Figure 2.14. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Equal-Sized Clusters Estimating an Average or a Total

In this Figure 2.14, we extend the idea of equiprobable selecting (EPS) to situations where the selected units are not *individual* respondent population elements but rather collections (*i.e.*, *clusters*) of L such elements; the response of *all* elements in each selected cluster is recorded. Thus, we are dealing with the case of *equiprobable selecting of clusters of equal size* but the more usual description is (the possibly misleading) [*one-stage equiprobable (simple random)*] *cluster sampling with equal-sized clusters*; examples of such clusters are a carton of cans or packages in the food industry, a box of manufactured components or articles and, more generally, a netful of fish, a page of a book, a class of students or a household of people. Notation for, and relationships among, numbers of elements and of clusters are shown in Table 2.14.1 at the right.

Table 2.14.1:

	Elements	Clusters	Relationships
Respondent population	$\mathbf{N}$	$\mathbf{M}$	$\mathbf{N} = \mathbf{M}\mathrm{L}$, $\mathrm{L} = \mathbf{N}/\mathbf{M}$
Sample	n	m	$\mathrm{n} = \mathrm{mL}$, $\mathrm{L} = \mathrm{n}/\mathrm{m}$

Also: $\overline{\mathbf{Y}}_{\mathrm{i}} = \frac{1}{\mathrm{L}}\sum_{\mathrm{k}=1}^{\mathrm{L}}\mathbf{Y}_{\mathrm{ik}}$ is the average response of the ith population cluster,

$\overline{y}_j = \frac{1}{\mathrm{L}}\sum_{l=1}^{\mathrm{L}} y_{jl}$ is the average response of the jth sampled cluster,

Selecting clusters is, in practice, usually undertaken for reasons of *convenience*, when a decision is made that the gain in the ease of carrying out the survey more than compensates for the common *statistical* disadvantage of higher imprecision for a given sample size. The practical advantages of selecting clusters include the following:

+ *not* having to obtain a frame listing all the *elements* of the study population – a frame (*e.g.*, a list) of *clusters* is sufficient;
+ *not* having (although it may be desirable) to know $\mathbf{N}$, the total number of *elements* in the respondent population, to be able to estimate $\overline{\mathbf{Y}}$, the respondent population average, and ${}_{\mathrm{T}}\mathbf{Y}$, the respondent population total;
+ assuming that there is complete response, gathering data from *all* the elements in the selected clusters may mean that a large sample can be obtained relatively easily and cheaply;
+ if clusters are defined *geographically*, there would usually be lower travel *costs* associated with data collection than would be the case for the more dispersed sample of *individual* elements representing the same sample size.

The main *dis*advantage of selecting clusters is:

– for a given sample size, precision of estimators is usually *lower* than EPS of units consisting of *individual* elements; starting with the discussion of efficiency in Section 2 on page 2.108 of this Figure 2.14, we find that greater imprecision becomes more pronounced as the *intracluster variation in the response becomes smaller in relation to the overall respondent population variation.*

1. Estimating $\overline{\mathbf{Y}}$, the Respondent Population Average

Equation (2.14.1) at the right, for the sample average under EPS of units each consisting of *one* element, involves the following matters (recall Table 2.3.5 in Figure 2.3):

$$\overline{y} = \frac{1}{\mathrm{n}}\sum_{j=1}^{\mathrm{n}} y_j \qquad \text{-----(2.14.1)}$$

* the $\binom{\mathbf{N}}{\mathrm{n}}$ samples of size n are all equally probable;
* the divisor is the *same* as the upper limit of the index of summation;
* y_j is the response of a respondent population unit consisting of *one element* selected from that population;
* EPS theory shows that the standard deviation of the sample average under EPS is given by equation (2.14.2) [equation (2.3.9) in Figure 2.3] at the right.

$$s.d.(\overline{Y}) = \mathbf{S}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \qquad \text{-----(2.14.2)}$$

To develop the theory for one-stage EPS of equal-sized clusters (denoted *ec*), it is convenient to argue by analogy from these ideas for EPS of *individual* elements (denoted by a subscript *in* when appropriate); to do so, we need to obtain, in an appropriate form, an expression for the average of a sample of m clusters of size L.

The average response for m clusters, each of size L, obtained by EPS is given by:

$$\overline{y}_{ec} = \frac{1}{\mathrm{mL}}\sum_{j=1}^{\mathrm{mL}} y_j$$

[Equiprobable selecting theory for individual elements does *not* apply to this expression because all the $\binom{\mathbf{M}\mathrm{L}}{\mathrm{mL}}$ samples of size mL are not equally probable; y_j is the value of the response for the jth selected element]

$$= \frac{1}{\mathrm{mL}}\sum_{j=1}^{\mathrm{m}}\sum_{l=1}^{\mathrm{L}} y_{jl}$$

[Equiprobable selecting theory for individual elements does *not* apply to this expression because it is not of the appropriate algebraic form; y_{jl} is the value of the response for the lth element in the jth selected cluster]

$$= \frac{1}{\mathrm{mL}}\sum_{j=1}^{\mathrm{m}} {}_{\mathrm{T}}y_j$$

[Equiprobable selecting theory for individual elements does *not* apply to this expression because the divisor (mL) is not the same as the upper limit of the index of summation; ${}_{\mathrm{T}}y_j$ is the *total* of the jth of the m selected clusters]

$$= \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}} \overline{y}_j$$

[Equiprobable selecting theory for individual elements *can* be applied to this expression but the response is now the *average* ($\overline{y}_j$) of each of the m selected clusters].

Hence: $s.d.(\overline{Y}_{ec}) = \mathbf{S}_{ec}\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}}$ where:

$$\mathbf{S}_{ec} = \sqrt{\frac{1}{\mathbf{M}-1}\sum_{\mathrm{i}=1}^{\mathbf{M}}(\overline{\mathbf{Y}}_{\mathrm{i}} - \overline{\mathbf{Y}})^2} \equiv \sqrt{\frac{1}{\mathbf{M}-1}\Big[\sum_{\mathrm{i}=1}^{\mathbf{M}}\overline{\mathbf{Y}}_{\mathrm{i}}^2 - \mathbf{M}\overline{\mathbf{Y}}^2\Big]}, \qquad \text{-----(2.14.3)}$$

and: $\overline{\mathbf{Y}}_{\mathrm{i}}$ is the *average* response of (all L elements in) *cluster* i in the respondent population [see equation; (2.14.10) on page 2.111];

recall from Table 2.3.5 on page 2.40 in Figure 2.3 that, when n *individual* elements are obtained by EPS from the $\mathbf{N}$ elements of the respondent population, we defined:

$$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\mathbf{Y}_{\mathrm{i}} - \overline{\mathbf{Y}})^2} \equiv \sqrt{\frac{1}{\mathbf{N}-1}\Big[\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_{\mathrm{i}}^2 - \mathbf{N}\overline{\mathbf{Y}}^2\Big]}, \qquad \text{-----(2.14.4)}$$

(*continued overleaf*)

where: $\mathbf{Y}_i$ is the (individual) response of *element* i in the respondent population.

To *estimate* the standard deviation of $\overline{Y}_{ec}$, $\mathbf{S}_{ec}$ is estimated by s_{ec}, the standard deviation of the cluster sample, given by equation (2.14.5) at the right; the estimated standard deviation of $\overline{Y}_{ec}$ is then given by equation (2.14.6).

$$s_{ec} = \sqrt{\frac{1}{\mathrm{m}-1}\sum_{j=1}^{\mathrm{m}}(\overline{y}_j - \overline{y}_{ec})^2} \equiv \sqrt{\frac{1}{\mathrm{m}-1}\Big[\sum_{j=1}^{\mathrm{m}}\overline{y}_j^2 - \mathrm{m}\overline{y}_{ec}^2\Big]} \qquad \text{-----(2.14.5)}$$

$$\hat{s.d.}(\overline{Y}_{ec}) = s_{ec}\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}} \qquad \text{-----(2.14.6)}$$

Comparing equations (2.14.5) and (2.14.6) with equations (2.3.3) and (2.3.17) from the bottoms of pages 2.37 and 2.41 in Figure 2.3, we see that EPS of units consisting of *clusters* of elements is like selecting individual elements except that:

$$s = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^2} \qquad \text{-----(2.3.3)}$$

$$\hat{s.d.}(\overline{Y}_{in}) = s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \qquad \text{-----(2.3.17)}$$

- m of $\mathbf{M}$ clusters are selected rather than n of $\mathbf{N}$ elments (although both samples contain the *same* number of elements),
- the *averages* of the selected clusters of L elements replace the individual element responses.

Thus, on the basis of the approximate normality of $\overline{Y}_{ec}$ as a consequence of the Central Limit Theorem, and arguing in a general way from the use of the *t* distribution in normal theory when the respondent population standard deviation is estimated by the sample standard deviation, the theory developed overleaf on page 2.105 and above leads to an in interval involving random variables:

$$I = [\overline{Y}_{ec} - {}_{\alpha}t^*_{m-1}\,\hat{s.d.}(\overline{Y}_{ec}),\ \overline{Y}_{ec} + {}_{\alpha}t^*_{m-1}\,\hat{s.d.}(\overline{Y}_{ec})],$$

such that $\Pr(I \ni \overline{\mathbf{Y}}) \simeq 100(1-\alpha)\%$, where ${}_{\alpha}t^*_{\mathrm{m}-1}$ is the $100(1-\alpha/2)th$ percentile of the $t_{\mathrm{m}-1}$ distribution. For calculating an approximate $100(1-\alpha)\%$ *confidence interval* for $\overline{\mathbf{Y}}$, we use:

$$\overline{y}_{ec} \pm {}_{\alpha}t^*_{m-1}\,\hat{s.d.}(\overline{Y}_{ec}) = \Big[\overline{y}_{ec} - {}_{\alpha}t^*_{m-1}s_{ec}\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}},\ \overline{y}_{ec} + {}_{\alpha}t^*_{m-1}s_{ec}\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}}\Big]. \qquad \text{-----(2.14.7)}$$

Notes 1, 2, 4, 5, 6, 7, 8 and 9 on pages 2.41 and 2.42 in Figure 2.3, near the expressions [equations (2.3.18) and (2.3.19)] for a confidence interval for $\overline{\mathbf{Y}}$ when the units consist of *individual* elements, are also generally relevant to equation (2.14.7) above.

NOTES: 1. Expressions for calculating a confidence interval for ${}_T\mathbf{Y}$, the respondent population *total*, can be obtained by multiplying those given above for $\overline{\mathbf{Y}}$ [in equation (2.14.7)] by $\mathbf{M}$L (= $\mathbf{N}$).

2. It was established in equation (2.3.8) at the top of page 2.41 in Figure 2.3 that the random variable ($\overline{Y}_{in}$) [representing the sample average for EPS of units consisting of *individual* elements] is an unbiased estimator of $\overline{\mathbf{Y}}$; it therefore follows that the random variable $\overline{Y}_{ec}$ (representing the *cluster* sample average for EPS of units each consisting of L elements of the respondent population) is also an unbiased estimator of $\overline{\mathbf{Y}}$, the respondent population average.

Example 2.14.1: In a student residence which houses 160 people, there are separate kitchen facilities for 40 groups of four residents. A sample survey was taken to estimate, among other things, the average amount spent per person on food during the week of June 5-11, 1994. For 8 of the groups of four residents obtained by EPS, each person was asked to keep a diary of the amount they spent on food (expenditures to be included were carefully defined for the purposes of the sample survey); the results (to the nearest dollar) were as follows:

Table 2.14.2:

Group	1	2	3	4	5	6	7	8
Amounts	115	32	75	90	45	81	62	97
	106	67	82	82	53	85	64	75
	79	65	59	75	62	78	63	105
	84	63	62	69	41	74	67	105
Total	384	227	278	316	201	318	256	382
Average	96.00	56.75	69.50	79.00	50.25	79.50	64.00	95.50.

(a) Find an approximate 95% confidence interval for the *average* amount spent on food per student in the residence during the specified week.

(b) Find an approximate 95% confidence interval for the *total* amount spent on food in the residence during the specified week.

(c) Describe briefly the advantage(s) of the method of data collection (*i.e.*, keeping a diary of food expenditures), and also its *dis*advantage(s) in the terms of limitation(s) on the sample survey's answers.

Solution: (a) We have: $\mathbf{N} = 160$, $\mathrm{n} = 32$, $\mathbf{M} = 40$, $\mathrm{m} = 8$, ${}_{.05}t^*_7 = 2.36462$ for 95% confidence;

for the 8 averages: $\sum_{j=1}^{\mathrm{m}}\overline{y}_j = 5{,}90.5$, $\sum_{j=1}^{\mathrm{m}}\overline{y}_j^2 = 45{,}569.375$,

so that: $\overline{y}_{ec} = 5{,}90.5/8 = \$73.8125$ (= $\overline{y}_{ec}$), $s_{ec} = \$16.831\,491\,53$ (= s_{ec}) [equation (2.14.5)].

Then: $\hat{s.d.}(\overline{Y}_{ec}) = s_{ec}\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}} = \$16.831\,492\sqrt{\frac{1}{8} - \frac{1}{40}} = \$5.322\,584\,966.$

Hence, an approximate 95% confidence interval for $\overline{\mathbf{Y}}$, the *average* amount spent on food per student in the residence during the specified week. is:

$\overline{y}_{ec} \pm 2.36462 \times \hat{s.d.}(\overline{Y}_{ec}) = 73.8125 \pm 2.36462 \times 5.322\,585 \implies$ (\$61.23, \$86.40) or about (\$61, \$87).

Figure 2.14. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Equal-Sized Clusters Estimating an Average or a Total (continued 1)

Solution: (cont.) (b) To convert the result in (a) for an *average* to that for a *total*, we multiply through by $\mathbf{N} = 160$ (*not* by $\mathbf{M} = 40$), to obtain:

$$\$11{,}810.00 \pm 2{,}013.74 \;\Rightarrow\; (\$9{,}796,\ \$13{,}824) \text{ or about } (\$9{,}700,\ \$13{,}900).$$

(c) An *advantage* of keeping a diary is that, if done conscientiously, it reduces mistakes arising from poor recall of relevant activities, food purchases in this instance.

A *dis*advantage is that the act of recording behaviour may cause the behaviour to change; this would be an example of the measuring process *altering* the value of the variate being measured.

Both matters affect *measurement* error – the first makes it smaller, the second may make it larger, with possibly these same effects, respectively, on measuring inccuracy and measuring imprecision.

Example 2.14.2: A newspaper has 39,800 subscribers served by carrier routes. A computer file contains one record per subscriber; the records of each carrier are consecutive in geographical order and neighbouring carrier routes follow each other in the file. The number of subscribers per carrier varies between about 50 and 200.

One question in the sample survey is to estimate how many subscribers own their home. There are resources for a sample of size 400 and, to use interviewer travel time efficiently, clusters of 10 subscribers obtained by EPS will be surveyed, because an interviewer can generally complete ten interviews in one neighbourhood in half a day.

We have: number of *elements* in the study population: $\mathbf{N} = 39{,}800$
number of *elements* in the sample: $n = 400$ (assuming complete response)
number of *clusters* in the study population: $\mathbf{M} = 3{,}980$, $L = 10$
number of *clusters* in the sample: $m = 40$ (obtained by EPS).

The data obtained for the number of people who own their home in each of the 40 selected clusters are:

Table 2.14.3:

10	8	6	5	9	8	8	5	9	9	9	10	4	3	1	2	3	4	0	6
3	5	0	3	0	0	4	0	8	0	10	5	6	1	3	3	1	5	5	4.

To carry out the calculations for cluster selecting for these data, it is convenient to summarize them as follows:

Table 2.14.4:

												Total
Cluster total	0	1	2	3	4	5	6	7	8	9	10	
Cluster size	10	10	10	10	10	10	10	10	10	10	10	--
Cluster average ($\overline{y}_j$)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	--
Frequency observed	6	3	1	6	4	6	3	0	4	4	3	40
$\sum_{j=1}^{m}\overline{y}_j$	0	0.3	0.2	1.8	1.6	3.0	1.8	0	3.2	3.6	3	18.5
$\sum_{j=1}^{m}\overline{y}_j^2$	0	0.03	0.04	0.54	0.64	1.50	1.08	0	2.56	3.24	3	12.63

Estimate the *total* number of the newspaper's subscribers served by carrier routes who own their home, and find an estimate of the standard deviation of the estimator you use.

Solution: From the numerical data summaries in the last two lines of Table 2.14.4 above, we find:

$$\overline{y}_{ec} = 18.5/40 = 0.4625\ (= \overline{y}_{ec}),\quad s_{ec} = 0.323\,195\,185\ (= s_{ec}) \quad \text{[equation (2.14.5)]}.$$

so that: $$\hat{s.d.}(\overline{Y}_{ec}) = s_{ec}\sqrt{\frac{1}{m} - \frac{1}{\mathbf{M}}} = 0.323\,195\,185\sqrt{\frac{1}{40} - \frac{1}{3{,}980}} = 0.050\,844\,205.$$

Multiplying both estimates by $\mathbf{N} = 39{,}800$, the answers for the number of the newspaper's subscribers who own their home is $18{,}407.5 \simeq$ **18,408** with an estimated standard deviation of $2{,}023.599 \simeq$ **2,024**.

NOTES: 3. Each cluster average (denoted $\overline{y}_j$) in Table 2.14.4 above is, of course, the *proportion* of subscribers in a cluster who own their home. Thus, what are actually *counts* have been treated as *measurements* so that the foregoing theory of EPS of equal-sized clusters can be applied to the situation in Example 2.14.2.

4. The source of Example 2.14.2 is Kish, L. *Survey Sampling.* John Wiley & Sons, New York, 1965. [HN29.K5 1965], pages 153-154, who comments that:

- in the (likely) situation where the population size $\mathbf{N}$ is not an exact multiple of the cluster size L, one possible convenient solution is to remove equiprobably from the file the fewer than L subscriber entries that will make the number remaining an exact multiple of L;
- a stratified or a systematic selecting process for the clusers might be preferable to EPS;
- inequalities in cluster sizes introduced by non-response could perhaps be managed by substitutions.

(*continued overleaf*)

2. Efficiency

We now compare the standard deviation of $\overline{Y}_{ec}$ with that of $\overline{Y}_{in}$ where the subscript *in* reminds us that the comparision is EPS of *in*dividual elements with EPS of *e*qual-sized *c*lusters of L elements. Such a comparison is usually referred to as investigating *efficiency*, and typically involves finding a *ratio* or a *difference* of standard deviations (or variances); we will use a difference here and we will work with the *squares* of standard deviation expressions to avoid algebraic complications that are introduced by the presence of square roots.

Remembering that a *positive* difference of the (squared) standard deviations of two estimators means that the *second* estimator is *more* precise, and that $\mathbf{N} = \mathbf{M}$L and n = mL, we have:

$$[s.d.(\overline{Y}_{ec})]^2 - [s.d.(\overline{Y}_{in})]^2 = (\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})\mathbf{S}_{ec}^2 - (\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})\mathbf{S}^2 = (\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})[\mathbf{S}_{ec}^2 - \tfrac{\mathbf{S}^2}{\mathrm{L}}] = (\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})[\mathbf{S}_{ec} + \tfrac{\mathbf{S}}{\sqrt{\mathrm{L}}}][\mathbf{S}_{ec} - \tfrac{\mathbf{S}}{\sqrt{\mathrm{L}}}]. \qquad \text{-----(2.14.8)}$$

Thus, equiprobable selecting (EPS) of clusters of size L will be *more* precise [*i.e.*, (2.14.8) is *negative*] when $\mathbf{S}_{ec} < \mathbf{S}/\sqrt{\mathrm{L}}$ (in the rightmost difference term). We recall that $\mathbf{S}_{ec}$ is the standard deviation for the *averages* of clusters of size L, and $\mathbf{S}$ is the standard deviation for *individual* elements; $\mathbf{S}_{ec}$ is therefore naturally smaller than $\mathbf{S}$ by a factor of (close to) $\sqrt{\mathrm{L}}$ (assuming $\mathrm{L} \ll \mathbf{N}$). Equation (2.14.8) shows that, for $s.d.(\overline{Y}_{ec})$ to be smaller than $s.d.(\overline{Y}_{in})$, the variation among the cluster averages must be small enough for $\mathbf{S}_{ec}$ to be smaller than $\mathbf{S}$ by *more* than its 'natural' decrease of a factor of (close to) $\sqrt{\mathrm{L}}$.

NOTES: 5. The foregoing interpretation of equation (2.14.8) is a particular instance of the general principle that *in*creased precision of estimating is favoured by *de*creased variation among the responses of the population elements (or units).

6. If clusters are formed *equiprobably* (*i.e.*, by EPS of groups of L respondent population elements), then, on average, $\mathbf{S}_{ec} = \mathbf{S}/\sqrt{\mathrm{L}}$ and equation (2.14.8) shows that, as expected, there would be *no* difference in precision between individual element and clustered sampling protocols.

7. The matter of fficiency when selecting clusters compared with selecting individual elements is sometimes considered from another perspective when we recognize that there is a *fixed* amount of variation among the responses of the elements of the respondent population; as a consequence, *de*creased *inter*cluster variation will result in *in*creased *intra*cluster variation, and conversely. Thus, the idea of selecting *clusters* leading to *more* precise estimating under low *inter*cluster variation can be equivalently expressed in terms of *high intra*cluster variation; the symbolic result [corresponding to equation (2.14.8)] for this perspective is:

$$[s.d.(\overline{Y}_{ec})]^2 - [s.d.(\overline{Y}_{in})]^2 = (\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})\frac{\mathbf{M}(\mathrm{L}-1)}{\mathrm{L}(\mathbf{M}-1)}[\mathbf{S} + \overline{\mathbf{S}}][\mathbf{S} - \overline{\mathbf{S}}] \qquad [\text{where:} \quad \overline{\mathbf{S}} = \sqrt{\tfrac{1}{\mathbf{M}}\sum_{i=1}^{\mathbf{M}}\mathbf{S}_i^2}]. \qquad \text{-----(2.14.9)}$$

The quantity denoted $\overline{\mathbf{S}}$ is the *square root of the average of the squares of the intracluster standard deviations* – we refer to it as the *average intracluster standard deviation* (*average* here denotes the *root average square*). Equation (2.14.9) is derived in Appendix 1 on pages 2.111 and 2.112 of this Figure 2.14.

From this alternative perspective, the precision of EPS of units which are clusters of size L, compared with the precision of EPS of units which are individual elements of the respondent population, is determined by the size of $\overline{\mathbf{S}}$ (the average intracluster standard deviation) relative to the size of $\mathbf{S}$ (the respondent population standard deviation). If $\overline{\mathbf{S}} < \mathbf{S}$ (which is commonly the case in practice), $s.d.(\overline{Y}_{ec}) > s.d.(\overline{Y}_{in})$, meaning that EPS of clusters is *less* precise.

- Equation (2.14.9) above allows, in principle, for the possibility that EPS of clusters is *more* precise than EPS of individual elements. However, conditions where this occurs (*i.e.*, where the average intracluster standard deviation is *greater than* the respondent population standard deviation) seldom arise in practice. Rather, it should be recognized that a clustered sampling protocol usually yields *lower* precision, and one aim of a good survey Plan should be to try to define the clusters in a way that will make their internal variation as *large* as feasible (to minimize the loss of precision) while exploiting, as effectively as possible, the *advantages* of selecting clusters. We say we want the Plan to yield clusters that are as *heterogeneous* as is feasible in the response(s) of interest.

8. In this discussion of efficiency, the random variable $\overline{Y}$, representing the average of n *in*dividual elements selected by EPS from the respondent population, has a subscript *in* to emphasize its distinction from $\overline{Y}_{ec}$. In the same vein, near the bottom of the third and at the top of the fourth side of Figure 4.3 in these STAT 332 Course Materials, the added subscript is *un* to distiguish $\overline{Y}$ for selecing from an *un*stratified population from a sampling protocol which involves stratification. Such notational adaptations can produce greater clarity in particular contexts

9. An *extension* of the foregoing discussion of efficiency, and of its further illustration in Example 2.14.3 on the facing page 2.109, is to compare the *widths of confidence intervals* (rather than the values of standard deviations) from EPS of clusters and EPS of individual elements, again assuming for simplicity the *same* sample size of n elements in both cases. Because the relevant *t* distribution will have *fewer* degrees of freedom (m – 1 compared with n – 1) when selecting clusters, this exacerbates its usual loss of precision by decreasing further the precision of an answer expressed as a confidence interval. The *magnitude* of the effect depends on the value of L – larger values (which mean that m is *much* smaller than n) work more strongly *against* precision when selecting clusters.

Figure 2.14. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Equal-Sized Clusters Estimating an Average or a Total (continued 2)

Example 2.14.3: A respondent population of $\mathbf{N} = 25$ elements has the following integer $\mathbf{Y}$-values for its response variate:

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5 (so that: $\overline{\mathbf{Y}} = 3$, $\mathbf{S} \simeq 1.4434$);

a sample of ten elements is chosen either by EPS of $n = 10$ *individual* elements or by EPS of $m = 2$ *clusters* of size $L = 5$ [so that $\mathbf{S}/\sqrt{L} = 0.6455$]. By defining the clusters in different ways, we illustrate below the complimentary effects of inter- and intracluster variation on the precision of $\overline{Y}_{ec}$ as an estimator of $\overline{\mathbf{Y}}$, the respondent population average.

	Case 1	**Case 2**	**Case 3**	**Case 4**	**Case 5**	**Case 6**
	Selecting clusters	Selecting clusters	Selecting clusters	*Selecting indi-*	Selecting clusters	Selecting clusters
	Clusters $\overline{\mathbf{Y}}_i$ $\mathbf{S}_i^2$	Clusters $\overline{\mathbf{Y}}_i$ $\mathbf{S}_i^2$	Clusters $\overline{\mathbf{Y}}_i$ $\mathbf{S}_i^2$	*vidual elements*	Clusters $\overline{\mathbf{Y}}_i$ $\mathbf{S}_i^2$	Clusters $\overline{\mathbf{Y}}_i$ $\mathbf{S}_i^2$
	(1, 1, 1, 1, 1) 1 0	(1, 1, 1, 1, 2) 1.2 0.2	(1, 2, 2, 2, 3) 2 0.5	*from the respon-*	(1, 2, 2, 3, 4) 2.4 1.3	(1, 2, 3, 4, 5) 3 2.5
	(2, 2, 2, 2, 2) 2 0	(2, 2, 2, 2, 3) 2.2 0.2	(2, 3, 3, 3, 4) 3 0.5	*dent population*	(2, 3, 3, 4, 5) 3.4 1.3	(1, 2, 3, 4, 5) 3 2.5
	(3, 3, 3, 3, 3) 3 0	(3, 3, 3, 3, 4) 3.2 0.2	(3, 4, 4, 4, 5) 4 0.5		(3, 4, 4, 5, 1) 3.4 2.3	(1, 2, 3, 4, 5) 3 2.5
	(4, 4, 4, 4, 4) 4 0	(4, 4, 4, 4, 5) 4.2 0.2	(4, 5, 5, 5, 1) 4 3		(4, 5, 5, 1, 2) 3.4 3.3	(1, 2, 3, 4, 5) 3 2.5
	(5, 5, 5, 5, 5) 5 0	(5, 5, 5, 5, 1) 4.2 3.2	(5, 1, 1, 1, 2) 2 3		(5, 1, 1, 2, 3) 2.4 2.8	(1, 2, 3, 4, 5) 3 2.5
$\mathbf{S}_{ec}$	1.5811	1.3038	1	[$\mathbf{S}/\sqrt{L} \simeq 0.6455$]	0.5477	0
	largest ←			intercluster variation		→ *smallest*
	smallest ←			intracluster variation ('heterogeneity')		→ *largest*
$\overline{\mathbf{S}}$	0	0.8944	1.2247	[$\mathbf{S} \simeq 1.4434$]	1.4832	1.5811
$s.d.(\overline{Y}_{ec})$	0.8660	0.7141	0.5477	[$s.d.(\overline{Y}_{in}) \simeq 0.3536$]	0.3	0
	lowest ←			precision of estimating $\overline{\mathbf{Y}}$		→ *highest*

$\mathbf{S}_{ec}$ is the standard deviation for the *averages* of clusters of size L, and $\mathbf{S}$ is the standard deviation for *individual* elements; $\mathbf{S}_{ec}$ is therefore naturally smaller than $\mathbf{S}$ by a factor of (close to) $\sqrt{L}$. The values in Case 4 above apply to this situation, which provides a reference point for the other five cases. To the *right* in Cases 5 and 6, the clusters are *heterogeneous* enough (with *high* values of $\overline{\mathbf{S}}$) to make the remaining *inter*cluster variation ($\mathbf{S}_{ec}$) *smaller* than $\mathbf{S}/\sqrt{L}$ and so these two clustered sampling protocols have *higher* precision than Case 4. To the *left* in Cases 1, 2 and 3, the clusters are *homogeneous* enough (with *low* values of $\overline{\mathbf{S}}$) to make the remaining *inter*cluster variation ($\mathbf{S}_{ec}$) *larger* than $\mathbf{S}/\sqrt{L}$ and so these three clustered sampling protocol have *lower* precision.

As well as illustrating the discussion of efficiency on the facing page 2.108, *looking* at the clusters in the five cases shows us numerical illustrations of what is meant by cluster homogeneity and cluster heterogeneity.

Example 2.14.4: In a student residence, there are separate kitchen and bathroom facilities for every six residents; in total, the residence contains 65 groups of six people who share such facilities. During midterm week, a sample survey is taken to estimate the proportion of residents who have had a cold in the preceding two weeks; the investigation Plan involves collecting this information (with a clear definition of what is meant by a cold) from 13 of the groups chosen by EPS. The data (as the number of people in each selected group who had a cold during the two-week period) were as follows:

6, 0, 2, 1, 0, 4, 5, 3, 0, 5, 1, 3, 0.

(a) Find an approximate 99% confidence interval for $\mathbf{P}$, the proportion of residents who had a cold during the two-week period.

(b) Compare and contrast the answer obtained in (a) with an approximate 99% confidence interval based on the same data but assuming they came from 78 residents chosen *individually* by EPS.

Solution: (a) This question involves finding a CI for a proportion under EPS of equal-sized clusters ($L = 6$); we convert the *counted* data in the question to *proportions* which we treat as *measured* data in the calculations.

Table 2.14.5:

														Sum	Sum sq.
Cluster frequency	6	0	2	1	0	4	5	3	0	5	1	3	0	30	126
Cluster proportion (average)	1	0	1/3	1/6	0	2/3	5/6	1/2	0	5/6	1/6	1/2	0	5	3.5

We have: $\mathbf{N} = 390$, $n = 78$, $\mathbf{M} = 65$, $m = 13$, $\overline{y}_{ec} \equiv p_{ec} = 5/13 \simeq 0.3846$ $(= \overline{y}_{ec})$, $_{.01}t_{12}^* = 3.05454$ for 99% confidence,

$$s_{ec}^2 = \frac{3.5 - 5^2/13}{12} = 0.362\,505\,5265^2 = 0.131\,410\,256 \; (= s_{ec}^2),$$

$z_{.01}^* = 2.57583$ for 99% confidence.

Then: $$\hat{s.d.}(\overline{Y}_{ec}) = s_{ec}\sqrt{\frac{1}{m} - \frac{1}{\mathbf{M}}} = 0.362\,5055\sqrt{\frac{1}{13} - \frac{1}{65}} = 0.089\,926\,553.$$

Hence, an approximate 99% confidence interval for $\overline{\mathbf{Y}} \equiv \mathbf{P}$, the respondent population proportion of residents who had a cold during the two-week period, is:

$$\overline{y}_{ec} \pm 3.05454 \times \hat{s.d.}(\overline{Y}_{ec}) = 0.3846 \pm 3.05454 \times 0.08993 \implies (0.1099,\ 0.6593) \text{ or about } (11\%,\ 66\%);$$

OR: $$\overline{y}_{ec} \pm 2.57583 \times \hat{s.d.}(\overline{Y}_{ec}) = 0.3846 \pm 2.57583 \times 0.08993 \implies (0.1530,\ 0.6163) \text{ or about } (15\%,\ 62\%).$$

(*continued overleaf*)

Solution: (cont.) (a) The confidence interval can be calculated using the *t* distribution (based on our treatment of the proportions as *measured* responses) or using the *normal* distribution (going back to the original *counted* data); either CI is acceptable but the second is more useful to compare with (b) below. Their substantial widths (around 50 percentage points) means that their difference is likely not practically important.

Solution: (b) This part of the question involves finding a CI for a proportion under EPS of *individual* elements using the same data as in (a); we use equation (2.10.6) on page 2.81 in Figure 2.10. Our main interest is to compare the answers from (a) and (b), to see the effect of clustering on precision.

We have: $\mathbf{N} = 390$, $\mathrm{n} = 78$, $\mathrm{p} = 5/13 \simeq 0.3846\ (= p)$, $z^*_{.01} = 2.57583$ for 99% confidence,

so that: $$\hat{s.d.}(P) = \sqrt{\frac{\mathrm{n}}{\mathrm{n}-1}pq\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)} = \sqrt{\frac{78}{77}(5/13 \times 8/13)\left(\frac{1}{78} - \frac{1}{390}\right)} = 0.049\,589\,105.$$

Hence, an approximate 99% confidence interval for $\mathbf{P}$, the respondent population proportion of residents who had a cold during the two-week period, is:

$p \pm 2.57583 \times \hat{s.d.}(P) = 0.3846 \pm 2.57582 \times 0.049\,5891 \implies (0.2569,\ 0.5123)$ or about (25%, 52%).

We see that the width (27 percentage points) of the CI in (b), where the clustering is ignored, is much *smaller* that the width (47 percentage points) in (a); thus, the clustered sampling protocol gives appreciably *lower* precision.

A possible explanation is that colds are *infectious* and so are more likely to spread with*in* the groups of people who share kitchen and bathroom facilities than to spread *among* these groups of 6 people. A binomial model with $\mathrm{n} = 6$ and $\pi = 5/13$ gives an *expected* frequency of about 0.71 for 0 colds in 13 trials, compared with a *much* higher *observed* frequency of 4 in the data; similarly, the expected frequency for 5 or 6 colds is about 0.45 compared with 3 observed. This shows how the clusters are *more homogeneous* than the population and so accounts for the lower precision under the clustered sampling protocol.

NOTES: 10. Anticipating results from Part 4 of these Course Materials, there is a noteworthy contrast between *clustered* sampling protocols (*i.e.*, EPS of units which are clusters of elements from an *un*stratified respondent population) and *stratified* sampling protocols (*i.e.*, EPS of individual elements from a stratified respondent population) – increased precision of estimators is favoured by *heterogeneity of clusters* but by *homogeneity of strata*.

11. This Figure 2.14 deals with EPS of clusters of *equal* size; the more complicated theory of *un*equal-sized clusters is presented in Figure 2.16 of the Course Materials. However, ideas from this Figure 2.14 (*e.g.*, the advantages of cluster selecting, the effect of intracluster variation on precision) are still applicable.

12. The distinction between *measured* data (represented by *real* numbers) and *counted* data (represented by *integers*) is important in statistics; *proportions* provide a link between the two situations. Some aspects of this matter are involved in Examples 2.14.2 and 2.14.4 of this Figure 2.14; additional aspects are encountered in a discussion of *ratios* in Figure 2.15 – see particularly Note 7 on pages 2.118 and 2.119 – also recall the schema near the top of page 2.88 in Appendix 1 in Figure 2.10.

13. An *essential* part of any data set is *how it was generated*, because the method of *analysis* depends on the method of collection. This matter is illustrated at the right by the schema comparing of EPS of units consisting of:
   - ∗ *individual* elements of the respondent population (the left side);
   - ∗ *clusters* of L elements of the respondent population (the right side).

   The differences among the two sets of expressions reflect the difference between individual element and clustered sampling protocols.

Respondent Population

$\mathbf{N}$ units of 1 element ≡ $\mathbf{M}$ units of L elements ($\mathbf{N} = \mathbf{M}$L)

n of $\mathbf{N}$ — **EPS** — m of $\mathbf{M}$

(n = mL)

Individual elements	Clusters
$\bar{y} = \frac{1}{\mathrm{n}}\sum_{j=1}^{\mathrm{n}} y_j$	$\bar{y}_{ec} = \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}} \bar{y}_j$
$s.d.(\bar{Y}) = \mathbf{S}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}$	$s.d.(\bar{Y}_{ec}) = \mathbf{S}_{ec}\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}}$
$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \bar{\mathbf{Y}})^2}$	$\mathbf{S}_{ec} = \sqrt{\frac{1}{\mathbf{M}-1}\sum_{i=1}^{\mathbf{M}}(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})^2}$

1 In Example 2.14.1 (on the second side of this Figure 2.14, page 106), find an approximate 95% confidence interval for the average amount spent on food per student in the residence during the specified week, *ignoring* the clustering and assuming the data had been obtained from 32 residents chosen *individually* by EPS. [*Answer*: (\$67.0555, \$79.3095) or about (\$67, \$80).]

- Briefly compare and contrast the answer you obtain with that from the analysis which takes into account the clustered sampling protocol actually used in the sample survey. [In Example 2.14.1: $\Sigma y_j = 2.362$; $\Sigma y_j^2 = 185{,}365$.]

(*continued*)

Figure 2.14. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Equal-Sized Clusters Estimating an Average or a Total (continued 3)

2 In Example 2.14.2 (on the third side of this Figure 2.14, page 2.107), estimate the *total* number of the newspaper's subscribers served by carier routes who own their home, and find an estimate of the standard deviation of the estimator you use, *ignoring* the clustering and assuming the data had been obtained from 400 subscribers selected *individually* by EPS.
[*Answer*: 18,408 ≃ 18,400 subscribers with an estimated standard deviation of 988.4 ≃ 990 subscribers.]

- Briefly compare and contrast the answer you obtain with that from the analysis which takes into account the clustered sampling protocol actually used in the sample survey.

3 Enhance your understanding of the relevant concepts for EPS from an unstratified population by verifying an appropriate selection of the 69 numerical values given in Example 2.14.3 on the fifth side of this Figure 2.14, page 2.109. (You may wish to use results given at the right for your calculations.)

- Outline, in point form, the features of equation (2.14.9) [near the middle of page 2.108] that are illustrated by the results of Example 2.14.3 on page 2.109.

$$\overline{\mathbf{Y}}_i = \frac{1}{L}\sum_{k=1}^{L}\mathbf{Y}_{ik} \qquad \text{-----(2.14.10)}$$

$$\mathbf{S}_i^2 = \frac{1}{L-1}\sum_{k=1}^{L}(\mathbf{Y}_{ik}-\overline{\mathbf{Y}}_i)^2 \equiv \frac{1}{L-1}\Big[\sum_{k=1}^{L}\mathbf{Y}_{ik}^2 - L\overline{\mathbf{Y}}_i^2\Big] \qquad \text{-----(2.14.11)}$$

4 Two protocols are defined for selecting a sample of n = 2 elements from a respondent population of $\mathbf{N}$ = 6 elements:
∗ EPS of two *individual* elements; ∗ EPS of one of three *clusters* of size L = 2.
For each protocol, list *all* the samples that can be selected.

- For each protocol, find the probability that any respondent population element is selected for the sample.
  - Use your answers to the previous parts of the question to explain briefly why it is *wrong* to say that EPS of individual elements is the selecting process under which *all elements of the respondent population are equally likely to be selected for the sample.*
    - ○ Compare your answer with the discussion of EPS and systematic selecting on the upper half of page 2.39 in Figure 2.3.

5 In Example 2.14.2 on page 107 of this Figure 2.14, suppose the data obtained had been such as to yield the following summary:

Table 2.14.6:

												Total
Cluster total	0	1	2	3	4	5	6	7	8	9	10	
Cluster size	10	10	10	10	10	10	10	10	10	10	10	--
Cluster average ($\overline{y}_i$)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	--
Frequency observed	0	0	0	7	11	12	10	0	0	0	0	40
$\sum_{j=1}^{m}\overline{y}_j$	0	0	0	2.1	4.4	6.0	6.0	0	0	0	0	18.5
$\sum_{j=1}^{m}\overline{y}_j^2$	0	0	0	0.63	1.76	3.0	3.6	0	0	0	0	8.99

- Estimate the *total* number of the newspaper's subscribers served by carrier routes who own their home, and find an estimate of the standard deviation of the estimator you use.
[*Answer*: 18,408 ≃ 18,400 subscribers with an estimated standard deviation of 660.3 ≃ 660 subscribers.]
- *Ignoring* the clustering and assuming the (revised) data had been obtained from 400 subscribers selected *individually* by EPS, estimate the *total* number of the newspaper's subscribers served by carrier routes who own their home, and find an estimate of the standard deviation of the estimator you use.
- By calculating estimates of $\overline{\mathbf{S}}$ (the average intracluster standard deviation) and $\mathbf{S}$ (the repondent population standard deviation), show that your answers to the previous two parts of the question are in accordance with equation (2.14.9) [near the middle of page 2.108 in this Figure 2.14].
  - Outline the essential *difference(s)* between the two data sets [Tables 2.14.4 and 2.14.6, (on page 2.107 of this Figure 2.14 and above)] for Example 2.14.2 that lead to the substantial differences in the *precision* of the two answers.

3. Appendix 1: Derivation of Equation (2.14.9)

To derive equation (2.14.9) near the middle of page 2.108 of this Figure 2.14, we first develop a result based on partitioning the sum of squares represented by $(\mathbf{N}-1)\mathbf{S}^2 \equiv (\mathbf{M}L-1)\mathbf{S}^2$ – recall equation (2.14.4) at the bottom of page 2.105; this partitioning illustrates part of a process that is usually (unhelpfully) called *analysis of variance.*

We have:

$$(\mathbf{M}L-1)\mathbf{S}^2 = \sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i-\overline{\mathbf{Y}})^2 = \sum_{i=1}^{\mathbf{M}}\sum_{k=1}^{L}(\mathbf{Y}_{ik}-\overline{\mathbf{Y}})^2 = \sum_{i=1}^{\mathbf{M}}\sum_{k=1}^{L}(\mathbf{Y}_{ik}-\overline{\mathbf{Y}}_i+\overline{\mathbf{Y}}_i-\overline{\mathbf{Y}})^2$$

$$= \sum_{i=1}^{\mathbf{M}}\sum_{k=1}^{L}(\mathbf{Y}_{ik}-\overline{\mathbf{Y}}_i)^2 + \sum_{i=1}^{\mathbf{M}}\sum_{k=1}^{L}(\overline{\mathbf{Y}}_i-\overline{\mathbf{Y}})^2 + 2\sum_{i=1}^{\mathbf{M}}(\overline{\mathbf{Y}}_i-\overline{\mathbf{Y}})\sum_{k=1}^{L}(\mathbf{Y}_{ik}-\overline{\mathbf{Y}}_i)$$

$$= \sum_{i=1}^{\mathbf{M}}(L-1)\mathbf{S}_i^2 + \sum_{i=1}^{\mathbf{M}}L(\overline{\mathbf{Y}}_i-\overline{\mathbf{Y}})^2 \qquad \text{[using equation (2.14.11) above; also: } \sum_{k=1}^{L}(\mathbf{Y}_{ik}-\overline{\mathbf{Y}}_i) \equiv 0]$$

$$= \mathbf{M}(L-1)\overline{\mathbf{S}}^2 + L\sum_{i=1}^{\mathbf{M}}(\overline{\mathbf{Y}}_i-\overline{\mathbf{Y}})^2 \qquad \text{[where: } \overline{\mathbf{S}}^2 = \frac{1}{\mathbf{M}}\sum_{i=1}^{\mathbf{M}}\mathbf{S}_i^2 \text{ – as in equation (2.14.9)] on page 2.108],}$$

(*continued overleaf*)

3. Appendix 1: Derivation of Equation (2.14.9) (continued)

so that: $\mathrm{L}\sum_{i=1}^{\mathbf{M}}(\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}})^2 \equiv \mathrm{L}(\mathbf{M}-1)\mathbf{S}_{ec}^2 = (\mathbf{M}\mathrm{L}-1)\mathbf{S}^2 - \mathbf{M}(\mathrm{L}-1)\overline{\mathbf{S}}^2$ [using equation (2.14.3) on page 2.105]. -----(2.14.12)

Then, remembering that a *positive* difference of the (squared) standard deviations of two estimators means that the *second* estimator is *more* precise, we have:

$$[s.d.(\overline{Y}_{ec})]^2 - [s.d.(\overline{Y}_{in})]^2 = \left(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}}\right)\frac{1}{\mathbf{M}-1}\sum_{i=1}^{\mathbf{M}}(\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}})^2 - \left(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}}\right)\frac{1}{\mathrm{L}}\mathbf{S}^2 \qquad \text{[because: } \mathrm{n} = \mathrm{mL},\ \mathbf{N} = \mathbf{M}\mathrm{L}]$$

$$= \left(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}}\right)\frac{1}{\mathrm{L}(\mathbf{M}-1)}\left[\mathrm{L}\sum_{i=1}^{\mathbf{M}}(\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}})^2 - (\mathbf{M}-1)\mathbf{S}^2\right]$$

$$= \left(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}}\right)\frac{1}{\mathrm{L}(\mathbf{M}-1)}\left[(\mathbf{M}\mathrm{L}-1)\mathbf{S}^2 - \mathbf{M}(\mathrm{L}-1)\overline{\mathbf{S}}^2 - (\mathbf{M}-1)\mathbf{S}^2\right] \qquad \text{[using equation (2.14.12) above]}$$

$$= \left(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}}\right)\frac{\mathbf{M}(\mathrm{L}-1)}{\mathrm{L}(\mathbf{M}-1)}[\mathbf{S} + \overline{\mathbf{S}}][\mathbf{S} - \overline{\mathbf{S}}], \qquad \text{which is equation (2.14.9) on page 2.108.}$$

4. Appendix 2: Standard Deviations

Four respondent population (data) standard deviations arise in this Figure 2.14; they are:

- $\mathbf{S}$: the standard deviation of the responses of the $\mathbf{N}$ *individual* elements [equation (2.14.4)],
- $\mathbf{S}_{ec}$: the standard deviation of the *average* responses of the $\mathbf{M}$ *clusters* (the *inter*cluster s.d.) [equation (2.14.3)],
- $\mathbf{S}_i$: the standard deviation of the responses of the L elements with*in* cluster i (the *intra*cluster s.d.) [equation (2.14.11)],
- $\overline{\mathbf{S}}$: the square root of the *average* of the $\mathbf{S}_i^2$s (the *average* intracluster s.d.) [equation (2.14.9)],

Also, rearranging the rightmost three terms of equation (2.14.12) above, we obtain equation: (2.14.13):

$$(\mathbf{M}\mathrm{L}-1)\mathbf{S}^2 = \mathrm{L}(\mathbf{M}-1)\mathbf{S}_{ec}^2 + \mathbf{M}(\mathrm{L}-1)\overline{\mathbf{S}}^2, \qquad \text{-----(2.14.13)}$$

which shows how the (*fixed* amount of) variation among the responses of the elements of the respondent population can be partitioned into *two* components – the variation *among* the cluster averages (quantified by $\mathbf{S}_{ec}$) and the variation with*in* the clusters (quantified by $\overline{\mathbf{S}}$ or the set of $\mathbf{S}_i$s). As Example 2.14.3 on page2.109 reminds us, how much of the overall variation goes into each component depends on how the clusters are formed.

Precision of estimates from a clustered sampling protocol is favoured over that of selecting individual elements by *heterogeneity* of clusters – higher *intra*cluster variation – because the *remaining* variation – the *inter*cluster variation – will be *lower*; we are then selecting from a set of clusters with relatively *similar* averages and so obtain higher precision.

Unfortunately, the factors which favour the formation of clusters (like households) in the real world have a tendency to promote *homogeneity* of clusters, which works *against* precision.

This is what is implied when we say clustered sampling protocols are often used for *convenience* – we can get access to a clustered frame relatively *cheaply* – but at the the 'cost' of lower precision.

A diagrammatic summary of these ideas is shown in the schema at the right:

	Population structure	Variation	Quantified by....
Respondent population of $\mathbf{N}$ $\mathbf{Y}_i$s →	$\mathbf{N}$ individual elements →	among the $\mathbf{N}$ individual elements	$\mathbf{S}$
	$\mathbf{M}$ clusters of size L →	*among* the $\mathbf{M}$ cluster averages	$\mathbf{S}_{ec}$
		within the $\mathbf{M}$ clusters	$\overline{\mathbf{S}}$

5. Appendix 3: Notation – Managing Falsehood

As discussed in Appendix 1 on page 2.44 in Figure 2.3, *we* help maintain the key distinction between the real world and the model by distinguishing an *observed* (roman) y fron a *model* (italic) *y*; this leads us to distinguish the observed sample average and standard deviation, $\overline{\mathrm{y}}$ and s, from the model quantities $\overline{y}$ and s – see Table 2.14.7 at the right below. This distinction in turn reminds us of the importance in the Design stage of the FDEAC cycle of developing a Plan that makes it reasonable to treat $\overline{\mathrm{y}}$ as $\overline{y}$ and s as s. Failure to observe the foregoing distinctions makes it easier to overlook the question to be asked about *all* data: *How were they generated?* If this question is *not* asked, it becomes more likely that statistical methods will be applied in situations where their underlying assumptions (like probability selecting, accurate measuring processes, independent measuring, distributional assumptions) are violated, leading to answers with such severe limitations that they are likely to embody falsehoods.

- Table 2.14.7 also shows the notation for the (respondent) population *we* use to help maintain the population-sample distinction.
- Table 2.14.7 includes notation for the contexts of Figures 2.10 and 2.14 (as well as for 2.3) – see also Table 2.10.9 on page 2.88.

Table 2.14.7:

		Figure 2.3	Figure 2.10	Figure 2.14
Population	Response	$\mathbf{Y}_i$	C_i or ${}^cC_i$	$\overline{\mathbf{Y}}_i$
	Attribute(s)	$\overline{\mathbf{Y}}$, $\mathbf{S}$	$\mathbf{P}$	$\overline{\mathbf{Y}}$, $\mathbf{S}_{ec}$
	Size	$\mathbf{N}$	$\mathbf{N}$	$\mathbf{M}$
Model	Random variable(s)	$\overline{Y}$, S	P	$\overline{Y}_{ec}$, (S_{ec})
	Value(s)	$\overline{y}$, s	p	$\overline{y}_{ec}$, s_{ec}
Sample	Response	y_j	C_j or ${}^cC_j$	$\overline{\mathrm{y}}_j$
	Attribute(s)	$\overline{\mathrm{y}}$, s	p	$\overline{\mathrm{y}}_{ec}$, s_{ec}
	Size	n	n	m

6. ACKNOWLEDGEMENT: Information and examples for this Figure 2.14 were kindly provided by Dr. M.E. Thompson.