Ratio Estimation

Survey Sampling
Statistics 4234/5234
Fall 2018

October 2, 2018

(Section 4.1)

Consider the population $\mathcal{U} = \{1, 2, \ldots, N\}$

The quantity of primary interest is the variable $\{y_1, y_2, \ldots, y_N\}$.

There also exists an *auxiliary variable* $\{x_1, x_2, \ldots, x_N\}$ for which the quantity

$$t_x = \sum_{i=1}^{N} x_i$$

is known.

Here we are assuming that $y_i > 0, x_i > 0$ for $i = 1, 2, \ldots, N$.

In **ratio estimation** we exploit the fact that t_x is known to obtain more precise estimation of

$$t_y = \sum_{i=1}^{N} y_i$$

A couple more definitions.

Let

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

and

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

Example: Consider a population of agricultural fields.

Let x_i = acreage of field i (known for all units),

Let y_i = yield (bushels of grain) for field i, only known for units in the sample.

Then t_y = total yield in bushels.

Also $\bar{y}_U$ = average yield, in bushels per field.

And B = average yield in bushels per acre.

The ratio estimator

Given $\mathcal{S}$, a simple random sample of size n for $\mathcal{U} = \{1, 2, \ldots, N\}$.

Let

$$\bar{x} = \frac{1}{n} \sum_{i \in \mathcal{S}} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i \quad \text{and} \quad B = \frac{\bar{y}}{\bar{x}}$$

The ratio estimators are

$$\widehat{\bar{y}}_r = \widehat{B} \bar{x}_U \qquad \text{and} \qquad \widehat{t}_{yr} = \widehat{B} t_x$$

Why do ratio estimation?

1. If the ratio $B = \frac{\bar{y}_U}{\bar{x}_U}$ is itself of interest.

2. If the population size N is unknown, estimate it by $\widehat{N} = \frac{t_x}{\bar{x}}$ and $\widehat{N}\bar{y} = \widehat{B}t_x = \widehat{t}_{yr}$.

3. To increase precision of estimated means and totals.

4. To adjust estimates from sample so they reflect demographic totals — **poststratification** is actually a special case of ratio estimation.

Example: Estimate the total acres of farmland in 1992.

Let y_i = millions of acres of farmland in the ith county, for $i = 1, 2, \ldots, N = 3078$ counties in the United States.

We observe y_i for a random sample of $n = 300$ counties, and obtain

$$\bar{y} = 0.2979 \qquad \text{and} \qquad s = 0.34455$$

Thus we estimate

$$\hat{t}_y = N\bar{y} = 916.9 \text{ million acres of farmland}$$

with a standard error of

$$\text{SE}(\hat{t}) = N\frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} = 58.17 \text{ million acres}$$

Now let x_i = millions of acres of farmland in 1987 for the ith county, $i = 1, 2, \ldots, N = 3078$.

It is known that $t_x = 964.47$.

Thus $\bar{x}_U = \frac{t_x}{N} = 0.31334$.

For the sample data, $\bar{x} = 0.30195$.

We can also estimate t_y by

$$\hat{t}_{yr} = \hat{B} t_x = \frac{\bar{y}}{\bar{x}} t_x = \frac{.298}{.302}(964.47) = .9866(964.47) = 951.51$$

We see that $\hat{t}_{yr} > \hat{t}$.

Why does ratio estimation adjust upward in this problem?

Well

$$\hat{t}_{yr} = \frac{t_x}{\bar{x}}\bar{y} = \frac{\bar{x}_U}{\bar{x}}N\bar{y}$$

and here we have

$$\frac{\bar{x}_U}{\bar{x}} = \frac{.31334}{.30195} = 1.038$$

We *know* that $\bar{x}$ underestimates $\bar{x}_U$.

Which *suggests* that $\bar{y}$ underestimates $\bar{y}_U$.

So we adjust it upward, by 3.8%.

$$\bar{y} = 0.2979 \quad \text{and} \quad \widehat{\bar{y}}_r = 0.3091 = 1.038\bar{y}$$

Bias and MSE

(Section 4.1.2)

The ratio estimator

$$\hat{t}_{yr} = \hat{B} t_x = \frac{\bar{y}}{\bar{x}} t_x$$

is a *biased* estimator of t_y.

Bias can be OK if it leads to lower variance.

(Recall MSE = Var + Bias^2.)

Which ratio estimation does.

Can be shown

$$\text{Bias}\left(\hat{\bar{y}}_r\right) = E\left(\hat{\bar{y}}_r - \bar{y}_U\right) \approx \frac{1}{n\bar{x}_U}\left(BS_x^2 - RS_xS_y\right)\left(1 - \frac{n}{N}\right)$$

Also

$$\begin{aligned}\text{MSE}\left(\hat{\bar{y}}_r\right) &= \mathrm{E}\left[\left(\hat{\bar{y}}_r - \bar{y}_U\right)^2\right] \\ &\approx E\left[(\bar{y} - B\bar{x})^2\right] = V\left[\frac{1}{n}\sum_{i\in\mathcal{S}}(y_i - Bx_i)\right] \\ &= \frac{1}{n}\left(S_y^2 - 2BRS_xS_y + B^2S_x^2\right)\left(1 - \frac{n}{N}\right)\end{aligned}$$

Okay, how about something that's actually useful for analyzing survey data?

Sure.

Let

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

the sample variance of the $e_i = y_i - \widehat{B} x_i$ for $i \in \mathcal{S}$.

The estimated variance of $\widehat{B} = \bar{y}/\bar{x}$ is:

$$\widehat{V}\left(\widehat{B}\right) = \widehat{V}\left(\frac{\bar{y}}{\bar{x}}\right) = \frac{1}{\bar{x}^2} \frac{s_e^2}{n} \left(1 - \frac{n}{N}\right)$$

From there it's easy enough to get

$$\widehat{V}\left(\widehat{t}_{yr}\right) = \widehat{V}\left(\widehat{B}t_x\right) = t_x^2\widehat{V}\left(\widehat{B}\right)$$

and

$$\widehat{V}\left(\widehat{\bar{y}}_r\right) = \widehat{V}\left(\widehat{B}\bar{x}_U\right) = \bar{x}_U^2\widehat{V}\left(\widehat{B}\right)$$

For each $\theta \in \{B, t_y, \bar{y}_U\}$ let

$$\mathsf{SE}\left(\widehat{\theta}\right) = \sqrt{\widehat{V}\left(\widehat{\theta}\right)}$$

and an approximate 95% confidence interval for θ is

$$\widehat{\theta} \pm 1.96 \times \mathsf{SE}\left(\widehat{\theta}\right)$$

Example: $t_y =$ total farmland in 1992

$$\widehat{t}_y = N\bar{y} = 916.9 \qquad \text{and} \qquad \mathrm{SE}\left(\widehat{t}_y\right) = 58.17$$

Using ratio estimation we get

$$\widehat{t}_{yr} = \frac{\bar{y}}{\bar{x}} t_x = N \frac{\bar{x}_U}{\bar{x}} \bar{y} = 951.5$$

and get a standard error of

$$\mathrm{SE}\left(\widehat{t}_{yr}\right) = 5.55$$

less than one-tenth the standard error of $N\bar{y}$!