Figure 2.16. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Unequal-Sized Clusters Estimating an Average, a Total or a Proportion

In Figure 2.14, we introduced the idea of *equiprobable selecting* (EPS) *of clusters of equal size*, sometimes called (the possibly misleading) [*one-stage simple random*] *cluster sampling with equal-sized clusters*; in this situation, the units (obtained by EPS from the respondent population) are not *individual* population elements but rather are *clusters* of L such elements; the response of *all* elements in each selected cluster is measured. In *this* Figure 2.16, we extend this idea to *un*equal-sized clusters (abbreviated *uc* in the notation shown below), but with two notable differences:

- ✻ the theory of *equal*-sized clusters is based on cluster *averages*, that of *un*equal-sized clusters uses cluster *totals*;
- ✻ under equiprobable selecting, the *un*equal cluster *sizes* must be modelled as *random variables*, which necessitates *bi*variate estimating – hence the involvement of Figure 2.15.

The theory for estimating $\overline{\mathbf{Y}}$ (the respondent population *average*), ${}_{\mathrm{T}}\mathbf{Y}$ (the corresponding *total*) and $\mathbf{P}$ (the respondent population *proportion*) when selecting *un*equal-sized clusters is based on the *bi*variate theory developed in Figure 2.15 for estimating a ratio of averages, but we approach the matter from a different perspective;

- ● in Figure 2.15, we deal with *two* responses $\mathbf{Y}$ and $\mathbf{X}$ which are of *equal* importance – the main distinction we make is that $\mathbf{Y}$ occurs in the *numerator*, and $\mathbf{X}$ in the *denominator*, of the ratio we wish to estimate;
- ● in this Figure 2.16, our primary concern is with some (quantitative or qualitative) response $\mathbf{Y}$; the *cluster size*, represented by $\mathbf{L}$, is of only *incidental* interest but must be considered in the theory.

Table 2.16.1 at the right above compares the notation used in the two situations; expressions for the estimating bias and the standard deviation of R, and for the corresponding estimates, from Figure 2.15 are also given at the right as equations (2.16.1) to (2.16.4); the (model) random variable R from Figure 2.15 represents the sample ratio under EPS. Table 2.16.2 below summarizes the (new) notation for this Figure 2.16.

Table 2.16.1:

Figure 2.15	$\mathbf{R} = \dfrac{\overline{\mathbf{Y}}}{\overline{\mathbf{X}}} = \dfrac{{}_{\mathrm{T}}\mathbf{Y}}{{}_{\mathrm{T}}\mathbf{X}} = \dfrac{\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i}{\sum_{i=1}^{\mathbf{N}}\mathbf{X}_i}$	$\mathbf{R}$	$\mathbf{Y}_i$	$\mathbf{X}_i$	$\overline{\mathbf{Y}}$	$\mathbf{N}$
	↓	↓	↓	↓	↓	↓
Figure 2.16	$\overline{\mathbf{Y}} = \dfrac{\sum_{i=1}^{\mathbf{M}}{}_{\mathrm{T}}\mathbf{Y}_i}{\sum_{i=1}^{\mathbf{M}}\mathbf{L}_i} = \dfrac{\text{sum of cluster totals}}{\text{sum of cluster sizes}}$	$\overline{\mathbf{Y}}$	${}_{\mathrm{T}}\mathbf{Y}_i$	$\mathbf{L}_i$	${}_{\mathrm{T}}\overline{\mathbf{Y}}$	$\mathbf{M}$

$$e.b.(R) \simeq \frac{1}{(\mathbf{N}-1)\overline{\mathbf{X}}^2}\sum_{i=1}^{\mathbf{N}}\mathbf{X}_i(\mathbf{R}\mathbf{X}_i - \mathbf{Y}_i)(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}). \qquad \text{-----(2.16.1)}$$

$$s.d.(R) \simeq \sqrt{\frac{1}{(\mathbf{N}-1)\overline{\mathbf{X}}^2}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \mathbf{R}\mathbf{X}_i)^2(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})} \qquad \text{-----(2.16.2)}$$

$$\hat{e}.b.(R) \simeq \frac{1}{(\mathrm{n}-1)\bar{x}^2}\sum_{j=1}^{\mathrm{n}}x_j(rx_j - y_j)(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) \qquad \text{-----(2.16.3)}$$

$$\hat{s.d.}(R) \simeq \sqrt{\frac{1}{(\mathrm{n}-1)\bar{x}^2}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})} \qquad \text{-----(2.16.4)}$$

Table 2.16.2:

......QUANTITY......	RESPONDENT POPULATION	...............SAMPLE [MODEL]................
Size: units ≡ clusters	$\mathbf{M}$	m
elements	$\mathbf{N} = \sum_{i=1}^{\mathbf{M}}\mathbf{L}_i$	$\mathrm{n} = \sum_{j=1}^{\mathrm{n}}\mathrm{l}_j$
Cluster size	$\mathbf{L}_i \ (i = 1, 2, \ldots\ldots, \mathbf{M})$	$\mathrm{l}_j \ (j = 1, 2, \ldots\ldots, \mathrm{m})$ [r.v. is L_j with value l_j]
Average cluster size	$\overline{\mathbf{L}} = \mathbf{N}/\mathbf{M} = \frac{1}{\mathbf{M}}\sum_{i=1}^{\mathbf{M}}\mathbf{L}_i$	$\bar{\mathrm{l}} = \mathrm{n}/\mathrm{m} = \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}}\mathrm{l}_j$ [r.v. is $\overline{L}$ with value $\bar{l}$]
Response	$\mathbf{Y}_{ik} \ (i = 1, 2, \ldots\ldots, \mathbf{M};\ k = 1, 2, \ldots\ldots, \mathbf{L}_i)$	$\mathrm{y}_{jk} \ (j = 1, 2, \ldots\ldots, \mathrm{m};\ k = 1, 2, \ldots\ldots, l_j)$ [r.v. is Y_{jk} with value y_{jk}]
Cluster average	$\overline{\mathbf{Y}}_i = \frac{1}{\mathbf{L}_i}\sum_{k=1}^{\mathbf{L}_i}\mathbf{Y}_{ik} = \frac{1}{\mathbf{L}_i}{}_{\mathrm{T}}\mathbf{Y}_i$	$\bar{\mathrm{y}}_j = \frac{1}{\mathrm{l}_j}\sum_{k=1}^{l_j}\mathrm{y}_{jk} = \frac{1}{\mathrm{l}_j}{}_{\mathrm{T}}\mathrm{y}_j$ [r.v. is $\overline{Y}_j$ with value $\bar{y}_j$]
Cluster total	${}_{\mathrm{T}}\mathbf{Y}_i = \sum_{k=1}^{\mathbf{L}_i}\mathbf{Y}_{ik} = \mathbf{L}_i\overline{\mathbf{Y}}_i$	${}_{\mathrm{T}}\mathrm{y}_j = \sum_{k=1}^{l_j}\mathrm{y}_{jk} = \mathrm{l}_j\bar{\mathrm{y}}_j$ [r.v. is ${}_{\mathrm{T}}Y_j$ with value ${}_{\mathrm{T}}y_j$]
Average cluster total	${}_{\mathrm{T}}\overline{\mathbf{Y}} = \frac{1}{\mathbf{M}}\sum_{i=1}^{\mathbf{M}}{}_{\mathrm{T}}\mathbf{Y}_i$	${}_{\mathrm{T}}\bar{\mathrm{y}} = \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j$ [r.v. is ${}_{\mathrm{T}}\overline{Y}$ with value ${}_{\mathrm{T}}\bar{y}$]
Average	$\overline{\mathbf{Y}} = \sum_{i=1}^{\mathbf{M}}{}_{\mathrm{T}}\mathbf{Y}_i \Big/ \sum_{i=1}^{\mathbf{M}}\mathbf{L}_i$	$\bar{\mathrm{y}}_{uc} = \sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j \Big/ \sum_{j=1}^{\mathrm{m}}\mathrm{l}_j$ [r.v. is $\overline{Y}_{uc}$ with value $\bar{y}_{uc}$]
Total	${}_{\mathrm{T}}\mathbf{Y} = \sum_{i=1}^{\mathbf{M}}{}_{\mathrm{T}}\mathbf{Y}_i = \mathbf{N}\overline{\mathbf{Y}} = \mathbf{M}\,{}_{\mathrm{T}}\overline{\mathbf{Y}}$	${}_{\mathrm{T}}\mathrm{y} = \mathbf{N}\bar{\mathrm{y}}_{uc}$ (if $\mathbf{N}$ is known) ${}_{\mathrm{T}}\mathrm{y} = \mathbf{M}\,{}_{\mathrm{T}}\bar{\mathrm{y}}$ (if $\mathbf{N}$ is *un*known) [r.v. is ${}_{\mathrm{T}}Y$ with value ${}_{\mathrm{T}}y$]

From equations (2.16.1) to (2.16.5) and Table 2.16.1 above, the expressions for EPS of unequal-sized clusters in this Figure 2.16 are equations (2.16.5) to (2.16.8) at the right and overleaf on page 2.126.

$$e.b.(\overline{Y}_{uc}) \simeq \frac{1}{(\mathbf{M}-1)\overline{\mathbf{L}}^2}\sum_{i=1}^{\mathbf{M}}\mathbf{L}_i(\overline{\mathbf{Y}}\mathbf{L}_i - {}_{\mathrm{T}}\mathbf{Y}_i)(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}}) \qquad \text{-----(2.16.5)}$$

$$s.d.(\overline{Y}_{uc}) \simeq \sqrt{\frac{1}{(\mathbf{M}-1)\overline{\mathbf{L}}^2}\sum_{i=1}^{\mathbf{M}}({}_{\mathrm{T}}\mathbf{Y}_i - \overline{\mathbf{Y}}\mathbf{L}_i)^2(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})} \qquad \text{-----(2.16.6)}$$

$$\hat{e}.b.(\overline{Y}_{uc}) \simeq \frac{1}{(\mathrm{m}-1)\bar{l}^2}\sum_{j=1}^{\mathrm{m}}l_j(\bar{y}_{uc}l_j - {}_{\mathrm{T}}y_j)(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}}) = \frac{1}{(\mathrm{m}-1)\bar{l}^2}(\bar{y}_{uc}\sum_{j=1}^{\mathrm{m}}l_j^2 - \sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j\,l_j)(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}}) \qquad \text{-----(2.16.7)}$$

(*continued overleaf*)

$$\hat{s.d.}(\overline{Y}_{uc}) \simeq \sqrt{\frac{1}{(\mathrm{m}-1)\bar{l}^2}\sum_{j=1}^{\mathrm{m}}({}_{\mathrm{T}}y_j - \bar{y}_{uc}l_j)^2(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})} = \sqrt{\frac{\mathrm{m}^2}{(\mathrm{m}-1)\mathrm{n}^2}(\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j^2 - 2\bar{y}_{uc}\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j l_j + \bar{y}_{uc}^2\sum_{j=1}^{\mathrm{m}}l_j^2)(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})} \quad \text{-----(2.16.8)}$$

[See also Note 10 at the bottom of page 2.130]

To estimate ${}_{\mathrm{T}}\mathbf{Y}$ when $\mathbf{N}$ in known, we use:

$${}_{\mathrm{T}}y = \mathbf{N}\bar{y}_{uc} \quad \text{and:} \quad \hat{s.d.}({}_{\mathrm{T}}Y) \simeq \hat{s.d.}(\mathbf{N}\overline{Y}_{uc}) = \mathbf{N}\times\hat{s.d.}(\overline{Y}_{uc}) \quad \text{-----(2.16.9)}$$

To estimate ${}_{\mathrm{T}}\mathbf{Y}$ when $\mathbf{N}$ in *un*known, we use:

$${}_{\mathrm{T}}y = \mathbf{M}_{\mathrm{T}}\bar{y} = \frac{\mathbf{M}}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j \quad \text{-----(2.16.10)}$$

$$\hat{s.d.}(\mathbf{M}_{\mathrm{T}}\overline{Y}) = \sqrt{\frac{\mathbf{M}^2}{\mathrm{m}-1}\sum_{j=1}^{\mathrm{m}}({}_{\mathrm{T}}y_j - {}_{\mathrm{T}}\bar{y})^2(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})} = \sqrt{\frac{\mathbf{M}^2}{\mathrm{m}-1}[\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j^2 - (\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j)^2/\mathrm{m}](\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})} \quad \text{-----(2.16.11)}$$

NOTES: 1. Figure 2.15 is concerned with the ratio (or quotient) of the average of *two* variates; its focus is *bi*variate estimating. When the theory developed there is used in *this* Figure 2.16 to deal with unequal-sized clusters, the different context brings in statistical issues that are easily overlooked.

- The focus is *uni*variate – estimating the average or total of response variate $\mathbf{Y}$; bivariate *estimating* is involved because the fixed number (m) of clusters selected results in a varying number (n) of *elements* in the sample.
  - This is our first encounter in these STAT 332 Course Materials of the sample size not being a fixed quantity.
- When estimating the population total ${}_{\mathrm{T}}\mathbf{Y}$, the two situations where the population (element) size $\mathbf{N}$ is known or *un*known involve *different* theory – the latter situation (surprisingly) involves the *uni*variate theory of Section 5 on pages 2.40 and 2.41 in Figure 2.3 to derive equation (2.16.11) above.
  - This contrasts with the *bi*variate theory from Figure 2.15 needed for equation (2.16.9).
  - Because $E({}_{\mathrm{T}}\overline{Y}) = \sum_{\mathrm{i}=1}^{\mathbf{M}}\mathbf{L}_{\mathrm{i}}\overline{\mathbf{Y}}_{\mathrm{i}}\bullet\frac{1}{\mathbf{M}} = {}_{\mathrm{T}}\overline{\mathbf{Y}}$, $\mathbf{M}_{\mathrm{T}}\overline{Y}$ is an *un*biased estimator of ${}_{\mathrm{T}}\mathbf{Y}$ (*un*like $\mathbf{N}\overline{Y}_{uc}$). -----(2.16.12)
- Two situations for estimating ${}_{\mathrm{T}}\mathbf{Y}$ implies (the often ignored) two situations for estimating $\overline{\mathbf{Y}}$.
  - When $\mathbf{N}$ in *un*known, the relevant estimator is $\overline{Y}_{uc}$ (usually treated as the default situation).
  - When $\mathbf{N}$ in known, the relevant estimator is $(\mathbf{M}/\mathbf{N})_{\mathrm{T}}\overline{Y}$; as pursued in Note 3 and Table 2.16.4 at the bottom of page 2.128, its *sample* version, $(\mathrm{m}/\mathrm{n})_{\mathrm{T}}\overline{Y}$, is $\overline{Y}_{uc}$ because it has numerator $\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}Y_j$ and denominator $\sum_{j=1}^{\mathrm{m}}L_j$.

  The seeming reversal of the $\mathbf{N}$ known and unknown situations when estimating $\overline{\mathbf{Y}}$ and ${}_{\mathrm{T}}\mathbf{Y}$ arises from fortuitous cancellation of $\mathbf{N}$ in numerator and denominator in the $\mathbf{N}$-unknown estimator for ${}_{\mathrm{T}}\mathbf{Y}$.
- From Figure 2.15, we recall the following matters about bivariate estimating.
  - The danger of *estimating bias* is the limitation it imposes on answers by making the *actual* coverage probability (or confidence level) of a confidence interval for a population attribute *lower* than its *stated* coverage probability (or confidence level); this limitation is made less severe by a stronger relationship between the two variates – here, the cluster size being a (good) predictor of the cluster total response; we can check the strength of such a relationship with an appropriate scatter diagram of the sample data.
    - ○ A second check on the limitation imposed on an answer by estimating bias is an (estimated) value below 0.1 (or 10%) for the coefficient of variation of the denominator variate average – here, the average cluster size.
  - Decreased *imprecison* of *bi*variate estimating is likewise favoured by a relationship between the two variates – this again involves visual assessment of the scatter diagram of sample data.
    - ○ Using the relationship in equation (2.16.13) at the right and the notation in equations (2.16.14) and (2.16.15) [which are like equations (2.15.2) and (2.15.4) on page 2.113 in Figure 2.15], reasoning as for equation (2.15.11) near the top of page 2.115 in Figure 2.15, we can write the standard deviations of the bi- and univariate estimators of ${}_{\mathrm{T}}\mathbf{Y}$ as the respective equations (2.16.16) and (2.16.17) at the right.

$$\mathbf{M} = \mathbf{N}/\overline{\mathbf{L}} \quad \text{-----(2.16.13)}$$

$$\mathbf{S}_{\mathbf{V}} = \sqrt{\frac{1}{\mathbf{M}-1}\sum_{\mathrm{i}=1}^{\mathbf{M}}(\mathbf{V}_{\mathrm{i}} - \overline{\mathbf{V}})^2} \quad \text{-----(2.16.14)}$$

$$\mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}\mathbf{L}} = \frac{1}{\mathbf{M}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}({}_{\mathrm{T}}\mathbf{Y}_{\mathrm{i}} - {}_{\mathrm{T}}\overline{\mathbf{Y}})(\mathbf{L}_{\mathrm{i}} - \overline{\mathbf{L}}) \quad \text{-----(2.16.15)}$$

$$s.d.(\mathbf{N}\overline{Y}_{uc}) = \sqrt{\frac{\mathbf{N}^2}{(\mathbf{M}-1)\overline{\mathbf{L}}^2}\left[\mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}}^2 - 2{}_{\mathrm{T}}\overline{\mathbf{Y}}\mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}\mathbf{L}} + {}_{\mathrm{T}}\overline{\mathbf{Y}}^2\mathbf{S}_{\mathbf{L}}^2\right](\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})} \quad \text{-----(2.16.16)}$$

$$s.d.(\mathbf{M}_{\mathrm{T}}\overline{Y}) = \sqrt{\frac{\mathbf{M}^2}{\mathbf{M}-1}\mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}}^2(\tfrac{1}{\mathrm{m}} - \tfrac{1}{\mathbf{M}})} \quad \text{-----(2.16.17)}$$

Comparing the terms in these two equations, we

$${}_{\mathrm{T}}\overline{\mathbf{Y}}^2\mathbf{S}_{\mathbf{L}}^2 < 2{}_{\mathrm{T}}\overline{\mathbf{Y}}\mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}\mathbf{L}} \quad \text{OR:} \quad {}_{\mathrm{T}}\overline{\mathbf{Y}} < 2\mathbf{B}_1 \quad \text{OR:} \quad {}_{\mathrm{T}}\overline{\mathbf{Y}}\mathbf{S}_{\mathbf{L}}/\mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}} < 2\rho_{{}_{\mathrm{T}}\mathbf{Y}\mathbf{L}} \quad \text{-----(2.16.18)}$$

see that $s.d.(\mathbf{N}\overline{Y}_{uc})$ will be *smaller* than $s.d.(\mathbf{M}_{\mathrm{T}}\overline{Y})$ (*i.e.*, the *b*ivariate estimator will be *more* precise) if the three equivalent inequalities (2.16.18) hold, where $\mathbf{B}_1 = \mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}\mathbf{X}}/\mathbf{S}_{\mathbf{L}}^2$ and $\rho_{{}_{\mathrm{T}}\mathbf{Y}\mathbf{L}} = \mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}\mathbf{X}}/\mathbf{S}_{{}_{\mathrm{T}}\mathbf{Y}}\mathbf{S}_{\mathbf{L}}$; $\mathbf{B}_1$ (represented by β_1 in the model) is the **slope of the least squares regression of cluster total response on cluster size** in the respondent population and $\rho_{\mathbf{Y}\mathbf{X}}$ is their **correlation coefficient** [recall equation (2.15.5) near the bottom of page 2.113 in Figure 2.15]. The (simplest) middle version of the three inequalities in equation (2.16.18) is usually the most convenient to use when comparing the precision of different estimators.

It is noteworthy that the relationship in the respondent population between cluster total and cluster size, as visualized by their (sample) scatter diagram, is involved in assessing:

+ the effect of estimating bias on the *accuracy* of the stated confidence level of a confidence interval used to answer a question about a respondent population attribute, AND:

(*continued*)

Figure 2.16. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Unequal-Sized Clusters Estimating an Average, a Totasl or a Proportioon (continued 1)

NOTES: (cont.) 1. + the relative *precision* of bi- and unievariate estimating of a respondent population attribute.

However, the focus of the two assessments differs: for estimating bias, concern is with scatter about the line (quantified, for instance, by *correlation*) whereas, for precision, the simplest concern [in equation (2.16.18)] is with *slope*.

The foregoing discussion of the second matter can be thought of as supplementing that of precision of selecting individual elements or clusters in Section 2 on pages 2.108 to 2.111 in Figure 2.14.

2. To estimate $\overline{\mathbf{Y}}$, we might think of using the average of the cluster averages, $\overline{\overline{y}}$ of equation (2.16.19) at the righr, but we note the following:

$$\overline{\overline{y}} = \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}}\overline{y}_j \qquad \text{-----(2.16.19)}$$

- *Un*like $\overline{y}_{uc}$, $\overline{\overline{y}}$ is not exactly correct when m = $\mathbf{M}$ unless the cluster sizes are all equal (all the $\mathbf{L}_\mathrm{i} = \mathbf{L}$).
- Under EPS, the random variable $\overline{\overline{Y}}$ is an unbiased estimator, *not* of $\overline{\mathbf{Y}}$ but of $\frac{1}{\mathbf{M}}\sum_{\mathrm{i}=1}^{\mathbf{M}}\overline{\mathbf{Y}}_\mathrm{i}$, so $\overline{\overline{y}}$ is not recommended for use under EPS.
- $\overline{\overline{Y}}$ *is* an unbiased estimator of $\overline{\mathbf{Y}}$ if clusters are not selected *equi*probably but rather with probabilitiy proportional to their size $\mathbf{L}_\mathrm{i}$, because we then have:

$$E(\overline{Y}_j) = \sum_{\mathrm{i}=1}^{\mathbf{M}}\frac{1}{\mathbf{L}_\mathrm{i}}{}_{\mathrm{T}}\mathbf{Y}_\mathrm{i}\bullet\frac{\mathbf{L}_\mathrm{i}}{\mathbf{N}} = \overline{\mathbf{Y}} \quad \text{so that:} \quad E(\overline{\overline{Y}}) = E[\frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}}\overline{Y}_j] = \frac{1}{\mathrm{m}}E[\sum_{j=1}^{\mathrm{m}}\overline{Y}_j] = \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}}E(\overline{Y}_j) = \frac{1}{\mathrm{m}}\mathrm{m}\overline{\mathbf{Y}} = \overline{\mathbf{Y}}. \qquad \text{-----(2.16.20)}$$

Example 2.16.1: Interviews were conducted in each of 25 city blocks obtained by equiprobable selecting from a city map divided into 415 such blocks; one purpose of the sample survey was to obtain an estimate of the average and the total income for all adult males in the city. The data obtained were as follows – the data on the number renting, in the fourth and eighth lines of Table 2.16.3, are for Example 2.16.2 on page 2.129:

Table 2.16.3:

Block number (j)	1	2	3	4	5	6	7	8	9	10	11	12
Number of adult males (l_j)	8	12	4	5	6	6	7	5	8	3	2	6
Total income \$000 (${}_{\mathrm{T}}\mathrm{y}_j$)	96	121	42	65	52	40	75	65	45	50	85	43
Number renting (${}_{\mathrm{T}}\mathrm{v}_j$)	4	7	1	3	3	4	4	2	3	2	1	3

Block number (j)	13	14	15	16	17	18	19	20	21	22	23	24	25
Number of adult males (l_j)	5	10	9	3	6	5	5	4	6	8	7	3	8
Total income \$000 (${}_{\mathrm{T}}\mathrm{y}_j$)	54	49	53	50	32	22	45	37	51	30	39	47	41
Number renting (${}_{\mathrm{T}}\mathrm{v}_j$)	2	5	4	1	4	2	3	1	3	3	4	0	3

for these data: $\sum_{j=1}^{\mathrm{m}}\mathrm{l}_j = 151$, $\sum_{j=1}^{\mathrm{m}}\mathrm{l}_j^2 = 1{,}047$; $\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j = 1{,}329{,}000$, $\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j^2 = 82{,}039{,}000{,}000$; $\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j\mathrm{l}_j = 8{,}403{,}000$, $\mathrm{m} = 25$;

Find an approximate 95% confidence interval for: $\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{v}_j = 72$, $\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{v}_j^2 = 262$, $\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{v}_j\mathrm{l}_j = 511$; $\mathbf{M} = 415$.

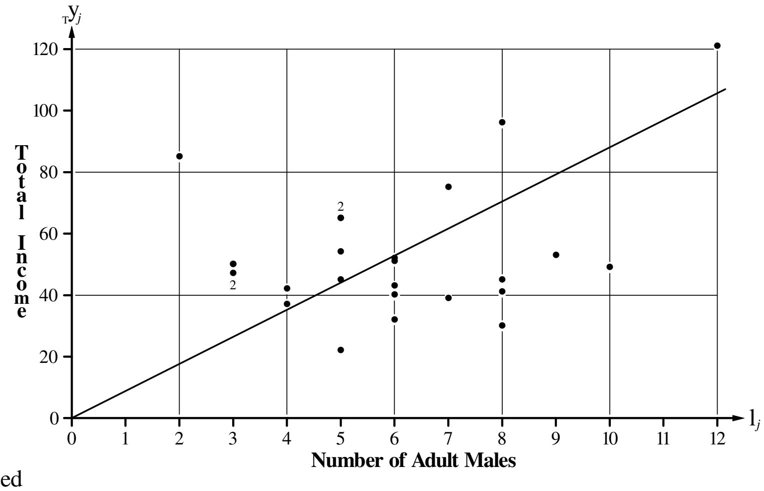
(a) the *average* income per adult male in the city;

(b) the *total* income of all adult males in the city if it is known that $\mathbf{N} = 2{,}500$;

(c) the *total* income of all adult males in the city if $\mathbf{N}$ *un*known.

Solution: (a) This question involves interval estimating of $\overline{\mathbf{Y}}$; we first assess the likely effect of the estimating bias of $\overline{Y}_{uc}$ on the accuracy of the *stated* (here, 95%) confidence level of the confidence interval for $\overline{\mathbf{Y}}$.

Scatter Diagram of Total Income vs. Number of Adult Males in 25 Blocks



The scatter diagram of the sample ${}_{\mathrm{T}}\mathrm{y}_j$s and l_js is shown at the right; we see that the points show *appreciable* scatter about the line with slope $\overline{y}_{uc}$ (estimating $\overline{\mathbf{Y}}$) through the origin. Hence, we do *not* seem to have particularly favourable conditions in this question for lower estimating bias and higher precision with respect to the relationship between ${}_{\mathrm{T}}\mathbf{Y}_\mathrm{i}$ and $\mathbf{L}_\mathrm{i}$. However, the sample of 25 clusters, with an average size about 6 adult males and involving a total of of 151 males, is of *moderate* size and so it will probably *not* appreciably increase the severity of the limitation imposed on the answer by model error (arising from estimating bias) and sample error (arising from imprecision).

We can also do the check involving *c.v.*$(\overline{L})$.

(*continued overleaf*)

Solution: (cont.) (a) We have: $\mathbf{M} = 415.$ $\mathrm{m} = 25,$ $\bar{\mathrm{l}} = 6.04\ (= \bar{l}\,),$ $\mathrm{s}_{\mathrm{l}} = \sqrt{\frac{1}{\mathrm{m}-1}\sum_{j=1}^{\mathrm{m}}(\mathrm{l}_j - \bar{\mathrm{l}})^2} = 2.371\,356\,855\ (= s_l),$

so that: $\widehat{s.d.}(\bar{L}) = s_l\sqrt{\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}} = 0.459\,764\,214,$

and: $\widehat{c.v.}(\bar{L}) = 0.459\,764\,214/6.04 = 0.076\,119\,902 \simeq 0.076 = 7.6\%.$

This estimate is *below* the conventional upper bound of 10% for the magnitude of the ratio $e.b.(\bar{Y}_{uc}]/s.d.(\bar{Y}_{uc}]$ so the *actual* confidence level of the confidence interval for $\bar{\mathbf{Y}}$ should not be too far below its stated level. This implies that the component of model error arising from the estimating bias of $\bar{Y}_{uc}$ should not impose undue limitation of the answer in this question context, so we now calculate the 95% confidence interval for $\bar{\mathbf{Y}}$.

We have: $\bar{\mathrm{y}}_{uc} = \sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j \Big/ \sum_{j=1}^{\mathrm{m}}\mathrm{l}_j = 1{,}329.000/151 = 8{,}801.32\ (= \bar{y}_{uc}),$ $\quad {}_{.05}t_{24}^{*} = 2.06\,390$ for 95% confidence;

also: $\sum_{j=1}^{\mathrm{m}}({}_{\mathrm{T}}\mathrm{y}_j - \bar{\mathrm{y}}_{uc}\mathrm{l}_j)^2 = \sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j^2 - 2\bar{\mathrm{y}}_{uc}\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j\mathrm{l}_j + \bar{\mathrm{y}}_{uc}^2\sum_{j=1}^{\mathrm{m}}\mathrm{l}_j^2$

$$= 82{,}039{,}000{,}000 - 2\times 8{,}801.3245\times 6{,}403{,}000 + 8{,}801.3245^2\times 1{,}047$$

$$= 163{,}143{,}088{,}720 - 147{,}915{,}059{,}500 = 15{,}228{,}029{,}220\ \Big[= \sum_{j=1}^{\mathrm{m}}({}_{\mathrm{T}}y_j - \bar{y}_{uc}l_j)^2\Big];$$

so that: $\widehat{s.d.}(\bar{Y}_{uc}) \simeq \sqrt{\frac{1}{(\mathrm{m}-1)\bar{l}^{\,2}}\sum_{j=1}^{\mathrm{m}}({}_{\mathrm{T}}y_j - \bar{y}_{uc}l_j)^2\left(\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}\right)} = \sqrt{\frac{1}{24\times 6.04^2}(15{,}228{,}029{,}220)\left(\frac{1}{25} - \frac{1}{415}\right)} = 808.569\,8477.$

Hence, an approximate 95% confidence interval for $\bar{\mathbf{Y}}$. the respondent population average income per adult male in the city. is:

$\bar{\mathrm{y}}_{uc} \pm 2.06390\times\widehat{s.d.}(\bar{Y}_{uc}) = 8{,}801.32 \pm 2.06390\times 808.570 \implies$ (\$7.132.51, \$10,470.13)
or about (\$7,100, \$10,500).

Solution: (b) This question involves estimating a population total when $\mathbf{N}$ is known and so uses results from (a).

We have: $\mathbf{N} = 2{,}500,$ $\bar{\mathrm{y}}_{uc} = 8{,}801.324\,503\ (= \bar{y}_{uc}),$ $\widehat{s.d.}(\bar{Y}_{uc}) = 808.569\,8477;$

so that: $\mathbf{N}\bar{\mathrm{y}}_{uc} = 22{,}003{,}311.26,\ (= \mathbf{N}\bar{y}_{uc}),$ $\widehat{s.d.}(\mathbf{N}\bar{Y}_{uc}) = \mathbf{N}\times\widehat{s.d.}(\bar{Y}_{uc}) \simeq 2{,}021{,}424.62.$

Hence, an approximate 95% confidence interval for ${}_{\mathrm{T}}\mathbf{Y}$, the respondent population total income of all adult males in the city. is:

${}_{\mathrm{T}}y \pm 2.06390\times\widehat{s.d.}(\mathbf{N}\bar{Y}_{uc}) = 22{,}003{,}311.26 \pm 2.06390\times 2{,}021{,}424.62 \implies$ (\$17,831,292.90, \$26,175,329.53)
or, in millions of dollars, about (\$17.8, \$26.2).

Solution: (c) This question involves estimating a population total when $\mathbf{N}$ is *un*known.

We have: $\mathbf{M} = 415,$ ${}_{\mathrm{T}}\bar{\mathrm{y}} = 1{,}329.000/151 = 53{,}160\ (= {}_{\mathrm{T}}\bar{y});$

so that: $\mathbf{M}\,{}_{\mathrm{T}}\bar{\mathrm{y}} = 22{,}061{,}400.00\ (= \mathbf{M}\,{}_{\mathrm{T}}\bar{y}),$

$\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j^2 - \Big(\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}\mathrm{y}_j\Big)^2\Big/\mathrm{m} = 82{,}039{,}000{,}000 - 1{,}329{,}000^2/25 = 11{,}389{,}360{,}000\ \Big[= \sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j^2 - \Big(\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j\Big)^2\Big/\mathrm{m}\Big];$

and: $\widehat{s.d.}(\mathbf{M}_{\mathrm{T}}\bar{Y}) = \sqrt{\frac{\mathbf{M}^2}{\mathrm{m}-1}\Big[\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j^2 - \Big(\sum_{j=1}^{\mathrm{m}}{}_{\mathrm{T}}y_j\Big)^2\Big/\mathrm{m}\Big]\left(\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}\right)} = \sqrt{\frac{415^2}{24}(11{,}389{,}360{,}000)\left(\frac{1}{25} - \frac{1}{415}\right)} = 1{,}752{,}792.018.$

Hence, an approximate 95% confidence interval for ${}_{\mathrm{T}}\mathbf{Y}$, the respondent population total income of all adult males in the city. is:

${}_{\mathrm{T}}y \pm 2.06390\times\widehat{s.d.}(\mathbf{M}_{\mathrm{T}}\bar{Y}) = 22{,}061{,}400.00 \pm 2.06390\times 1{,}752{,}792.02 \implies$ (\$18,443,812.55, \$25,678,987.45)
or, in millions of dollars, about (\$18.4, \$25,7).

NOTE: 3. Table 2.16.4 below shows the values (rounded to the nearest dollar) that produced the confidence intervals which are the three answers in Example 2.16.1 above; a fourth pair of entries (in the second line of the Table) is for the case of estimating the respondent population average $\bar{\mathbf{Y}}$ when $\mathbf{N}$ is *known*.

∗ As can be verified from the entries in Table 2.16.2 on page 2.125 for the average cluster size and the average, $\bar{y}_{uc}$ is the sample analogue of this fourth estimate, with m/n replacing $\mathbf{M}/\mathbf{N}$.

In the second column of Table 2.16.4, we see that, in the case of the data for Example 2.16.1, the cluster-to-element ratios in the sample and the respondent population agree to within about one quarter of one percent.

- As a consequence, the differences between the two (point) estimates of the respondent population average and of its total are likely too small to be practically important – that is, whether or not $\mathbf{N}$ is known has little effect on the values of these estimates in Example 2.16.1.

Table 2.16.4:

	. . RATIO OF SIZES . .	**. . . . . AVERAGE . . . .**		**. . . . . TOTAL . . . .**	
		Estimate	**Estimated S.d.**	**Estimate**	**Estimated S.d.**
$\mathbf{N}$ *un*known	m/n = 25/151 ≃ 0.165 563	$\bar{y}_{uc} = 8{,}801$	$\widehat{s.d.}(\bar{Y}_{uc}) = 809$	$\mathbf{M}\,{}_{\mathrm{T}}\bar{y} = 22{,}061{,}400$	$\widehat{s.d.}(\mathbf{M}_{\mathrm{T}}\bar{Y}) = 1{,}752{,}792$
$\mathbf{N}$ known	$\mathbf{M}/\mathbf{N}$ = 415/2,500 = 0.166	$(\mathbf{M}/\mathbf{N})\,{}_{\mathrm{T}}\bar{y} = 8{,}825$	$\widehat{s.d.}(\mathbf{M}_{\mathrm{T}}\bar{Y}/\mathbf{N}) = 701$	$\mathbf{N}\bar{y}_{uc} = 22{,}003{,}311$	$\widehat{s.d.}(\mathbf{N}\bar{Y}_{uc}) = 2{,}021{,}245$

(*continued*)

Figure 2.16. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Unequal-Sized Clusters Estimating an Average, a Totasl or a Proportioon (continued 2)

NOTE: (cont.) 3. However, the situation is different for the two pairs of estimated standard deviations in Table 2.16.4 – both *uni*variate estimators are *more* precise.

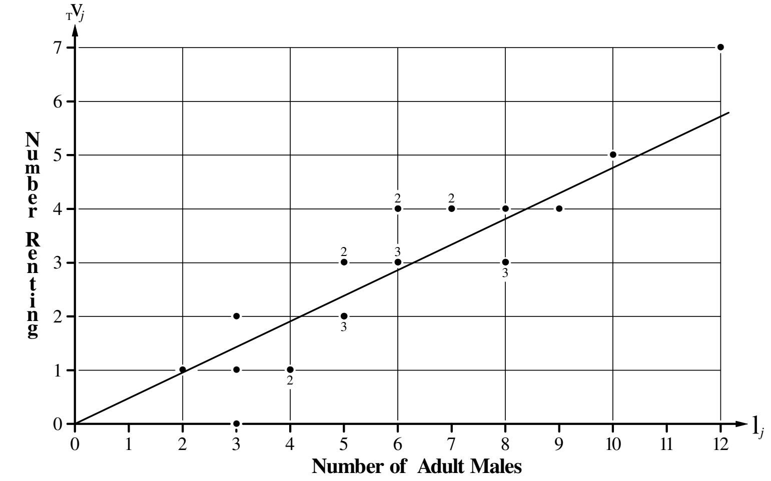
- The situation in Example 2.16.1 is *un*favourable to *bi*variate estimating because cluster size is a poor predictor of cluster total, as reflected in the scatter about the line in the scatter diagram on the lower half of page 2.127.
  - ○ The visual assessment of scatter is (of course) supported by a low (sample) correlation of 0.303 143 78, calculated from $s_{_Ty1} = 15{,}660$, $s_{_Ty} = 21{,}784.321\,58$ and $s_1 = 2.371\,356\,855$.

Example 2.16.2: In the sample survey described in Example 2.16.1, data were also collected on the number of adult males in the selected blocks who were renting their accommodation. Find an approximate 90% confidence interval for the proportion of all adult males in the city who were renting their accommodation. [Using v instead of y in the three data summaries for males who rent (on page 2.127) is only for convenience – we still *think* of v as a response y]

Solution: This question involves *counted* data but we apply the theory of unequal-sized clusters to them as though they were *measured* data. Interval estimating of $\bar{\mathbf{Y}} \equiv \mathbf{P}$ is involved; we first assess the likely effect of the estimating bias of $\overline{Y}_{uc}$ on the accuracy of the *stated* (here, 90%) confidence level of the confidence interval for $\bar{\mathbf{Y}} \equiv \mathbf{P}$.

The scatter diagram of the sample $_Tv_j$s and l_js is shown at the right; compared with Example 2.16.1, we see appreciably *less* scatter of the points about the line with slope $\overline{y}_{uc} \equiv p_{uc}$ (estimating $\bar{\mathbf{Y}} \equiv \mathbf{P}$) through the origin. Hence, we seem to have *favourable* conditions in this situation for lower estimating bias and lower imprecision with respect to the relationship between $\mathbf{Y}$ and $\mathbf{L}$. In addition, the sample of 25 clusters, with an average size of about 6 adult males and involving a total of 151 males, is of *moderate* size and so will likely also *not* appreciably increase the severity of the limitation imposed on the answer by model error (arising from estimating bias) and sample error (arising from imprecision).

Scatter Diagram of Number Renting vs. Number of Adult Males in 25 Blocks



As in Example 2.16.1, we can also (re-)do the check involving $c.v.(\overline{L})$.

We have: $\mathbf{M} = 415.$ $\quad$ m = 25, $\quad \bar{l} = 6.04$ (= $\bar{l}$), $\quad s_l = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(l_j - \bar{l})^2} = 2.371\,356\,855$ (= s_l),

so that: $\quad \hat{s.d.}(\overline{L}) = s_l\sqrt{\frac{1}{m} - \frac{1}{\mathbf{M}}} = 0.459\,764\,214,$

and: $\quad \hat{c.v.}(\overline{L}) = 0.459\,764\,214/6.04 = 0.076\,119\,902 \simeq 0.076 = 7.6\%.$

This estimate is *below* the conventional upper bound of 10% for the magnitude of the ratio $e.b.(\overline{Y}_{uc}]/s.d.(\overline{Y}_{uc}]$ so the *actual* confidence level of the confidence interval for $\bar{\mathbf{Y}} \equiv \mathbf{P}$ should not be too far below its stated level. This implies that the component of model error arising from the estimating bias of $\overline{Y}_{uc}$ should not impose undue limitation of the answer in this question context, so we now calculate the 90% confidence interval for $\bar{\mathbf{Y}} \equiv \mathbf{P}$.

We have: $\quad p_{uc} = \sum_{j=1}^{m} {_Tv_j} \Big/ \sum_{j=1}^{m} l_j = 72/151 = 0.476\,821\,192$ (= p_{uc}), $\qquad {_{.1}t^*_{24}} = 1.64485$ for 90% confidence;

also:
$$\sum_{j=1}^{m}({_Tv_j} - p_{uc}l_j)^2 = \sum_{j=1}^{m}{_Tv_j^2} - 2p_{uc}\sum_{j=1}^{m}{_Tv_j}l_j + p_{uc}^2\sum_{j=1}^{m}l_j^2$$
$$= 262 - 2 \times 0.476\,821\,192 \times 511 + 0.476\,821\,192^2 \times 1{,}047$$
$$= 500.044\,2963 - 487.311\,2583 = 12.733\,0380 \;\Big[= \sum_{j=1}^{m}({_Tv_j} - p_{uc}l_j)^2\Big];$$

so that:
$$\hat{s.d.}(P_{uc}) \simeq \sqrt{\frac{1}{(m-1)\bar{l}^2}\sum_{j=1}^{m}({_Tv_j} - p_{uc}l_j)^2\left(\frac{1}{m} - \frac{1}{\mathbf{M}}\right)} = \sqrt{\frac{1}{24 \times 6.04^2}(12.733\,0380)\left(\frac{1}{25} - \frac{1}{415}\right)} = 0.023\,380\,926.$$

Hence, an approximate 90% confidence interval for $\bar{\mathbf{Y}} \equiv \mathbf{P}$, the respondent population proportion of adult males in the city who were renting their accommodation. is:

$$p_{uc} \pm 1.64485 \times \hat{s.d.}(P_{uc}) = 0.476\,821 \pm 1.64485 \times 0.023\,3809 \implies (0.4383,\ 0{,}5153)$$

or about (43%, 52%).

(*continued overleaf*)

NOTES: 4. The scatter diagram overleaf on page 2.129 for Example 2.16.2 has a (sample) correlation of 0.886 404 129, calculated from $s_{vl} = 3.1716$, $s_v = 1.508\,862\,706$ and $s_l = 2.371\,356\,855$. The high value (much higher than the correlation of about 0.303 in Example 2.16.1) is (of course) consistent with the substantially reduced scatter of the points about the line through the origin in the diagram overleaf (compared with the diagram on page 2.127).

5. Example 2.16.2 formally involves estimating a *proportion* but is actually estimating a *ratio of totals* using the bivariate theory developed in Fiure 2.15; in this regard, it is like Example 2.15.2(a) and assignment question **A4 - 3**. It reminds us of the relationship between proportions and ratios – recall Note 19 on page 2.122 in Figure 2.15.

6. Pursuing the idea from Example 2.16.1(b) in the middle of page 2.128 of knowing that $\mathbf{N} = 2{,}500$, we have:

$$(\mathbf{M}/\mathbf{N})_{\mathrm{T}}\bar{\mathrm{v}} = (415/2{,}500)\times(72/25) = 0.478\,08\ [= (\mathbf{M}/\mathbf{N})_{\mathrm{T}}\bar{v}],$$

$$\sum_{j=1}^{m}{}_{\mathrm{T}}\mathrm{v}_j^2 - \Big(\sum_{j=1}^{m}{}_{\mathrm{T}}\mathrm{v}_j\Big)^2\Big/\mathrm{m} = 262 - 72^2/25 = 54.64\ \Big[= \sum_{j=1}^{m}{}_{\mathrm{T}}v_j^2 - \Big(\sum_{j=1}^{m}{}_{\mathrm{T}}v_j\Big)^2\Big/\mathrm{m}\Big],$$

and:

$$\widehat{s.d.}\Big(\frac{\mathbf{M}}{\mathbf{N}}{}_{\mathrm{T}}\overline{V}\Big) = \sqrt{\frac{\mathbf{M}^2}{(\mathrm{m}-1)\mathbf{N}^2}\Big[\sum_{j=1}^{m}{}_{\mathrm{T}}v_j^2 - \Big(\sum_{j=1}^{m}{}_{\mathrm{T}}v_j\Big)^2\Big/\mathrm{m}\Big]\Big(\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}\Big)} = \sqrt{\frac{415^2}{24\times 2{,}500^2}(54.64)\Big(\frac{1}{25} - \frac{1}{415}\Big)} = 0.048\,561\,943.$$

These values are discussed in Note 8 below.

7. If we *ignore* the clustering in Example 2.16.2 and assume the data on males who were renting their accommodation had been obtained from 151 (*individual*) males obtained from the respondent population by EPS, we have:

$$\mathrm{p}_{in} = 72/151 = 0.476\,821\,192\ (= p_{in}) \qquad \text{[the \textit{same} point estimate as in Example 2.16.2 overleaf on page 2.129],}$$

and, using equation (2.10.6) near the bottom of page 2.81 in Figure 2.10:

$$\widehat{s.d.}(P_{in}) = \sqrt{\frac{\mathrm{n}}{\mathrm{n}-1}p_{in}q_{in}\Big(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\Big)} \qquad [\text{where } q_{in} = 1 - p_{in} = 0.523\,178\,807]$$

$$= \sqrt{\frac{151}{150}(0.476\,821\times 0.523\,179)\Big(\frac{1}{151} - \frac{1}{2{,}500}\Big)} = 0.039\,530\,173.$$

These values, together with those in Note 6 above, are also discussed in Note 8 below.
The calculation of $\widehat{s.d.}(P_{in})$ here involves knowing $\mathbf{N} = 2{,}500$, as is also the case in the preceding Note 6.

8. Table 2.16.5 at the right summarizes relevant values from the three foregoing approaches to estimating a proportion.

Table 2.16.5: Three Approaches to Estimating a Respondent Population Proportion

Context	EPS of ...	Estimating	Estimate	Estimated S.d.	Ratio
Ex. 2.16.2, p. 2.129	Clusters	Bivariate	$\mathrm{p}_{uc} = 0.4768$	$\widehat{s.d.}(P_{uc}) = 0.02338$	---
Note 7, page 2.130	Elements	Univariate	$\mathrm{p}_{in} = 0.4768$	$\widehat{s.d.}(P_{in}) = 0.03953$	1.68
Note 6, page 2.130	Clusters	Univariate	$(\mathbf{M}/\mathbf{N})_{\mathrm{T}}\bar{\mathrm{v}} = 0.4781$	$\widehat{s.d.}(\frac{\mathbf{M}}{\mathbf{N}}{}_{\mathrm{T}}\overline{V}) = 0.04856$	2.08

For the context of the data for Example 2.16.2, we see that:

- the approach has no or little effect on the value for the point estimate,
- the approach has a substantial effect on the precision of the estimate.
  - The *most* precise is bivariate estimating based on the theory of unequal-sized clusters; this likely reflects the favourable conditions shown by the scattter diagram overleaf on page 2.129 – see also Note 4 above.
  - The *least* precise (with an estimated s.d. more than *twice* as large) is the *uni*variate clustered value (which involves a *known* value of $\mathbf{N}$); it reminds us (as discussed in Section 2 on pages 2.108 to 2.110 in Figure 2.14) that the convenience of a Plan using a clustered frame may come at the 'cost' of lower precision.
    - ○ The intermediate precision of the approach which ignores the clustering is consistent with this idea.

Table 2.16.5 reminds us that, throughout these Course Materials, the notation uses lower-case Roman letters (like y, v and p) for data values and the corresponding lower-case *italic* letters (like *y*, *v* and *p*) for the values of model random variables, which are represented by *upper*-case italic letters (like *Y*, *V* and *P*).

∗ It is (of course) investigators' responsibility to implement a Plan for an investigation that makes it reasonable to treat data values as values of random variables, as the additions in brackets ( ) to calculations of sample estimates remind us in the solutions of Examples 2.16.1 and 2.16.2.

9. Equations (2.16.13), (2.16.16) and (2.16.17) (on the lower half of page 2.126) show that, when the clusters are all of *equal* size ($\mathbf{L}$, say) [so ($\mathbf{L}_i - \overline{\mathbf{L}}$) is zero], the two standard deviation estimators $s.d.(\mathbf{N}\overline{Y}_{uc})$ and $s.d.(\mathbf{M}_{\mathrm{T}}\overline{Y})$ are equivalent.

- The point estimators $\mathbf{N}\overline{Y}_{uc}$ and $\mathbf{M}_{\mathrm{T}}\overline{Y}$ are similarly equivalent, because:

$$\mathbf{N}\overline{Y}_{uc} = \mathbf{N}\sum_{j=1}^{m}{}_{\mathrm{T}}Y_j\Big/\mathrm{m}\mathbf{L} = (\mathbf{N}/\mathbf{L})\sum_{j=1}^{m}{}_{\mathrm{T}}Y_j\Big/\mathrm{m} = \mathbf{M}\sum_{j=1}^{m}{}_{\mathrm{T}}\overline{Y}. \qquad \text{-----(2.16.21)}$$

10. Equation (2.16.8) at the top of page 2.126 can also be written as shown at the right.

$$\widehat{s.d.}(\overline{Y}_{uc}) \simeq \sqrt{\frac{1}{(\mathrm{m}-1)\bar{l}^2}\sum_{j=1}^{m}l_j^2(\bar{y}_j - \bar{y}_{uc})^2\Big(\frac{1}{\mathrm{m}} - \frac{1}{\mathbf{M}}\Big)} \qquad \text{-----(2.16.8)}$$