Figure 2.17. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Ratio Estimating for an Average or a Total

1. Background

In Figure 2.15, for equiprobable selecting (EPS) of units consisting of individual elements of the respondent population, we developed the theory for estimating a response which is the *ratio* of two random variables: *viz.* $R = \overline{Y}/\overline{X}$; the standard deviation of R and its estimated standard deviation [equations (2.15.11) and (2.15.13) on page 2.115 in Figure 2.15] are shown at the right as equations (2.17.1) and (2.17.2).

$$s.d.(R) \simeq \sqrt{\frac{1}{\overline{\mathbf{X}}^2}\left[\mathbf{S}_{\mathbf{Y}}^2 - 2\mathbf{R}\mathbf{S}_{\mathbf{YX}} + \mathbf{R}^2\mathbf{S}_{\mathbf{X}}^2\right]\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)} \qquad \text{-----(2.17.1)}$$

$$\hat{s.d.}(R) \simeq \sqrt{\frac{1}{(\mathrm{n}-1)\overline{x}^2}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)} \qquad \text{-----(2.17.2)}$$

This theory now provides a basis for so-called *ratio estimating* – measuring the response variate $\mathbf{Y}$ and a *related* response variate $\mathbf{X}$ on each unit selected; under favourable circumstances [which are stated below and discussed in more detail on the third side (page 2.134) of this Figure 2.17], we can then use information on $\mathbf{X}$ (which we can think of as an 'auxiliary' response) to manage the imprecision of estimating $\overline{\mathbf{Y}}$ (the respondent population *average*) and (the closely related theory for estimating) ${}_{\mathrm{T}}\mathbf{Y}$ (the *total*) under EPS. The conditions required are:

- ⁎ $\mathbf{X}$ should be strongly positively correlated with $\mathbf{Y}$ in the respondent population;
- ⁎ there should be information about $\mathbf{X}$ (such as $\overline{\mathbf{X}}$ or ${}_{\mathrm{T}}\mathbf{X}$) already available.

On page 2.115 in Figure 2.15, we defined the notation shown in Table 2.17.1 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 2.17.1: 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. Ratio Estimating: Point Estimates and Standard Deviations

Relevant average and total estimates and estimators in Table 2.17.1 above are now denoted $\overline{y}_R$, ${}_{\mathrm{T}}y_R$, $\overline{Y}_R$ and ${}_{\mathrm{T}}Y_R$.

We have: $\overline{\mathbf{Y}} = \mathbf{R}\overline{\mathbf{X}}$; ${}_{\mathrm{T}}\mathbf{Y} = \mathbf{R}\,{}_{\mathrm{T}}\mathbf{X}$, ${}_{\mathrm{T}}\mathbf{X} = \mathbf{N}\overline{\mathbf{X}}$. -----(2.17.3)

We estimate: $\overline{\mathbf{Y}}$ by $\overline{y}_R = \dfrac{\overline{\mathbf{X}}}{\overline{x}}\overline{y} \equiv \dfrac{{}_{\mathrm{T}}\mathbf{X}}{{}_{\mathrm{T}}x}\overline{y} = \dfrac{\overline{y}}{\overline{x}}\overline{\mathbf{X}} = r\overline{\mathbf{X}}$ [requires knowing $\overline{\mathbf{X}}$ or $\mathbf{N}$ and ${}_{\mathrm{T}}\mathbf{X}$], -----(2.17.4)

and: ${}_{\mathrm{T}}\mathbf{Y}$ by ${}_{\mathrm{T}}y_R = \dfrac{\overline{\mathbf{X}}}{\overline{x}}\,{}_{\mathrm{T}}y \equiv \dfrac{{}_{\mathrm{T}}\mathbf{X}}{{}_{\mathrm{T}}x}\,{}_{\mathrm{T}}y = \dfrac{\overline{y}}{\overline{x}}\,{}_{\mathrm{T}}\mathbf{X} = r\,{}_{\mathrm{T}}\mathbf{X} = r\mathbf{N}\overline{\mathbf{X}} = \mathbf{N}\overline{y}_R$ [requires knowing ${}_{\mathrm{T}}\mathbf{X}$ or $\mathbf{N}$ and $\overline{\mathbf{X}}$]. -----(2.17.5)

Hence: $s.d.(\overline{Y}_R) \simeq \sqrt{\left[\mathbf{S}_{\mathbf{Y}}^2 - 2\mathbf{R}\mathbf{S}_{\mathbf{YX}} + \mathbf{R}^2\mathbf{S}_{\mathbf{X}}^2\right]\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)}$ [using equations (2.17.4) and (2.17.1) and (2.17.2)], -----(2.17.6)

and: $\hat{s.d.}(\overline{Y}_R) \simeq \dfrac{\overline{\mathbf{X}}}{\overline{x}}\sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)}$ where: $\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2 = \sum_{j=1}^{\mathrm{n}}y_j^2 - 2r\sum_{j=1}^{\mathrm{n}}y_jx_j + r^2\sum_{j=1}^{\mathrm{n}}x_j^2$; -----(2.17.7)

also: $s.d.({}_{\mathrm{T}}Y_R) = \mathbf{N}\times s.d.(\overline{Y}_R) \simeq \mathbf{N}\sqrt{\left[\mathbf{S}_{\mathbf{Y}}^2 - 2\mathbf{R}\mathbf{S}_{\mathbf{YX}} + \mathbf{R}^2\mathbf{S}_{\mathbf{X}}^2\right]\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)}$, -----(2.17.8)

and: $\hat{s.d.}({}_{\mathrm{T}}Y_R) = \mathbf{N}\times \hat{s.d.}(\overline{Y}_R) \simeq \dfrac{\mathbf{N}\overline{\mathbf{X}}}{\overline{x}}\sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)} = \dfrac{{}_{\mathrm{T}}\mathbf{X}}{\overline{x}}\sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right)}$. -----(2.17.9)

3. Ratio Estimating: Estimating bias

Equations (2.17.10) and (2.17.11) at the right are equations (2.15.7) and (2.15.8) for the (approximate) estimating bias and *relative* estimating bias of R from page 2.114 in Figure 2.15. Using the expression for $\overline{y}_R$ given in equation (2.17.4), the equivalent expressions for the ratio estimator of $\overline{\mathbf{Y}}$ (the respondent population *average*) are as shown at the right as equations (2.17.12) and (2.17.15); note that the expressions for the *relative* estimating bias of R and of $\overline{Y}_R$ are the *same.* Also, equation (2.15.22) can be written as equation (2.17.16) (where, we recall, the *equality* holds when $\rho_{\mathbf{YX}}^2 = 1$).

$$e.b.(R) \simeq \frac{1}{\overline{\mathbf{X}}^2}\left[\mathbf{R}\mathbf{S}_{\mathbf{X}}^2 - \mathbf{S}_{\mathbf{YX}}\right]\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right) \equiv \frac{1}{(\mathbf{N}-1)\overline{\mathbf{X}}^2}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{X}_{\mathrm{i}}(\mathbf{R}\mathbf{X}_{\mathrm{i}} - \mathbf{Y}_{\mathrm{i}})\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right) \qquad \text{-----(2.17.10)}$$

$$r.e.b.(R) \equiv \frac{e.b.(R)}{\mathbf{R}} \simeq \frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}\frac{\mathbf{X}_{\mathrm{i}}}{\overline{\mathbf{X}}}\left(\frac{\mathbf{X}_{\mathrm{i}}}{\overline{\mathbf{X}}} - \frac{\mathbf{Y}_{\mathrm{i}}}{\overline{\mathbf{Y}}}\right)\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right) \qquad \text{-----(2.17.11)}$$

$$e.b.(\overline{Y}_R) \simeq \frac{1}{\overline{\mathbf{X}}}\left[\mathbf{R}\mathbf{S}_{\mathbf{X}}^2 - \mathbf{S}_{\mathbf{YX}}\right]\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right) = \frac{1}{(\mathbf{N}-1)\overline{\mathbf{X}}}\sum_{\mathrm{i}=1}^{\mathbf{N}}\mathbf{X}_{\mathrm{i}}(\mathbf{R}\mathbf{X}_{\mathrm{i}} - \mathbf{Y}_{\mathrm{i}})\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right) \qquad \text{-----(2.17.12)}$$

$$\hat{e.b.}(\overline{Y}_R) \simeq \frac{\overline{\mathbf{X}}}{\overline{x}^2}(rs_x^2 - s_{yx})\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right) = \frac{\overline{\mathbf{X}}}{(\mathrm{n}-1)\overline{x}^2}\sum_{j=1}^{\mathrm{n}}x_j(rx_j - y_j)\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right) \qquad \text{-----(2.17.13)}$$

$$\text{[where: } e.b.(\overline{Y}_R) = E(\overline{Y}_R) - \overline{\mathbf{Y}} \text{ and } \sum_{j=1}^{\mathrm{n}}x_j(rx_j - y_j) = r\sum_{j=1}^{\mathrm{n}}x_j^2 - \sum_{j=1}^{\mathrm{n}}y_jx_j\text{]} \qquad \text{-----(2.17.14)}$$

$$r.e.b.(\overline{Y}_R) \equiv \frac{e.b.(\overline{Y}_R)}{\overline{\mathbf{Y}}} \simeq \frac{1}{\mathbf{N}-1}\sum_{\mathrm{i}=1}^{\mathbf{N}}\frac{\mathbf{X}_{\mathrm{i}}}{\overline{\mathbf{X}}}\left(\frac{\mathbf{X}_{\mathrm{i}}}{\overline{\mathbf{X}}} - \frac{\mathbf{Y}_{\mathrm{i}}}{\overline{\mathbf{Y}}}\right)\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}\right) \qquad \text{-----(2.17.15)}$$

$$\frac{|e.b.(\overline{Y}_R)|}{s.d.(\overline{Y}_R)} \equiv \frac{|e.b.({}_{\mathrm{T}}Y_R)|}{s.d.({}_{\mathrm{T}}Y_R)} \leq c.v.(\overline{X}) \qquad \text{-----(2.17.16)}$$

(*continued overleaf*)

NOTE: 1. By analogy with *estimating a ratio* in Figure2.15, we incorporate into *ratio estimating* the following checks:

- $*$ 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 ratio estimating;
- $*$ if the value found for the estimate of $c.v.(\overline{X})$ in the previous check is *greater* than 0.1, we use equation (2.17.13) overleaf on page 2.131 to estimate the estimating bias of $\overline{Y}_R$ – the previous check only involves an (estimated) *upper bound* for the magnitude of the ratio $e.b.(\overline{Y}_R)/s.d.(\overline{Y}_R)$, 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).

Example 2.17.1: A company that sells television sets has 123 district offices; we define the response $\mathbf{Y}_i$ to be the earnings in office i in the first quarter of *this* year and $\mathbf{X}_i$ to be the earnings in office i in the first quarter of *last* year. It is known that the total earnings of the company in the first quarter of last year were ${}_T\mathbf{X} = 128{,}200$ in suitable units (so that $\overline{\mathbf{X}} = 128{,}200/123 \simeq 1{,}042.276\,423$ units).

For n = 13 of the company's offices obtained by equiprobable selecting, the sales data (in the same units) for the two periods were:

Table 2.17.2:

j	1	2	3	4	5	6	7	8	9	10	11	12	13
y_j	610	780	1,600	1,030	600	1,050	977	1,440	1,570	2,210	980	865	1,710
x_j	550	720	1,500	1,020	620	980	928	1,200	1,350	1,750	670	729	1,530;

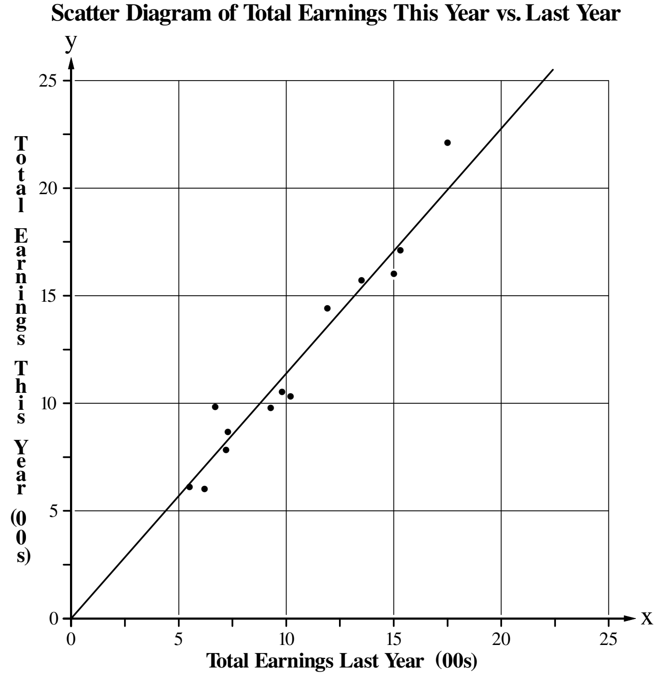
for these data: $\sum_{j=1}^{n} y_j = 15{,}422$, $\sum_{j=1}^{n} y_j^2 = 21{,}073{,}754$; $\sum_{j=1}^{n} x_j = 13{,}547$, $\sum_{j=1}^{n} x_j^2 = 15{,}963{,}525$; $\sum_{j=1}^{n} y_j x_j = 18{,}273.841$, n = 13.

Find an approxiate 99% confidence interval for ${}_T\mathbf{Y}$, the total earnings of the company in the first quarter of this year, using: (a) ratio (*i.e.*, *bi*variate) estimating; (b) *uni*variate estimating based *only* on the y values.

Solution: (a) This question involves ratio estimating of ${}_T\mathbf{Y}$; we first assess the effect of the estimating bias of ${}_T Y_R$ on the likely accuracy of the *stated* (here, 99%) confidence level of the confidence interval for ${}_T\mathbf{Y}$.

We have: $\mathrm{r} = 15{,}422/13{,}547 = 1.138\,407\,027$ $(= r)$, $\mathbf{N} = 123$, ${}_T\mathbf{X} = 128{,}200$, $\overline{\mathbf{X}} = 1{,}042.276\,423 \simeq \overline{\mathrm{x}} = 1{,}042.076\,923$.

Scatter Diagram of Total Earnings This Year vs. Last Year



The scatter diagram of the sample y_js and x_js is shown at the right; we see that the points all lie quite close to the line with slope r (estimating $\mathbf{R}$) through the origin. Hence, we seem to have 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 13 is likely to have an adverse effect, particularly on precision with its $\mathrm{n}^{-½}$ dependence (compared with n^{-1} for the estimating bias).

To estimate $c.v.(\overline{X})$, we have:

$\overline{\mathrm{x}} = 1{,}042.076\,923$, $\mathrm{s_x} = 392.269\,9982$,

$$\hat{s.d.}(\overline{X}) = \mathrm{s_x}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 102.886\,2182,$$

$$\hat{c.v.}(\overline{X}) = \frac{102.886}{1{,}042.08} = 0.098\,732 \simeq 9.9\%.$$

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

We have: ${}_T y_R = \mathrm{r}\,{}_T\mathbf{X} = 145{,}943.7809 \simeq 145{,}944$ units $(= {}_T y_R)$, ${}_{.01}t_{12}^{*} = 3.05\,454$ for 99% confidence;

also: $$\sum_{j=1}^{n}(\mathrm{y}_j - \mathrm{r}\mathrm{x}_j)^2 = \sum_{j=1}^{n}\mathrm{y}_j^2 - 2\mathrm{r}\sum_{j=1}^{n}\mathrm{y}_j\mathrm{x}_j + \mathrm{r}^2\sum_{j=1}^{n}\mathrm{x}_j^2 = 21{,}073{,}754 - 2\times1.138407\times18{,}723.841 + 1.138407^2\times15{,}963{,}525$$

$$= 41{,}762{,}012.43 - 41{,}606{,}138.02 = 155{,}874.41 \ [= \sum_{j=1}^{n}(y_j - r x_j)^2];$$

(*continued*)

Figure 2.17. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Ratio Estimating of an Average or a Totasl (continued 1)

Solution: (a) so that: $\hat{s.d.}({}_{\mathrm{T}}Y_R) \simeq \frac{{}_{\mathrm{T}}\mathbf{X}}{\bar{x}}\sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} = \frac{128{,}200}{1{,}042.077}\sqrt{\frac{1}{12}(155{,}874.41)(\frac{1}{13} - \frac{1}{123})} = 3{,}677.537\,803.$
(cont.)

Hence, an approximate 99% confidence interval for ${}_{\mathrm{T}}\mathbf{Y}$, the total earnings of the company in the first quarter of this year, is:

$${}_{\mathrm{T}}y_R \pm 3.05454 \times \hat{s.d.}({}_{\mathrm{T}}Y_R) = 145{,}943.78 \pm 3.05454 \times 3{,}677.5378 \implies (134{,}714.59,\ 157{,}176.97) \text{ units}$$
$$\text{or about } (134.770,\ 157{,}200) \text{ units.}$$

Solution: (b) This question involves *uni*variate estimating of ${}_{\mathrm{T}}\mathbf{Y}$ based on only the sample y values – the calculation is like that in Examples 2.3.4 and 2.3.5 on page 2.44 in Figure 2.3.

We have: ${}_{\mathrm{T}}\mathrm{y} = \mathbf{N}\bar{\mathrm{y}} = 123 \times 15{,}422/13 = 145{,}915.8462\ (= {}_{\mathrm{T}}y)$, ${}_{.01}t^*_{12} = 3.05\,454$ for 99% confidence,

$$\mathrm{s_y} = \sqrt{\frac{21{,}073{,}754 - 15{,}422^2/13}{12}} = 481.189\,2187\ (= s_y),$$

so that: $\hat{s.d.}(\bar{Y}) = \mathrm{s_y}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 481.189\,2187\sqrt{\frac{1}{13} - \frac{1}{123}} = 126.208\,3238,$

and: $\mathbf{N} \times \hat{s.d.}(\bar{Y}) = 15{,}523.623\,83$ kg.

Hence, an approximate 99% confidence interval for ${}_{\mathrm{T}}\mathbf{Y}$, the total earnings of the company in the first quarter of this year, is:

$${}_{\mathrm{T}}\mathrm{y} \pm 3.05454 \times \hat{s.d.}({}_{\mathrm{T}}Y) = {}_{\mathrm{T}}\mathrm{y} \pm 3.05454 \times \mathbf{N} \times \hat{s.d.}(\bar{Y}) = 145{,}915.85 \pm 47{,}417.5299 \implies$$
$$(98{,}498.32,\ 193{,}333.38) \text{ units or about } (98{,}400,\ 193{,}400) \text{ units.}$$

NOTES: 2. Comparing the answers in parts (a) and (b) of Example 2.17.1, we see that the ratio $\mathbf{N} \times \hat{s.d.}(\bar{Y}) / \hat{s.d.}({}_{\mathrm{T}}Y_R) \simeq 4.22$, reflecting a substantial *in*crease in precision for *bi*variate (or ratio) over *uni*variate estimating. The reason, apparent from the scatter diagram on the facing page 2.132, is that the points all lie close to the line with slope r (estimating $\mathbf{R}$) through the origin.

- Example 2.17.1 shows that, under favourable conditions [$\mathbf{X}$ positively correlated with $\mathbf{Y}$ in the respondent population, information about $\mathbf{X}$ (such as $\bar{\mathbf{X}}$ or ${}_{\mathrm{T}}\mathbf{X}$) already available], the additional expenditure of resources in collecting information on a second response on each selected unit, in conjunction with taking proper account of this information in the Analysis stage of the FDEAC cycle, can lead to a substantial gain in precision and so reduce the limitation on the answer from this source of model error.

3. As indicated above the scatter diagram on the facing page 2.132, the values of $\bar{\mathbf{X}} = 1{,}042.276\,423$ and $\bar{\mathrm{x}} = 1{,}042.076\,923$ are (very) close in Example 2.17.1; this is why the point estimates of ${}_{\mathrm{T}}\mathbf{Y}$ [${}_{\mathrm{T}}y_R = 145{,}944$ in (a) and ${}_{\mathrm{T}}y = 145{,}916$ in (b)] are also close in value.

4. Efficiency

As in Section 2 on pages 2.108 and 2.109 in Figure 2.14, we can compare formally the precison of ratio (*i.e.*, *bi*variate) estimating with the corresponding *uni*variate estimating based only on the sample y values. Such discussions of efficiency typically involve a *ratio* or *difference* of standard deviations; here, we use a difference and, for algebraic convenience, work with the squares of standard deviation expressions to avoid the algebraic complications introduced by square roots. Using the 'univariate' expression (2.17.17) [given at the right above, which is equation (2.3.9) near the top of page 2.41 in Figure 2.3 but with notation appropriate for this Figure 2.17)] and the 'bvariate' expression (2.17.6) for $s.d.(\bar{Y}_R)$ [near the middle of the first side (page 2.131) of this Figure 2.17], we have:

$$s.d.(\bar{Y}_{un}) = \mathbf{S_Y}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \qquad \text{-----(2.17.17)}$$

$$[s.d.(\bar{Y})]^2 - [s.d.(\bar{Y}_R)]^2 \equiv [s.d.(\bar{Y}_{un})]^2 - [s.d.(\bar{Y}_{bi})]^2 \simeq \big[\mathbf{S}^2_{\mathbf{Y}} - \mathbf{S}^2_{\mathbf{Y}} + 2\mathbf{R}\mathbf{S}_{\mathbf{YX}} - \mathbf{R}^2\mathbf{S}^2_{\mathbf{X}}\big](\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) \qquad \text{-----(2.17.18)}$$

$$= \mathbf{R}\big[2\mathbf{S}_{\mathbf{YX}} - \mathbf{R}\mathbf{S}^2_{\mathbf{X}}\big](\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) = \mathbf{R}\mathbf{S}^2_{\mathbf{X}}\big[2\frac{\mathbf{S}_{\mathbf{YX}}}{\mathbf{S}^2_{\mathbf{X}}} - \mathbf{R}\big](\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) = \mathbf{R}\mathbf{S}^2_{\mathbf{X}}\big[2\mathbf{B}_1 - \mathbf{R}\big](\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) \qquad \text{-----(2.17.19)}$$

$$\equiv \mathbf{R}\big[2\mathbf{S}_{\mathbf{YX}} - \mathbf{R}\mathbf{S}^2_{\mathbf{X}}\big](\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) = \mathbf{R}\mathbf{S}_{\mathbf{X}}\big[2\frac{\mathbf{S}_{\mathbf{YX}}}{\mathbf{S}_{\mathbf{Y}}\mathbf{S}_{\mathbf{X}}} - \mathbf{R}\frac{\mathbf{S}_{\mathbf{X}}}{\mathbf{S}_{\mathbf{Y}}}\big](\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}}) = \mathbf{R}\mathbf{S}_{\mathbf{X}}\big[2\rho_{\mathbf{YX}} - c.v.(\bar{X})/c.v.(\bar{Y})\big]. \qquad \text{-----(2.17.20)}$$

As in equation (2.15.5) near the bottom of page 2.113 in Figure 2.15, we define additional terminology in equation (2.17.21):

$$\rho_{\mathbf{YX}} = \mathbf{S}_{\mathbf{YX}}/\mathbf{S}_{\mathbf{Y}}\mathbf{S}_{\mathbf{X}} \quad \text{and:} \quad \mathbf{B}_1 = \mathbf{S}_{\mathbf{YX}}/\mathbf{S}^2_{\mathbf{X}}, \qquad \text{-----(2.17.21)}$$

where $\rho_{\mathbf{YX}}$ is the **correlation coefficient** of response variates $\mathbf{Y}$ 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;
also, from equations (2.15.14) and (2.15.15) on page 2.115 in Figure 2.15, we obtain equation (2.17.22) at the right.

$$c.v.(\bar{X})/c.v.(\bar{Y}) = \frac{s.d.(\bar{X})}{\bar{\mathbf{X}}}\bigg/\frac{s.d.(\bar{Y})}{\bar{\mathbf{Y}}} = \mathbf{R}\frac{\mathbf{S}_{\mathbf{X}}}{\mathbf{S}_{\mathbf{Y}}} \qquad \text{-----(2.17.22)}$$

When the difference of squared standard deviations [the LHS of equation (2.17.18)] is *positive*, $s.d.(\bar{Y}) \equiv s.d.(\bar{Y}_{un})$ is *greater*

(*continued overleaf*)

than $s.d.(\overline{Y}_R)$, meaning that *bi*variate estimating is *more precise* than *uni*variate estimating. The condition for this situation to arise is shown in several equivalent versions by equations (2.17.19) and (2.17.20). Table 2.17.3 below gives three of these versions, ordered roughly by compactness or interpretability; a fourth version is derived algebraically from the second. There is also an explanatory comment at the right of each version.

Table 2.17.3:

1.	$2\mathbf{S}_{\mathbf{YX}} > \mathbf{R}\mathbf{S}_{\mathbf{X}}^2$	A compact version of the condition but with no appealing interpretation
2.	$2\mathbf{B}_1 > \mathbf{R}$	A compact version of the condition with a geometric interpretation; $\mathbf{B}_1$ (represented by β_1 in the model) is the slope of the least squares regression line of the $\mathbf{Y}_i$ on the $\mathbf{X}_i$
3.	$2\rho_{\mathbf{YX}} > c.v.(\overline{X})/c.v.(\overline{Y})]$	A compact version of the condition in terms of the correlation coefficient of the $\mathbf{Y}_i$ and $\mathbf{X}_i$ – it implies that we need $c.v.(\overline{Y})$ to be (appreciably) *larger* than $c.v.(\overline{X})$
4.	$2(\sum_i \mathbf{X}_i)(\sum_i \mathbf{Y}_i\mathbf{X}_i - \overline{\mathbf{X}}\sum_i \mathbf{Y}_i) > (\sum_i \mathbf{Y}_i)(\sum_i \mathbf{X}_i^2 - \overline{\mathbf{X}}\sum_i \mathbf{X}_i)$	A convenient calculational, but the least compact, version of the condition

Thus, focusing on the *second* version of the condition, we can say that, under equiprobable selecting, ratio (or *bi*variate) estimating is *more* precise than *uni*variate estimating when $2\mathbf{B}_1$ (twice the slope of the least squares regression line of $\mathbf{Y}$ on $\mathbf{X}$ in the respondent population) is greater than $\mathbf{R}$ (the ratio of $\overline{\mathbf{Y}}$ to $\overline{\mathbf{X}}$ in the respondent population).

NOTES: 4. When using a ratio estimator, *in addition to* the two or three checks already described in Note 1 at the top page 2.132, we check (usually on the basis of sample data) that the precision *will* be greater than that of univariate estimating using the *fourth* version given in Table 2.17.3 above, or the *second* version that $2\mathbf{B}_1$ is greater the $\mathbf{R}$; alternatively, a *uni*variate calculation of $[s.d.(\overline{Y})]^2$ (or its estimate) can be carried out.

- The *advantage* of direct univariate calculation of $[s.d.(\overline{Y})]^2$ (or its estimate) is that it shows explicitly the *magnitude* of the difference in precision between the univariate and bivariate calculations; unfortunately, it seems that this magnitude must be inferred with caution from the values on the two sides of any of the four versions in Table 2.17.3 above for the condition for bivariate estimating to be more precise than univariate estimating.
  – Among the four versions, their two numbers in Note 5 below are of greatly differing magnitudes; their *common* feature [(likely reflecting their common origin in equations (2.17.19) and (2.17.20) overleaf on page 2.133] is a *ratio* of 2.095 9545. This ratio being substantially greater than 1 *may* be indicative of the appreciable gain in precison for *bi*variate estimating in Example 2.17.1.

5. For the sample data given in Example 2.17.1, we find that: $s_{yx} = 183{,}577.558$, $s_y^2 = 231{,}543.064$, $s_x^2 = 153{,}875.744$, $r = 1.138\,407\,027$;
hence, estimates of the quantities involved in the four versions in Table 2.17.3 above for the condition for *bi*variate estimating to be more precise than *uni*variate estimating are as follows:

1.	$2s_{yx} > r\,s_x^2$	367,155.12 > 175,173.23
2.	$2\hat{\beta}_1 > r$	2.386 049 33 > 1.138 407 03
3.	$2\hat{\rho}_{\mathbf{YX}} > \hat{c.v.}(\overline{X})/\hat{c.v.}(\overline{Y})$	1.945 129 91 > 0.928 040 142
4.	$2(\sum_j x_j)(\sum_j y_j x_j - \overline{x}\sum_j y_j) > (\sum_j y_j)(\sum_j x_j^2 - \overline{x}\sum_j x_j)$	59,686,204,223 > 28,476,860,641

These values show, as already demonstrated in part (b) of Example 2.17.1, that *bi*variate estimating is more precise than *uni*variate estimating under the conditions prevailing in Example 2.17.1 with respect to the strength of the positive correlation between $\mathbf{Y}$ and $\mathbf{X}$ (here, estimated to be 0.9726).

EXERCISE: Show that versions 2 and 3 of the condition given in Table 2.17.3 above are equivalent to version 1, and derive version 4 from version 2.

Example 2.17.2: To estimate the total sugar content of a truck load of oranges, ten oranges were obtained by equiprobable selecting from the load; these oranges were then weighed, their juice was extracted and its sugar content was measured. The numerical summaries of the data were as follows:

sugar content – sum: 0.1116 kg,	sum of squares: 0.001 280 kg^2;	sum of cross	
weight – sum: 1.973 kg,	sum of squares: 0.3915 kg^2;	products: 0.02229 kg^2.	

If the total weight of the load of oranges is 820 kg, find an approximate 95% confidence interval for ${}_T\mathbf{Y}$, the *total* sugar content of the load, using:

(a) ratio (*i.e.*, *bi*variate) estimating;
(b) *uni*variate estimating based only on the data for the sugar content.

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

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 above, we *can* do the check involving $c.v.(\overline{X})$.

(*continued*)

Figure 2.17. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Ratio Estimating of an Average or a Totasl (continued 2)

Solution: (cont.) (a) This question does not provide a value for $\mathbf{N}$ but we can calculate an approximate value as $10 \times 820/1.973 = 4{,}156.107\,45 \simeq 4{,}156$ oranges in the load.

We have: $\mathbf{N} \simeq 4{,}156.$ $\quad \mathrm{n} = 10,$ $\quad \bar{\mathrm{x}} = 0.1973$ kg $(= \bar{x})$, $\quad \mathrm{s_x} = 0.015\;730\;720\,12$ kg $(= s_x)$,

so that: $\hat{s.d.}(\bar{X}) = \mathrm{s_x}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 0.004\;968\;502\;32$ kg,

and: $\hat{c.v.}(\bar{X}) = 0.004\;968\;515/0.1973 = 0.025\;182\,475\;03 \simeq 0.025 = 2.5\%$.

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

We have: $\mathrm{r} = \bar{\mathrm{y}}/\bar{\mathrm{x}} = \sum_{j=1}^{\mathrm{n}}\mathrm{y}_j \Big/ \sum_{j=1}^{\mathrm{n}}\mathrm{x}_j = 0.1116/1.973 = 0.056\;563\;6087\;(= r)$, $\quad {}_{.05}t_9^* = 2.26\;216$ for 95% confidence,

${}_{T}\mathbf{X} = 820$ kg, $\quad {}_{T}\mathrm{y}_R = \mathrm{r}\,{}_{T}\mathbf{X} = 46.382\;159\;14 \simeq 46{,}382$ kg $(= {}_{T}y_R)$;

also: $\sum_{j=1}^{\mathrm{n}}(\mathrm{y}_j - \mathrm{rx}_j)^2 = \sum_{j=1}^{\mathrm{n}}\mathrm{y}_j^2 - 2\mathrm{r}\sum_{j=1}^{\mathrm{n}}\mathrm{y}_j\mathrm{x}_j + \mathrm{r}^2\sum_{j=1}^{\mathrm{n}}\mathrm{x}_j^2 = 0.001\,280 - 2 \times 0.056\;5636 \times 0{,}02229 + 0.056\;5636^2 \times 0.3915$

$= 0.002\;532\;581\,477 - 0.002\;521\,605\;677 = 0.000\;010\;975\;800\;[= \sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2]$;

so that: $\hat{s.d.}({}_{T}Y_R) \simeq \frac{{}_{T}\mathbf{X}}{\bar{x}}\sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - rx_j)^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} = \frac{820}{0{,}1973}\sqrt{\frac{1}{9}(0.000010975)(\frac{1}{10} - \frac{1}{4{,}156})} = 1.449\;641\;114$ kg.

Hence, an approximate 95% confidence interval for ${}_{T}\mathbf{Y}$, the total sugar content of the load of oranges, is:

${}_{T}\mathrm{y}_R \pm 2.26216 \times \hat{s.d.}({}_{T}Y_R) = 46.382159 \pm 2.26216 \times 1.449641 \;\Rightarrow\; (43.1028,\; 49.66157)$ kg
or about $(43.1,\; 49.7)$ kg.

Solution: (b) This question involves *uni*variate estimating of ${}_{T}\mathbf{Y}$ based on only the sample y values – the calculation is like that in Examples 2.3.4 and 2.3.5 on page 2.44 in Figure 2.3.

We have: ${}_{T}\mathrm{y} = \mathbf{N}\bar{\mathrm{y}} = 4{,}156.10745 \times 0.1116/10 = 46.382\;159\;14$ kg $(= {}_{T}y)$, $\quad {}_{.05}t_9^* = 2.26\;216$ for 95% confidence,

$\mathrm{s_y} = \sqrt{\frac{0.001\,280\; - 0{,}1116^2/10}{9}} = 0.001\;959\;138\;132$ kg $(= s_y)$,

so that: $\hat{s.d.}(\bar{Y}) = \mathrm{s_y}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} = 0.001\,959\,138\sqrt{\frac{1}{10} - \frac{1}{4{,}156}} = 0.000\;618\;788\,2759$ kg,

and: $\mathbf{N} \times \hat{s.d.}(\bar{Y}) = 2.571\;845\;495$ kg.

Hence, an approximate 95% confidence interval for ${}_{T}\mathbf{Y}$, the total sugar content of the load of oranges, is:

${}_{T}\mathrm{y} \pm 3.05454 \times \hat{s.d.}({}_{T}Y) = 46.382159 \pm 2.26216 \times 2.571\,845 \;\Rightarrow\; (40.5642,\; 52.2001)$ kg
or about $(40.5,\; 52.2)$ kg.

NOTE: 6. Comparing the answers in parts (a) and (b) of Example 2.17.2, we see that the ratio $\mathbf{N} \times \hat{s.d.}(\bar{Y})/\hat{s.d.}({}_{T}Y_R) \simeq 1.77$, reflecting an *in*crease in precision for *bi*variate (or ratio) over *uni*variate estimating. This implies that, if the individual sample data had been provided. their scatter diagram would show the points lying (reasonably) close to the line with slope r (estimating $\mathbf{R}$) through the origin.

- From the sample summary data in Example 2.17.2, we find that: $\quad \mathrm{s_{yx}} = 0.000\;030\;14\dot{6}$, $\quad \mathrm{s_y^2} = 0.000\;003\;838\dot{2}$, $\quad \mathrm{s_x^2} = 0.000\;247\;4\dot{5}$, $\quad \mathrm{r} = 0.056\;563\;6087$;

  hence, estimates of the quantities involved in the four versions in Table 2.17.3 on the facing page 2.134 for the condition for *bi*variate estimating to be more precise than *uni*variate estimating are as follows:

  1. $2\mathrm{s_{yx}} > \mathrm{rs_x^2}$ $\qquad 0.000\;060\;29\dot{3} > 0.000\;013\;996\;9792$
  2. $2\hat{\beta}_1 > \mathrm{r}$ $\qquad 0.243\;653\;181 > 0.056\;563\;6087$
  3. $2\hat{\rho}_{\mathbf{YX}} > \hat{c.v.}(\bar{X})/\hat{c.v.}(\bar{Y})$ $\qquad 1.956\;390\,909 > 0.454\;172\;181$
  4. $2(\sum_j \mathrm{x}_j)(\sum_j \mathrm{y}_j\mathrm{x}_j - \bar{\mathrm{x}}\sum_j \mathrm{y}_j) > (\sum_j \mathrm{y}_j)(\sum_j \mathrm{x}_j^2 - \bar{\mathrm{x}}\sum_j \mathrm{x}_j)$ $\qquad 0.001\;070\;628\;72 > 0.000\;248\;544\;36$

  These values show, as already demonstrated in part (b) of Example 2.17.2, that *bi*variate estimating is more precise than *uni*variate estimating under the conditions prevailing in this example with respect to the strength of the positive correlation between $\mathbf{Y}$ and $\mathbf{X}$ (here, estimated to be 0.9782).

  – The common right-left ratio of the numbers in the four inequalities at the right above is 4.307 5961 and the s.d. ratio for (a) and (b) in Example 2.17.2 is 1.77; the corresponding figures for Example 2.17.1 are 2.095 9545 and 4.22. This reminds us of the caution issued in the second paragraph of Note 4 on page 2.134.

(*continued overleaf*)

NOTE: (cont.) 6. In Example 2.17.2, lack of an (explicit) value for $\mathbf{N}$ in the statement of the question is reminiscent of Example 2.16.1 on pages 2.127 and 2.128 in Figure 2.16. In that situation involving *un*equal-sized clusters, whether $\mathbf{N}$ is known or *un*known affected values for point estimates and their estimated standard deviations for bi- and univariate estimating – see Table 2.16.4 at the bottom of page 2.128.

The situation in Example 2.17.2 above differs from that in Example 2.16.1 in that:

- an approximate value of $\mathbf{N}$ can be calculated and, being more compact, is more convenient to use than its source: $10 \times 820/1.973$;
- use of this approximate value for $\mathbf{N}$ makes the univariate point estimate of ${}_{\mathrm{T}}\mathbf{Y}$ in (b) the *same* value as in (a).

Example 2.17.3: A company wishes to estimate $\overline{\mathbf{Y}}$, the average amount of money paid to its employees for dental expenses during the current quarter. From the previous year's annual report, it is known that the 1,000 employees of the company received a total of \$125,000 (${}_{\mathrm{T}}\mathbf{X}$). From the health records of 100 employees obtained by equiprobable selecting, data for dental expenses (in dollars) during the two quarters gave the following numerical summaries:

$$\sum_{j=1}^{n} y_j = 17{,}500 \qquad \sum_{j=1}^{n} y_j^2 = 3{,}165{,}000; \qquad \sum_{j=1}^{n} x_j = 12{,}000, \qquad \sum_{j=1}^{n} x_j^2 = 1{,}562{,}000; \qquad \sum_{j=1}^{n} y_j x_j = 2{,}205{,}935.$$

Find an approximate 90% confidence interval for $\overline{\mathbf{Y}}$, the *average* amount paid by the company to its employees for dental expenses, using:

(a) ratio (*i.e.*, *bi*variate) estimating;

(b) *uni*variate estimating based only on the data for the *current* quarter.

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

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 above, we *can* do the check involving $c.v.(\overline{X})$.

We have: $\mathbf{N} = 1{,}000, \quad n = 100, \quad \overline{x} = 12{,}000/100 = \$120.00\ (= \overline{x}),$

$$s_x = \sqrt{\frac{1{,}562{,}000 - 12{,}000^2/100}{99}} = \$35.104\,461\,72\ (= s_x),$$

so that: $\hat{s.d.}(\overline{X}) = s_x\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = \$3.330\,301\,652,$

and: $\hat{c.v.}(\overline{X}) = 3.330\,301\,652/1{,}200 = 0.027\,752\,513 \simeq 0.028 = 2.8\%.$

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

We have: $r = \overline{y}/\overline{x} = \sum_{j=1}^{n} y_j \Big/ \sum_{j=1}^{n} x_j = 17{,}500/12{,}000 = 1.458\dot{3}\ (= r), \qquad {}_{.1}t^*_{99} = 1.66\,039$ for 90% confidence,

$${}_{\mathrm{T}}\mathbf{X} = \$125{,}000, \qquad \overline{y}_R = r\overline{\mathbf{X}} = 1.458\dot{3} \times 125 = \$182.291\dot{6}\ (= \overline{y}_R);$$

also:

$$\sum_{j=1}^{n}(y_j - rx_j)^2 = \sum_{j=1}^{n} y_j^2 - 2r\sum_{j=1}^{n} y_j x_j + r^2\sum_{j=1}^{n} x_j^2 = 3{,}165{,}080 - 2 \times 0.458\dot{3} \times 3{,}205{,}935 + 1.458\dot{3}^2 \times 1{,}562{,}000$$

$$= 6{,}486{,}961.804 - 6{,}433{,}977.082 = 52{,}984.7221\ [= \sum_{j=1}^{n}(y_j - rx_j)^2],$$

so that:

$$\hat{s.d.}(\overline{Y}_R) \simeq \frac{\overline{\mathbf{X}}}{\overline{\overline{x}}}\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - rx_j)^2\left(\frac{1}{n} - \frac{1}{\mathbf{N}}\right)} = \frac{125}{120}\sqrt{\frac{1}{99}(52{,}984.7221)\left(\frac{1}{100} - \frac{1}{1{,}000}\right)} = \$2.286\,165\,964.$$

Hence, an approximate 90% confidence interval for $\overline{\mathbf{Y}}$, the *average* amount paid by the company to its employees for dental expenses, is:

$$\overline{y}_R \pm 1.66039 \times \hat{s.d.}(\overline{Y}_R) = 182.2917 \pm 1.66039 \times 2.286166 \implies (178.4958,\ 186.0876)$$

or about (\$178, \$187).

Solution: (b) This question involves *uni*variate estimating of $\overline{\mathbf{Y}}$ based on only the sample y values – the calculation is like that in Examples 2.3.2 and 2.3.3 on page 2.43 in Figure 2.3.

We have: $\overline{y} = 17{,}500/100 = \$175.00\ (= \overline{y}), \qquad {}_{.1}t^*_{99} = 1.66\,039$ for 90% confidence;

$$s_y = \sqrt{\frac{3{,}165{,}000 - 17{,}500^2/100}{9}} = \$32.176\,909\,97\ (= s_y),$$

so that: $\hat{s.d.}(\overline{Y}) = s_y\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = 32.176\,909\,97\sqrt{\frac{1}{100} - \frac{1}{1{,}000}} = \$3.052\,569\,707.$

(*continued*)

Figure 2.17. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Ratio Estimating of an Average or a Totasl (continued 3)

Solution: (cont.) (b) Hence, an approximate 90% confidence interval for $\overline{\mathbf{Y}}$, the *average* amount paid by the company to its employees for dental expenses, is:

$$\overline{y} \pm 1.66039 \times \hat{s.d.}(\overline{Y}) = 175.00 \pm 1.66039 \times 3.052\,570 \;\Rightarrow\; (169.9315,\ 180.0685)$$

or about ($169, $181).

NOTES: 7. Comparing the answers in parts (a) and (b) of Example 2.17.3, we see that the ratio $\hat{s.d.}(\overline{Y})/\hat{s.d.}(\overline{Y}_R) \simeq 1.34$, reflecting an *in*crease in precision for *bi*variate (or ratio) over *uni*variate estimating. This implies that, if the individual sample data had been provided. their scatter diagram would show the points generally lying (reasonably) close to the line with slope r (estimating $\mathbf{R}$) through the origin.

- From the sample summary data in Example 2.17.3, we find that: $s_{yx} = 1{,}070.0\dot{0}\dot{5}$, $s_x^2 = 1{,}232.\dot{3}\dot{2}$, $s_y^2 = 1{,}035.\dot{3}\dot{5}$, $r = 1.458\dot{3}$;
  hence, estimates of the quantities involved in the four versions in Table 2.17.3 on page 2.134 for the condition for *bi*variate estimating to be more precise than *uni*variate estimating are as follows:

1.	$2s_{yx} > rs_x^2$	$2{,}140.\dot{1}\dot{0} > 1{,}797.138\,046$
2.	$2\hat{\beta}_1 > r$	$1.736\,639\,345 > 1.458\dot{3}$
3.	$2\hat{\rho}_{\mathbf{YX}} > \hat{c.v.}(\overline{X})/\hat{c.v.}(\overline{Y})$	$1.894\,644 > 1.591\,016\,869$
4.	$2(\sum_j x_j)(\sum_j y_j x_j - \overline{x}\sum_j y_j) > (\sum_j y_j)(\sum_j x_j^2 - \overline{x}\sum_j x_j)$	$2{,}542{,}440{,}000 > 2{,}135{,}000{,}000$

  These values show, as already demonstrated in part (b) of Example 2.17.3, that *bi*variate estimating is somewhat more precise than *uni*variate estimating under the conditions prevailing in this example with respect to the strength of the positive correlation between $\mathbf{Y}$ and $\mathbf{X}$ (here, estimated to be 0.9473).

  – The common right-left ratio of the numbers in the four inequalities at the right above is 1.190 8384 and the s.d. ratio for (a) and (b) in Example 2.17.3 is 1.34; the corresponding figures for Example 2.17.1 are 2.095 9545 and 4.22 and for Example 2.17.2 they are 4.307 5961 and 1.77. This again reminds us of the caution issued in the second paragraph of Note 4 on page 2.134.

8. As well as a difference in precision, the point estimates for the average in Example 2.17.3 differ: in (a), $\overline{y}_R = \$182.291\dot{6}$, in (b), $\overline{y} = \$175.00$. These values differ more than those in Example 2.17.1 because the values of $\overline{\mathbf{X}} = 125$ and $\overline{x} = 120$ differ more – recall the second term in equation (2.17.4) and Note 3 near the middle of page 2.133.

5. Appendix 1: How Does Ratio Estimating Increase Precision?

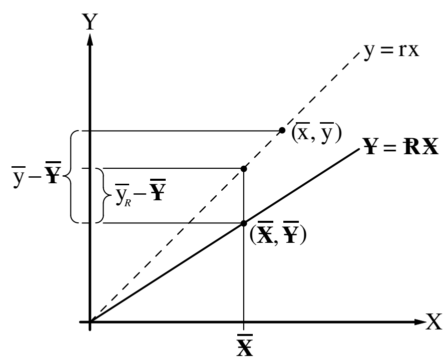
Equation (2.17.4) [shown again at the right] is notable for two reasons:

$$\overline{y}_R = \frac{\overline{\mathbf{X}}}{\overline{x}}\overline{y} \equiv \frac{{}_T\mathbf{X}}{{}_Tx}\overline{y} = \frac{\overline{y}}{\overline{x}}\overline{\mathbf{X}} = r\overline{\mathbf{X}} \qquad \text{-----(2.17.4)}$$

* on its right-hand side, the four different arrangements of its three components summarize why and how ratio estimating of an average works,
* the first arrangement shows *why* it works [when $\mathbf{Y}$ (the response variate of primary interest) and $\mathbf{X}$ (an 'auxilliary' response variate for which relevant population attribute values are known) have roughly a straight-line relationship through the origin].

Informally, the ratio $\overline{\mathbf{X}}/\overline{x}$ is taken as a measure (in aggregate) of how well the sample x-values reflect the population $\mathbf{X}$-values and, because of the $\mathbf{X}$-$\mathbf{Y}$ relationship, of how well the sample y-values are likely to reflect the population $\mathbf{Y}$-values. Whether $\overline{x}$ is greater than or less than $\overline{\mathbf{X}}$, multiplying $\overline{y}$ (the *uni*variate sample average) by the ratio $\overline{\mathbf{X}}/\overline{x}$ is likely to bring it closer to $\overline{\mathbf{Y}}$. [In the (unlikely) situation that $\overline{\mathbf{X}} = \overline{x}$, $\overline{y}_R = \overline{y}$.]

Formally, we use the diagram at the right, in which the solid line shows the [idealized (model)] $\mathbf{X}$-$\mathbf{Y}$ relationship in the population and the dashed line the corresponding x-y relationship in the sample – the substantial difference in the slopes of the two lines is only to make the diagram clearer. Both lines (of course) go through their respective point of averages.



The third arrangement of three components on the right-hand side of equation (2.17.4) shows that the point on the (dashed) sample line at X = $\overline{\mathbf{X}}$ has ordinate $\overline{y}_R$ so that the difference $\overline{y}_R - \overline{\mathbf{Y}}$ is *smaller* than $\overline{y} - \overline{\mathbf{Y}}$; this is why, under repetition, $\overline{Y}_R$ is a more precise estimator of $\overline{\mathbf{Y}}$ than is (the univariate) $\overline{Y}$.

The diagram at the right is drawn with the *sample* line having greater slope so $\overline{x}$ is larger than $\overline{\mathbf{X}}$ and $\overline{y}$ larger than $\overline{\mathbf{Y}}$. The demonstration from the diagram is (of course) *un*affected had it been drawn with the sample line having the *smaller* slope.

6. Appendix 2: The Sample Ratio as a Linear Model Slope Parameter

The following theory is relevant to this Figure 2.17 and also to Figures 2.15 and 2.16.

(*continued overleaf*)

The response model of equation (2.17.23) at the right is a straight line with slope β through the origin; it is contingent on the sample x_js so x is *not* a random variable. Y_j (the response) and R_j (the residual) *are* random variables, the latter being modelled by a normal distribution with mean 0 and s.d. $\sqrt{x_j}\sigma$.

$$Y_j = \beta x_j + R_j, \quad j = 1, 2, ..., n, \quad R_j \sim N(0, \sqrt{x_j}\sigma), \text{ prob. independent, EPS.} \qquad \text{-----(2.17.23)}$$

A linear estimator $B \equiv \tilde{\beta}$ (a random variable) of the model slope parameter β in equation (2.17.23) is equation (2.17.24), in which c_j depends on x_j. From the model (2.17.23), we obtain equations (2.17.25) and (2.17.26) for the mean and standard deviation of $\tilde{\beta}$ – for later algebraic convenience, we work with the *square* of the latter..
The last term of equation (2.17.25) shows that *un*biasedness of $\tilde{\beta}$ requires the constraint of equation (2.17.27).

$$B \equiv \tilde{\beta} = \sum_{j=1}^{n} c_j Y_j \qquad \text{-----(2.17.24)}$$

$$E(\tilde{\beta}) = \sum_{j=1}^{n} c_j E(Y_j) = \beta \sum_{j=1}^{n} c_j x_j \qquad \text{-----(2.17.25)}$$

$$[s.d.(\tilde{\beta})]^2 = \sum_{j=1}^{n} c_j^2 [s.d.(Y_j)]^2 = \sigma^2 \sum_{j=1}^{n} c_j^2 x_j \qquad \text{-----(2.17.26)}$$

$$\sum_{j=1}^{n} c_j x_j = 1 \qquad \text{-----(2.17.27)}$$

To minimize the standard deiation of $\tilde{\beta}$, we use the method of Lagrange multipliers and the function in equation (2.17.28), which is equation (2.17.26) plus λ times the constraint (2.17.27).

$$f(c_j, \lambda) = \sigma^2 \sum_{j=1}^{n} c_j^2 x_j + \lambda\Big(\sum_{j=1}^{n} c_j x_j - 1\Big) \qquad \text{-----(2.17.28)}$$

The partial derivative with respect to c_j of $f(c_j, \lambda)$, and the result (which no longer involves subscript j) that follows from setting it to zero, are shown in equation (2.17.29).

$$\frac{\partial f(c_j, \lambda)}{\partial c_j} = 2\sigma^2 c_j x_j + \lambda x_j \text{ equals 0 when } c_j = -\lambda/2\sigma^2 \qquad \text{-----(2.17.29)}$$

The partial derivative with respect to λ of $f(c_j, \lambda)$ is shown in equation (2.17.30) [= (2.17.27)]. Setting this to zero, substituting for c_j from equation (2.17.29), solving for λ and substituting this back in the expesssion for c_j in equation (2.17.29), yields the results in equation (2.17.31).

$$\frac{\partial f(c_j, \lambda)}{\partial \lambda} = \sum_{j=1}^{n} c_j x_j - 1 \qquad \text{-----(2.17.30)}$$

$$-\lambda/2\sigma^2 \sum_{j=1}^{n} x_j = 1 \quad \text{so that} \quad -\lambda = 2\sigma^2 \Big/ \sum_{j=1}^{n} x_j \quad \text{and} \quad c_j = 1 \Big/ \sum_{j=1}^{n} x_j \qquad \text{-----(2.17.31)}$$

Using the last expression in equation (2.17.31) in equation (2.17.24), we obtain equation (2.17.32), which shows that, under EPS, $\tilde{\beta}$ is an unbiased minimum standard deviation estimator of random variable R, the sample ratio of the averages $\overline{Y}$ and $\overline{x}$.

$$\tilde{\beta} = \sum_{j=1}^{n} c_j Y_j = \sum_{j=1}^{n} Y_j \Big/ \sum_{j=1}^{n} x_j = \sum_{j=1}^{n} Y_j/n \Big/ \sum_{j=1}^{n} x_j/n = \overline{Y}/\overline{x} = R \qquad \text{-----(2.17.32)}$$

These matters are also discussed by Cochran on pages 158-160, with the acronym BLUE for *best linear unbiased estimator.*

The foregoing derivation shows that optimum estimating conditions for ratio estimating occur when the deviations of the (sample) data from the line with slope r through the origin are proportional to $\sqrt{x_j}$, the weight of standard deviation σ in the model (2.17.23); the resulting spread of the points would appear 'trumpet-shaped' (*i.e.*, with *curved* 'boundaries') about the line.

- The unhelpful traditional emphasis in statistics teaching on variance rather than standard deviation makes it easy to *mis*interpret the optimum shape of deviations about the line as *wedge-shaped* (with *straight* 'boundaries') rather than trumpet-shaped.

NOTES: 9. The derivation in Appendix 2 above is a rare instance in these Materials of the same symbol – here, upper-case *italic R* – being used for the random variable representing two *different* quantities:

- the residual R_j in the model (2.17.23) AND:
  the sample ratio $R = \overline{Y}/\overline{x}$ in equation (2.17.32).

Confusing these entities here is unlikely (particularly with the residual being subscripted) and is outweighed by:

+ the evocative R for 'residual,'
+ the evocative 'residual' for the stochastic term in a regression model, instead of the more usual 'error,' which:
  – invites confusion with our (evocative) use of 'error' – recall Figure 1.4 in these Course Materials, AND:
  – is often denoted ε which (without the ~ of STAT 231 notation) looks like a model parameter, *not* a random variable.

10. Regression modelling like equation (2.17.23) is usually taken as contingent on the x-values of the particular sample, which are regarded as *fixed* and so not modelled by a random variable. This assumption is reflected by use throughout the derivation above of (lower-case Roman) x for the x_js, in contrast to the (upper case *italic* random variable) Y_js.
This basis for regression modelling *differs* from that in Figure 2.15, which also provides the theory for Figures 2.16 and 2.17 – in these three Figures, the sample y- *and* x-values are modelled by random variables.

- In Figure 2.15, **Y** and **X** are of *equal* interest in the ratio of their averages.
- In Figures 2.16 and 2.17, the primary interest is **Y** with **X** involved for other reasons.
  – In this Figure 2.17, there is also the matter of information about the *population* **X**-values, in the form of $\overline{\mathbf{X}}$ or $_{T}\mathbf{X}$.

7. REFERENCE:

Cochran, W. G. *Sampling Techniques.* John Wiley & Sons, Inc., New York, 3rd Edition, 1977.