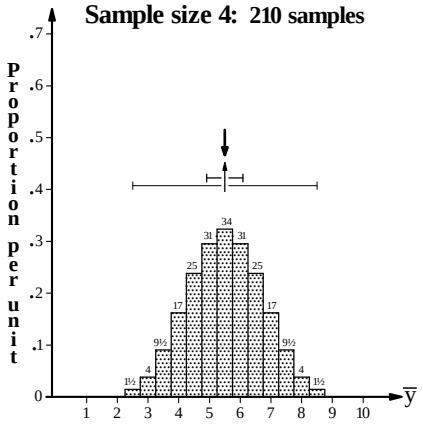
Above each histogram, except the one at the bottom right of the page:

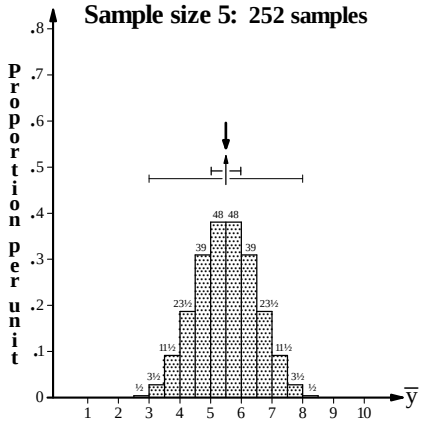
- the upper bold arrow (↓) indicates the value of the *population* average $\bar{Y} = 5.5$;
- the lower arrow (↑) indicates the *average* of the set of *sample* averages ($\bar{\bar{y}} = 5.5$);
- the upper bar (⊢—⊣) crossing the lower arrow has a length equal to the value of the (data) *standard deviation*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

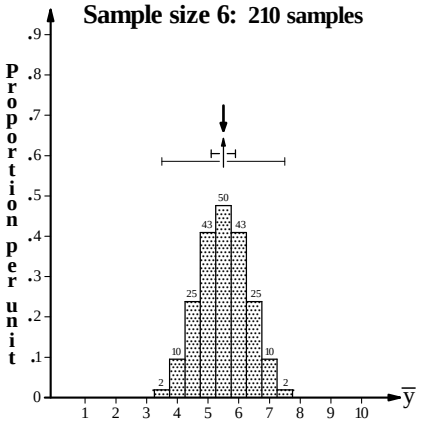


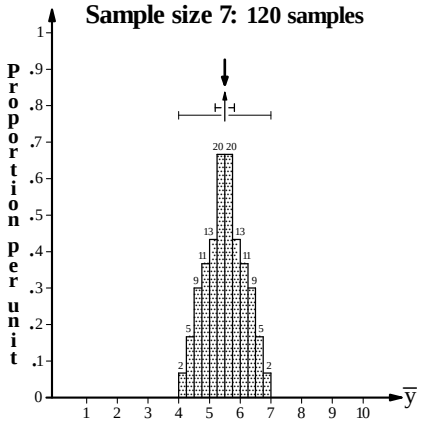


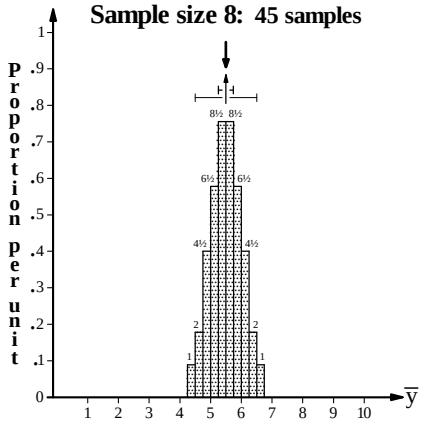


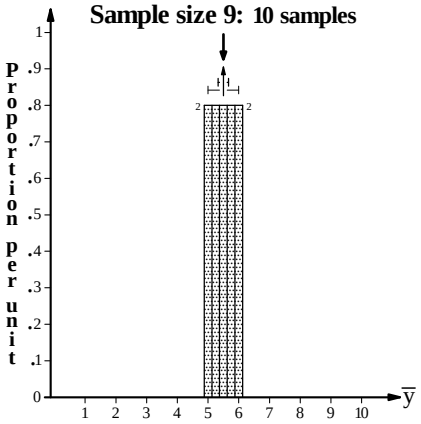


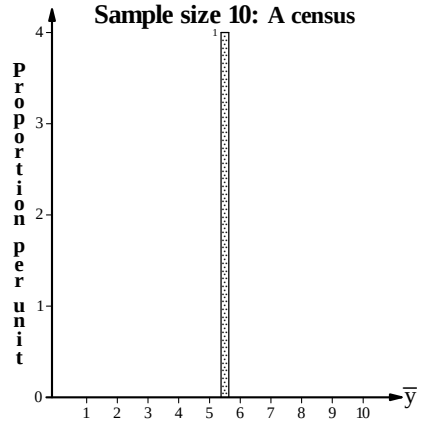












(*continued*)

University of Waterloo W. H. Cherry

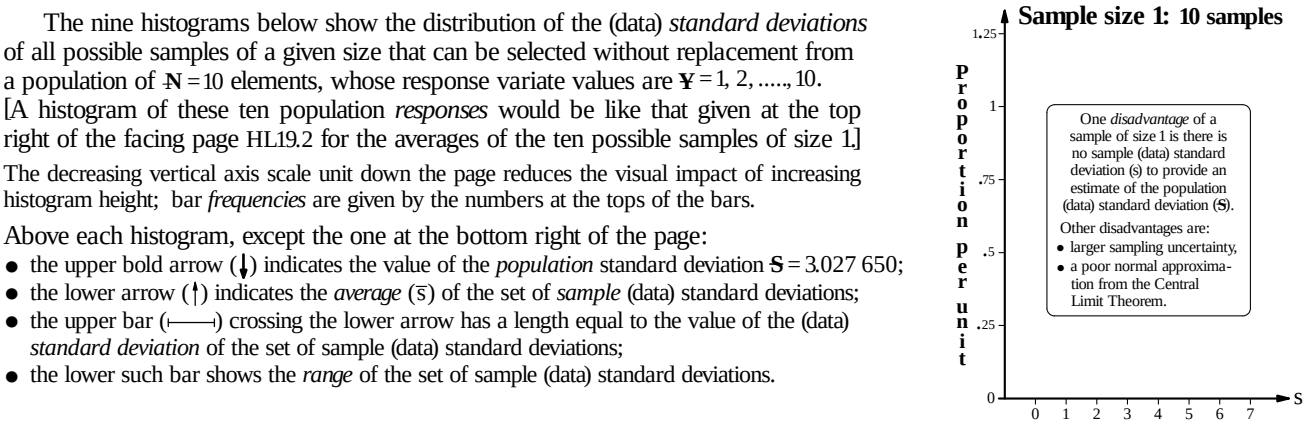
ERROR: Sample Error – Illustrations from the Set of All Possible Samples from a Population (continued 1)

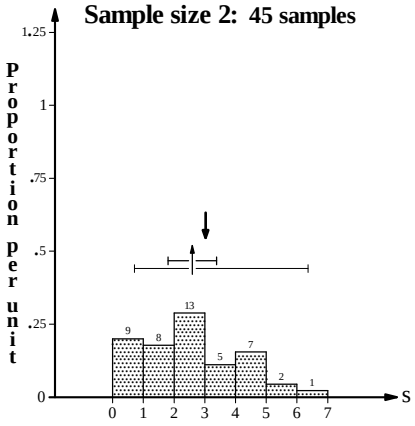
The nine histograms below show the distribution of the (data) *standard deviations* of all possible samples of a given size that can be selected without replacement from a population of $N = 10$ elements, whose response variate values are $Y = 1, 2, \ldots, 10$. [A histogram of these ten population *responses* would be like that given at the top right of the facing page HL19.2 for the averages of the ten possible samples of size 1.]

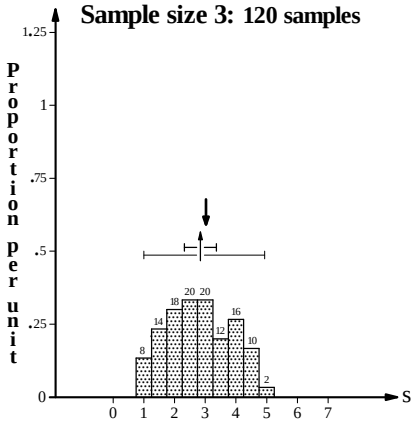
The decreasing vertical axis scale unit down the page reduces the visual impact of increasing histogram height; bar *frequencies* are given by the numbers at the tops of the bars.

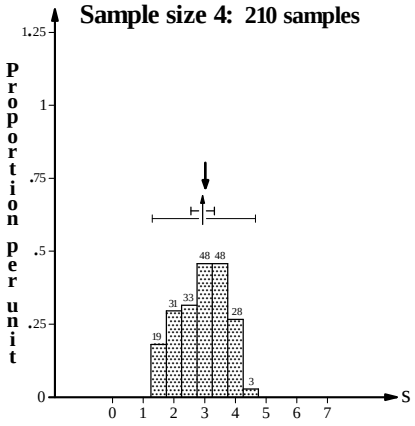
Above each histogram, except the one at the bottom right of the page:

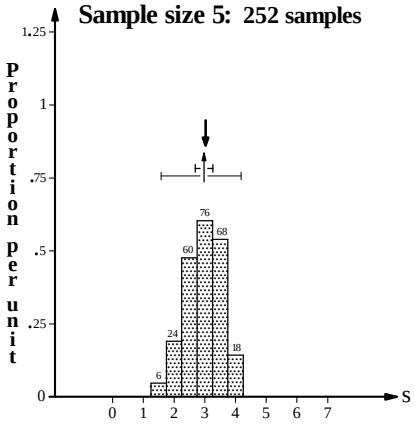
- the upper bold arrow (↓) indicates the value of the *population* standard deviation $S = 3.027\,650$;
- the lower arrow (↑) indicates the *average* ($\bar{s}$) of the set of *sample* (data) standard deviations;
- the upper bar (⊢—⊣) crossing the lower arrow has a length equal to the value of the (data)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

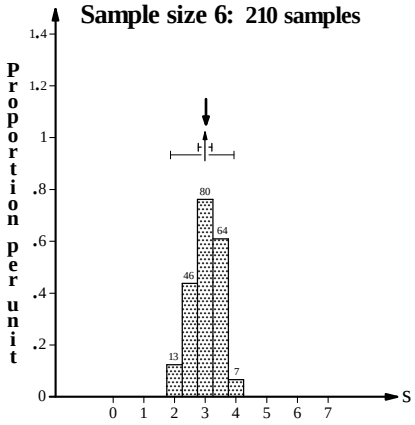


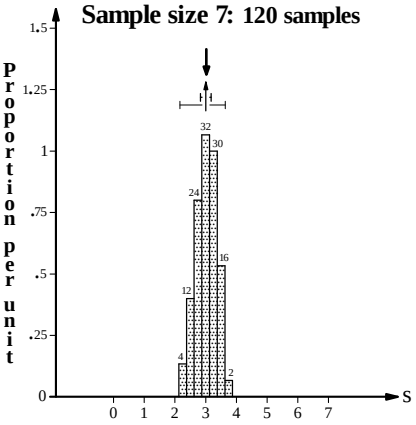


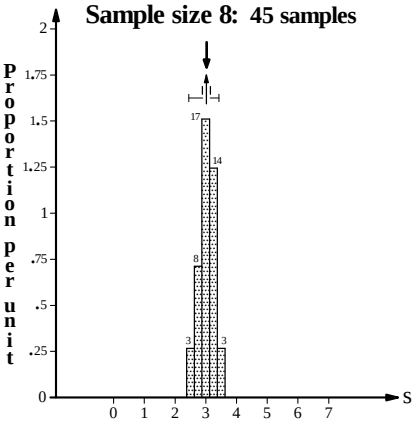


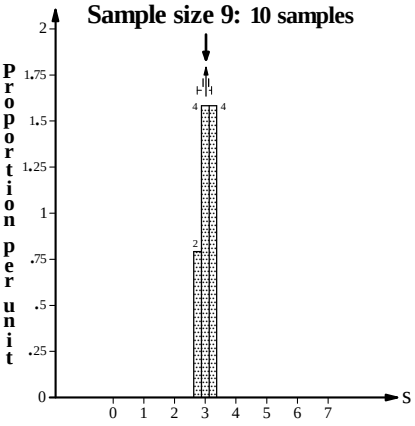


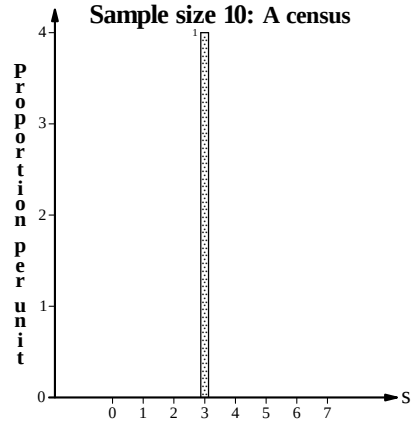












(*continued overleaf*)

University of Waterloo W. H. Cherry

Appendix: 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 HL19.2 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 HL19.2: A respondent population of $\mathbf{N} = 4$ elements has the following integer $\mathbf{Y}$-values for its responses:

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 HL19.2
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 HL19.2.

Table HL19.3a
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 HL19.3b
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 HL19.3c
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 HL19.4a
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 HL19.4b
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 HL19.4c
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 HL19.2 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 HL19.2, 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].