Domain Estimation

Survey Sampling
Statistics 4234/5234
Fall 2018

October 4, 2018

Often we want separate estimates for subpopulations, or **domains**.

Example: Want to estimate average salary for female lawyers, and male lawyers.

But sampling frame is just a list of lawyers.

Sample will include men and women, of course, but the resulting sample sizes are *random variables*, not fixed by the study design.

The appropriate estimate is fairly obvious — it's just what you think it should be. But standard error requires special care.

This is called **domain estimation**. It is a special case of **ratio estimation**, as we will demonstrate.

Domain estimation

The set-up:

Suppose the population $\mathcal{U}$ of N units has D domains:

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

with N_d units in domain d, satisfying

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

Any particular random sample $\mathcal{S}$ from $\mathcal{U}$, of n units, can be similarly partitioned

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

and

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

Estimator

Estimate the domain d population mean

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

by the domain d sample mean

$$\bar{y}_d = \frac{1}{n_d} \sum_{i \in \mathcal{S}_d} y_i$$

Note! Here n_d is a random variable.

That's what makes it 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

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

and

$$\bar{x}_U = \frac{t_x}{N} = \frac{N_d}{N}$$

and

$$t_u = \sum_{i=1}^{N} u_i = t_{yd}$$

and

$$\bar{y}_{U_d} = \frac{t_u}{N_d} = \frac{t_u}{t_x} = B$$

and

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

and

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

Thus

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

where

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

Thus we have

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

Interesting to note that

$$\mathsf{SE}\,(\bar{y}_d) = \left[\frac{n(n_d-1)}{n_d(n-1)}\right]^{1/2} \times (\text{SE for } \bar{y}_d \text{ based on SRS of } n_d)$$

Estimating the domain total

Wish to estimate $t_u = t_{yd}$.

Case 1: If N_d is known, estimate t_{yd} by $N_d \bar{y}_d$.

Standard error is $N_d \times \text{SE}\,(\bar{y}_d)$.

Case 2: If N_d is unknown, estimate $t_{yd} = t_u$ by

$$\hat{t}_{yd} = \hat{t}_u = N\bar{u}$$

and

$$\text{SE}\left(\hat{t}_{yd}\right) = N\text{SE}(\bar{u}) = N\frac{s_u}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$