Small Area Estimation

1 - Small area estimation problem

2 - Estimation for domains - Direct estimators – estimation for planned domains

3 – Coefficient of Variation and Minimum level of precision

4- Estimation for unplanned domains and/or where the sample size is not enough for the minimum level of precision

Recap

- Target parameters in the domain:

- Total of the study variable $t = \sum_{k \in U} y_k$
- Mean of the study variable $\bar{y} = \sum_{k \in U} y_k / N$
- At risk of poverty rate $P_0 = \frac{1}{N} \sum_{i=1}^{N} I(y_i < z).$
- Poverty gap $P_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{G_i}{z}.$

Horvitz-Thompson estimator of domain totals

Horvitz-Thompson (HT) estimator (*expansion estimator*) is the basic *design-based direct* estimator of the domain total $t_d = \sum_{k \in U_d} y_k$, $d = 1, \ldots, D$:

$$\hat{t}_{dHT} = \sum_{k \in U_d} I_k y_k / \pi_k = \sum_{k \in s_d} y_k / \pi_k = \sum_{k \in s_d} a_k y_k \qquad (1)$$

HT estimates of domain totals are additive: they sum up to the HT estimator $\hat{t}_{HT} = \sum_{k \in s} a_k y_k$ of the population total $t = \sum_{k \in U} y_k$

As $E(I_k) = \pi_k$, the HT estimator is design unbiased for t_d

$$\hat{t}_{dHT} = \sum_{U} I_k y_k / \pi_k$$

RANDOM VARIABLE $I_k \begin{cases} 1 & k \in s \\ 0 & k \notin s \end{cases}$

$$\pi_k = Pr(I_k = 1)$$ PROBABILITY OF INCLUSION of K

is $\hat{t}_{dHT}$ unbiased for t ?

is $E[\hat{t}_{dHT}] = t = \sum_{k=1}^{N} y_k$?

$\Longrightarrow$ LET'S SEE!

$$E[\hat{t}] = E\left[\sum I_k y_k / \pi_k\right]$$

$$= \sum \left[E(I_k y_k / \pi_k)\right]$$

$$= E\left[\sum I_k y_k / \pi_k\right]$$

RANDOM VARIABLE: BERNOULLI R.V. ITS EXP VALUE $= \pi_k$

$$= \sum \left[E(I_k) \cdot y_k / \pi_k\right]$$

$$= \sum \left[\cancel{\pi_k} \cdot y_k / \cancel{\pi_k}\right]$$

$$= \sum y_k = t$$

$\hat{t}$ IS UNBIASED FOR t

Variance estimation for HT - 3

Variance estimation for planned domains in practice

$$\hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d - 1)} \sum_{k \in s_d} \left(n_d a_k y_k - \hat{t}_{dHT}\right)^2 \qquad (4)$$

For example, SAS Procedure SURVEYMEANS uses (4)

- The variance of HT estimator can be too "large"
- That is the "sampling error" associated with the estimator too large to consider it reliable for estimating the total

$$t = \sum_{k \in U} y_k$$

- …….this even if it is unbiased

EXAMPLE

Simple random sampling

n_d : sample size in the domain

$$\pi_k = \frac{n_d}{N_d} \qquad \hat{t} = \sum \frac{y_k}{n_d} \cdot N_d = $$

$$= \bar{y}_d \cdot N_d$$

APPROX ESTIMATED VARIANCE $\hat{V}(\hat{t}) \doteq$

$$\frac{1}{n_d(n_d-1)} \sum \Big(n_d \cdot \overbrace{\frac{N_d}{n_d}}^{a_k} y_k - \hat{t} \Big)^2$$

ESTIMATED VARIANCE

$$\boxed{\hat{V}(\hat{t}) = \sum_{k=1}^{n} \left(\frac{1-\pi_k}{\pi_k^2}\right) y_k^2 + \sum\sum_{k \neq l} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l}\right) \frac{y_k y_l}{\pi_{kl}}}$$

when $\pi_k = \frac{n_d}{N_d}$, $\pi_{kl} = \frac{n_d}{N_d} \cdot \frac{n_d - 1}{N_d - 1}$

$$\hat{V}(\hat{t}) = N_d^2 \cdot \frac{S^2}{n_d} \left(\frac{N_d - n_d}{N_d - 1}\right)$$

where $S^2 = \frac{1}{n_d - 1} \sum (y_i - \bar{y}_d)^2$ estimated variance of the study variable in the domain

$$\simeq N_d^2 \frac{S^2}{n_d}\left(1 - \frac{n_d}{N_d}\right)$$

$\Rightarrow N_d^2,\ S^2$: out of our control

$\Rightarrow n_d$

VALUE of SAMPLE SIZE IS CRUCIAL
TOO $\downarrow \hat{V}(\hat{t})$

This is one of the reasons why it is important to plan the domains at the design stage of the survey. In this case the allocation of the sample to the domains is controlled by the researcher.
Instead, in case of “unplanned” domains, the distribution of the whole sample by Domain is the result of a random allocation and the n_d may be “too” small to reduce the value of the sampling error $\hat{V}(\hat{t})$

There is a minimum level of precision required (or a maximum level of sampling error accepted) for an estimate to be “statistically sound”!!!!!

3 – Coefficient of Variation and Minimum level of precision

Definition and interpretation of the Coefficient of Variation

We want a “statistically sound” estimate even for unplanned domains and/or where the sample size is not enough for the minimum level of precision

“Statistically sound estimate” 1

In descriptive statistics: the coefficient of variation (CV) is the ratio of the standard deviation to the value of the mean

Coefficient of Variation = (Standard Deviation/ mean) * 100.

For example, the expression “The standard deviation is 15% of the mean” is a coefficient of variation.

"Statistically sound estimate" 2

In descriptive statistics:

the CV is particularly useful when you want to compare variability of two different groups or populations.

For example: Income in Pop A has CV=15%, Income in Pop B has CV=30%...the distribution of income in Pop B has more dispersion (is more variable)

“Statistically sound estimate” 3

In Statistical Inference: the coefficient of variation (CV) is the ratio of the standard error of an estimate to the value of the estimate

Coefficient of Variation = (Standard Error / Estimate) * 100.

For example, the expression “The standard error is 15% of the estimate” is a coefficient of variation.

“Statistically sound estimate” 3

In Statistical Inference:

For example: estimator A has CV=15%, estimator B has CV=30%...the sampling distribution of estimator B has more dispersion (is more variable) and the estimator B is less efficient than A

“Statistically sound estimate” 4

In sample survey (Inference)

The CV is particularly useful when you want to assess the accuracy (efficiency + unbiasdeness) of the results of a survey (estimate):

The MSE (Mean Squared Error) is equal to Variance + Bias^2

MSE(estimator) = Variance(estimator)+bias(estimator)^2

Coefficient of Variation = square root(MSE(estimate)) /(Estimate) * 100.

For example, the expression “The sqrt(MSE) is 15% of the estimate” is a coefficient of variation and it is a measure of the accuracy of the estimate

“Statistically sound estimate” 5

It means accurate, with a low CV.

When I say low it means that its value should not exceed the 20-30% of the value of the estimate itself.

Many Official Statistical Agencies do not publish estimates with CV higher than 20%

Estimators for Domain Totals

- The sample size is enough and the CV associated with HT estimator is less than 20%
  - Accurate estimate

- The sample size is not enough and the CV associated with HT estimator is greater than 20%
  - Not accurate estimate: cannot publish it
  - This is often the case when the domain is unplanned

4 – Estimation for unplanned domains and/or where the sample size is not enough for the minimum level of precision

SAE models