# Friday, October 10

## Re-weighting Revisited

Recall that estimators of $\tau_y$ can often be written as

$$\hat{\tau}_y = \sum_{i \in \mathcal{S}} w_i y_i,$$

where $w_i$ is the *survey weight* for the $i$-th sampled element. Estimators of $\mu_y$ can be written as

$$\hat{\mu}_y = \frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i},$$

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

In practice, survey weights usually depend on two things.

1. The *sampling design* used. We have already seen examples of this.
   - For *simple random sampling*, $w_i = N/n$.
   - For *stratified random sampling*, $w_i = N_j/n_j$ if the $i$-th element is in the $j$-th stratum.
2. If and how information from *auxiliary variables* is incorporated into estimation.

Dependence of weights on the *sampling design* and *auxiliary variables* can be done by defining the weights as

$$w_i = d_i a_i$$

where $d_i$ is the *design weight* and $a_i$ is an *adjustment weight*.

**Design weights** depend on the *sampling design*. Typically they are the reciprocal of the inclusion probability $\pi_i$ so that $d_i = 1/\pi_i$.

**Adjustment weights** depend on *other information* about the elements, such as auxiliary variables.

Different sampling designs can result in different design weights, and different estimators can result in different adjustment weights.

Suppose we have a simple random sampling design so that the *design weights* are $d_i = N/n$ for all sampled elements. What are the sampling weights ($w_i$) for different choices of adjustments weights ($a_i$), and what is the resulting estimator of $\tau_y$?

1. Suppose we define $a_i = 1$ (i.e., no adjustment). Then $w_i = N/n$ and $\hat{\tau}_y = N\bar{y}$.
2. Suppose we define $a_i = \frac{\tau_x}{N\bar{x}}$. Then $w_i = \frac{\tau_x}{n\bar{x}}$ and $\hat{\tau}_y = \tau_x \bar{y}/\bar{x}$.
3. Suppose we define $a_i$ as
$$a_i = 1 + \frac{(\tau_x - \hat{\tau}_x)x_i}{\sum_{i \in \mathcal{S}} d_i x_i^2}$$
where $\hat{\tau}_x = N\bar{x}$. Then it can be shown that $\hat{\tau}_y = N\bar{y} + b(\tau_x - N\bar{y})$.
4. Suppose we define $a_i$ as
$$a_i = \frac{N_j n}{N n_j},$$

if the $i$-th element is in the $j$-th stratum. Then

$$\hat{\tau} = N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L,$$

which is the estimator we use for *post-stratification*.

## Calibration

Suppose we have determined a set of weights $(w_i)$ for the elements in a sample, and suppose we were to estimate $\tau_x$ using

$$\hat{\tau}_x = \sum_{i \in \mathcal{S}} w_i x_i.$$

The sample is said to be *calibrated with respect to the auxiliary variable* if $\tau_x = \hat{\tau}_x$ (i.e., the sample produces a *perfect* estimate of $\tau_x$). To *calibrate* our sample means that we are finding/adjusting weights so that it is calibrated in some way.

**Example**: Suppose we use a ratio estimator for a simple random sampling design. It can be shown that $\tau_x = \hat{\tau}_x$ where

$$\hat{\tau}_x = \sum_{i \in \mathcal{S}} w_i x_i,$$

and $w_i = \frac{\tau_x}{n\bar{x}}$.

**Example**: For post-stratification with a simple random sampling design, let $x_i$ be defined as

$$x_i = \begin{cases} 1, & \text{if the } i\text{-th element is in the first stratum,} \\ 0, & \text{otherwise.} \end{cases}$$

It can be shown that $\tau_x = \hat{\tau}_x$ where

$$\hat{\tau}_x = \sum_{i \in \mathcal{S}} w_i x_i,$$

and $w_i = N_j/n_j$ if the $i$-th element is in the $j$-th stratum.

### General Approach to Calibration

The objective is to select weights $(w_i)$ for the elements in the sample to meet the following criteria.

1. The weights are close (in some sense) to the original design weights.
2. The sample is calibrated with respect to the auxiliary variable(s).

Different approaches arise due to (a) the choice of auxiliary variable(s), and (b) what we mean by "close." This can be viewed as a constrained optimization problem.

## Raking

**Raking** is a calibration method based on *marginal totals* of two or more categorical auxiliary variables.

**Example**: Suppose we can post-stratify the sample into $2 \times 4 = 8$ strata such that we know the following totals for a population of $N = 2000$ elements.

| | Farthing | | | | |
|---|---|---|---|---|---|
| Age | North | South | East | West | Total |
| Juvenile | ? | ? | ? | ? | 500 |
| Adult | ? | ? | ? | ? | 1500 |
| Total | 200 | 700 | 800 | 300 | 2000 |

If we knew the total number of elements in each of the eight *combinations* of age and farthing, we could use regular post-stratification with the eight strata. But here we only know the number of elements for the two stratification variables *separately* (i.e., the *marginal* totals).

The goal of raking is to adjust the weights of the elements in the sample so that when we sum them by Farthing we get to totals reported in the table above, *and* when we sum them by age we get the age totals reported in the table above. This goal is similar to what we do in post-stratification except here the totals are the marginal totals.

There is no direct way to compute the necessary weights to calibrate the sample with respect to the known totals, but the weights can be obtained using an algorithm.[1] This is sometimes called **raking**. Raking can be generalized to more than two stratification variables.

[1] One simple algorithm that can do this is called *iterative proportional fitting*. It is a relatively simple algorithm and can even be done "by hand" with a simple calculator.