Ratio Estimation

Survey Sampling
Statistics 4234/5234
Fall 2018

October 9, 2018

Ratio Estimation

Suppose the population consists of $\{(x_i, y_i) : i = 1, 2, \ldots, N\}$.

Here we require $x_i \geq 0$ and $y_i \geq 0$.

Population quantities

$$\bar{x}_u \quad \text{and} \quad t_x \quad \text{and} \quad S_x$$

and

$$\bar{y}_u \quad \text{and} \quad t_y \quad \text{and} \quad S_y$$

as usual.

Also

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

and

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

The data are a simple random sample of size n, denote by $\mathcal{S}$.

1. Estimate B by

$$\widehat{B} = \frac{\bar{y}}{\bar{x}}$$

Standard error is

$$\text{SE}(\widehat{B}) = \frac{1}{\bar{x}} \frac{s_e}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

where

$$s_e^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(y_i - \widehat{B} x_i \right)^2$$

Thus s_e is the sample standard deviation of the $e_i = y_i - \widehat{B} x_i$.

2. Estimate $\bar{y}_U$ by

$$\hat{\bar{y}}r = \hat{B}\bar{x}_U$$

Standard error is

$$\text{SE}(\hat{\bar{y}}r) = \bar{x}_U \text{SE}(\hat{B})$$

3. Estimate t_y by

$$\hat{t}_{yr} = \hat{B}t_x$$

Standard error is

$$\text{SE}(\hat{t}_{yr}) = t_x \text{SE}(\hat{B})$$

Weights

(4.1.4)

Recall, under SRS, selection probability for any unit $i \in \mathcal{U} = \{1, 2, \ldots, N\}$ is $\pi_i = n/N$.

The *sampling weight* for unit i is $w_i = 1/\pi_i$.

Then

$$\widehat{t}_y = N\bar{y} = \sum_{i \in \mathcal{S}} w_i y_i \tag{1}$$

Now

$$\hat{t}_{yr} = \hat{B} t_x = \frac{\bar{y}}{\bar{x}} N \bar{x}_U = \frac{\bar{x}_U}{\bar{x}} \hat{t}_y = \sum_{i \in \mathcal{S}} w_i^* y_i \tag{2}$$

where

$$w_i^* = \frac{\bar{x}_U}{\bar{x}} w_i = g_i w_i$$

With the ordinary estimator $\hat{t}_y$ in (1), the weights satisfy

$$\sum_{i \in \mathcal{S}} w_i = N$$

With the ratio estimator (2), the adjusted weights satisfy

$$\sum_{i \in \mathcal{S}} w_i^* x_i = t_x$$

The weight adjustments $g_i = \bar{x}_U/\bar{x}$ are called the *calibration factors*; they calibrate the weights to the auxiliary variable, rather than to the population size N.

Example: Farmland

Population values $N = 3078$ and $t_x = 964.47$.

Sample values $n = 300$ and $\bar{x} = .30195$ and $\bar{y} = .2979$.

Then

$$\hat{t}_y = N\bar{y} = \frac{3078}{300}\sum_{i\in\mathcal{S}} y_i = \sum_{i\in\mathcal{S}} w_i y_i$$

with $w_i = 10.26$, so each county in the sample represents 10.26 U.S. counties, we get

$$\sum_{i\in\mathcal{S}} w_i = 3078 = N$$

Meanwhile

$$\hat{t}_{yr} = \frac{t_x}{\bar{x}}\bar{y} = \frac{964.47}{.30195}\cdot\frac{1}{300}\sum_{i\in\mathcal{S}} y_i = \sum_{i\in\mathcal{S}} w_i^* y_i$$

with $w_i^* = 10.65$, so each county in the sample represents 10.65 U.S. counties, we see that

$$\sum_{i\in\mathcal{S}} w_i^* x_i = \frac{964.47}{.30195}(.30195) = 964.47 = t_x$$

Note

$$\sum_{i \in \mathcal{S}} w_i^* = 3194 > 3078 = N$$

For $\widehat{t}_y$ the weights are calibrated to N.

For $\widehat{t}_{yr}$ the weights are calibrated to t_x.

When does ratio estimation help?

(4.1.5)

Recall that

$$\text{MSE}\,(\bar{y}) = \frac{W_y^2}{n}\left(1 - \frac{n}{N}\right)$$

and

$$\text{MSE}\left(\hat{\bar{y}}_r\right) \approx \frac{1}{n}\left(S_y^2 - 2BRS_xS_y + B^2S_x^2\right)\left(1 - \frac{n}{N}\right)$$

and thus

$$\text{MSE}\left(\hat{\bar{y}}_r\right) - \text{MSE}\,(\bar{y}) \approx \frac{1}{n}\left(1 - \frac{n}{N}\right)BS_x\,(BS_x - 2RS_y)$$

A simple condition for exactly when we will have

$$\mathsf{MSE}\left(\hat{\bar{y}}_r\right) < \mathsf{MSE}\left(\bar{y}\right)$$

is

$$R > \frac{BS_x}{2S_y} = \frac{\mathsf{CV}(x)}{2\mathsf{CV}(y)}$$

If the CVs are approximately equal, then it pays to use ratio estimation when the correlation between x and y is larger than $1/2$.

Domain Estimation

(Section 4.2)

Suppose we want to estimate the average salary of female lawyers.

But our sampling frame is just a list of lawyers.

We take a SRS of n lawyers, and let

$$\bar{y}_d = \text{ average salary of women in sample}$$

just like you'd think.

This approach just requires a minor tweak to the standard error.

Note that $n_d =$ number of women in sample is *random*, not fixed by the sampling design.

The general set-up:

Write the population as

$$\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2 \cup \cdots \cup \mathcal{U}_D$$

And suppose domain d has N_d units, so

$$N = N_1 + N_2 + \cdots + N_D$$

The domain d population mean is

$$\bar{y}_{U_d} = \frac{1}{N_d} \sum_{i \in \mathcal{U}_d} y_i$$

Let $\mathcal{S}$ denote our SRS of size n from $\mathcal{U}$.

Then

$$\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_D$$

and

$$n = n_1 + n_2 + \cdots + n_D$$

The goal is inference about $\bar{y}_{U_d}$.

Domain estimation is ratio estimation!

Let

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

and

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

Then

$$\bar{y}_d = \frac{1}{n_d} \sum_{i \in \mathcal{S}_d} y_i = \frac{\sum\limits_{i \in \mathcal{S}} u_i}{\sum\limits_{i \in \mathcal{S}} x_i} = \frac{\bar{u}}{\bar{x}} = \widehat{B}$$

The standard error is thus

$$\mathsf{SE}\,(\bar{y}_d) = \mathsf{SE}(\widehat{B}) = \frac{1}{\bar{x}}\frac{s_e}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

Here $\bar{x} = \frac{n_d}{n}$ and

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

And so we have $\mathsf{SE}(\bar{y}_d) = \sqrt{\widehat{V}\,(\bar{y}_d)}$ where

$$\widehat{V}\,(\bar{y}_d) = \frac{n^2}{n_d^2}\frac{1}{n}\frac{(n_d - 1)s_{yd}^2}{n-1}\left(1 - \frac{n}{N}\right) = \frac{n(n_d - 1)}{n_d(n-1)}\frac{s_{yd}^2}{n_d}\left(1 - \frac{n}{N}\right)$$

Estimating domain totals

Case 1: If N_d is known, use $\hat{t}_{yd} = N_d \bar{y}_d$.

In this case,

$$\text{SE}(\hat{t}_{yd}) = N_d \text{SE}(\bar{y}_d)$$

Case 2: If N_d is unknown, note that $t_{yd} = t_u$, and use $\hat{t}_u = N\bar{u}$.

In this case,

$$\text{SE}(\hat{t}_u) = N \frac{s_u}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$