Calibration Estimators in Survey Sampling

Deville and Särndal

JASA, 1992, 376-382

Presented by Glen Meeden

The pdf file for this talk and its tex file are available on the class web page.

Sample Survey

From one point of view sample survey is the fundamental problem of statistics.

Sample survey is unusual because it introduces **unnecessary** randomization into a problem through the sampling design.

In the standard frequentist approach inferences are based on the sampling design.

Calibration is an attempt to improve inferences when bad samples are selected.

Does it work?

The usual simple setup

$\mathcal{U} = \{1, \ldots, k \ldots, N\}$ is a finite population.

y_k is the value of a single characteristic for unit k.

$y = (y_1, \ldots, y_N) \in \Theta \subseteq \mathcal{R}^N$ is the unknown parameter.

$s \subset \mathcal{U}$ is a sample.

$p(\cdot)$, a probability distribution defined on subsets of $\mathcal{U}$, is a sampling design.

$\mathbf{y}(s) = \{y_i : i \in s\}$ are the "seen"

$\mathbf{y}(s') = \{y_j : j \notin s\}$ are the "unseen"

The problem

To Estimate the Population Total

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

Under simple random sampling (srs) of size n given the sample s the usual estimator is

$$t_{y,s} = \frac{N}{n} \sum_{i \in s} y_i$$

Each unit in the sample is given the same weight.

This estimator is unbiased under srs.

Using a general sampling design $p(\cdot)$

For unit k let

$$\pi_k = \sum_{s\,:\,k\in s} p(s)$$

be its inclusion probability.

Under srs of size n, $\pi_k = n/N$.

Assume $\pi_k > 0$ for each k. Then

$$d_k = 1/\pi_k$$

is the design weight associated with unit k.

Under the design $p(\cdot)$ an unbiased estimator of the population total is

$$\widehat{y}_{y,\pi} = \sum_{i\in s} y_i/\pi_i = \sum_{i\in s} d_i y_i = \widehat{t}_{y,d}$$

Standard theory

Wants to calculate the variance of an estimator and find an unbiased estimator of this variance.

These facts can then be used to construct approximate confidence intervals for the parameter of interest.

Justifications are often asymptotic.

The underlying theory is based on repeated sampling from the sampling design $p(\cdot)$.

Adding an auxiliary variable

x_k is the value of an auxiliary variable at unit k.

$x = (x_1, \ldots, x_k, \ldots, x_N)$ may or may not be known a prior.

We assume

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

is known a priori and that we learn x_s, the values of the auxiliary variable for units in the sample.

Suppose we get a "bad" sample s where

$$\sum_{i \in s} d_i x_i \quad \text{is far from} \quad t_x$$

We calibrate!

Choosing other sets of weights.

Given a bad sample s we want weights w_i for $i \in s$ such that

$$\hat{t}_{x,w} = \sum_{i \in s} w_i x_i = t_x \quad \text{calibration}$$

and the w_i's are close to the d_i's.

Our new estimate will be

$$\hat{t}_{y,w} = \sum_{i \in s} w_i y_i$$

What distance measure should we use?

One possible family of distance measures

Let $1/q_i > 0$ be a set of know weights unrelated to the d_i.

Subject to

$$\hat{t}_{x,w} = \sum_{i \in s} w_i x_i = t_x \quad \text{calibration}$$

we wish to minimize

$$\sum_{i \in s} \frac{(w_i - d_i)^2}{d_i q_i}$$

Method of Lagrange multipliers considers

$$\sum_{i \in s} \frac{(w_i - d_i)^2}{d_i q_i} - 2\lambda \sum_{i \in s} w_i x_i$$

Differentiating with respect to w_i and setting it equal to 0

Solving for λ

yields

$$w_i = d_i + \lambda d_i q_i x_i$$

Multiplying by x_i we have

$$w_i x_i = d_i x_i + \lambda\, d_i q_i x_i^2$$

and summing over the sample we get

$$\sum_{i \in s} w_i x_i = \sum_{i \in s} d_i x_i + \lambda \sum_{i \in s} d_i q_i x_i^2$$

Finally we use the calibration constraint to solve for λ.

Form of the solution

$$\lambda = \frac{\sum_{i \in s} w_i x_i - \sum_{i \in s} d_i x_i}{\sum_{i \in s} d_i q_i x_i^2}$$
$$= \frac{t_x - \hat{t}_{x,d}}{\sum_{i \in s} d_i q_i x_i^2}$$

Now

$$w_i = d_i + \frac{t_x - \hat{t}_{x,d}}{\sum_{i \in s} d_i q_i x_i^2} q_i x_i d_i$$

and so

$$\hat{t}_{y,w} = \sum_{i \in s} w_i y_i$$
$$= \hat{t}_{y,d} + (t_x - \hat{t}_{x,d}) \frac{\sum_{i \in s} d_i q_i x_i y_i}{\sum_{i \in s} d_i q_i x_i^2}$$

A special case yields the ratio estimator

If $q_i = 1/x_i$ then

$$\hat{t}_{y,w} = t_x \frac{\hat{t}_{y,d}}{\hat{t}_{x,d}}$$

which in the case of srs becomes

$$\hat{t}_{y,w} = \left(\sum_{i=1}^{N} x_i \right) \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i}$$

which is the usual ratio estimator. This estimator is based on the idea

$$\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} \simeq \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i}$$

Some theoretical results

They show for a large class of distance measures and a vector of auxiliary variables that

1. λ_s has a unique solution with probability tending to one as $n \to \infty$.

2. λ_s tends to 0 in design probability.

3. $\widehat{t}_{y,w}$ is design-consistent and

$$N^{-1}(\widehat{t}_{y,w} - \widehat{t}_{y,d}) = \mathcal{O}_p(n^{-1/2})$$

4. Variance estimating is not so clear.

Some remarks

No guidelines for selecting a distance measure!

Note we always have

$$t_y = \sum_{i \in s} y_i + \sum_{i \notin s} y_i$$

The basic question in finite population sampling is:

How to relate the seen to the unseen?

Predictive approach

Assumes a model that relates y and x like simple linear regression. Uses the sample to estimate the unknown parameters of the model and uses the estimated model to predict the unseen y_i's not in the sample. Variance estimation uses the model not the sampling design.

"Finite population sampling and inference" by Valliant, Dorfman and Royall, Wiley (2000)

Bayesian approach

Find a joint prior distribution for the population

$$P\Big(y_1, y_2, \ldots, y_N\Big)$$

After observing sample must find

$$P\Big(y_j \ : \ j \notin s \ \Big| \ y_i \ : \ i \in s\Big)$$

the conditional distribution of the unseen given the seen.

Ericson (1969) JRSSB

Simulate from the posterior to get completed copies of the entire population. Inferences do not depend of the design only on the prior.

Hard to do.