Lecture 4: Ratio Estimation

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

What is a ratio estimation?

Ratio takes advantage of the correlation of x and y in the population; the higher the correlation, the better they work.

For ratio estimation to apply, two quantities y_i and x_i must be measured on each sample unit; x_i is often called an auxiliary variable or subsidiary variable. In the population of size N

$$t_y = \sum_{i=1}^{N} y_i, \quad t_x = \sum_{i=1}^{N} x_i$$

and their ratio is

$$B = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U}.$$

In the simplest use of ratio estimation, a simple random sample (SRS) of size n is taken, and the information in both x and y is used to estimate B, t_y, or $\bar{y}_U$.

Define the population correlation coefficient of x and y to be

$$R = \frac{\sum_{i=1}^{N} (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}.$$

Here, S_x is the population standard deviation of the x_i's, S_y is the population standard deviation of the y_i's, and R is simply the Pearson correlation coefficient of x and y for the N units in the population.

Why Use Ratio Estimation?

- Sometimes we simply want to estimate a ratio.

  Example: x_i= total number of pages in issue i;
  y_i = total number of pages in issue i that contain at least one advertisement.
  The proportion of interest can be estimated as

$$\hat{B} = \frac{\hat{t}_y}{\hat{t}_x}.$$

- Estimate a population total, but the population size N is unknown.

$$N = t_x / \bar{x}_U.$$

- Increase the precision of estimated means and totals.
- Ratio estimation is used to adjust estimates from the sample so that they reflect demographic totals.

  An SRS of 400 students taken at a university with 4000 students may contain 240 women and 160 men, with 84 of the sampled women and 40 of the sampled men planning to follow careers in teaching. Using only the information from the SRS, you would estimate that

$$\frac{4000}{400} \times 124 = 1240$$

  students plan to be teachers. Knowing that the college has 2700 women and 1300 men, a better estimate of the number of students planning teaching careers might be

$$\frac{84}{240} \times 2700 + \frac{40}{160} \times 1300 = 1270.$$

Example

For this example, suppose we know the population totals for 1987, but only have 1992 information on the SRS of 300 counties. When the same quantity is measured at different times, the response of interest at an earlier time often makes an excellent auxiliary variable. Let

$$y_i = \text{ total acreage of farms in county } i \text{ in 1992}$$
$$x_i = \text{ total acreage of farms in county } i \text{ in 1987.}$$

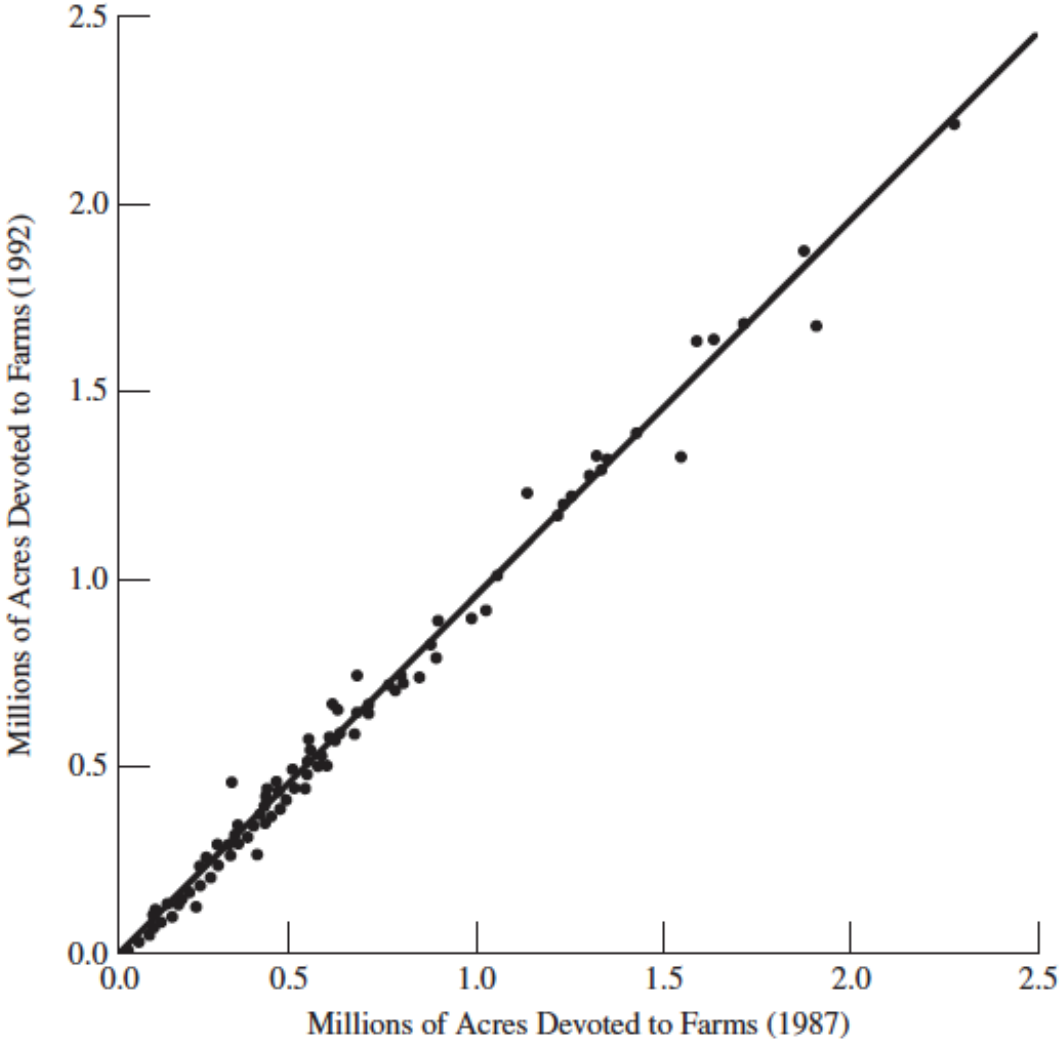
In 1987 a total of $t_x = 964{,}470{,}625$ acres were devoted to farms in the United States. The average acreage per county for the population is then $\bar{x}_U = 964{,}470{,}625/3078 = 313{,}343.3$ acres of farms per county.

Example

	A	B	C	D	E
1	County	State	*acres92* (y)	*acres87* (x)	Residual ($y - \hat{B}x$)
2					
3	Coffee County	AL	175209	179311	−1693.00
4	Colbert County	AL	138135	145104	−5019.56
5	Lamar County	AL	56102	59861	−2954.78
6	Marengo County	AL	199117	220526	−18446.29
⋮	⋮	⋮	⋮	⋮	⋮
299	Rock County	WI	343115	357751	−9829.70
300	Kanawha County	WV	19956	21369	−1125.91
301	Pleasants County	WV	15650	15716	145.14
302	Putnam County	WV	55827	55635	939.44
303					
304	Column sum		89369114	90586117	3.96176E-09
305	Column average		297897.0467	301953.7233	
306	Column standard deviation		344551.8948	344829.5964	31657.21817
307	$\hat{B}$ = C305/D305=		0.986565237		

Example

The plot of acreage, 1992 vs. 1987, for an SRS of 300 counties. The line in the plot goes through the origin and has slope $\hat{B}$= 0.9866. Note that the variability about the line increases with x.



Example

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U = (\hat{B})(313{,}343.283) = 309{,}133.6,$$

and

$$\hat{t}_{yr} = \hat{B}t_x = (\hat{B})(964{,}470{,}625) = 951{,}513{,}191.$$

Note that $\bar{y}$ for these data is 297,897.0, so $\hat{t}_{y\text{SRS}} = (3078)(\bar{y}) = 916{,}927{,}110$. In this example, $\bar{x}_S = 301{,}953.7$ is smaller than $\bar{x}_U = 313{,}343.3$. This means that our SRS of size 300 slightly underestimates the true population mean of the x's. Since the x's and y's are positively correlated, we have reason to believe that $\bar{y}_S$ may also underestimate the population value $\bar{y}_U$. Ratio estimation gives a more precise estimate of $\bar{y}_U$ by expanding $\bar{y}_S$ by the factor $\bar{x}_U/\bar{x}_S$.

Bias and MSE

- Estimate

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U = \frac{\bar{y}}{\bar{x}}\bar{x}_U.$$

- Variance

$$\mathrm{Var}(\hat{\bar{y}}_r) = (\bar{x}_U)^2 \,\mathrm{Var}(\hat{B}).$$

$$\hat{\mathrm{Var}}(\hat{B}) = \frac{N-n}{N}\frac{s_e^2}{n\bar{x}^2},$$

where

$$s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \hat{B}x_i\right)^2.$$

- Total

$$\hat{\mathrm{Var}}(\hat{t}_{yr}) = \hat{\mathrm{Var}}(\hat{B}t_x) = \left(1-\frac{n}{N}\right)\left(\frac{t_x}{\bar{x}}\right)^2\frac{s_e^2}{n\bar{x}^2}.$$

Bias and MSE

We can then show that

$$\begin{aligned}\text{Bias}\,[\hat{\bar{y}}_r] = E[\hat{\bar{y}}_r - \bar{y}_U] &\approx \frac{1}{\bar{x}_U}\left[B\,V(\bar{x}) - \text{Cov}(\bar{x},\bar{y})\right] \\ &= \left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U}(BS_x^2 - RS_xS_y),\end{aligned}$$

with R the correlation between x and y. The bias of $\hat{\bar{y}}_r$ is thus small if:

- the sample size n is large
- the sampling fraction n/N is large
- $\bar{x}_U$ is large
- S_x is small
- the correlation R is close to 1.

For estimating the MSE of $\hat{\bar{y}}_r$,

$$\begin{aligned}E[(\hat{\bar{y}}_r - \bar{y}_U)^2] &= E\left[\left\{(\bar{y}-B\bar{x})\left(1-\frac{\bar{x}-\bar{x}_U}{\bar{x}}\right)\right\}^2\right] \\ &= E\left[(\bar{y}-B\bar{x})^2 + (\bar{y}-B\bar{x})^2\left\{\left(\frac{\bar{x}-\bar{x}_U}{\bar{x}}\right)^2 - 2\frac{\bar{x}-\bar{x}_U}{\bar{x}}\right\}\right].\end{aligned}$$

It can be shown that the second term is generally small compared with the first term, so the variance and MSE are approximated by

$$\text{MSE}\,(\hat{\bar{y}}_r) = E[(\hat{\bar{y}}_r - \bar{y}_U)^2] \approx E\left[(\bar{y}-B\bar{x})^2\right].$$

The term $E\left[(\bar{y}-B\bar{x})^2\right]$ can also be written as

$$E\left[(\bar{y}-B\bar{x})^2\right] = V\left[\frac{1}{n}\sum_{i\in\mathcal{S}}(y_i - Bx_i)\right] = \left(1-\frac{n}{N}\right)\frac{S_y^2 - 2BRS_xS_y + B^2S_x^2}{n}.$$

Confidence interval

- Normal approximation

$$\hat{\bar{y}}_r \pm z_{\alpha/2} SE(\hat{\bar{y}}_r),\ \hat{B}_r \pm z_{\alpha/2} SE(\hat{B});$$

- t approximation

$$\hat{\bar{y}}_r \pm t_{\alpha/2,n-1} SE(\hat{\bar{y}}_r),\ \hat{B}_r \pm t_{\alpha/2,n-1} SE(\hat{B}).$$

Example:

$$\mathrm{SE}(\hat{t}_{yr}) = 3078\sqrt{1-\frac{300}{3078}}\left(\frac{313{,}343.283}{301{,}953.723}\right)\frac{31{,}657.218}{\sqrt{300}} = 5{,}546{,}162.$$

An approximate 95% CI for the total farm acreage, using the ratio estimator, is

$$951{,}513{,}191 \pm 1.96(5{,}546{,}162) = [940{,}642{,}713,\ 962{,}383{,}669].$$

Estimate Proportion

Santa Cruz Island Seedling Data

Tree	$x =$ Number of Seedlings, 3/92	$y =$ Seedlings Alive, 2/94
1	1	0
2	0	0
3	8	1
4	2	2
5	76	10
6	60	15
7	25	3
8	2	2
9	1	1
10	31	27
Total	206	61
Average	20.6	6.1
Standard deviation	27.4720	8.8248

Estimate the proportion of seedlings still alive in 1994.

Bias and MSE

$y_i =$ number of seedlings near tree i that are alive in 1994

$x_i =$ number of seedlings near tree i that are alive in 1992.

Then the ratio estimate of the proportion of seedlings still alive in 1994 is

$$\hat{B} = \hat{p} = \frac{\bar{y}}{\bar{x}} = \frac{6.1}{20.6} = 0.2961.$$

Ignoring the finite population correction (fpc),

$$\begin{aligned} \mathrm{SE}(\hat{B}) &= \sqrt{\frac{1}{(10)(20.6)^2} \frac{\sum_{i \in S} (y_i - 0.2961165 x_i)^2}{9}} \\ &= \sqrt{\frac{56.3778}{(10)(20.6)^2}} \\ &= 0.115. \end{aligned}$$

Estimation in Domains

Estimating domain means is a special case of ratio estimation.

Example: Suppose we want to estimate the mean salary for the domain of women.

Suppose we want to estimate the mean salary for the domain of women,

$$\bar{y}_{U_d} = \frac{\sum_{i \in \mathcal{U}_d} y_i}{N_d} = \frac{\text{total salary for all women in population}}{\text{number of women in population}}.$$

A natural estimator of $\bar{y}_{U_d}$ is

$$\bar{y}_d = \frac{\sum_{i \in \mathcal{S}_d} y_i}{n_d} = \frac{\text{total salary for women in sample}}{\text{number of women in sample}},$$

This is in fact a ratio estimation.

Let

$$x_i = \begin{cases} 1 & \text{if } i \in \mathcal{U}_d \\ 0 & \text{if } i \notin \mathcal{U}_d, \end{cases}$$

$$u_i = y_i x_i = \begin{cases} y_i & \text{if } i \in \mathcal{U}_d \\ 0 & \text{if } i \notin \mathcal{U}_d. \end{cases}$$

Then $t_x = \sum_{i=1}^{N} x_i = N_d, \bar{x}_U = N_d/N, t_u = \sum_{i=1}^{N} u_i, \bar{y}_{U_d} = t_u/t_x = B, \bar{x} = n_d/n$, and

$$\bar{y}_d = \hat{B} = \frac{\bar{u}}{\bar{x}} = \frac{\hat{t}_u}{\hat{t}_x}.$$

Because we are estimating a ratio, we calculate the standard error:

$$\begin{aligned} \text{SE}(\bar{y}_d) &= \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}^2} \frac{\sum_{i \in \mathcal{S}} (u_i - \hat{B}x_i)^2}{n-1}} \\ &= \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}^2} \frac{\sum_{i \in \mathcal{S}_d} (y_i - \hat{B})^2}{n-1}} \\ &= \sqrt{\left(1 - \frac{n}{N}\right) \frac{n}{n_d^2} \frac{(n_d - 1)s_{yd}^2}{n-1}}, \end{aligned}$$

where

$$s_{yd}^2 = \frac{\sum_{i \in \mathcal{S}_d} (y_i - \bar{y}_d)^2}{n_d - 1}$$

- If N_d is known,

$$t_u = N_d \bar{y}_d = N_d \hat{B}.$$

- If N_d is unknown,

$$\hat{t}_u = N\bar{u} = N \sum_{i \in \mathcal{S}_d} \frac{y_i}{n}, \quad SE(\hat{t}_u) = N * SE(\bar{u}) = N\sqrt{(1 - n/N)\frac{s_u^2}{n}}.$$

Example: In the SRS of size 300 from the Census of Agriculture, 39 counties are in western states. What is the estimated total number of acres devoted to farming in the West?

The sample mean of the 39 counties is $\bar{y}_d = 598{,}681$, with sample standard deviation $s_{yd} = 516{,}157.7$.

$$\text{SE}(\bar{y}_d) = \sqrt{\left(1 - \frac{300}{3078}\right)\frac{300}{39}\frac{38}{299}}\frac{516{,}157.7}{\sqrt{39}} = 77{,}637.$$

Thus, $\widehat{\text{CV}}[\bar{y}_d] = 0.1297$, and an approximate 95% CI for the mean farm acreage for counties in the western United States is [445,897, 751,463].

For estimating the total number of acres devoted to farming in the West, suppose we do not know how many counties in the population are in the western United States. Define

$$x_i = \begin{cases} 1, & \text{if county } i \text{ is in the western United States} \\ 0, & \text{otherwise} \end{cases}$$

and define $u_i = y_i x_i$. Then

$$\hat{t}_{yd} = \hat{t}_u = \sum_{i \in \mathcal{S}} \frac{3078}{300} u_i = 239{,}556{,}051.$$

The standard error is

$$\text{SE}(\hat{t}_{yd}) = 3078\sqrt{\left(1 - \frac{300}{3078}\right)}\frac{273005.4}{\sqrt{300}} = 46{,}090{,}460.$$

Regression in the simple random sample

- Regression Estimate

Let $\hat{B}_1$ and $\hat{B}_0$ be the ordinary least squares regression coefficients of the slope and intercept. For the straight line regression model,

$$\hat{B}_1 = \frac{\sum_{i \in \mathcal{S}} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in \mathcal{S}} (x_i - \bar{x})^2} = \frac{r s_y}{s_x},$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x},$$

and r is the sample correlation coefficient of x and y.

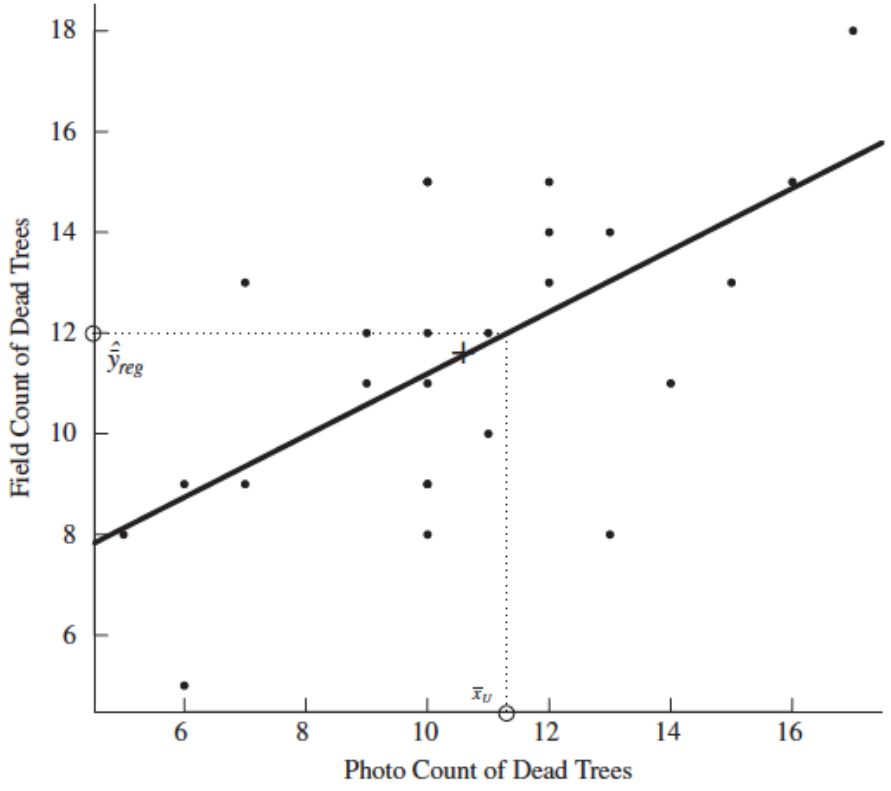
Suppose we know $\bar{x}_U$, the population mean for the x's. Then the regression estimator of $\bar{y}_U$ is the predicted value of y from the fitted regression equation when $x = \bar{x}_U$:

$$\hat{\bar{y}}_{\text{reg}} = \hat{B}_0 + \hat{B}_1 \bar{x}_U = \bar{y} + \hat{B}_1 (\bar{x}_U - \bar{x}).$$

Example

To estimate the number of dead trees in an area, we divide the area into 100 square plots and count the number of dead trees on a photograph of each plot. Photo counts can be made quickly, but sometimes a tree is misclassified or not detected. So we select an SRS of 25 of the plots for field counts of dead trees.

The plot of photo and field tree-count data, along with the regression line. Note that $\hat{\bar{y}}_{\text{reg}}$ is the predicted value from the regression equation when $x = \bar{x}_U$. The point $(\bar{x}, \bar{y})$ is marked by "+" on the graph.



Example

For these data, $\bar{x} = 10.6$, $\bar{y} = 11.56$, $s_y^2 = 9.09$, and the sample correlation between x and y is $r = 0.62420$. Fitting a straight line regression model gives

$$\hat{y} = 5.059292 + 0.613274x$$

with $\hat{B}_0 = 5.059292$ and $\hat{B}_1 = 0.613274$. In this example, x and y are positively correlated so that $\bar{x}$ and $\bar{y}$ are also positively correlated.

The regression estimate of the mean is

$$\hat{\bar{y}}_{reg} = 5.059292 + 0.613274(11.3) = 11.99.$$

The standard error is

$$SE(\hat{\bar{y}}_{reg}) = \sqrt{\left(1 - \frac{25}{100}\right)(9.09)(1 - 0.62420^2)} = 0.408.$$

The standard error of $\hat{\bar{y}}_{reg}$ is less than that for $\bar{y}$:

$$SE[\bar{y}] = \sqrt{\left(1 - \frac{25}{100}\right)\frac{s_y^2}{25}} = 0.522.$$

We expect regression estimation to increase the precision in this example because the variables photo and field are positively correlated. To estimate the total number of dead trees, use

$$\hat{t}_{yreg} = (100)(11.99) = 1199;$$

$$SE(\hat{t}_{yreg}) = (100)(0.408) = 40.8.$$

SAS ouput

Analysis of Estimable Functions

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Interval	
Total field trees	1198.92920	42.7013825	28.08	<.0001	1110.79788	1287.06053
Mean field trees	11.98929	0.4270138	28.08	<.0001	11.10798	12.87061

Ratio Estimation with Stratified Samples

- Combined ratio estimator

$$\hat{t}_{yrc} = \hat{B} t_x,$$

where

$$\hat{B} = \frac{\hat{t}_{y,\text{str}}}{\hat{t}_{x,\text{str}}}.$$

$$\hat{t}_{y,\text{str}} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj},$$

where the sampling weight is $w_{hj} = (N_h/n_h)$, and

$$\hat{t}_{x,\text{str}} = \sum_{h=1}^{H} N_h \bar{x}_h = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} x_{hj}.$$

- Separate ratio estimator

$$\hat{t}_{yrs} = \sum_{h=1}^{H} \hat{t}_{yhr} = \sum_{h=1}^{H} t_{xh} \frac{\hat{t}_{yh}}{\hat{t}_{xh}},$$

When to use?

Seperate estimator: It can improve efficiency if the $\hat{t}_{yh}/\hat{t}_{xh}$ vary from stratum to stratum, but should not be used when strata sample sizes are small because each ratio is biased, and the bias can propagate through the strata.

Combined estimator: has less bias when the sample sizes in some of the strata are small. When the ratios vary greatly from stratum to stratum, however, the combined estimator does not take advantage of the extra efficiency afforded by stratification as does the separate ratio estimator.

Example

A sampling plan was devised for estimating the total assessed value of all 940 claims. Since it was expected that the assessed value would be highly correlated with the incurred costs, ratio estimation is desirable here. Two strata were sampled: Stratum 1 consisted of the claims in which the incurred cost exceeded \$25,000, and stratum 2 consisted of the smaller claims (incurred cost less than \$25,000). Summary statistics for the strata are given in the following table, with r_h the sample correlation in stratum h:

Stratum	N_h	n_h	$\bar{x}_h$	s_{xh}	$\bar{y}_h$	s_{yh}	r_h
1	102	70	\$59,549.55	\$64,047.95	\$38,247.80	\$32,470.78	0.62
2	838	101	\$5,718.84	\$5,982.34	\$3,833.16	\$5,169.72	0.77

$$\hat{t}_{x,\text{str}} = \sum_{h=1}^{2} \hat{t}_{xh} = (102)(59{,}549.55) + (838)(5{,}718.84) = 10{,}866{,}442.02$$

$$\hat{t}_{y,\text{str}} = \sum_{h=1}^{2} \hat{t}_{yh} = (102)(38{,}247.80) + (838)(3{,}833.16) = 7{,}113{,}463.68$$

Example

and

$$\hat{B} = \frac{\hat{t}_{y,\text{str}}}{\hat{t}_{x,\text{str}}} = \frac{7{,}113{,}463.68}{10{,}866{,}442.02} = 0.654626755.$$

Using formulas for variances of stratified samples,

$$\begin{aligned}\hat{V}(\hat{t}_{x,\text{str}}) &= \left(1 - \frac{70}{102}\right)(102)^2 \frac{(64{,}047.95)^2}{70} + \left(1 - \frac{101}{838}\right)(838)^2 \frac{(5982.34)^2}{101} \\ &= 410{,}119{,}750{,}555,\end{aligned}$$

$$\begin{aligned}\hat{V}(\hat{t}_{y,\text{str}}) &= \left(1 - \frac{70}{102}\right)(102)^2 \frac{(32{,}470.78)^2}{70} + \left(1 - \frac{101}{838}\right)(838)^2 \frac{(5169.72)^2}{101} \\ &= 212{,}590{,}045{,}044,\end{aligned}$$

and

$$\begin{aligned}\widehat{\text{Cov}}\,(\hat{t}_{x,\text{str}}, \hat{t}_{y,\text{str}}) &= \left(1 - \frac{70}{102}\right)(102)^2 \frac{(32{,}470.78)(64{,}047.95)(0.62)}{70} \\ &\quad + \left(1 - \frac{101}{838}\right)(838)^2 \frac{(5169.72)(5982.34)(0.77)}{101} \\ &= 205{,}742{,}464{,}829.\end{aligned}$$

Example

Using the combined ratio estimator, the total assessed value of the claims is estimated by

$$\hat{t}_{yrc} = (9.407 \times 10^6)(0.654626755) = \$6.158 \text{ million}$$

with standard error

$$\begin{aligned} \text{SE}(\hat{t}_{yrc}) &= \frac{10.866}{9.407}\sqrt{[2.126 + (0.6546)^2(4.101) - 2(0.6546)(2.057)] \times 10^{11}} \\ &= \$0.371 \text{ million.} \end{aligned}$$

We use 169 = (number of observations) − (number of strata) degrees of freedom for the CI. An approximate 95% CI for the total assessed value of the claims is $6.158 \pm 1.97(0.371)$, or between \$5.43 and \$6.89 million.