3 Ratio and regression estimators

3.1 Motivating examples

Frequently, we are interested in measuring the *ratio* of a matched pair of variables. This occurs when the *sampling unit* comprises a *group* or *cluster* of individuals, and our interest is in the population mean per individual.

For example, to estimate average income/adult in the population in a household survey, we record for the ith household ($i = 1, \cdots, n$) the number of adults who live there, x_i, and the household income, y_i.

Then the parameter, average income per adult in the population,

$$R = \frac{\text{household income}}{\text{total no. of adults}} = \frac{\sum\limits_{i=1}^{N} Y_i}{\sum\limits_{i=1}^{N} X_i}$$

can be estimated by the *ratio* estimator

$$\boxed{\widehat{R} = r = \frac{\sum\limits_{i=1}^{n} y_i}{\sum\limits_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}.}$$

Relationship between estimates
Ratio Mean Total

$$\begin{tikzcd} R \arrow[r, "\times \overline{X}"] & \overline{Y} \arrow[r, "\times N"] & Y \\ R \arrow[rr, "\times X"] & & Y \end{tikzcd}$$

3.2 Two characteristics per unit in SRS

Theorem: If X_i and Y_i are a pair of numerical characteristics defined on every unit of the population, and $\bar{y}$ and $\bar{x}$ are the corresponding means from a SRS without replacement of size n, then

$$\mathrm{Cov}\,(\bar{x},\bar{y}) = \left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})(X_i-\bar{X})}{N-1}\right] = \left(1-\frac{n}{N}\right)\frac{S_{xy}}{n} \tag{1}$$

and

$$E\left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{n-1}\right) = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})(X_i-\bar{X})}{N-1}. \tag{2}$$

Proof. Consider $U_i = X_i + Y_i$ and the corresponding sample values are $u_i = x_i + y_i$. Clearly

$$\begin{aligned}\mathrm{Var}\,(\bar{u}) &= \left(1-\frac{n}{N}\right)\frac{S_U^2}{n} = \left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{\sum_{i=1}^{N}(X_i-\bar{X}+Y_i-\bar{Y})^2}{N-1}\right]\\ &= \left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2+\sum_{i=1}^{N}(Y_i-\bar{Y})^2+2\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{N-1}\right]\\ &= \mathrm{Var}\,(\bar{x})+\mathrm{Var}\,(\bar{y})+\frac{2}{n}\left[\frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{N-1}\right]\left(1-\frac{n}{N}\right).\end{aligned}$$

Since $\mathrm{Var}\,(\bar{u}) = \mathrm{Var}\,(\bar{x}+\bar{y}) = \mathrm{Var}\,(\bar{x})+\mathrm{Var}\,(\bar{y})+2\mathrm{Cov}(\bar{x},\bar{y})$, (1) is proved. (2) can be proved in a similar way.

Theorem: For large sample,

(a) $E(r)-R\approx 0$, *approximately unbiased*,

(b) $$\mathrm{Var}(r)\approx\frac{1}{\bar{X}^2}\left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{\sum_{i=1}^{N}(Y_i-RX_i)^2}{N-1}\right]=\frac{1}{\bar{X}^2}\left(1-\frac{n}{N}\right)\frac{S_r^2}{n}.$$

Proof:

(a) Recall $E(\bar{y}) = \bar{Y}$, $E(\bar{x}) = \bar{X}$ and $\text{Var}(\bar{x}) = O(n^{-1})$ (order of n^{-1}). Thus for large sample,

$$E(r) = E\left(\frac{\bar{y}}{\bar{x}}\right) \approx \frac{E(\bar{y})}{\bar{X}} = R.$$

(b) Note that

$$r - R = \frac{\bar{y}}{\bar{x}} - R \approx \frac{\bar{y} - R\bar{x}}{\bar{X}}.$$

Thus, for large sample,

$$\text{Var}(r) = E[(r-R)^2] \approx \frac{1}{\bar{X}^2} E[(\bar{y} - R\bar{x})^2] = \frac{E(\bar{d}^2)}{\bar{X}^2} = \frac{\text{Var}(\bar{d})}{\bar{X}^2}$$

where $\bar{d} = \bar{y} - R\bar{x}$ is the sample mean of $d_i = y_i - Rx_i, i = 1, \cdots, n$, drawn from the population of $D_i = Y_i - RX_i,\ i = 1, \cdots, N$ with

$$E(\bar{d}) = E(\bar{y} - R\bar{x}) = E(\bar{y}) - RE(\bar{x}) = \bar{Y} - R\bar{X} = \bar{Y} - \frac{\bar{Y}}{\bar{X}}\bar{X} = 0.$$

For a SRS of d_i,

$$\text{Var}(\bar{d}) = \left(1 - \frac{n}{N}\right)\frac{S_r^2}{n}$$

where

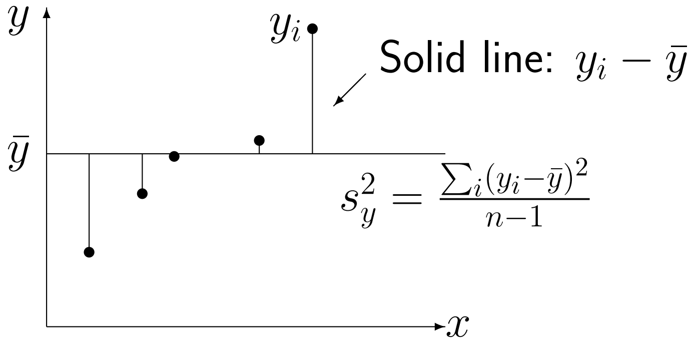
$$S_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}(D_i - \bar{D})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2.$$

Hence

$$\text{Var}(r) \approx \frac{1}{\bar{X}^2}\left(1 - \frac{n}{N}\right)\frac{1}{n}\left[\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2\right] = \frac{1}{\bar{X}^2}\left(1 - \frac{n}{N}\right)\frac{S_r^2}{n},$$

$$\boxed{\text{var}(r) \approx \frac{1}{\bar{X}^2}\left(1 - \frac{n}{N}\right)\frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2\right] = \frac{1}{\bar{X}^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}}$$

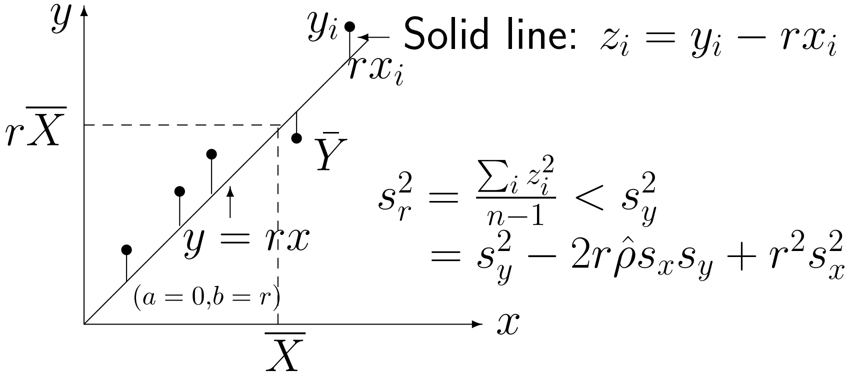
1. Ordinary: x not related to y $\quad \hat{\bar{Y}} = \bar{y}$ & $\text{var}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}$



Solid line: $y_i - \bar{y}$

$s_y^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$

2. Ratio: x positively related to y $\quad \hat{\bar{Y}}_r = \bar{y}\frac{\bar{X}}{\bar{x}}$ & $\text{var}(\hat{\bar{Y}}_r) = \left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}$



Solid line: $z_i = y_i - rx_i$

$$s_r^2 = \frac{\sum_i z_i^2}{n-1} < s_y^2$$
$$= s_y^2 - 2r\hat{\rho}s_x s_y + r^2 s_x^2$$

Calculation of s_r^2:

$$\begin{aligned}
s_r^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[(y_i - \bar{y}) - r(x_i - \bar{x})]^2 \quad \text{since } \bar{y} - r\bar{x} = \bar{y} - \frac{\bar{y}}{\bar{x}}\bar{x} = 0 \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2r\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + r^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= s_y^2 - 2r\,s_{xy} + r^2\,s_x^2 = s_y^2 - 2r\,\hat{\rho}s_x s_y + r^2\,s_x^2
\end{aligned}$$

$$\text{or } s_r^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2 - 2r\sum_{i=1}^{n}x_i y_i + r^2\sum_{i=1}^{n}x_i^2\right).$$

Remark:

1. If X_i and Y_i are positively related, we have $s_r^2 \ll s_y^2$. Hence X_i can be used as an *auxiliary* variable which provides additional information and hence improves the precision of the estimate $\bar{Y}$.

2. When $\overline{X}$ is replaced by $\overline{x}$ if it is unknown, ordinary estimator results.

3. When ratio estimation is used, estimates of variance and sample size are quite sensitive to data points that do not fit the ideal pattern called *influential observation*. It is important to plot the data and look for these unusual data points before proceeding with an analysis.

4. The *'ratio of means'* $\widehat{R} = \frac{\overline{y}}{\overline{x}}$ is biased and can be almost unbiased if n is large. Another ratio estimator is the *'mean of ratios'* $\widehat{R}^* = \overline{r}^* = \frac{1}{n}\sum\limits_{i=1}^{n} \frac{y_i}{x_i}$ where $r_i^* = \frac{y_i}{x_i}$ is unbiased for $R^* = \frac{1}{N}\sum\limits_{i=1}^{N} \frac{y_i}{x_i}$. However $\widehat{R}^*$ gives equal weight to each cluster which may vary greatly in size. Unlike $\widehat{R}^*$, $\widehat{R}$ is weighed by the cluster size which is an advantage over $\widehat{R}^*$.

3.3 Ratio estimate for population mean and total

The ratio estimator of the *population total* Y is

$$\boxed{\widehat{Y}_r = \frac{\bar{y}}{\bar{x}} X = rX}$$

Similarly, the *ratio* estimator of *population mean* is

$$\boxed{\widehat{\bar{Y}}_r = \frac{\bar{y}}{\bar{x}} \bar{X} = r\bar{X}}$$

These ratio estimates use extra information of x_i, $i = 1, \cdots, n$ and the true total and mean X or $\bar{X}$, thus improving the *precision* of ratio estimates over the ordinary estimates $\widehat{Y} = N\bar{y}$ and $\widehat{\bar{Y}} = \bar{y}$ respectively.

From the previous result,

(a) $E(\widehat{\bar{Y}}_r) = \bar{X}E(r) \approx \bar{X}R = \bar{Y}$.
Similarly $E(\widehat{Y}_r) = XE(r) \approx XR = Y$.

(b) Since $\mathrm{Var}(\widehat{\bar{Y}}_r) \approx \left(1 - \frac{n}{N}\right) \frac{S_r^2}{n}$ and $\mathrm{Var}(\widehat{Y}_r) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{S_r^2}{n}$,

$$\boxed{\mathrm{var}(\widehat{\bar{Y}}_r) = \left(1 - \frac{n}{N}\right) \frac{s_r^2}{n} \quad \text{and} \quad \mathrm{var}(\widehat{Y}_r) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_r^2}{n}.}$$

The estimator r for R is generally ***biased***, so $\widehat{Y}_r$ and $\widehat{\bar{Y}}_r$ are also biased for Y and $\bar{Y}$ respectively.

Bias:

$$\mathrm{Cov}(r, \bar{x}) = E(r\bar{x}) - E(r)E(\bar{x}) = E\left(\frac{\bar{y}}{\bar{x}}\bar{x}\right) - E(r)E(\bar{x})$$

so

$$E(r) = \frac{E(\bar{y})}{E(\bar{x})} - \frac{\mathrm{Cov}(r, \bar{x})}{E(\bar{x})} = R - \frac{\rho_{r,\bar{x}}\, \sigma_r\, \sigma_{\bar{x}}}{\bar{X}}.$$

Therefore for any ratio estimates,

$$\frac{|\text{bias}\, r|}{\sigma_r} = \frac{|R - E(r)|}{\sigma_r} = \frac{\rho_{r,\bar{x}}\, \sigma_{\bar{x}}}{\bar{X}} \leq \frac{\sigma_{\bar{x}}}{\bar{X}} = \text{cv}(\bar{x}) \qquad (3)$$

since $|\rho_{r,\bar{x}}| \leq 1$. Thus if the $\text{CV}(\bar{x})$ is small, the bias of $\widehat{R} = r$ is small relative to $\text{SE}(r)$ of $\widehat{R}$. But if n is small, the bias can be large.

Efficiency:

The ratio estimator is more efficient than the ordinary estimator, that is $\text{var}(\widehat{\bar{Y}}) > \text{var}(\widehat{\bar{Y}}_r)$, if

$$\hat{\rho} > \frac{\text{cv}(x)}{2\text{cv}(y)} \qquad (4)$$

where $\text{cv}(y)$ is the sample cv for Y defined as

$$\text{cv}(y) = \frac{s_y}{\bar{y}}.$$

Then

$$\begin{aligned}
\text{var}(\widehat{\bar{Y}}) - \text{var}(\widehat{\bar{Y}}_r) > 0 \;&\Rightarrow\; \left(1 - \frac{n}{N}\right)\frac{1}{n}[s_y^2 - s_r^2] > 0 \\
&\Rightarrow\; [s_y^2 - (s_y^2 - 2r\hat{\rho}s_x s_y + r^2 s_x^2)] > 0 \\
&\Rightarrow\; r s_x(2\hat{\rho}s_y - r s_x) > 0 \\
&\Rightarrow\; 2\hat{\rho}s_y - r s_x > 0 \quad \text{since } r > 0 \;\&\; s_x > 0 \\
&\Rightarrow\; \hat{\rho} > \frac{\bar{y}}{\bar{x}}\frac{s_x}{2s_y} = \frac{\text{cv}(x)}{2\text{cv}(y)} \quad \text{since } r = \frac{\bar{y}}{\bar{x}}
\end{aligned}$$

and the equality holds when $\hat{\rho} = \dfrac{\text{cv}(x)}{2\text{cv}(y)}$.

Example: (7-11) The manager of 7-11 is interested in estimating the total sale in thousands for all of its 300 branches. From last year record, the total sale in thousands for all the 300 branches is 21300. Careful check of this year records are obtained for a SRS of 15 branches with the following results:

Branch	Last year sale x	This year sale y	Branch	Last year sale x	This year sale y
1	50	56	9	100	165
2	35	48	10	250	409
3	12	22	11	50	73
4	10	14	12	50	70
5	15	18	13	150	95
6	30	26	14	100	55
7	9	11	15	40	83
8	25	30			

$$\sum_{i=1}^n x_i = 926, \ \sum_{i=1}^n x_i^2 = 117400, \ \sum_{i=1}^n y_i = 1175, \ \sum_{i=1}^n y_i^2 = 231815, \ \sum_{i=1}^n x_i y_i = 155753$$
$$s_y^2 = 9983.81$$

The *ordinary* estimate of the total sale this year in thousands is

$$\widehat{Y} = N\overline{y} = 300\left(\frac{1175}{15}\right) = 23500$$

with

$$\text{se}(\widehat{Y}) = N\sqrt{(1-\frac{n}{N})\frac{s_y^2}{n}} = 300\sqrt{\left(1-\frac{15}{300}\right)\frac{9983.81}{15}} = 7543.72.$$

The *ratio* estimate and its se for the total sale this year in thousands are

$$\widehat{Y}_r = Xr = 21300\left(\frac{1175}{926}\right) = 27027.54$$

$$\begin{aligned}\text{se}(\widehat{Y}_r) &= N\sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - 2r\sum_{i=1}^{n} x_iy_i + r^2\sum_{i=1}^{n} x_i^2\right)} \\ &= 300\sqrt{\left(1-\frac{15}{300}\right)\frac{1}{15\times 14}\left(231815 - 2\cdot\frac{1175}{926}\cdot 155753 + (\frac{1175}{926})^2\cdot 117400\right)} \\ &= 3226.66\end{aligned}$$

which is much smaller than $\text{se}(\widehat{Y}) = 7543.72$ thousands.

Read Tutorial 11 Q2a,b, Q3a,b.

3.4 Regression estimator

Since $\widehat{\overline{Y}}_r = \overline{X}\widehat{R} = \overline{X}\frac{\overline{y}}{\overline{x}}$, the line $y = mx$ with slope $m = \frac{\overline{y}}{\overline{x}}$ passes through the origin $(0,0)$ and $(\overline{X}, \widehat{\overline{Y}}_r)$. However, the linear relationship between X and Y may not pass through the origin. A more general estimator, the *regression* estimator fits a *regression line*:

$$y = A + Bx = \overline{y} - B\overline{x} + Bx = \overline{y} + B(x - \overline{x}) \qquad (5)$$

to the sample data where the least square estimate of B is

$$B = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{N}(y_i - \overline{Y})(x_i - \overline{X})}{\sum_{i=1}^{N}(x_i - \overline{X})^2} = \frac{\sum_{i=1}^{N} x_iy_i - N\overline{XY}}{\sum_{i=1}^{N} x_i^2 - N\overline{X}^2} = \frac{S_{xy}}{S_x^2} = \frac{\rho S_y}{S_x}.$$

and $A = \overline{y} - B\overline{x}$.

Note: $\text{Cov}(X,Y) = S_{xy} = SS_{xy}/(N-1)$, $\text{Var}(X) = S_x^2 = SS_{xx}/(N-1)$, $\text{cov}(X,Y) = s_{xy} = ss_{xy}/(n-1)$ and $\text{var}(X) = s_x^2 = ss_{xx}/(n-1)$.

Then the regression estimator of the population mean $\overline{Y}$ is to substitute $x = \overline{X}$ to (5) to obtain

$$\boxed{\widehat{\overline{Y}}_{reg} = \overline{y} + b(\overline{X} - \overline{x})}$$

where

$$b = \frac{ss_{xy}}{ss_{xx}} = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n}x_iy_i - n\overline{xy}}{\sum_{i=1}^{n}x_i^2 - n\overline{x}^2} = \frac{s_{xy}}{s_x^2}. \qquad (6)$$

Since

$$\widehat{\overline{Y}}_{reg} = \overline{y} + b(\overline{X}-\overline{x}) \simeq \overline{y} + B(\overline{X}-\overline{x}) = \overline{z}'$$

the sample mean of the variable $z_i' = y_i + B(\overline{X} - x_i)$, we have

$$E(\widehat{\overline{Y}}_{reg}) \simeq E[\overline{y}+B(\overline{X}-\overline{x})] = E(\overline{y})+B[\overline{X}-E(\overline{x})] = \overline{Y} \text{ Approx. unbiased}$$

and

$$\begin{aligned}\text{Var}(\widehat{\overline{Y}}_{reg}) &\simeq \text{Var}(\overline{z}') = \text{Var}[\overline{y}+B(\overline{X}-\overline{x})] = \text{Var}(\overline{y}-B\overline{x})\\ &= \text{Var}(\bar{y}) + B^2\,\text{Var}(\bar{x}) - 2B\,\text{Cov}(\bar{y},\bar{x})\\ &= \left(1-\frac{n}{N}\right)\frac{S_y^2}{n} + \rho^2\frac{S_y^2}{S_x^2}\left(1-\frac{n}{N}\right)\frac{S_x^2}{n} - 2\rho\frac{S_y}{S_x}\left(1-\frac{n}{N}\right)\frac{\rho S_xS_y}{n}\\ &= \left(1-\frac{n}{N}\right)\frac{S_y^2}{n}\left(1-\rho^2\right).\end{aligned}$$

Hence

$$\boxed{\text{var}(\widehat{\overline{Y}}_{reg}) = \left(1-\frac{n}{N}\right)\frac{s_{reg}^2}{n} = \left(1-\frac{n}{N}\right)\frac{s_y^2(1-\hat{\rho}^2)}{n}}$$

where s_{reg}^2 is the sample variance of $z_i' = y_i + b(\overline{X} - x_i)$.

The regression estimator for the population total Y is

$$\boxed{\widehat{Y}_{reg} = N[\overline{y} + b(\overline{X}-\overline{x})]}$$

and its variance estimate is

$$\boxed{\text{var}(\widehat{Y}_{reg}) = N^2\left(1-\frac{n}{N}\right)\frac{s_{reg}^2}{n} = N^2\left(1-\frac{n}{N}\right)\frac{s_y^2(1-\hat{\rho}^2)}{n}}$$

Bias:

$$\begin{aligned}\text{Bias in } \hat{\bar{Y}}_{reg} &= E(\hat{\bar{Y}}_{reg}) - \bar{Y} = E(\bar{y}) + E[b(\bar{X} - \bar{x})] - \bar{Y} \\ &= E[b(\bar{X} - \bar{x})] = -\text{Cov}(b, \bar{x}).\end{aligned}$$

Efficiency:

1. The regression estimator is *at least as efficient as* the ordinary estimator, that is $\text{var}(\hat{\bar{Y}}) \geq \text{var}(\hat{\bar{Y}}_{reg})$ since

$$\begin{aligned}\text{var}(\hat{\bar{Y}}) - \text{var}(\hat{\bar{Y}}_{reg}) &= \left(1 - \frac{n}{N}\right)\frac{1}{n}[s_y^2 - s_{reg}^2] \\ &= \left(1 - \frac{n}{N}\right)\frac{1}{n}s_y^2\hat{\rho}^2 \geq 0\end{aligned}$$

   where the equality holds when $\hat{\rho} = 0$, i.e. there is no association between Y and X.

2. The regression estimator is *more efficient* than the ratio estimator, that is $\text{var}(\hat{\bar{Y}}_r) \geq \text{var}(\hat{\bar{Y}}_{reg})$ unless

$$b = r = \frac{\bar{y}}{\bar{x}}$$

   in which case they are equivalent and the regression of y on x is linear through the origin and the variance of y is proportional to x.

$$\begin{aligned}\text{var}(\hat{\bar{Y}}_r) - \text{var}(\hat{\bar{Y}}_{reg}) &= \left(1 - \frac{n}{N}\right)\frac{1}{n}[s_r^2 - s_{reg}^2] \\ &= \left(1 - \frac{n}{N}\right)\frac{1}{n}[s_y^2 - 2r\hat{\rho}s_xs_y + r^2s_x^2 - s_y^2(1 - \hat{\rho}^2)] \\ &= \left(1 - \frac{n}{N}\right)\frac{1}{n}(r^2s_x^2 - 2r\hat{\rho}s_xs_y + s_y^2\hat{\rho}^2) \\ &= \left(1 - \frac{n}{N}\right)\frac{1}{n}(rs_x - \hat{\rho}s_y)^2\end{aligned}$$

$$\begin{aligned} &= \left(1-\frac{n}{N}\right)\frac{1}{n}(rs_x - bs_x)^2 \\ &= \left(1-\frac{n}{N}\right)\frac{s_x^2}{n}(r-b)^2 > 0 \Rightarrow (r-b)^2 > 0 \end{aligned}$$

where $\hat{\rho}s_y = \dfrac{s_{xy}}{s_x s_y}s_y = \dfrac{s_{xy}}{s_x} = \dfrac{s_{xy}}{s_x^2}s_x = bs_x$.

3. Since $\widehat{\overline{Y}}_{reg} = \overline{y} + b(\overline{X} - \overline{x})$, the regression estimator adjusts the $\overline{y}$ up or down by an amount $b(\overline{X} - \overline{x})$.

   (a) When the slope $b = 0$, the regression estimator $\widehat{\overline{Y}}_{reg} = \overline{y}$ becomes the *ordinary* estimator $\widehat{\overline{Y}}$.

   (b) When the y-intercept $a = \overline{y} - b\overline{x} = 0 \Leftrightarrow b = \frac{\overline{y}}{\overline{x}} = r$, the slope b becomes the ratio estimate r and the regression estimator

$$\widehat{\overline{Y}}_{reg} = \overline{y} + \frac{\overline{y}}{\overline{x}}(\overline{X} - \overline{x}) = \overline{y} + \frac{\overline{y}}{\overline{x}}\overline{X} - \overline{y} = \frac{\overline{y}}{\overline{x}}\overline{X} = \overline{X}r = \widehat{\overline{Y}}_r$$

   becomes the *ratio* estimator $\widehat{\overline{Y}}_r$.

Example: (7-11) Estimate the total sale using the regression estimator.

Solution: The *regression* estimate of the total sale this year in thousands is

$$\begin{aligned} ss_{xy} &= \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} = 155753 - 15 \times \frac{926}{15} \times \frac{1175}{15} = 83216.33, \\ ss_{xx} &= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 117400 - 15 \times \left(\frac{926}{15}\right)^2 = 60234.93, \\ ss_{yy} &= \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = 231815 - 15 \times \left(\frac{1175}{15}\right)^2 = 139773.33. \end{aligned}$$

We have

$$b = \frac{ss_{xy}}{ss_{xx}} = \frac{83216.33}{60234.93} = 1.3815$$

and

$$\hat{\rho} = \frac{ss_{xy}}{\sqrt{ss_{xx}ss_{yy}}} = \frac{83216.33}{\sqrt{60234.93 \times 139773.33}} = 0.9069.$$

It follows that

$$\begin{aligned}\widehat{Y}_{reg} &= N[\overline{y} + b(\overline{X} - \overline{x})] \\ &= 300\left[\frac{1175}{15} + 1.3815\left(\frac{21300}{300} - \frac{926}{15}\right)\right] = 27340.65\end{aligned}$$

as compared with $\widehat{Y} = 23500$ and $\widehat{Y}_r = 27027.54$. The s.e. estimate is

$$\begin{aligned}\text{se}(\widehat{Y}_{reg}) &= N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_y^2(1 - \hat{\rho}^2)}{n}} \\ &= 300\sqrt{\left(1 - \frac{15}{300}\right)\frac{9983.81(1 - 0.9069^2)}{15}} = 3178.52\end{aligned}$$

which is $<$ se($\widehat{Y}_r$) = 3226.66 $<<$ se($\widehat{Y}$) = 7543.72. This shows that the dropping of zero y-intercept assumption improves the estimate slightly. Note that the y-intercept estimate is

$$a = \overline{y} - b\overline{x} = \frac{1175}{15} - 1.3815 \times \frac{926}{15} \approx -6.9531$$

which is quite close to zero.

Read Tutorial 11 Q2c,d, & 3c,d.

3.5 The Hartley - Ross Estimator

Since the ratio estimator r for R is biased, the following leads to an unbiased estimator of R.

Theorem: Let $Z = f(X, Y)$ be a fixed function of two variables. Define $Z_i = f(X_i, Y_i)$ and $z_i = f(x_i, y_i)$. Then

$$E\left(\bar{z} + \frac{N-1}{N\bar{X}} \frac{\sum_{i=1}^n z_i(x_i - \bar{x})}{n-1}\right) = \frac{\sum_{i=1}^N Z_i X_i}{N\bar{X}}. \qquad (7)$$

Proof: The LHS is

$$E(\bar{z}) + \frac{N-1}{N\bar{X}} E\left(\frac{\sum_{i=1}^n z_i(x_i - \bar{x})}{n-1}\right)$$

$$= \bar{Z} + \frac{N-1}{N\bar{X}} \frac{\sum\limits_{i=1}^{N} Z_i(X_i - \bar{X})}{N-1} = \bar{Z} + \frac{\sum\limits_{i=1}^{N} Z_i X_i - \bar{X}\sum\limits_{i=1}^{N} Z_i}{N\bar{X}} = \frac{\sum\limits_{i=1}^{N} Z_i X_i}{N\bar{X}}.$$

For the problem of estimation of R from sample (x_i, y_i), $i = 1, \cdots, n$, we assume $X_i > 0$, $i = 1, \cdots, N$ and define the function

$$z_i = f(x_i, y_i) = y_i/x_i = r_i^*, \quad i = 1, \cdots, n$$

and $Z_i = Y_i/X_i$, $i = 1, 2, \cdots, N$, so from (7)

$$E\left\{\bar{r}^* + \frac{N-1}{N\bar{X}} \frac{n(\bar{y} - \bar{x}\bar{r}^*)}{n-1}\right\} = \frac{\sum\limits_{i=1}^{N} \frac{Y_i}{X_i} X_i}{N\bar{X}} = \frac{\bar{Y}}{\bar{X}} = R$$

since

$$\sum_{i=1}^n z_i(x_i - \bar{x}) = \sum_{i=1}^n \frac{y_i}{x_i}(x_i - \bar{x}) = \sum_{i=1}^n y_i - \bar{x}\sum_{i=1}^n \frac{y_i}{x_i} = n(\bar{y} - \bar{x}\bar{r}^*).$$

Thus the *Hartley-Ross* estimator as an *unbiased* estimator of R is

$$\boxed{\hat{R}_{hr} = \bar{r}^* + \frac{N-1}{N\bar{X}} \frac{n(\bar{y} - \bar{r}^*\bar{x})}{n-1}}$$

for which we need to know $\bar{X}$ (or $X = N\bar{X}$). This estimator contains a *mean of ratio* estimate and an adjustment for unbiasness.

The Hartley-Ross estimators for mean and total are

for the population mean: $\boxed{\widehat{\bar{Y}}_{hr} = \bar{X}\bar{r}^* + \frac{N-1}{N}\frac{n(\bar{y} - \bar{r}^*\bar{x})}{n-1}}$ and

for the population total: $\boxed{\widehat{Y}_{hr} = X\bar{r}^* + (N-1)\frac{n(\bar{y} - \bar{r}^*\bar{x})}{n-1}.}$

Remarks:

1. So far, we have $\widehat{R} = \dfrac{\bar{y}}{\bar{x}}$ biased for R, $\widehat{R}^* = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \frac{y_i}{x_i}$ biased for R & unbiased for $R^* = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \frac{y_i}{x_i}$ and $\widehat{R}_{hr}$ unbiased for R. Finally, could we just use
$$\widehat{R}_o = \bar{y}/\bar{X} \ ?$$
This is the ordinary estimator $E\left(\dfrac{\bar{y}}{\bar{X}}\right) = \dfrac{E(\bar{y})}{\bar{X}} = \dfrac{\bar{Y}}{\bar{X}} = R$ which does not use the information from the sample $\{x_i\}$ but is unbiased for R.

2. For small samples we might expect the Hartley-Ross estimator to be better. There is no general result on the comparison of the variances of $r = \dfrac{\bar{y}}{\bar{x}}$, $r_o = \dfrac{\bar{y}}{\bar{X}}$, and $r_{hr} = \bar{r}^* + \dfrac{N-1}{N\bar{X}} \dfrac{n(\bar{y} - \bar{r}^*\bar{x})}{n-1}$ for *all sample* sizes.
See Cochran (2nd Ed) Theorem 6.3 §6.15.

Summary of estimators and variance estimates based on 1 SRS

	Ord.	Ratio	Regression	Hartley-Ross
Ratio R	$\frac{\bar{y}}{\bar{X}}$ $\frac{1}{\bar{X}^2}(1-\frac{n}{N})\frac{s_y^2}{n}$	$\frac{\bar{y}}{\bar{x}}$ $\frac{1}{\bar{X}^2}(1-\frac{n}{N})\frac{s_r^2}{n}$	- -	$\bar{r}^* + \frac{N-1}{N\bar{X}}\frac{n(\bar{y}-\bar{r}^*\bar{x})}{n-1}$ -
Mean $\bar{Y}$	$\bar{y}$ $(1-\frac{n}{N})\frac{s_y^2}{n}$	$\frac{\bar{y}}{\bar{x}}\bar{X}$ $(1-\frac{n}{N})\frac{s_r^2}{n}$ $\text{var}(\hat{\bar{Y}}_r) < \text{var}(\hat{\bar{Y}})$ if $\hat{\rho} > \frac{\bar{y}s_x}{2\bar{x}s_y}$	$\bar{y} + \frac{s_{xy}}{s_x^2}(\overline{X}-\overline{x})$ $(1-\frac{n}{N})\frac{s_y^2(1-\hat{\rho}^2)}{n}$ $\text{var}(\hat{\bar{Y}}_{reg}) < \text{var}(\hat{\bar{Y}}_r)$ equal if $b = r = \frac{\bar{y}}{\bar{x}}$	$\bar{X}\bar{r}^* + \frac{N-1}{N}\frac{n(\bar{y}-\bar{r}^*\bar{x})}{n-1}$ -
Total Y	$N\bar{y}$ $N^2(1-\frac{n}{N})\frac{s_y^2}{n}$	$\frac{\bar{y}}{\bar{x}}X$ $N^2(1-\frac{n}{N})\frac{s_r^2}{n}$	$N\bar{y} + \frac{s_{xy}}{s_x^2}(X-N\bar{x})$ $N^2(1-\frac{n}{N})\frac{s_y^2(1-\hat{\rho}^2)}{n}$	$X\bar{r}^* + (N-1)\frac{n(\bar{y}-\bar{r}^*\bar{x})}{n-1}$ -

$$s_y^2 = \frac{1}{n-1}\Big(\sum_{i=1}^n y_i^2 - n\bar{y}^2\Big),$$

$$s_r^2 = \frac{1}{n-1}\Big(\sum_{i=1}^n y_i^2 - 2r\sum_{i=1}^n x_iy_i + r^2\sum_{i=1}^n x_i^2\Big) = s_y^2 - 2r\hat{\rho}s_xs_y + r^2s_x^2,\ r = \frac{\bar{y}}{\bar{x}},$$

Example: (7-11)

Solution: The ratios and their summary are given below:

i	x_i	y_i	$r'_i = y_i/x_i$	i	x_i	y_i	$r'_i = y_i/x_i$
1	50	56	1.120	9	100	165	1.650
2	35	48	1.371	10	250	409	1.636
3	12	22	1.833	11	50	73	1.460
4	10	14	1.400	12	50	70	1.400
5	15	18	1.200	13	150	95	0.633
6	30	26	0.867	14	100	55	0.550
7	9	11	1.222	15	40	83	2.075
8	25	30	1.200	Total			19.618

We have $\bar{r}^* = \frac{1}{n}\sum_{i=1}^{n} r_i^* = \frac{19.618}{15} = 1.3079$, $\bar{x} = 61.7333$ and $\bar{y} = 78.3333$.

The *Hartley-Ross* estimate of the total sale this year in thousands is

$$\begin{aligned}\widehat{Y}_{hr} &= X\bar{r}^* + (N-1)\frac{n(\bar{y} - \bar{r}^*\bar{x})}{n-1} \\ &= 21300(1.3079) + (300-1)\frac{15[78.333 - 1.3079(61.7333)]}{15-1} \\ &= 27086.9\end{aligned}$$

Read Tutorial 12 Q1(a).

Example: In a survey of family size (x_1), weekly income (x_2) and weekly expenditure on food (y), we want to estimate the average weekly expenditure on food per family in the most efficient way. A simple random sample of 27 families yields the following data:

$$\sum_i x_{1i} = 109,\ \sum_i x_{2i} = 16277,\ \sum_i y_i = 2831,\ \hat{\rho}_{x_1,y} = 0.925,\ \hat{\rho}_{x_2,y} = 0.573$$

The sample covariance matrix for y, x_1 and x_2 is

$$\begin{pmatrix} 547.8234 & 26.5057 & 1796.5541 \\ 26.5057 & 1.4986 & 80.1595 \\ 1796.5541 & 80.1595 & 17967.0541 \end{pmatrix}.$$

From the census data $\bar{X}_1 = 3.91$ and $\bar{X}_2 = 542$.

(a) Estimate the standard errors of the ratio estimators for $\bar{Y}$ using x_1 and using x_2. Compare the standard errors with the s.e. for the simple estimate ignoring the covariates. Which estimator has the smallest estimated s.e.?

(b) Calculate the best available estimate of the average weekly expenditure on food per family and give an approximate 95% confidence interval for this average.

Solution:

(a) The standard errors of the ratio estimators for $\bar{Y}$ using x_1 and using x_2 are

$$
\begin{aligned}
r_1 &= \frac{\sum_i y_i}{\sum_i x_{1i}} = \frac{2831}{109} = 25.97 \\
s_{r1}^2 &= s_y^2 - 2r\, s_{x_1 y} + r^2\, s_{x_1}^2 \\
&= 547.8234 - 2(25.97)(26.5057) + 25.97^2(1.4986) \\
&= 181.896 \\
r_2 &= \frac{\sum_i y_i}{\sum_i x_{2i}} = \frac{2831}{16277} = 0.1739 \\
s_{r2}^2 &= s_y^2 - 2r\, s_{x_2 y} + r^2\, s_{x_2}^2 \\
&= 547.8234 - 2(0.1739)(1796.5541) + 0.1739^2(17967.0541) \\
&= 475.7895 \\
\text{se}(\hat{\bar{Y}}_{r1}) &= \sqrt{\frac{s_{r1}^2}{n}} = \sqrt{\frac{181.896}{27}} = 2.5956 \\
\text{se}(\hat{\bar{Y}}_{r2}) &= \sqrt{\frac{s_{r2}^2}{n}} = \sqrt{\frac{475.7895.193}{27}} = 4.1978 \\
\text{se}(\hat{\bar{Y}}) &= \sqrt{\frac{s_y^2}{n}} = \sqrt{\frac{547.8234}{27}} = 4.5044
\end{aligned}
$$

The first ratio estimator $\hat{\bar{Y}}_{r1}$ has the lowest s.e. due to the higher correlation $\hat{\rho}_{y,x_1} = 0.925$. The second ratio estimator only has marginal improvement as the correlation $\hat{\rho}_{y,x_x} = 0.573$ is weak but

$$\rho_{x2,y} = 0.573 > \frac{\bar{y} s_x}{2\bar{x} s_y} = \frac{2831 \cdot \sqrt{17967.0541}}{2 \cdot 16277 \cdot \sqrt{547.8234}} = 0.4980.$$

Note that fpc is ignored because the population size N is unknown.

(b) The estimate of the average weekly expenditure on food per family

$$
\begin{aligned}
\hat{\bar{Y}}_{r1} &= r_1 \bar{X} = 25.97(3.91) = 101.5524 \\
95\% \text{ CI for } \bar{Y} &= 101.5524 \mp 1.96(2.5956) = (96.4651,\ 106.6397)
\end{aligned}
$$

3.6 Ratio estimate for subpopulation in poststratification

For some C_l, we want to estimate:

$$R_l = \sum_{i \in C_l} Y_i \Bigg/ \sum_{i \in C_l} X_i = \sum_{i=1}^{N} Y_i' \Bigg/ \sum_{i=1}^{N} X_i', = R'$$

if we define

$$\begin{aligned} (Y_i',\, X_i') &= (Y_i, X_i) \text{ if } i \in C_l \\ &= (0,0) \text{ if } i \notin C_l\,. \end{aligned}$$

Note: $X' = X_l$, i.e. the sum of X_i' over all population equals to the sum of X_i over C_l. Hence the natural estimator of ratio and its variance estimate is

$$\boxed{r' = \frac{\sum_{i=1}^{n} y_i'}{\sum_{i=1}^{n} x_i'} = \frac{\sum_{i \in C_j} y_i}{\sum_{i \in C_l} x_i} = r_l \quad \text{and} \quad \text{var}(r_l) \approx \frac{1}{(\bar{X}')^2}\left(1 - \frac{n}{N}\right)\frac{{s_{rl}'}^2}{n}}$$

where

$$\begin{aligned} (x_i',\, y_i') &= (x_i, y_i) \text{ if } i \in C_l \\ &= (0,0) \text{ if } i \notin C_l, \end{aligned}$$

$\bar{X}' = \dfrac{X'}{N}$ can be estimated by $\bar{x}' = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i' = \dfrac{1}{n}\sum_{i \in C_l} x_i$ and

$${S_{rl}'}^2 = \frac{1}{N-1}\sum_{i-1}^{N}(Y_i' - R'X_i')^2$$

can be estimated by

$${s_{rl}'}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i' - r'x_i')^2 = \frac{1}{n-1}\sum_{i \in C_l}(y_i - r_l x_i)^2.$$

The ratio estimator of mean in C_l and its variance estimate are

$$\boxed{\hat{\bar{Y}}_{rl} = \bar{X}_l\, r_l \quad \text{and} \quad \text{var}(\hat{\bar{Y}}_{rl}) \approx \frac{1}{W_l^2}\left(1-\frac{n}{N}\right)\frac{{s'_{rl}}^2}{n}}$$

since

$$\begin{aligned}\text{var}(\hat{\bar{Y}}_{rl}) &= \bar{X}_l^2\text{var}(r_l) = \frac{\bar{X}_l^2}{\bar{X}'^2}\left(1-\frac{n}{N}\right)\frac{{s'_{rl}}^2}{n} \\ &= \frac{X'^2}{N_l^2}\frac{N^2}{X'^2}\left(1-\frac{n}{N}\right)\frac{{s'_{rl}}^2}{n} = \frac{1}{W_l^2}\left(1-\frac{n}{N}\right)\frac{{s'_{rl}}^2}{n}.\end{aligned}$$

Similarly, the ratio estimator of total in C_l and its variance estimate are

$$\boxed{\hat{Y}_{rl} = X_l\, r_l \quad \text{and} \quad \text{var}(\hat{Y}_{rl}) \approx N^2\left(1-\frac{n}{N}\right)\frac{{s'_{rl}}^2}{n}}$$

since

$$\text{var}(\hat{Y}_{rl}) = X_l^2\text{var}(r_l) = \frac{X_l^2}{\bar{X}'^2}\left(1-\frac{n}{N}\right)\frac{{s'_{rl}}^2}{n} = X'^2\frac{N^2}{X'^2}\left(1-\frac{n}{N}\right)\frac{{s'_{rl}}^2}{n}.$$

Note that these estimators correspond to method 1 in Section 1.5 for *poststratification* and n_l does not come into any of these calculations.

Read Tutorial 12 Q1b,c.

3.7 Ratio Estimation for Stratified SRS

In a stratified SRS, a SRS of a specified sample size n_l is taken in each of the L strata with known size N_l, e.g. the 6 states of Australia. There are *two types of ratio estimates* depending on the *order of taking ratio and summing* over strata.

1. Take ratios $r_l = \bar{y}_l/\bar{x}_l$ first and sum over $\widehat{Y}_l = X_l r_l$ to obtain $\widehat{R}_s = \sum_{l=1}^{L} \widehat{Y}_l / X$.

2. Sum over $\widehat{Y}_l$ and $\widehat{X}_l$ first to obtain $\widehat{Y}$ and $\widehat{X}$ and then take ratio $\widehat{R}_c = \widehat{Y}/\widehat{X}$.

3.7.1 The 'Separate' Ratio Estimate

Suppose the stratum *totals* $X_l,\ l = 1, \cdots, L$ are known so $X = \sum\limits_{l=1}^{L} X_l$ is known also. Then

$$\boxed{\widehat{R}_s = \frac{\widehat{Y}}{X} = \frac{1}{X}\sum_{l=1}^{L} \widehat{Y}_l = \frac{1}{X}\sum_{l=1}^{L} X_l\, r_l = \frac{1}{\bar{X}}\sum_{l=1}^{L} W_l \bar{X}_l\, r_l}$$

since $\dfrac{X_l}{X} = \dfrac{N}{X}\dfrac{N_l}{N}\dfrac{X_l}{N_l} = \dfrac{1}{\bar{X}} W_l \bar{X}_l$ and $r_l = \dfrac{\bar{y}_l}{\bar{x}_l}$. Then

$$E(\widehat{R}_s) = \sum_{l=1}^{L} \frac{X_l}{X} E(r_l) \approx \sum_{l=1}^{L} \frac{X_l}{X}\frac{Y_l}{X_l} = \frac{1}{X}\sum_{l=1}^{L} Y_l = \frac{Y}{X} = R,$$

since $E(r_l) \approx R_l = \dfrac{Y_l}{X_l}$ and $\boxed{\mathrm{var}(\widehat{R}_s) = \dfrac{1}{\bar{X}^2}\sum\limits_{l=1}^{L} W_l^2 \left(1 - \dfrac{n_l}{N_l}\right)\dfrac{s_{srl}^2}{n_l}}$

where

$$s_{rl}^2 = s_{y_l}^2 - 2r_l s_{x_l y_l} + r_l^2 s_{x_l}^2 = \frac{1}{n_l - 1}\left[\sum_{i=1}^{n_l} y_{il}^2 - 2r_l \sum_{i=1}^{n_l} x_{il} y_{il} + r_l^2 \sum_{i=1}^{n_l} x_{il}^2\right].$$

Similarly the *separate ratio* estimate for the mean is

$$\boxed{\widehat{\bar{Y}}_{st,s} = \sum_{l=1}^{L} W_l \bar{X}_l r_l} \text{ and } \boxed{\text{var}(\widehat{\bar{Y}}_{st,s}) = \sum_{l=1}^{L} W_l^2 \left(1 - \frac{n_l}{N_l}\right) \frac{s_{srl}^2}{n_l}}$$

Bias:

For *large* stratum sample sizes, r_l will be approximately unbiased for R_l and var(r_l) will approximate Var(r_l) reasonably well.

For moderate and small samples, bias is important, and we should consider it here. We know that in a single stratum

$$\frac{|\text{bias } r_l|}{\sigma_{r_l}} \le \frac{\sigma_{\bar{x}_l}}{\bar{X}_l} = \text{cv}(\bar{x}_l)$$

Consider the bias of $\widehat{R}_s$:

$$\begin{aligned}
|\text{bias}\,(\widehat{R}_s)| &= E(\widehat{R}_s - R) \\
&= E\left(\sum_{l=1}^{L} \frac{X_l}{X}(r_l - R_l)\right) = \sum_{l=1}^{L} \frac{X_l}{X} E(r_l - R_l) \\
&= \sum_{l=1}^{L} \frac{X_l}{X} |\text{bias } r_l| \le \max_l |\text{bias } r_l| \sum_{l=1}^{L} \frac{X_l}{X} \\
&\le \max_l |\text{bias } r_l| \le \max_l \left(\frac{\sigma_{\bar{x}_l}\sigma_{r_l}}{\bar{X}_l}\right)
\end{aligned}$$

Hence

$$\frac{|\text{bias}\,(\widehat{R}_s)|}{\text{s.e.}\widehat{R}_s} \le \frac{\max_l \sigma_{r_l}}{\text{s.e.}(\widehat{R}_s)} \max_l \left(\frac{\sigma_{\bar{x}_l}}{\bar{X}_l}\right) \le \sqrt{L}\left(\frac{\max_l \sigma_{r_l}}{\min_l \sigma_{r_l}}\right) \max_l \left(\frac{\sigma_{\bar{x}_l}}{\bar{X}_l}\right)$$

since

$$\begin{aligned}\text{s.e.}(\widehat{R}_s) &= \sqrt{\sum_{l=1}^{L}\left(\frac{X_l}{X}\right)^2 \text{var}(r_l)} \geq \min_l \sigma_{r_l}\sqrt{\sum_{l=1}^{L} p_l^2} \\ &\geq \min_l \sigma_{r_l}\sqrt{\sum_{l=1}^{L}\left(\frac{1}{L}\right)^2} \geq \min_l \sigma_{r_l}\sqrt{\frac{L}{L^2}} \geq \frac{1}{\sqrt{L}}\min_l \sigma_{r_l}\end{aligned}$$

where $X_l = p_l X$. The sum of squares of unequal proportions is higher than that from equal proportions in general. This is due to the convexity property of the function $f(p) = p^2$. For example, when $L = 2$ with cases $(1-p, p)$ and $(\frac{1}{2}, \frac{1}{2})$,

$$(1-p)^2 + p^2 - 2(\frac{1}{2})^2 = 2p^2 - 2p + \frac{1}{2} = \frac{1}{2}(2p-1)^2 \geq 0.$$

Therefore the ratio on the LHS can be $\sqrt{L}$ times as large as the $\sigma_{\bar{x}_l}/\bar{X}_l$ bound on individual relative biases. Even if the biases are individually small, the *overall bias* can be large.

3.7.2 The 'Combined' Ratio Estimate

It is defined as

$$\boxed{\widehat{R}_c = \frac{\sum_{l=1}^{L} W_l \bar{y}_l}{\sum_{l=1}^{L} W_l \bar{x}_l} = \frac{\bar{y}_{st}}{\bar{x}_{st}} = \frac{\widehat{\bar{Y}}}{\widehat{\bar{X}}} = \frac{\widehat{Y}}{\widehat{X}}}$$

and in contrast to $\widehat{R}_s$, it does not require the knowledge of individual X_l's. Note that

$$E(\widehat{R}_c) = E\left(\frac{\bar{y}_{st}}{\bar{x}_{st}}\right) \approx E\left(\frac{\bar{y}_{st}}{\bar{X}}\right) \approx \frac{1}{\bar{X}}\sum_{l=1}^{L} W_l E(\bar{y}_l) = \frac{Y}{X} = R \text{ Approx. unbiased}$$

Theorem:

$$\text{Var}(\widehat{R}_c) \approx \frac{1}{\bar{X}^2}\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{1}{n_l}\left(\frac{\sum_{i=1}^{N_l}[Y_{il}-\bar{Y}_l-R(X_{il}-\bar{X}_l)]^2}{N_l-1}\right).$$

Proof: First

$$\begin{aligned}\widehat{R}_c - R &= \frac{\bar{y}_{st}}{\bar{x}_{st}} - R = \frac{1}{\bar{x}_{st}}(\bar{y}_{st}-R\bar{x}_{st}) = \frac{1}{\bar{x}_{st}}\sum_{l=1}^{L} W_l(\bar{y}_l - R\bar{x}_l) \\ &= \frac{1}{\bar{x}_{st}}\sum_{l=1}^{L} W_l\bar{d}_l = \frac{1}{\bar{x}_{st}}\bar{d}_{st} \approx \frac{1}{\bar{X}}\bar{d}_{st}\end{aligned}$$

where $d_{li} = y_{li} - Rx_{li},\ i = 1,\cdots,n_l$ estimates $D_{li} = Y_{li} - RX_{li}$ and $\bar{d}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} d_{il}$. Note that typically $\bar{D}_l \neq 0$. Hence

$$\text{Var}(\widehat{R}_c) \approx \frac{1}{\bar{X}^2}\text{Var}(\bar{d}_{st}) \approx \frac{1}{\bar{X}^2}\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{S_{crl}^2}{n_l}$$

where

$$\begin{aligned}S_{crl}^2 &= \frac{1}{N_l-1}\sum_{i=1}^{N_l}(D_{li}-\bar{D}_l)^2 = \frac{1}{N_l-1}\sum_{i=1}^{N_l}[Y_{li}-\bar{Y}_l-R(X_{li}-\bar{X}_l)]^2 \\ &= S_{y_l}^2 - 2RS_{x_ly_l} + R^2S_{x_l}^2,\end{aligned}$$

and this can be estimated by

$$s_{crl}^2 = \frac{1}{n_l-1}\sum_{i=1}^{n_l}[y_{li}-\bar{y}_l-r_c(x_{li}-\bar{x}_l)]^2 = s_{y_l}^2 - 2r_cs_{x_ly_l} + r_c^2s_{x_l}^2$$

as compared to

$$s_{srl}^2 = \frac{1}{n_l - 1}\left[\sum_{i=1}^{n_l} y_{li}^2 - 2r_l \sum_{i=1}^{n_l} x_{li}y_{li} + r_l^2 \sum_{i=1}^{n_l} x_{il}^2\right] = s_{y_l}^2 - 2r_l s_{x_l y_l} + r_l^2 s_{x_l}^2$$

for separate ratio estimator.

There is less risk of bias in $\hat{R}_c$ than in $\hat{R}_s$. We can show that

$$\frac{|E(\widehat{R}_c - R)|}{\text{s.e.}\widehat{R}_c} \leq \max_l \left(\frac{\sigma_{\bar{x}_l}}{\bar{X}_l}\right)$$

in contrast to

$$\frac{|E(\widehat{R}_s - R)|}{\text{s.e.}\widehat{R}_s} \leq \sqrt{L}\left(\frac{\max_l \sigma_{r_l}}{\min_l \sigma_{r_l}}\right) \max_l \left(\frac{\sigma_{\bar{x}_l}}{\bar{X}_l}\right)$$

for the separate ratio estimator $\widehat{R}_s$.

Similarly the *combine ratio* estimate for the *mean* and its variance are

$$\boxed{\widehat{\bar{Y}}_{st,c} = \bar{X}\widehat{R}_c} \quad \text{and} \quad \boxed{\text{var}(\widehat{\bar{Y}}_{st,c}) = \sum_{l=1}^{L} W_l^2 \left(1 - \frac{n_l}{N_l}\right)\frac{s_{crl}^2}{n_l}.}$$

and for the *total* are

$$\boxed{\widehat{Y}_{st,c} = X\widehat{R}_c} \quad \text{and} \quad \boxed{\text{var}(\widehat{Y}_{st,c}) = N^2 \sum_{l=1}^{L} W_l^2 \left(1 - \frac{n_l}{N_l}\right)\frac{s_{crl}^2}{n_l}.}$$

Read Tutorial 12 Q2.

Estimators and variance estimates for stratified SRS (Ch.2)

Parameter	Estimator	Variance
Ordinary/naive estimator $s_{yl}^2 = \frac{1}{n_l-1}\big(\sum_{i=1}^{n_l} y_{li}^2 - n_l\bar{y}_l^2\big), \quad W_l = \frac{N_l}{N}$		
Ratio R	$\widehat{R}_{st} = \frac{1}{\bar{X}}\sum_{l=1}^{L} W_l\bar{y}_l$	$\text{var}(\widehat{R}_{st}) = \frac{1}{\bar{X}^2}\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{yl}^2}{n_l}$
Mean $\bar{Y}$	$\widehat{\bar{Y}}_{st} = \sum_{l=1}^{L} W_l\bar{y}_l$	$\text{var}(\widehat{\bar{Y}}_{st}) = \sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{yl}^2}{n_l}$
Total Y	$\widehat{Y}_{st} = N\sum_{l=1}^{L} W_l\bar{y}_l$	$\text{var}(\widehat{Y}_{st}) = N^2\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{yl}^2}{n_l}$
Separate ratio estimator $s_{sr,l}^2 = s_{yl}^2 - 2\,r_l\hat{\rho}s_{xl}s_{yl} + r_l^2 s_{xl}^2, \quad r_l = \frac{\bar{y}_l}{\bar{x}_l}$		
Ratio R	$\widehat{R}_{st,sr} = \frac{1}{\bar{X}}\sum_{l=1}^{L} W_l\bar{X}_l r_l$	$\text{var}(\widehat{R}_{st,sr}) = \frac{1}{\bar{X}^2}\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{sr,l}^2}{n_l}$
Mean $\bar{Y}$	$\widehat{\bar{Y}}_{st,sr} = \sum_{l=1}^{L} W_l\bar{X}_l r_l$	$\text{var}(\widehat{\bar{Y}}_{st,sr}) = \sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{sr,l}^2}{n_l}$
Total Y	$\widehat{Y}_{st,sr} = N\sum_{l=1}^{L} W_l\bar{X}_l r_l$	$\text{var}(\bar{Y}_{st,sr}) = N^2\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{sr,l}^2}{n_l}$
Combine ratio estimator $s_{cr,l}^2 = s_{yl}^2 - 2\,r_c\hat{\rho}s_{xl}s_{yl} + r_c^2 s_{xl}^2, \quad r_{st,cr} = \frac{\sum_{l=1}^{L} W_l\bar{y}_l}{\sum_{l=1}^{L} W_l\bar{x}_l}$		
Ratio R	$\widehat{R}_{st,cr} = \frac{\sum_{l=1}^{L} W_l\bar{y}_l}{\sum_{l=1}^{L} W_l\bar{x}_l} = r_{st,cr}$	$\text{var}(\widehat{R}_{st,cr}) = \frac{1}{\bar{X}^2}\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{cr,l}^2}{n_l}$
Mean $\bar{Y}$	$\widehat{\bar{Y}}_{st,cr} = \bar{X}r_{st,cr}$	$\text{var}(\widehat{\bar{Y}}_{st,cr}) = \sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{cr,l}^2}{n_l}$
Total Y	$\widehat{Y}_{st,cr} = N\bar{X}r_{st,cr}$	$\text{var}(\bar{Y}_{st,cr}) = N^2\sum_{l=1}^{L} W_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_{cr,l}^2}{n_l}$