Figure 2.15. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating the Ratio of Two Averages or Totals

Figure 2.3 of these Course Materials develops the theory for estimating $\overline{\mathbf{Y}}$ (the respondent population *average*) and (the closely related theory for estimating) ${}_{\mathrm{T}}\mathbf{Y}$ (the *total*) under equiprobable selecting (EPS); these ideas are adapted in Figure 2.10 to estimating $\mathbf{P}$ and $\mathbf{NP}$ (the respondent population *proportion* and *frequency*). These Figures are concerned with *one* response variate ($\mathbf{Y}$) for population elements. This Figure 2.15 extends Figures 2.3 and 2.10 to estimating $\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}}$, the *ratio* of the *averages* of *two* response variates but we should recognize from the beginning that:

- a ratio of population *totals* is equivalent to the ratio of the corresponding two averages; *i.e.*, we also have: $\mathbf{R} = {}_{\mathrm{T}}\mathbf{Y}/{}_{\mathrm{T}}\mathbf{X}$;
- there is more than one type of ratio, which we must learn to distinguish.

1. Illustrations:

Suppose that n units (or elements) are obtained by EPS from a respondent population of $\mathbf{N}$ units (or elements) to estimate the following attributes:

1. average cigarette consumption per person:

   estimate from sample is $\frac{1}{\mathrm{n}}\sum_{j=1}^{\mathrm{n}} y_j$ (where y_j is the cigarette consumption for person j)

   $= \overline{\mathrm{y}}$ so this case involves only *uni*variate estimating under EPS;

2. average fuel consumption per vehicle for a particular car model:

   estimate from sample is $\frac{1}{\mathrm{n}}\sum_{j=1}^{\mathrm{n}} \frac{y_j}{x_j}$ (where y_j is the amount of fuel used by vehicle j, x_j is the distance covered by vehicle j)

   $= \overline{\mathrm{v}}$ (say) so this case also involves only *uni*variate estimating under EPS;

3. average cigarette consumption per *smoker*:

   estimate from sample is $\dfrac{\sum_{j=1}^{\mathrm{n}} y_j}{\sum_{j=1}^{\mathrm{n}} x_j}$ (where y_j is the cigarette consumption for smoker j) (where x_j is the number of smokers in the sample – *e.g.*, $x_j = 1$ for a smoker, 0 for a non-smoker)

   $= \dfrac{\overline{\mathrm{y}}}{\overline{\mathrm{x}}}$ (dividing both the numerator and the denominator by n).

Our main concern in this Figure 2.15 is with ratios of the type in Illustration 3, where *both* the numerator *and* the denominator can be represented by distinct *random variables*; *we* refer to this case as involving *bi*variate estimating under EPS.

Notation is defined in Table 2.15.1 at the right for the two response variates $\mathbf{Y}$ and $\mathbf{X}$ of each element of the respondent population; the model random variables are the *estimators* of the respondent population attributes and the sample values are their *estimates.* It is (of course) investigators' responsibility to develop and execute a Plan for an investigation that makes it reasonable to take the (sample) *data* $\overline{\mathrm{y}}$ and $\overline{\mathrm{x}}$ as (the *model*) $\overline{y}$ and $\overline{x}$ – recall Appendix 3 at the bottom of page 2.112 in Figure 2.14.

Table 2.15.1: **Quantity**	**Respondent Population**		**Sample** (Estimate) (= model *value*)		**Model** (*Estimator*)	
Average	$\overline{\mathbf{Y}}$	$\overline{\mathbf{X}}$	$\overline{\mathrm{y}}$ (= $\overline{y}$)	$\overline{\mathrm{x}}$ (= $\overline{x}$)	$\overline{Y}$	$\overline{X}$
Total	${}_{\mathrm{T}}\mathbf{Y}$	${}_{\mathrm{T}}\mathbf{X}$	${}_{\mathrm{T}}\mathrm{y} = \mathbf{N}\overline{\mathrm{y}}$ (= ${}_{\mathrm{T}}y$)	${}_{\mathrm{T}}\mathrm{x} = \mathbf{N}\overline{\mathrm{x}}$ (= ${}_{\mathrm{T}}x$)	${}_{\mathrm{T}}Y$	${}_{\mathrm{T}}X$
Ratio	$\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}} \equiv {}_{\mathrm{T}}\mathbf{Y}/{}_{\mathrm{T}}\mathbf{X}$		$\mathrm{r} = \overline{\mathrm{y}}/\overline{\mathrm{x}} \equiv {}_{\mathrm{T}}\mathrm{y}/{}_{\mathrm{T}}\mathrm{x}$ (= r)		R	

2. Statistical Properties of $R = \overline{Y}/\overline{X}$ – Preliminaries

We now derive results for the *estimating bias* and the *standard deviation* of R, a random variable that is the *ratio* of two random variables which each represent an *average.* As a first step, we obtain an expression for the *covariance of* $\overline{Y}$ *and* $\overline{X}$.

We have: $\overline{Y+X} = \overline{Y}+\overline{X}$ so that: $[s.d.(\overline{Y+X})]^2 = [s.d.(\overline{Y}+\overline{X})]^2 = [s.d.(\overline{Y})]^2 + [s.d.(\overline{X})]^2 + 2cov(\overline{Y}, \overline{X}).$ -----(2.15.1)

Then, using the two results at the right for $\mathbf{S}_{\mathbf{V}}$ and $s.d.(\overline{V})$, which are based on Table 2.3.5 and equation (2.3.9) on pages 2.40 and 2.41 of Figure 2.3, we have:

$$\mathbf{S}_{\mathbf{V}} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\mathbf{V}_{\mathrm{i}}-\overline{\mathbf{V}})^2} \qquad s.d.(\overline{V}) = \mathbf{S}_{\mathbf{V}}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \qquad \text{-----(2.15.2)}$$

$$\mathbf{S}^2_{\mathbf{Y}+\mathbf{X}}(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) = [\mathbf{S}^2_{\mathbf{Y}} + \mathbf{S}^2_{\mathbf{X}} + 2\mathbf{S}_{\mathbf{YX}}](\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) = \mathbf{S}^2_{\mathbf{Y}}(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) + \mathbf{S}^2_{\mathbf{X}}(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) + 2\mathbf{S}_{\mathbf{YX}}(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}); \qquad \text{-----(2.15.3)}$$

comparing the rightmost three terms of equations (2.15.1) and (2.15.3), we obtain equation (2.15.4):

$$cov(\overline{Y}, \overline{X}) = \mathbf{S}_{\mathbf{YX}}(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) \quad \text{where:} \quad \mathbf{S}_{\mathbf{YX}} = \frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\mathbf{Y}_{\mathrm{i}}-\overline{\mathbf{Y}})(\mathbf{X}_{\mathrm{i}}-\overline{\mathbf{X}}). \qquad \text{-----(2.15.4)}$$

We introduce additional terminology in equation (2.15.5): $\rho_{\mathbf{YX}} = \mathbf{S}_{\mathbf{YX}}/\mathbf{S}_{\mathbf{Y}}\mathbf{S}_{\mathbf{X}}$ and: $\mathbf{B}_1 = \mathbf{S}_{\mathbf{YX}}/\mathbf{S}^2_{\mathbf{X}}$, -----(2.15.5)
where $\rho_{\mathbf{YX}}$ is the **correlation coefficient** of responses $\mathbf{Y}$ and $\mathbf{X}$ in the respondent population,
and $\mathbf{B}_1$ (represented by β_1 in the model) is the **slope of the least squares regression of $\mathbf{Y}$ on $\mathbf{X}$** in the respondent population.

We then rewrite the first part of equation (2.15.4) in three equivalent forms in equation (2.15.6):

$$cov(\overline{Y}, \overline{X}) = \mathbf{S}_{\mathbf{YX}}(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) = \mathbf{S}_{\mathbf{Y}}\mathbf{S}_{\mathbf{X}}\rho_{\mathbf{YX}}(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) = \mathbf{B}_1\mathbf{S}^2_{\mathbf{X}}(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}). \qquad \text{-----(2.15.6)}$$

Different presentations of these ideas vary in which of these three equivalent expressions for $cov(\overline{Y}, \overline{X})$ they use; these Course Materials use mainly the *first* one involving $\mathbf{S}_{\mathbf{YX}}$.

(*continued overleaf*)

NOTES: 1. We (reluctantly) use the notation ρ for correlation overleaf on page 2.113 in equations (2.15.5) and (2.15.6) because 'r' is needed for a ratio (as $\mathbf{R}$, r, R and r) in this Figure 2.15 and elsewhere in these Course Materials. Originally, 'c' would have been a better choice to denote correlation but 'r' is now too widespread to change. [Confusion of 'c' for correlation with the speed of light in physics – as in E = mc^2 for instance – would not be an issue.]

- This matter is like the unfortunate current widespread use of P for probability; *our* use of 'Pr' conveniently leaves 'p' for a proportion (as $\mathbf{P}$, p, P and p) – recall Table 2.10.9 on page 2.88 in Figure 2.10.
- When providing background for simple linear regression in these Materials – for example, in Figure 13.3 in STAT 220 and Figure 16.1 in STAT 231 – $\mathbf{R}_\mathrm{i}$ is the **residual** for element i of the study (or respondent) population. This notation seldom arises in these STAT 332 Course Materials and, when it does (*e.g.*, on page 2.138 in Figure 2.17), the advantage of the evocative '**R**' for both a ratio and a residual outweighs the (slight) opportunity for confusion.

2. Background for $\mathbf{B}_1$ and $\rho_{\mathbf{YX}}$ in equations (2.15.5) and (2.15.6) overleaf on page 2.113, and for $\mathbf{R}_\mathrm{i}$, is available in the STAT 231 Course Materials in Figure 16.1 in discussion of equation (16.1.3) on page 16.1 and the upper diagram on page 16.6, and in Figure 13.3 of the STAT 221 Course Materials in equations (13.3.3), (13.3.31) and (13.3.33) on pages 13.17, 13.19 and 13.20.

3. Statistical Properties of $R = \bar{Y}/\bar{X}$ – Estimating Bias

To find an approximate expression for the *estimating bias* of R, denoted here *e.b.*(R), we use a bivariate Taylor series expansion taken to *second* order, written as follows:

$$g(u,v) \simeq g(a,b) + (u-a)\frac{\partial g}{\partial u} + (v-b)\frac{\partial g}{\partial v} + \tfrac{1}{2}(u-a)^2\frac{\partial^2 g}{\partial u^2} + \tfrac{1}{2}(v-b)^2\frac{\partial^2 g}{\partial v^2} + (u-a)(v-b)\frac{\partial^2 g}{\partial u\partial v},$$

where the five partial derivatives are evaluated at (a, b), the point about which $g(u,v)$ is expanded. If $g(u,v) = \frac{u}{v}$, then:

$$\frac{\partial g}{\partial u} = \frac{1}{v}, \qquad \frac{\partial g}{\partial v} = -\frac{u}{v^2}, \qquad \frac{\partial^2 g}{\partial u^2} = 0, \qquad \frac{\partial^2 g}{\partial v^2} = \frac{2u}{v^3}, \qquad \frac{\partial^2 g}{\partial u\partial v} = -\frac{1}{v^2}.$$

We now expand $g(\bar{Y}, \bar{X}) = \frac{\bar{Y}}{\bar{X}}$ about the point ($\bar{\mathbf{Y}}$, $\bar{\mathbf{X}}$); we then take $E(\frac{\bar{Y}}{\bar{X}})$, and find $E(\frac{\bar{Y}}{\bar{X}}) - \mathbf{R}$, which is *e.b.*($R$);

$$\frac{\bar{Y}}{\bar{X}} \simeq \frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}} + (\bar{Y}-\bar{\mathbf{Y}})(\frac{1}{\bar{\mathbf{X}}}) + (\bar{X}-\bar{\mathbf{X}})(-\frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^2}) + \tfrac{1}{2}(\bar{Y}-\bar{\mathbf{Y}})^2\cdot 0 + \tfrac{1}{2}(\bar{X}-\bar{\mathbf{X}})^2(\frac{2\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^3}) + (\bar{Y}-\bar{\mathbf{Y}})(\bar{X}-\bar{\mathbf{X}})(-\frac{1}{\bar{\mathbf{X}}^2}),$$

$$\therefore \quad E(\frac{\bar{Y}}{\bar{X}}) \simeq \frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}} + (\frac{1}{\bar{\mathbf{X}}})E(\bar{Y}-\bar{\mathbf{Y}}) + (-\frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^2})E(\bar{X}-\bar{\mathbf{X}}) + (\frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^3})E(\bar{X}-\bar{\mathbf{X}})^2 + (-\frac{1}{\bar{\mathbf{X}}^2})E[(\bar{Y}-\bar{\mathbf{Y}})(\bar{X}-\bar{\mathbf{X}})]$$

$$= \mathbf{R} + (\frac{1}{\bar{\mathbf{X}}})\cdot 0 + (-\frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^2})\cdot 0 + (\frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^3})[s.d.(\bar{X})]^2 + (-\frac{1}{\bar{\mathbf{X}}^2})cov(\bar{Y}, \bar{X});$$

hence:

$$e.b.(R) \simeq \frac{1}{\bar{\mathbf{X}}^2}[\mathbf{R}\mathbf{S}_{\mathbf{X}}^2 - \mathbf{S}_{\mathbf{YX}}](\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}) \equiv \frac{1}{(\mathbf{N}-1)\bar{\mathbf{X}}^2}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{X}_\mathrm{i}(\mathbf{R}\mathbf{X}_\mathrm{i} - \mathbf{Y}_\mathrm{i})(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}). \qquad \text{-----(2.15.7)}$$

Also, the *relative* estimating bias of R is given by:

$$r.e.b.(R) \equiv \frac{e.b.(R)}{\mathbf{R}} \simeq \frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}\frac{\mathbf{X}_\mathrm{i}}{\bar{\mathbf{X}}}(\frac{\mathbf{X}_\mathrm{i}}{\bar{\mathbf{X}}} - \frac{\mathbf{Y}_\mathrm{i}}{\bar{\mathbf{Y}}})(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}). \qquad \text{-----(2.15.8)}$$

To obtain an approximate expression for the *estimator* of the estimating bias of $R = \bar{Y}/\bar{X}$, we replace the respondent population attributes in equation (2.15.7) by their estimators; the expression for the corresponding *estimate* is:

$$\hat{e}.b.(R) \simeq \frac{1}{\bar{x}^2}(rs_x^2 - s_{yx})(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}) = \frac{1}{(\mathrm{n}-1)\bar{x}^2}\sum_{j=1}^{\mathrm{n}}x_j(rx_j - y_j)(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}) = \frac{1}{(\mathrm{n}-1)\bar{x}^2}(r\sum_{j=1}^{\mathrm{n}}x_j^2 - \sum_{j=1}^{\mathrm{n}}y_jx_j)(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}), \qquad \text{-----(2.15.9)}$$

where:

$$s_v = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(v_j - \bar{v})^2} = \sqrt{\frac{1}{\mathrm{n}-1}(\sum_{j=1}^{\mathrm{n}}v_j^2 - \mathrm{n}\bar{v}^2)} \quad \text{and:} \quad s_{yx} = \frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \bar{y})(x_j - \bar{x}) = \frac{1}{\mathrm{n}-1}(\sum_{j=1}^{\mathrm{n}}y_jx_j - \mathrm{n}\bar{y}\bar{x}). \qquad \text{-----(2.15.10)}$$

We see from equation (2.15.8) that the estimating bias of R as an estimator of $\mathbf{R}$ becomes *smaller* relative to the magnitude of $\mathbf{R}$ as:

∗ n (the sample size) becomes *larger* – the functional *form* of the dependence is n^{-1};

- *de*creasing magnitude of the estimating bias of R with *in*creasing sample size contrasts sharply with other catetegories of (real-world) inaccuracy– studying, non-responding, selecting and measuring – which do *not* get smaller as sample size increases;

∗ the points ($\mathbf{Y}_\mathrm{i}$, $\mathbf{X}_\mathrm{i}$) lie *closer* to the line with slope $\mathbf{R}$ through the origin, which makes the term ($\mathbf{X}_\mathrm{i}/\bar{\mathbf{X}} - \mathbf{Y}_\mathrm{i}/\bar{\mathbf{Y}}$) smaller; the initial $\mathbf{X}_\mathrm{i}/\bar{\mathbf{X}}$ multiplying this term means that departures of the $\mathbf{X}_\mathrm{i}$ from the line make *more* contribution to increasing the estimating bias of R than do departures of the $\mathbf{Y}_\mathrm{i}$, and the larger the magnitude of an $\mathbf{X}_\mathrm{i}$, the greater the effect of its departure from the line.

We see from equation (2.15.7) that the estimating bias of R is *zero* if all the $\mathbf{X}_\mathrm{i}$ are *equal* (when $\mathbf{X}_\mathrm{i} = \bar{\mathbf{X}}$ for all i and $\mathbf{R}\mathbf{X}_\mathrm{i} = \bar{\mathbf{Y}}$).

4. Statistical Properties of $R = \bar{Y}/\bar{X}$ – Standard Deviation

To find an expression for the approximate *standard deviation* of R, we use a bivariate Taylor series expansion to *first* order:

$$\frac{\bar{Y}}{\bar{X}} \simeq \frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}} + (\bar{Y}-\bar{\mathbf{Y}})(\frac{1}{\bar{\mathbf{X}}}) + (\bar{X}-\bar{\mathbf{X}})(-\frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^2});$$

(*continued*)

Figure 2.15. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating the Ratio of Two Averages or Totals (continued 1)

hence: $$s.d.(\frac{\overline{Y}}{\overline{X}}) \simeq \sqrt{0 + (\frac{1}{\bar{\mathbf{X}}})^2[s.d.(\overline{Y})]^2 + (-\frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^2})^2[s.d.(\overline{X})]^2 + 2(\frac{1}{\bar{\mathbf{X}}})(-\frac{\bar{\mathbf{Y}}}{\bar{\mathbf{X}}^2})cov(\overline{Y}, \overline{X})}$$

$$= \sqrt{\frac{1}{\bar{\mathbf{X}}^2}[\mathbf{S}_{\mathbf{Y}}^2 - 2\mathbf{R}\mathbf{S}_{\mathbf{YX}} + \mathbf{R}^2\mathbf{S}_{\mathbf{X}}^2](\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} \equiv \sqrt{\frac{1}{(\mathbf{N}-1)\bar{\mathbf{X}}^2}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\mathbf{Y}_{\mathrm{i}} - \mathbf{R}\mathbf{X}_{\mathrm{i}})^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} \quad \text{[because } \bar{\mathbf{Y}} = \mathbf{R}\bar{\mathbf{X}}\text{].} \qquad \text{-----(2.15.11)}$$

Also, the *coefficient of variation* (*c.v.*) of *R* is: $$c.v.(R) \equiv \frac{s.d.(R)}{\mathbf{R}} \simeq \sqrt{\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\frac{\mathbf{Y}_{\mathrm{i}}}{\bar{\mathbf{Y}}} - \frac{\mathbf{X}_{\mathrm{i}}}{\bar{\mathbf{X}}})^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})}. \qquad \text{-----(2.15.12)}$$

To obtain an approximate expression for the *estimator* of the standard deviation of $R = \overline{Y}/\overline{X}$, we replace the respondent population attributes in (2.15.11) by their estimators; the expression for the corresponding *estimate* is:

$$\left.\begin{aligned} \widehat{s.d.}(R) &\simeq \sqrt{\frac{1}{\bar{x}^2}(s_y^2 - 2rs_{yx} + r^2s_x^2)(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} = \sqrt{\frac{1}{(\mathrm{n}-1)\bar{x}^2}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} \quad \text{[because } \bar{y} = r\bar{x}\text{]} \\ &= \sqrt{\frac{1}{(\mathrm{n}-1)\bar{x}^2}(\sum_{j=1}^{\mathrm{n}}y_j^2 - 2r\sum_{j=1}^{\mathrm{n}}y_jx_j + r^2\sum_{j=1}^{\mathrm{n}}x_j^2)(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})}. \end{aligned}\right\} \qquad \text{-----(2.15.13)}$$

We see from equation (2.15.12) above that the standard deviation of *R* as an estimator of **R** becomes *smaller* relative to the magnitude of **R** as:

∗ n (the sample size) becomes *larger* – this is the familiar increase in precision of estimating with increasing sample size but, in contrast to the n^{-1} of *estimating bias*, the functional form of the dependence is now the usual $\mathrm{n}^{-½}$; *i.e.*, increasing sample size decreases the *estimating bias* of *R faster* than it decreases the *standard deviation* (or increases the *precision*) of *R*.

∗ the points $(\mathbf{Y}_{\mathrm{i}}, \mathbf{X}_{\mathrm{i}})$ lie *closer* to the line with slope **R** through the origin, which makes the term $(\mathbf{Y}_{\mathrm{i}}/\bar{\mathbf{Y}} - \mathbf{X}_{\mathrm{i}}/\bar{\mathbf{X}})$ smaller.

NOTES: 3. Comparing equations (2.15.8) and (2.15.12), we see that the expressions for the relative magnitude of the estimating bias and the standard deviation of *R* are *similar*; this implies that conditions which make one of these quantities smaller (as is desirable statistically) will tend *also* to make the other smaller.

- A further advantage of larger sample sizes is that they improve the accuracy of the Taylor series approximations because, as n becomes larger, the value of $(u, v) = (\bar{y}, \bar{x})$ will usually become *closer* to $(\bar{\mathbf{Y}}, \bar{\mathbf{X}})$, the point about which $g(u, v)$ is expanded; thus, larger values of n tend to improve the accuracy of our approximate expressions for the estimating bias and the standard deviation of *R*.
- The derivation on the facing page 2.114 of the approximate expression (2.15.8) for the relative estimatng bias of *R* shows that it originates in the *last two* (second-order) terms in the Taylor series expansion; this contrasts with equation (2.15.12) for the approximate coefficient of variation which comes from only a *first-order* expansion.

4. The *coefficients of variation* of $\overline{Y}$ and $\overline{X}$, measures of *relative* precision of these two *uni*variate estimators, are shown at the right – $\overline{Y}$ and $\overline{X}$ are the random variables representing the average of the sample y_js (= y_js) and the sample x_js (= x_js) under EPS.

$$c.v.(\overline{Y}) \equiv \frac{s.d.(\overline{Y})}{\bar{\mathbf{Y}}} = \frac{\mathbf{S}_{\mathbf{Y}}}{\bar{\mathbf{Y}}}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\frac{\mathbf{Y}_{\mathrm{i}}}{\bar{\mathbf{Y}}} - 1)^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} \qquad \text{-----(2.15.14)}$$

$$c.v.(\overline{X}) \equiv \frac{s.d.(\overline{X})}{\bar{\mathbf{X}}} = \frac{\mathbf{S}_{\mathbf{X}}}{\bar{\mathbf{X}}}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\frac{\mathbf{X}_{\mathrm{i}}}{\bar{\mathbf{X}}} - 1)^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} \qquad \text{-----(2.15.15)}$$

- Comparing the approximate expression for the coefficient of variation of *R* in equation (2.15.8) [near the middle of the facing page 2.114] with the corresponding (exact) expressions for $\overline{Y}$ and $\overline{X}$ given at the right above, we note that they differ most obviously in the second term inside the brackets of the sigma sign.

5. Using equation (2.15.2) [on page 2.113] for $\mathbf{S}_{\mathbf{Y}}$ and equation (2.15.10) [on page 2.114] for s_v, and by analogy with $\mathbf{S}_{\mathbf{Y}}$ and $\mathbf{S}_{\mathbf{X}}$ [introduced in equation (2.15.3) on page 2.113] and with s_y and s_x [introduced in equation (2.15.13) above], equations (2.15.11) and (2.15.13) suggest defining respondent population and sample measures of variation, $\mathbf{S}_{\mathbf{R}}$ and s_r, as in equations (2.15.16) and (2.15.17) at the right. Equation (2.15.18) is for calculating $\mathrm{s_r}$ from sample data; this can be used as a value for s_r under an appropriate investigation Plan.

$$\mathbf{S}_{\mathbf{R}} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}(\mathbf{Y}_{\mathrm{i}} - \mathbf{R}\mathbf{X}_{\mathrm{i}})^2} \qquad \text{-----(2.15.16)}$$

$$s_r = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2} \qquad \text{-----(2.15.17)}$$

$$\mathrm{s_r} = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(\mathrm{y}_j - \mathrm{r}\mathrm{x}_j)^2} \qquad \text{-----(2.15.18)}$$

- In the real world, the information needed to calculate $\mathbf{S}_{\mathbf{R}}$ would rarely (if ever) be available.

5. Confidence Intervals for R and the Effect of Estimating Bias

The effect of estimating bias of *R* on the accuracy of an answer (*e.g.*, expressed as a confidence interval for **R**) depends on the magnitude of the bias relative to that of the standard deviation of *R*; we deal with this matter as follows:

(*continued overleaf*)

from equation (2.15.7): $e.b.(R) \simeq [c.v.(\overline{X})]^2(\mathbf{R} - \frac{\mathbf{S}_{\mathbf{YX}}}{\mathbf{S}^2_{\mathbf{X}}})$ (based on a *second*-order approximation) -----(2.15.19)

and from equation (2.15.10): $[s.d.(R)]^2 \simeq (\frac{\mathbf{S}_{\mathbf{X}}}{\overline{\mathbf{X}}})^2(\mathbf{R}^2 - 2\mathbf{R}\frac{\mathbf{S}_{\mathbf{YX}}}{\mathbf{S}^2_{\mathbf{X}}} + \frac{\mathbf{S}^2_{\mathbf{Y}}}{\mathbf{S}^2_{\mathbf{X}}})(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})$ (based on a *first*-order approximation)

$$= [c.v.(\overline{X})]^2(\mathbf{R}^2 - 2\mathbf{R}\frac{\mathbf{S}_{\mathbf{YX}}}{\mathbf{S}^2_{\mathbf{X}}} + \frac{\mathbf{S}^2_{\mathbf{Y}}}{\mathbf{S}^2_{\mathbf{X}}}); \qquad \text{-----(2.15.20)}$$

introducing a multiplier of $\rho^2_{\mathbf{YX}} \leq 1$ [from equation (2.15.5)] into the last term on the right-hand side of (2.15.20) gives:

$$[s.d.(R)]^2 \geq [c.v.(\overline{X})]^2(\mathbf{R}^2 - 2\mathbf{R}\frac{\mathbf{S}_{\mathbf{YX}}}{\mathbf{S}^2_{\mathbf{X}}} + \frac{\mathbf{S}^2_{\mathbf{YX}}}{\mathbf{S}^2_{\mathbf{Y}}\mathbf{S}^2_{\mathbf{X}}} \cdot \frac{\mathbf{S}^2_{\mathbf{Y}}}{\mathbf{S}^2_{\mathbf{X}}}) = [c.v.(\overline{X})]^2(\mathbf{R} - \frac{\mathbf{S}_{\mathbf{YX}}}{\mathbf{S}^2_{\mathbf{X}}})^2. \qquad \text{-----(2.15.21)}$$

Dividing equation (2.15.19) by the square root of equation (2.15.21) leads to: $\frac{|e.b.(R)|}{s.d.(R)} \leq c.v.(\overline{X})$ (the *equality* holds when $\rho^2_{\mathbf{YX}} = 1$). -----(2.15.22)

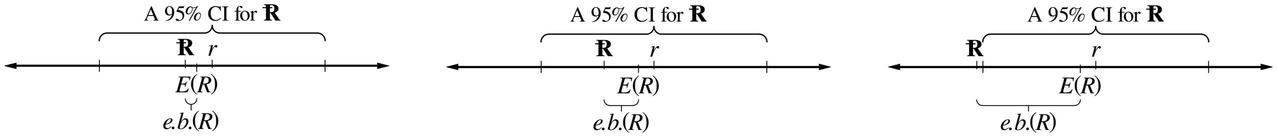
Use of the result (2.15.22) is discussed below, after the confidence interval expressions (2.15.23) for estimating $\mathbf{R}$.

NOTE: 6. The result (2.15.22), which provides an *upper bound* on the magnitude of the estimating bias of R in relation to its standard deviation, appears to be only an *approximate* result based on the respective *second*- and *first*-order approximations for these two quantities; surprisingly, as shown by the alternative derivation in Appendix 2 on page 2.123 of this Figure 2.15, it is actually an *exact* result.

On the basis of the approximate normality of R 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 on the preceding three sides of this Figure 2.15 leads to an interval involving random variables: $I = [R - {}_{\alpha}t^{*}_{\mathrm{n}-1}\,\hat{s.d.}(R),\ R + {}_{\alpha}t^{*}_{\mathrm{n}-1}\,\hat{s.d.}(R)]$, such that $\Pr(I \ni \mathbf{R}) \simeq 100(1-\alpha)\%$, where ${}_{\alpha}t^{*}_{\mathrm{n}-1}$ is the $100(1-\alpha/2)th$ percentile of the $t_{\mathrm{n}-1}$ distribution. For calculating an approximate $100(1-\alpha)\%$ *confidence interval* for $\mathbf{R}$, we use:

$$r \pm {}_{\alpha}t^{*}_{\mathrm{n}-1}\,\hat{s.d.}(R) = [r - {}_{\alpha}t^{*}_{\mathrm{n}-1}\,\hat{s.d.}(R),\ r + {}_{\alpha}t^{*}_{\mathrm{n}-1}\,\hat{s.d.}(R)]. \qquad \text{-----(2.15.23)}$$

The following diagrams illustrate why the ratio of the magnitude of the estimating bias of R to its standard deviation [as expressed in equation (2.15.22) above] is useful for assessing the likely accuracy of the stated confidence *level* (the coverage probability) of a confidence interval for $\mathbf{R}$.



In each diagram, the 95% confidence interval for $\mathbf{R}$ is centred on r, and the horizontal distance between $\mathbf{R}$ and $E(R)$] represents the estimating bias of R; because $E(R)$ is to the *right* of $\mathbf{R}$, the estimating bias is shown as being *positive* in these illustrations and, for graphical convenience, the r value at the centre of the confidence interval is likewise to the right of $E(R)$. Also, the diagrams have been constructed with the ratio $e.b.(R)/s.d.(R)$ taking successively the values 0.2, 0.6 and 1.8 from left to right across the page. We see that the displacement between the positions of $E(R)$ and $\mathbf{R}$ means that the coverage probability (or confidence level) of a confidence interval *de*creases progressively as this ratio increases – the small effect on the left when the ratio is 0.2 would likely not usually be of practical importance, but the substantial effect on the right when the ratio is 1.8 would likely often impose a severe limitation on the answer provided by the confidence interval.

In practice, to try to ensure that the *actual* confidence level of a confidence interval for $\mathbf{R}$ *is* close to its *stated* level, a common requirement is that the magnitude of the estimating bias of R be less than *one-tenth* (or 10%) of the standard deviation of R; this is usually checked by calculating the estimate of $c.v.(\overline{X})$ based on the data for the sample x_js. Another illustration of the effect of estimating bias on coverage probability of a confidence interval is given in Appendix 3 which starts at the bottom of page 2.123 of this Figure 2.15.

NOTE: 7. In questions involving interval estimating of $\mathbf{R}$, we proceed as follows:

∗ if the individual sample y_js and x_js are given in the question statement, we plot them as a scatter diagram, to assess how close they lie to a line through the origin with slope r (estimating $\mathbf{R}$);

∗ we also estimate $c.v.(\overline{X})$ from the sample data – if its value is *less* than 0.1, we proceed with estimating $\mathbf{R}$;

∗ if the value found for the estimate of $c.v(\overline{X})$ in the previous check is *greater* than 0.1, we use equation (2.15.9) [on the lower half of page 2.114] to estimate the estimating bias of R – the previous check only involves an *upper bound* for the magnitude of the ratio $e.b.(R)/s.d.(R)$, and we may find it is *actually* less than 0.1 from this last calculation.

● The upper limit of 0.1 or 10% in the foregoing discussion should be recognized as a convenient *working* guideline, *not* a rigid standard; in this regard, it is like the value of 0.05 that is conventionally used to separate what is, or is not, statistically significant (but *not* to decide what is, or is not, practically important).

(*continued*)

Figure 2.15. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating the Ratio of Two Averages or Totals (continued 2)

Example 2.15.1: To investigate the change in assessed home values in a (respondent) population of 1,000 homes, twenty homes were obtained by equiprobable selecting. From tax records, the home values two years earlier (x_j) and at present (y_j) were obtained; the numerical summaries of these data are as shown at the right. Find an approximate 90% confidence interval for $\mathbf{R}$, the ratio of average present-to-past home values.

$$\sum_{j=1}^{n} y_j = 618.2, \qquad \sum_{j=1}^{n} y_j^2 = 19{,}380.80;$$
$$\sum_{j=1}^{n} x_j = 516.1, \qquad \sum_{j=1}^{n} x_j^2 = 13{,}497.73;$$
$$\sum_{j=1}^{n} y_j x_j = 16{,}171.81, \qquad n = 20.$$

Solution: This question involves interval estimating of $\mathbf{R}$; we first assess the likely effect of the estimating bias of R on the accuracy of the *stated* (here, 90%) confidence level of the confidence interval for $\mathbf{R}$.

We are not given the individual sample y_js and x_js in the statement of the question (only the usual five numerical data summaries), so we cannot make a scatter diagram to assess how close they lie to a line through the origin with slope r (estimating $\mathbf{R}$); however, using the two numerical summaries of the sample x data given at the right above, we *can* do the check involving $c.v.(\overline{X})$.

We have: $\mathbf{N} = 1{,}000$, $n = 20$, $\overline{x} = 25.805\ (= \overline{x})$, $s_x = 3.075\,963\,692\ (= s_x)$ [equation (2.15.10)],

so that: $\hat{s.d.}(\overline{X}) = s_x\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = 0.680\,893\,588$ [using equation (2.15.2) on page 2.113],

and: $\hat{c.v.}(\overline{X}) = 0.680\,893\,588/25.805 = 0.026\,386\,11 \simeq 0.026 = 2.6\%$ [see Note 4 on page 2.115].

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

We have: $r = \overline{y}/\overline{x} = \sum_{j=1}^{n} y_j \Big/ \sum_{j=1}^{n} x_j = 618.2/516.1 = 1.197\,829\,878\ (= r)$, ${}_{.1}t^*_{19} = 1.72\,913$ for 90% confidence;

also:
$$\sum_{j=1}^{n}(y_j - rx_j)^2 = \sum_{j=1}^{n} y_j^2 - 2r\sum_{j=1}^{n} y_j x_j + r^2\sum_{j=1}^{n} x_j^2 = 19{,}380.80 - 2\times 1.197830\times 16{,}171.81 + 1.197830^2\times 13{,}497.73$$
$$= 38{,}747.29463 - 38{,}742.1544 = 5.140\,237\ [= \sum_{j=1}^{n}(y_j - rx_j)^2];$$

so that:
$$\hat{s.d.}(R) \simeq \sqrt{\frac{1}{(n-1)\overline{x}^2}\sum_{j=1}^{n}(y_j - rx_j)^2\left(\frac{1}{n} - \frac{1}{\mathbf{N}}\right)} = \sqrt{\frac{1}{19\times 25.805^2}(5.140\,237)\left(\frac{1}{20} - \frac{1}{1{,}000}\right)} = 0.004\,461\,788.$$

Hence, an approximate 90% confidence interval for $\mathbf{R}$, the respondent population ratio of average present-to-past home values. is:

$$r \pm 1.72913\times\hat{s.d.}(R) = 1.197830 \pm 1.72913\times 0.004\,461\,788 \Rightarrow (1.1901,\ 1.2055) \text{ or about } (1.19,\ 1.21).$$

NOTE: 8. This approximate 90% confidence interval is narrow, representing a (surprisingly, in light of the small sample size) precise answer; this may reflect present-to-past home value ratios that are *less* variable than the y_js and the x_js. This is consistent with the values of the three sample standard deviation estimates at the right, where s_r [equation (2.15.18) on page 2.115] is appreciably smaller than the (similar) s_y and s_x values [based on equation (2.15.10) on page 2.114].

$s_y \simeq 3.785$
$s_x \simeq 3.076$
$s_r \simeq 0.520$

- We also bear in mind limitations on the answer imposed by the Taylor series *approximations* in Sections 3 amd 4 in deriving the relevant theory on pages 2.113 to 2.116.

Exercises: 1. Describe briefly how the solution of Example 2.15.1 would be affected if its last sentence were: *Find an approximate 90% confidence interval for the average ratio of present-to-past hone values.*

2. Estimate the estimating bias of R and so estimate the *actual value* of the ratio $e.b.(R)/s.d.(R)$ [*Answer*: 0.003 321 ≃ 0.33%]; compare your estimate with the upper limit provided by equation (2.15.22) on the upper half of page 2.116.

Example 2.15.2: From an urban area containing 270 blocks, 20 blocks were obtained by EPS; for each chosen block, the number of dwellings it contained, and the number of these that were rented, were found to be as follows:

Table 2.15.2:

Block number (j)	1	2	3	4	5	6	7	8	9	10
Number of dwellings (x_j)	5	9	18	68	32	48	11	1	1	4
Number rented (y_j)	3	5	5	52	21	34	3	0	0	0
Proportion rented (p_j)	0.6	$0.\dot{5}$	$0.2\dot{7}$	0.7647	0.65625	$0.708\dot{3}$	$0.\dot{2}\dot{7}$	0	0	0

Block number (j)	11	12	13	14	15	16	17	18	19	20
Number of dwellings (x_j)	29	31	5	2	4	102	20	15	1	29
Number rented (y_j)	17	14	0	0	2	54	11	11	0	23
Proportion rented (p_j)	0.5862	0.4516	0	0	0.5	0.5294	0.55	$0.7\dot{3}$	0	0.7931

(*continued overleaf*)

Example 2.15.2: (continued) for these data: $\sum_{j=1}^{n} y_j = 255$, $\sum_{j=1}^{n} y_j^2 = 8{,}545$; $\sum_{j=1}^{n} x_j = 435$, $\sum_{j=1}^{n} x_j^2 = 22{,}239$; $\sum_{j=1}^{n} y_j x_j = 13{,}518$, $n = 20$.

$\sum_{j=1}^{n} p_j = 7.961\,018$, $\sum_{j=1}^{n} p_j^2 = 4.863\,739$.

(a) Find an approximate 95% confidence interval for the proportion of rented dwellings in the urban area.

(b) Ignoring the actual sample survey Plan and assuming that the same number of rented dwellings had been found in 435 dwellings obtained by EPS from the whole urban area, find an approximate 95% confidence interval for the proportion of rented dwellings in the urban area.

(c) Find an approximate 95% confidence interval for the average proportion of rented dwellings per block in the urban area.

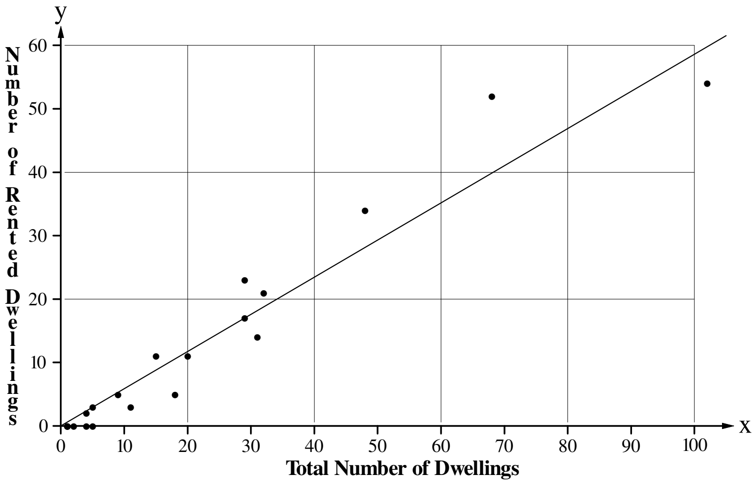
(d) Find an approximate 95% confidence interval for the total number of dwellings in the urban area.

(e) Find an approximate 95% confidence interval for the total number of *rented* dwellings in the urban area.

Solution: (a) This question involves interval estimating of $\mathbf{R}$; we first assess the likely effect of the estimating bias of R on the accuracy of the *stated* confidence level of the confidence interval for $\mathbf{R}$. [The question in part (a) is posed as a *proportion* so it may be convenient here, and also in (b) and (c), to think of $\mathbf{R}$ as $\mathbf{P}$.]

The scatter diagram of the sample y_js and x_js is shown at the right; we see that the points lie *fairly* close to the line with slope r (estimating $\mathbf{R}$) through the origin. Hence, we seem to have reasonably favourable conditions in this situation for lower estimating bias and higher precision with respect to the relationship between $\mathbf{Y}$ and $\mathbf{X}$. However, the relatively *small* sample size of 20 is likely to have an adverse effect, particularly on precision with its $n^{-½}$ dependence (compared with n^{-1} dependence for the estimating bias).

Scatter Diagram of Rented vs. Total Number of Dwellings in 20 Blocks



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

We have: $\mathbf{N} = 270$, $n = 20$, $\overline{x} = 21.75$ (= $\overline{x}$), $s_x = 25.932\,858\,65$ (= s_x) [equation (2.15.10)],

so that: $\hat{s.d.}(\overline{X}) = s_x\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = 5.579\,862\,76$ [using equation (2.15.2) on page 2.113],

and: $\hat{c.v.}(\overline{X}) = 5.579\,862\,76/21.75 = 0.256\,545\,414 \simeq 0.257 = 25.7\%$ [see Note 4 on page 2.115].

This estimate is well *above* the conventional upper bound of 10% for the magnitude of the ratio $e.b.(R)/s.d.(R)$ so we estimate the *actual value* of the ratio using equations (2.15.9) and (2.15.13) on pages 2.114 and 2.115.

We have: $r = \overline{y}/\overline{x} = \sum_{j=1}^{n} y_j \Big/ \sum_{j=1}^{n} x_j = 255/435 = 0.586\,206\,896$ (= r), ${}_{.05}t_{19}^{*} = 2.09302$ for 95% confidence;

also: $\sum_{j=1}^{n} x_j(rx_j - y_j) = r\sum_{j=1}^{n} x_j^2 - \sum_{j=1}^{n} y_j x_j = 0.5862069 \times 22{,}239 - 13{,}518 = -481.344\,840$ [$= \sum_{j=1}^{n} x_j(rx_j - y_j)$],

so that: $\hat{e.b.}(R) \simeq \frac{1}{(n-1)\overline{x}^2}(r\sum_{j=1}^{n} x_j^2 - \sum_{j=1}^{n} y_j x_j)(\frac{1}{n} - \frac{1}{\mathbf{N}}) = -0.002\,479\,307\,79$;

further: $\sum_{j=1}^{n}(y_j - rx_j)^2 = \sum_{j=1}^{n} y_j^2 - 2r\sum_{j=1}^{n} y_j x_j + r^2\sum_{j=1}^{n} x_j^2 = 8{,}545 - 2 \times 0.5862069 \times 13{,}518 + 0.5862069^2 \times 22{,}239$

$= 16{,}187.17717 - 15{,}848.68966 = 338.487\,514$ [$= \sum_{j=1}^{n}(y_j - rx_j)^2$];

so that: $\hat{s.d.}(R) \simeq \sqrt{\frac{1}{(n-1)\overline{x}^2}(\sum_{j=1}^{n} y_j^2 - 2r\sum_{j=1}^{n} y_j x_j + r^2\sum_{j=1}^{n} x_j^2)(\frac{1}{n} - \frac{1}{\mathbf{N}})} = \sqrt{0.001\,743\,479\,232} = 0.041\,754\,99$;

hence: $\frac{|\hat{e.b.}(R)|}{\hat{s.d.}(R)} \simeq \frac{0.002\,479\,308}{0.041\,754\,99} = 0.059\,3775$ or about 5.9%.

This estimate of about 5.9% is well *below* the conventional upper bound of 10% for the magnitude of the ratio $e.b.(R)/s.d.(R)$ so the *actual* confidence level of the confidence interval for $\mathbf{R}$ should not be too far below its stated level (here, 95%). This implies that the component of model error arising from the estimating bias of R should not impose undue limitation of our answer in this question context; the approximate 95% confidence interval for $\mathbf{R}$, representung the proportion of rented dwellings in the urban area, is:

$r \pm 2.09302 \times \hat{s.d.}(R) = 0.586207 \pm 2.09302 \times 0.041\,754\,99 \Rightarrow (0.4988, 0.6736)$ or about (49%, 68%).

(*continued*)

Figure 2.15. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating the Ratio of Two Averages or Totals (continued 3)

NOTES: 9. The expressions for the estimate of the estimating bias and the standard deviation of R used in estimating the magnitude of the ratio $e.b.(R)/s.d.(R)$ in the preliminary check in part (a) of Example 2.15.2 are only *approximate*; we are therefore unsure of the accuracy of the resulting estimate. However, the *ratio* is likely to be *more* accurate than the estimate of either quantity considered *by itself* because, as we see in Sections 3 and 4 on pages 2.114 and 2.115, the approximations involved in the derivations of the two expressions are influenced by essentially the *same* two factors – recall Note 3 on page 2.115.

10. Example 2.15.2 involves selecting equal-sized *clusters* (blocks) so notation consistent with that of Figure 2.14 would (strictly) involve using $\mathbf{M}$ and m rather than $\mathbf{N}$ and n. However, the latter have been used in the solution of part (a) on the facing page 2.118 because:

- ratios in this Figure 2.15 involve *averages* (although Figure 2.14 does too);
- the theory of this Figure 2.15 is demanding enough without a second notational version for clustering that is not strictly necessary.

The number of *dwellings* (rather than blocks) in the (respondent) population, which we denote $\mathbf{N}^*$, is unknown but we can calculate its approximate value as $\mathbf{N}\overline{\mathrm{x}}$ (or $\mathbf{M}\overline{\mathrm{x}}$) = 5,872.5 dwellings by assuming that $\mathrm{m}/\mathbf{M} \simeq \mathrm{n}/\mathbf{N}$, which is usually likely to be a good approximation [see part (d) below].

11. The data for Example 2.15.2 are taken from L. Kish: *Survey Sampling*, 1965 (John Wiley & Sons); the reader is referred to this source (*e.g.*, pages 42-43, 71, 189-190 and 205) for further discussion.

Solution: (cont.) (b) When the clustering is ignored, this question involves interval estimating of a proportion (of rented dwellings) under EPS of a sample of 435 dwellings from the urban area which we infer contains 2,872.5 dwellings – see Note 10 above; we use equation (2.10.6) for $\widehat{s.d.}(P)$ on page 2.81 in Figure 2.10.

We have: $\mathbf{N}^* = 5{,}872.5$ $\quad \mathrm{n} = 435,$ $\quad \mathrm{p} = \frac{255}{435} \simeq 0.586\,207\ (= p),$ $\quad \mathrm{q} = \frac{180}{435} \simeq 0.413\,793\ (= q),$ $\quad z^*_\alpha = 1.95996$ for 95% confidence,

so that: $\widehat{s.d.}(P) = \sqrt{\frac{\mathrm{n}}{\mathrm{n}-1}pq(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} = \sqrt{\frac{435}{434} \times \frac{255}{435} \times \frac{180}{435}(\frac{1}{435} - \frac{1}{5{,}872.5})} = 0.022\,748\,896.$

Hence, an approximate 95% confidence interval for $\mathbf{P}$, the respondent population proportion of rented dwellings, is:

$p \pm 1.95996 \times \widehat{s.d.}(P) = 0.586207 \pm 1.95996 \times 0.022\,7489 \implies (0.5416,\ 0.6308)$ or about (54, 64%).

NOTE: 12. The ratio of the two estimated standard deviations in (a) and (b) is 0.041755/0.022749 ≃ 1.835; the substantially higher value in (a) reminds us of the increased 'cost' we pay for the convenience of selecting clusters – recall the discussion of efficiency on page 2.108 and Example 2.14.3 on page 2.109 in Figure 2.14.

Solution: (cont.) (c) This question involves interval estimating of an *average* proportion which, in the context of this Figure 2.15, we would think of as an average *ratio*, even though the relevant theory is that of Figure 2.3 for *univariate* estimating, as in Illustration 2 on page 2.113. The designation 'per block' in part (c) means that the clustering is ignored [as is also the case in part (b) above].

We have: $\mathbf{N} = 270,$ $\quad \mathrm{n} = 20,$ $\quad \overline{\mathrm{p}} \equiv \overline{\mathrm{r}} = \frac{7{,}961018}{20} = 0.398\,0509\ (= \overline{p}),$ $\quad {}_{.05}t^*_{19} = 2.09302$ for 95% confidence,

$$\mathrm{s}_{\overline{\mathrm{p}}} = \sqrt{\frac{4{,}863739 - 7.961018^2/20}{19}} = 0.298\,667\,974\ (= s_{\overline{p}}),$$

so that: $\widehat{s.d.}(\overline{P}) \equiv \widehat{s.d.}(\overline{R}) = s_{\overline{p}}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 0.298668\sqrt{\frac{1}{20} - \frac{1}{270}} = 0.064\,263\,116.$

Hence, an approximate 95% confidence interval for $\overline{\mathbf{P}}$, the average proportion of rented dwellings per block, is:

$\overline{p} \pm 2.09302 \times \widehat{s.d.}(\overline{P}) = 0.39805 \pm 2.09302 \times 0.064\,263 \implies (0.2635,\ 0.5326)$ or about (26%, 54%).

NOTE 13. The wide (about 28 percentage points) confidence interval in part (c) reflects low precision from estimating an attribute based of a very variable variate – in the sample, there are five p_js of 0 and four of at least 0.7.

Solution: (cont.) (d) This question involves interval estimating of a *total* but where $\mathbf{N}$ (the number of *dwellings* in the respondent population) is *un*known; we use an approximate value $\mathbf{N}^* = 5{,}872.5$ – see Note 10 above.

We have: $\mathbf{N}^* = 5{,}872.5,$ $\quad \mathbf{N}$ (or $\mathbf{M}$) = 270, $\quad \widehat{s.d.}(\overline{X}) = 5.579\,862\,76$ [from part (a)], $\quad {}_{.05}t^*_{19} = 2.09302$ for 95% confidence,

so that: $\mathbf{N}\widehat{s.d.}(\overline{X}) = 270 \times 5.579\,862\,76 = 1{,}506.6$ dwellings.

Hence, an approximate 95% confidence interval for ${}_{\mathrm{T}}\mathbf{X}$, the total number of dwellings in the urban area, is:

$\mathbf{N}\overline{x} \pm 2.09392 \times \mathbf{N}\widehat{s.d.}(\overline{X}) = 5{,}872.5 \pm 2.09302 \times 1{,}506.6 \implies (2{,}719,\ 9{,}026)$ or about (2,700, 9,100).

(*continued overleaf*)

NOTE 14. The wide confidence intervals in part (d) overleaf at the bottom of page 2.119 and in part (e) below are other instances of the statistical issue described in Note 13 overleaf on paee 2.119.

15. Parts (d) and (e) actually involve estimating a total in *un*equal-sized clusters when $\mathbf{N}$ (but not $\mathbf{M}$) is *un*known, a matter pursued in the following Figure 2.16.

Solution: (cont.) (e) This question involves interval estimating of a *total* but where $\mathbf{N}$ (the number of *rented* dwellings in the respondent population) in *un*known; as in Note 10 overleaf near the top of page 2.119, we can calcalate its approximate value as $\mathbf{N}\overline{\mathrm{y}}$ (or $\mathbf{M}\overline{\mathrm{y}}$) = 3,442.5 rented dwellings by assuming that $\mathrm{m}/\mathbf{M} = \mathrm{n}/\mathbf{N}$, which is usually likely to be a good approximation. [3,442.5 is also the value of $\mathbf{N}^*\mathrm{r}$.]

We have: $\mathbf{N} = 270$, $\mathrm{n} = 20$, $\overline{\mathrm{y}} = \frac{255}{20} = 12.75$ $(= \overline{y})$, $\mathrm{s}_\mathrm{y} = 16.691\,866\,91$ $(= s_y)$ [equation (2.15.10)],

so that: $\mathbf{N}\hat{s.d.}(\overline{Y}) = \mathbf{N}s_y\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 969.709\,8386$ [using equation (2.15.2) on page 2.113].

Hence, an approximate 95% confidence interval for $_\mathrm{T}\mathbf{Y}$, the total number of *rented* dwellings in the urban area, is:

$$\mathbf{N}\overline{\mathrm{y}} \pm 2.09302 \times \mathbf{N}\hat{s.d.}(\overline{Y}) = 3{,}442.5 \pm 2.09302 \times 969.7 \implies (1{,}413,\ 5{,}472) \text{ or about } (1{,}400,\ 5{,}500).$$

Example 2.15.3: In a political poll of 1,100 Canadian voters, a sample was obtained by EPS consisting of voters who were prepared to respond to a telephone interview; this sample contained 250 Liberal supporters and 325 undecided voters. Find an approximate 99% confidence interval for $\mathbf{P}$, the proportion of Liberal supporters among *decided* voters in this (respondent) population.

Solution: The situation in this question is like Illustration 3 on page 2.113 because, in the sample, the number of Liberal supporters and the number of *decided* voters in the sample selected by EPS can be represented by *distinct* random variables; hence, the question involves interval estimating of $\mathbf{R}$ (or $\mathbf{P}$).

Reasoning as in Section 1 on page 2.81 in Figure 2.10, we define two indicator variates as follows:

$\mathbf{Y}_\mathrm{i} = 1$ if respondent population element i is a Liberal supporter,
$\mathbf{Y}_\mathrm{i} = 0$ if respondent population element i is *not* a Liberal supporter;
$\mathbf{X}_\mathrm{i} = 1$ if respondent population element i is a decided voter,
$\mathbf{X}_\mathrm{i} = 0$ if respondent population element i is an *un*decided voter;

the proportion of Liberal supporters among decided voters is then: $\mathbf{R}$ (or $\mathbf{P}$) $= \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{Y}_\mathrm{i}/\mathbf{N} \Big/ \sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{X}_\mathrm{i}/\mathbf{N}$.

The question does not give the value of $\mathbf{N}$. but we know that the number of Canadian voters exceeds 10 million; with a sample size of around 1,000, we can therefore take $1/\mathrm{n} - 1/\mathbf{N}$ as $1/\mathrm{n}$ to a good approximation.

We denote the random variables, for the model based on EPS of the sample of voters, by Y_j and X_j for unit (or element) j of the sample; R (or P) is the random variable representing the sample ratio, which is the estimator (and whose value is the estimate) of $\mathbf{P}$.

The sample information in the question is n = 1,100 and it also allows us to calculate:

$\overline{\mathrm{y}} \equiv \underline{\mathrm{p}}_y = \frac{250}{1{,}100} = 0.2\dot{2}\dot{7}$ $(= \overline{y} \equiv \underline{p}_y)$, $\overline{\mathrm{x}} \equiv \underline{\mathrm{p}}_\mathrm{x} = \frac{775}{1{,}100} = 0.70\dot{4}\dot{5}$ $(= \overline{x} \equiv \underline{p}_x)$, $z^*_\alpha = 2.57583$ for 99% confidence,

$\mathrm{r} \equiv \mathrm{p} = \overline{\mathrm{y}}/\overline{\mathrm{x}} = \frac{250}{775} = 0.322\,580\,645.$ $(= r \equiv p)$.

Equation (2.10.5) from the lower half of page 2.81 in Figure 2.10 is given below as equation: (2.15.24);

$$s_v = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(v_j - \overline{v})^2} = \sqrt{\frac{1}{\mathrm{n}-1}\Big[\sum_{j=1}^{\mathrm{n}}v_j^2 - \mathrm{n}\overline{v}^2\Big]} = \sqrt{\frac{1}{\mathrm{n}-1}[\mathrm{n}\underline{p}_v - \mathrm{n}\underline{p}_v^2]} = \sqrt{\frac{\mathrm{n}}{\mathrm{n}-1}\underline{p}_v(1-\underline{p}_v)} \qquad \text{-----(2.15.24)}$$

together with equation (2.15.10) on the lower half of page 2.114, this enables us to calculate:

$$s_y = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^2} = \sqrt{\frac{\mathrm{n}}{\mathrm{n}-1}\underline{p}_y(1-\underline{p}_y)} = \sqrt{0.175\,779\,634} = 0.419\,260\,819, \qquad \text{-----(2.15.25)}$$

$$s_x = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(x_j - \overline{x})^2} = \sqrt{\frac{\mathrm{n}}{\mathrm{n}-1}\underline{p}_x(1-\underline{p}_x)} = \sqrt{0.208\,350\,566} = 0.445\,454\,342, \qquad \text{-----(2.15.26)}$$

$$s_{yx} = \frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})(x_j - \overline{x}) = \frac{1}{\mathrm{n}-1}\Big(\sum_{j=1}^{\mathrm{n}}y_jx_j - \mathrm{n}\overline{y}\,\overline{x}\Big) = \frac{1}{\mathrm{n}-1}[\mathrm{n}\underline{p}_y - \mathrm{n}\underline{p}_y\underline{p}_x] = \frac{\mathrm{n}}{\mathrm{n}-1}\underline{p}_y(1-\underline{p}_x) \qquad \text{-----(2.15.27)}$$
$$= 0.067\,209\,86;$$

in equation (2.15.27), $\sum_{j=1}^{\mathrm{n}}y_jx_j = \mathrm{n}\underline{p}_y$ because $y_j = 1$ (a Liberal supporter) entails $x_j = 1$ (a decided voter).

Then using equation (2.15.13) on the upper half of page 2.115, we have:

$$\hat{s.d.}(R) \equiv \hat{s.d.}(P) \simeq \sqrt{\frac{1}{\overline{x}^2}(s_y^2 - 2r\,s_{yx} + r^2s_x^2)\Big(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\Big)}$$
$$\simeq \sqrt{\frac{1}{0.70\dot{4}\dot{5}^2}(0.175\,780 - 2 \times 0.322\,581 \times 0.067\,2099 + 0.322\,581^2 \times 0.208\,351)\frac{1}{1{,}100}} = 0.016\,799\,432.$$

(*continued*)

Figure 2.15. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating the Ratio of Two Averages or Totals (continued 4)

Solution: (cont.) Hence, an approximate 99% confidence interval for $\mathbf{P} \equiv \mathbf{R}$, the proportion of Liberal supporters among decided voters, is:

$$p \pm 2.57583 \times \hat{s.d.}(P) \;=\; 0.322581 \pm 2.57583 \times 0.016\,7994 \;\Longrightarrow\; (0.2793,\ 0.3658) \text{ or about } (28,\ 37\%).$$

NOTE 16. As in Examples 2.15.1 and 2.15.2, we can assess the likely effect of the estimating bias of R on the accuracy of the *stated* (here, 99%) confidence level of the confidence interval for $\mathbf{P} \equiv \mathbf{R}$.

From equation (2.15.15) on the lower half of page 2.115:

$$\hat{c.v.}(\overline{X}) = \frac{\hat{s.d.}(\overline{X})}{\overline{x}} = \frac{s_x}{\overline{x}}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \simeq \frac{0.455454}{0.70\dot{4}\dot{5}}\sqrt{\frac{1}{1{,}100}} = 0.019\,063\,289 \simeq 1.9\%;$$

this estimate is well *below* the conventional upper bound of 10% for the magnitude of the ratio $e.b.(R)/s.d.(R)$ so the *actual* confidence level of the confidence interval for $\mathbf{P} \equiv \mathbf{R}$ should not be too far below its stated level (here, 99%). This implies that the component of model error arising from the estimating bias of R should not impose undue limitation of our answer in this question context.

However, from equation (2.15.15) on the lower half of page 2.114, we have: $\hat{e.b.}(R) \simeq \frac{1}{\overline{x}^2}(rs_x^2 - s_{yx})(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})$, which is *zero* in this situation, as we see from the expression $\mathrm{r} \equiv \mathrm{p} = \overline{\mathrm{y}}/\overline{\mathrm{x}} = \frac{250}{775}$ $(= r \equiv p)$ and the symbolic expressions in equations (2.15.26) and (2.15.27) near the bottom of the facing page 2.120; we can confirm, of course, that the relevant values of r, s_x and s_{yx} given on the lower half of page 2.120 make $(rs_x^2 - s_{yx})$ zero to within the limits imposed by rounding errors – here, it is zero to 8 or 9 decimal places.

Thus, it seems that, in the context of Example 2.15.3, estimating bias actually imposes *no* limitation on the answer for $\mathbf{P}$ provided by the confidence interval.

NOTES: 17. Concern with the estimating bias of R in this Figure 2.15 supplements the discussion in Note 21 at the top of page 2.48 in Figure 2.3 and in its Appendix 7 on pages 2.48 and 2.49. We are reminded of the following key ideas.

∗ In these Materials, the word *error* has its statistical meaning of *the difference from the true value*, **not** its ordinary English meaning of *mistake*; error arises in the context of an *individual* case (*e.g.*, in a *particular* investigastion), *not* under repetition.

∗ Under conceptual or actual repetition, the two characteristics of the *sign* and *magnitude* of (numerical) error lead to what we call *inaccuracy* and *imprecision*; the *inverses* of these two ideas – accuracy and precision – provide more familiar terminology, but we must remember that statistical methods (try to) manage the (undesirable) *former* to achieve needed levels of their (desirable) inverses.

- In the classroom, we can do *actual* repetition (repeating over and over) to demonstrate, for example, the statistical behaviour of the values of sample attributes (*e.g.*, averages) and of the values which arise from measuring processes;
- *Out*side the classroom, *actual* repetition is usually not a viable option – we use instead *hypothetical* repetition based on an appropriate statistical *model* (for the selecting and measuring processes, for example).

We help maintain the distinction between the real world and the model by using bias and variability as the *model* quantities which represent inaccuracy and imprecision in the real world.

∗ Under EPS, the random variables $\overline{Y}$ [representing the sample average (in Figure 2.3)] and R [representing the sample ratio of averages (in this Figure 2.15)] are **consistent** estimators, meaning that as the sample size approaches the population size, the algebraic form of each estimator is such that it approaches the corresponding population attribute (so that when the sample consists of the whole population, the 'estimate' is *equal to* the attribute).

- This idea overlaps that of estimating bias decreasing in magnitude with increasing sample size, behaviour quite *un*like that of real-world inaccuracy.

18. Note 13 on the lower half of the sixth side (page 2.110) of Figure 2.14 compares selecting processes involving EPS of individual elements (on the left) and EPS of equal-sized clusters (on the right, *ec*) when estimating an average; Table 2.15.3 overleaf on page 2.122 broadens this comparison to include *un*equal-sized clusters and changes the attribute being estimated to a *proportion* rather than an average. As in Note 13, we see again in Table 2.15.3 how the expressions for the estimates and for calculating their estimated standard deviations change in response to the different methods of data collection represented by the different unit compositions (as specified in the relevant sampling protocol, denoted 1, 2 and 3 in Table 2.15.3). There are four comments about Table 2.15.3.

∗ The change from averages to proportions means that most 'y's become 'p's.

∗ The v_j in the numerator of the expression for p_{in} are *indicator variates.*

∗ The cluster proportions (p_j) in the numerator of the *first* expression for p_{ec} are calculated for each cluster by the method of protocol 1, thus providing a link between protocols 1 and 2; the *second* expression for p_{ec} is analogous to that for p_{uc}, thus linking protocols 2 and 3. These links suggest that we can regard the models

(*continued overleaf*)

NOTES: (cont.) 18. ∗ for protocols 2 and 3 as successive adaptations of the basic model for protocol 1.

∗ Because of its origin in the theory for estimating a ratio, equation (2.15.13) [on the upper half of the third side (page 2.115) of this Figure 2.15], which is the basis for modelling protocol 3, uses n to denote sample size; this has been changed to m (and, likewise, **N** changed to **M**) in the third row of Table 2.15.3 to agree with the more usual notation for selecting clusters – recall also Note 10 near the the top of page 2.115.

Table 2.15.3:

Sampling protocol	**Estimate of Ᵽ**	**Estimate of standard deviation of estimator**	**Reference**
1. EPS of *n* individual elements	$p_{in} = \frac{1}{\text{n}}\sum_{j=1}^{\text{n}} v_j$	$\sqrt{\frac{\text{n}}{\text{n}-1} p_{in} q_{in} (\frac{1}{\text{n}} - \frac{1}{\mathbf{N}})}$	Figure 2.10 [(equation (2.10.6)]
2. EPS of *m* clusters of size L elements	$p_{ec} = \frac{1}{\text{m}}\sum_{j=1}^{\text{m}} p_j = \sum_{j=1}^{\text{mL}} y_j \Big/ \text{mL}$	$\sqrt{\frac{1}{\text{m}-1}\sum_{j=1}^{\text{m}} (p_j - p_{ec})^2 (\frac{1}{\text{m}} - \frac{1}{\mathbf{M}})}$	Figure 2.14 [(equation (2.14.6)]
3. EPS of *m* clusters of *un*equal sizes	$p_{uc} = \sum_{j=1}^{\text{m}} y_j \Big/ \sum_{j=1}^{\text{m}} x_j$	$\sqrt{\frac{1}{(\text{m}-1)\bar{x}^2}\sum_{j=1}^{\text{m}} (y_j - p_{uc} x_j)^2 (\frac{1}{\text{m}} - \frac{1}{\mathbf{M}})}$	Figure 2.15 [(equation (2.15.13)]

19. The occurrence of (measured) *y* and *x* in the last line of Table 2.15.3 (protocol 3) reminds us that averages, proportions and ratios:

- are similar *mathematically* in being the quotient of two numbers, BUT:
- differ *statistically* depending on how the numerator and denominator of the quotient need to be modelled (under EPS of the sample, for example) – recall the three Illustrations on page 2.113.
  - Using an indicator variate to enable a proportion to be treated like an average is discussed in Figure 2.10 – see Section 1 on page 2.81, for example.
  - A proportion and a ratio are equivalent in, for instance, part (a) of Example 2.15.2 on page 2.118.
    - ○ The context of Example 2.15.2 also allows for an *average proportion* [in part (c) on page 2.119], where it is treated as an *average* ratio (and involves *uni*variate estimating).
- Expressing proportions as *percentages* introduces two further complications:
  - the need to distinguish percent from percentage points – for instance, a change from 35 percent to 38 percent in a party's support among voters in a political poll is an increase of 3 *percentage points*, **not** 3 *percent.*
  - a ratio expressed (perhaps for greater-sounding impact) as ... *an increase of 500 per cent* would be clearer if described as ... *a 6-fold increase.*
    - ○ It is also likely that such wording is often *mis*used and/or *mis*understood to mean a *5*-fold increase.

20. Having completed Example 2.15.3, it is useful to summarize the different *unit compositions* (a matter specified in the sampling protocol in the Plan at the Design stage of the FDEAC cycle) we have encountered so far in STAT 332; as part of this summary, we learn more about *uni*- and *bi*variate estimating.

∗ On the upper half of page 2.6 in Figure 2.1, we list five population attributes we may wish to estimte – averages, totals, ratios, proportions and frequencies. Associating these attributes with their corresponding data type (measured or counted) and considering also EPS at three different levels of aggregation of the respondent population (individual elements, clusters of *equal* size, clusters of *un*equal sizes), Table 2.15.4 at the right summarizes how the Examples in various Figures of these Course Materials cover the different unit compositions (numerical table entries in the last three columns are *Example numbers* from Figures 2.3, 2.10, 2.14 and 2.15).

Table 2.15.4		**Equiprobable selecting of**		
Type of Data	**Type of Attribute**	**individual elements**	**clusters of equal size**	**clusters of unequal sizes**
Measured	Average	3.1, 3.2, 15.2c	14.1a	*see Figure 2.16*
	Total	3.3, 3.4	14.1b	*see Figure 2.16*
	Ratio	15.1	*Not considered in STAT 332*	
Counted (in 2 categories)	Proportion	10.1, 10.2, 14.4b, 15.2b, 15.3	14.4a	15.2a
	Frequency	10.3, 10.4	14.2	15.2d, 15.2e

Noteworthy matters about Table 2.15.4 are as follows.

∗ Examples 14.4b and 15.2b remind us of the effect on precision of EPS of *individual elements* as compared with EPS of *clusters* from the respondent population.

∗ Example 15.2 involves three *other* situations which differ in the types of attributes being estimated.

- Examples 15.2a and 15.1 involve the *same* (*bi*variate) situation, although formally the former is estimating a *proportion* and the latter a *ratio*; the link between them is the equivalence of a ratio of averages and a ratio

(*continued*)

Figure 2.15. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating the Ratio of Two Averages or Totals (continued 5)

NOTES: 20. ∗ – of totals [see Table 2.15.1 of notation on the first side (page 2.113) of this Figure 2.15].
(cont.)

- Example 2.15.2c is concerned with a *ratio* but it involves only *uni*variate estimating (*cf.* Illustration 2 on page 2.113 of this Figure 2.15). [The equivalent calculation in Example 2.15.1 would be estimating the average ratio of present-to-past home values if this (different) question had been posed.]
  - ○ The essential difference between parts (a) and (c) of Example 2.15.2 lies in the *order* of the two operations of taking an *average* and taking a *ratio* – (a) involves a ratio of averages and (c) an average of ratios.
- Examples 15.2d and 15.2e also involve *uni*variate estimating; as we will see in Figure 2.16 which deals with selecting clusters of *un*equal sizes, both questions are concerned with estimating a total when the number of *elements* (but not the number of *clusters*) in the population is *un*known.

The foregoing discussion illustrates the close relationship between ratios and proportions in the context of sample surveys; it also reminds us of how the processes for obtaining answers in these different contexts (here, estimating a ratio of averages from EPS of individual elements, estimating a proportion from EPS of unequal-sized clusters) can have the *same* mathematical structure.

6. Appendix 1 – Approximations Based on Taylor Series Expansions

The following information is a reminder, in geometric terms, of what is involved in using a truncated Taylor series to approximate a function, as is done on the second and third sides (pages 2.114 and 2.115) of this Figure 2.15 to obtain approximate expressions for the estimating bias and the standard deviation of *R*.

∗ In *one dimension* (the *uni*variate case), we have a polynomial approximation to a function of *one* variable which we graph (as a *line*) in *two* dimensions; the expansion is carried out about a particular point. [When this point is 0, we have a *Maclaurin* series.]

- The *first-order* (or *linear*) approximation is the tangent *straight line* at the point of expansion, which agrees with the function in value (represented on its graph by the height) and in slope (represented by the derivative) at the point.
- The *second-order* approximation is a tangent *parabola* at the point of expansion, which agrees with the function in value, in slope *and* in curvature at the point.
- The *third-order* approximation is a cubic polynomial, and so on.

∗ In *two dimensions* (the *bi*variate case), we have an approximation to a function of *two* variables which we graph (as a *surface*) in *three* dimensions; the expansion is again carried out about a particular point.

- The *first-order* approximation is now the tangent *plane* at the point of expansion, which agrees with the function in value (represented on its graph by the height) and in the slopes in two perpendicular directions (represented by two partial derivatives) at the point.
- The *second-order* approximation is now a tangent *surface* at the point of expansion, which agrees with the function in value, in the slopes *and* in the curvatures at the point; we see the six aspects of this agreement in the six terms of the Taylor series approximation near the start of Section 3 in the middle of the second side (page 2.114) of this Figure 2.15.

The so-called *error term* in the Taylor series expansion, which is usually more difficult to work with, is not needed in our use of Taylor series in this Figure 2.15. [This terminology is reminiscent of our meaning of 'error' – recall Note 17 on page 2.121.]

7. Appendix 2 – Another Derivation of the Estimating Bias of $R = \overline{Y}/\overline{X}$

The result (2.15.22) on the upper half pf page 2.116 is derived from *approximate* expressions for the estimating bias and the standard deviation of *R*; the following alternative derivation shows it is actually an *exact* result:

$$cov(R,\overline{X}) = E(R\bullet\overline{X}) - E(R)\bullet E(\overline{X}) \qquad \text{[definition of covariance]}$$

$$= E(\frac{\overline{Y}}{\overline{X}}\bullet\overline{X}) - E(R)\bullet\overline{\mathbf{X}} \qquad \text{[because } E(\overline{X}) = \overline{\mathbf{X}} \text{ under EPS]}$$

$$= \overline{\mathbf{Y}} - E(R)\bullet\overline{\mathbf{X}} \qquad \text{[because } E(\overline{Y}) = \overline{\mathbf{Y}} \text{ under EPS]}$$

$$\therefore \quad E(R) = \overline{\mathbf{R}} - \frac{cov(R,\overline{X})}{\overline{\mathbf{X}}} \qquad \text{[because [}\overline{\mathbf{R}} = \frac{\overline{\mathbf{Y}}}{\overline{\mathbf{X}}}\text{],}$$

so that: $$e.b.(R) \equiv E(R) - \overline{\mathbf{R}} = -\frac{cov(R,\overline{X})}{\overline{\mathbf{X}}} = -\frac{\rho_{R\overline{X}}\bullet s.d.(R)\bullet s.d.(\overline{X})}{E(\overline{X})} = -\rho_{R\overline{X}}\bullet s.d.(R)\bullet c.v.(\overline{X}).$$

Hence: $$\frac{|e.b.(R)|}{s.d.(R)} \le c.v.(\overline{X}) \qquad \text{[because } |\rho_{R\overline{X}}| \le 1\text{].} \qquad \text{-----(2.15.22)}$$

8. Appendix 3 – Effect of Estimating Bias on Coverage Probability

Appendix 3 illustrates, for our discussion in this Figure 2.15 of estimating a ratio of averages, how (estimating) bias affects the coverage probability of a confidence interval. The following incidental background information may also be of interest.

∗ The context is that of repeated weighing of the U.S. ten-gram standard NB 10, which was acquired by the U.S. National

(*continued overleaf*)

Bureau of Standards in about 1940 and has been weighed by them on approximately a weekly basis ever since. A question of interest is: *what can be inferred about the true mass of NB 10 from the data set formed by the outcomes of these weighings?* Here, we think in terms of a confidence interval as a way of answering this question.

∗ The diagram at the right below shows results from a computer simulation of one hundred 95% confidence intervals for the true mass (μ) of NB 10; the simulation is based on $N(405, 5)$ and $N(405, 20)$ models representing 96% and 4%, respectively, of the measurements of the mass of NB 10.

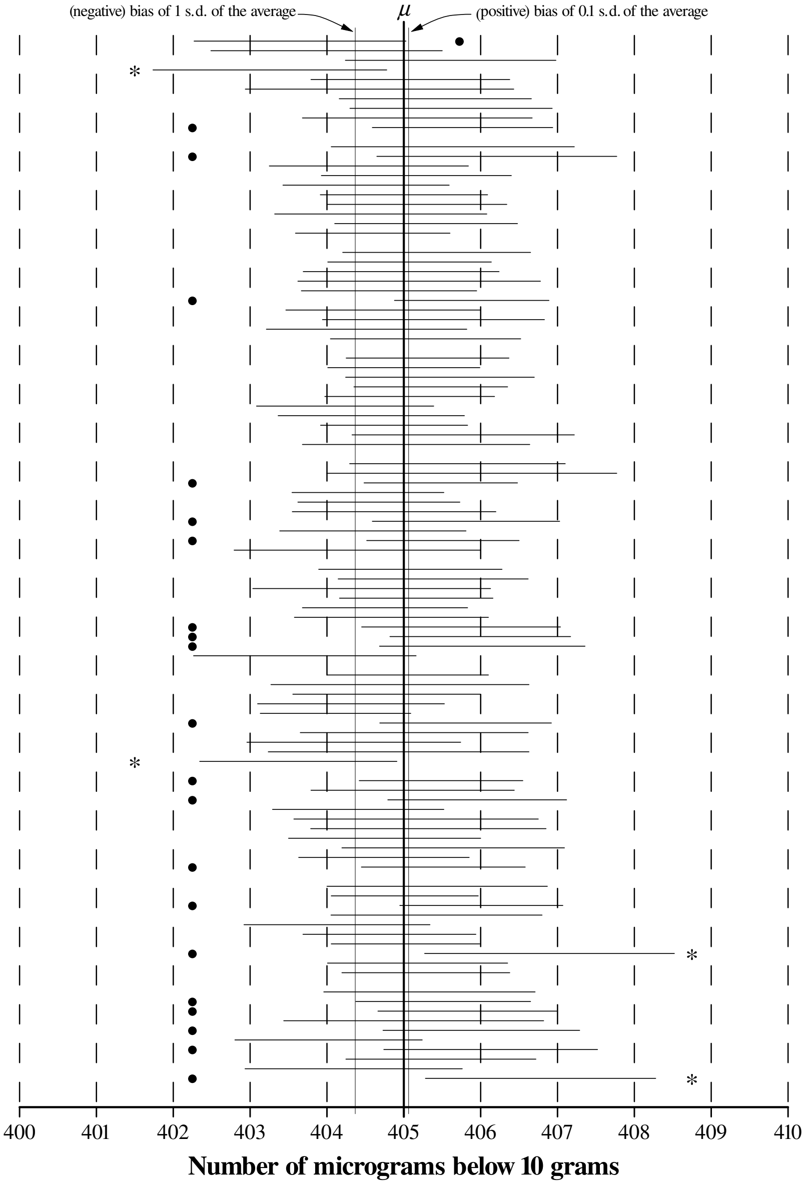
∗ Each approximate 95% CI is based on a simulated sample of 100 measurements; the average *and* the standard deviation vary from sample to sample (see also Figure 2.5 on pages 2.53 to 2.62), resulting in CIs with different centres and different widths (generated from relevant random variables) but subject to the constraints of the normal models; the standard deviation of the average of a sample of 100 measurements is $\sqrt{(0.96 \times 5^2 + 0.04 \times 20^2)/100} \simeq 0.632$ micrograms.

∗ The units of the scale on the horizontal axis at the bottom of the diagram are *micrograms below tens grams*; this coding of the weights has been adopted to avoid the inconvenience of working with numbers like 9.999 595 for the weights.

For the purpose of this Appendix 3, the diagram at the right shows that:

- The confidence level (approximately 95%) of the CIs means that about 95 of the 100 intervals should cover the true value of μ, set at 405 microrgams below ten grams in the models for the simulation and represented by the central heavy vertical line; we see in the diagram that actually 96 of the intervals cover μ – the four marked with an asterisk (∗) do not.
- 95 of the 100 intervals cover the vertical line just to the rght of μ at 405.063 micrograms, which represents a shift (*i.e.*, a *bias* in the present context) of 0.1 standard deviations of the average; the one addiional interval that does not cover the line is the first (uppermost) one, marked with a bullet (●) on its right. Thus, a bias whose magnitude is only *one-tenth* of the standard deviation of the estimator of μ (here, the sample average) does *not* appreciably affect the coverage probability. This is why a value of 0.1 or 10% for $c.v.(\overline{X})$ is commonly taken as an acceptable upper limit for the magnitude of the bias of R in relation to its standard deviation [recall equation (2.15.22) on the upper half (page 2.115) of this Figure 2.15 and also Appendix 2 overleaf on page 2.123].
- Only 80 of the 100 intervals cover the vertical line to the left of μ at 404.368 micrograms, which represents a shift (again, a *bias* in the present context) of 1 standard deviation of the average; the 20 intervals that do not cover this line are marked with a bullet (●) on their left. Thus, a bias whose magnitude is *one* standard deviation of the estimator *does* affect (*i.e.*, *de*crease) substantially the coverage probability.

Computer Simulation of One Hundred Approximate 95% Confidence Intervals for the True Mass (μ) of the U.S. ten-gram standard, NB 10



NOTE: 21. Information about NB 10 has been taken from Freedman, D., Pisani, R., Purves, R. and A. Adhikari: *Statistics.*. 2nd edition, 1991 (W. W. Norton & Co., Inc.), pages 92 and 93; the reader is referred to this source (particularly Chapters 6 and 24) for further information.