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 $P_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{G_i}{z}.$

More on direct estimators for domains

There are more complex formulations of the direct estimator, other than HT estimator:

- ratio estimator
- regression estimator
- calibration estimator

These estimators are based on the use of auxiliary information, the HT estimator is based only on the y_i values of the sampled units

Auxiliary information

Auxiliary: providing supplementary or additional help and support to the estimation process.

- The auxiliary information about the population in the domain may include one or more known variables to which the variable of interest is approximately related.
- The auxiliary information typically is easy to measure, whereas the variable of interest may be expensive to measure.

Examples in case of one aux variable -1

Population units: 1, 2, ... , N_d

variable of interest : y_1 , y_2, ... , y_{Nd} (expensive or costly to measure)

auxiliary variable : x_1 , x_2, ... , x_{Nd} (known)

A national park is partitioned into N_d units.

- y_i = the number of animals in unit i
- x_i = the size of unit i

Examples in case of one aux variable -2

Another example might be where a certain domain (city) has N_d bookstores.

- y_i = the sales of a given book title at bookstore i
- x_i = the size of the bookstore i

A third example would be a region that has N_d households.

- y_i = the consumption of the household
- x_i = the income of the household

Ratio estimator

For sake of simplicity let me indicate the domain population size by N (instead of by N_d)

If $\tau_y = \sum_{i=1}^{N} y_i$ and $\tau_x = \sum_{i=1}^{N} x_i$ then, $\frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}$ and $\tau_y = \frac{\mu_y}{\mu_x} \cdot \tau_x$

The ratio estimator, denoted as $\hat{\tau}_r$, is $\hat{\tau}_r = \frac{\bar{y}}{\bar{x}} \cdot \tau_x$

Ratio estimator

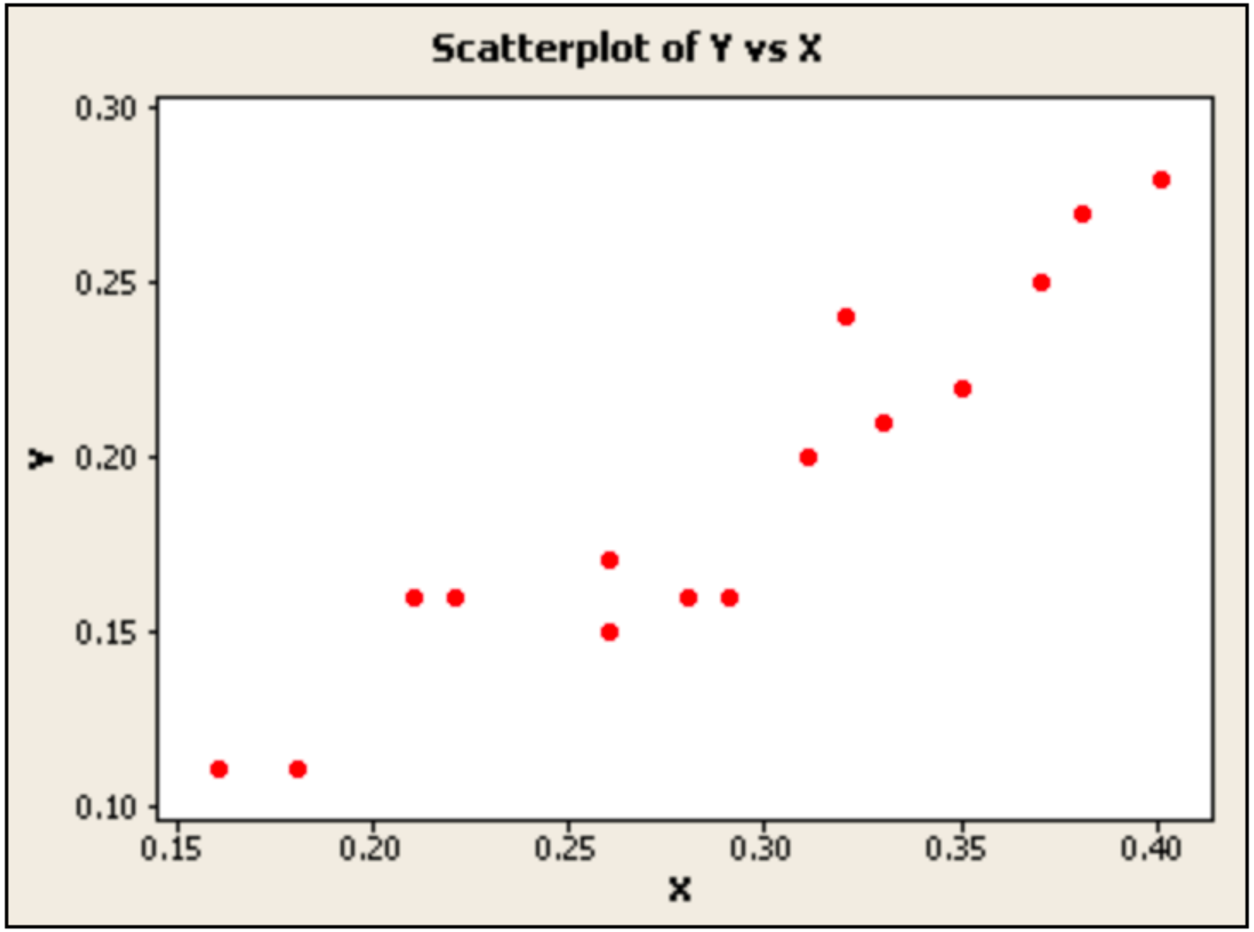
The estimator is useful in the following situations:

1 - when X and Y are highly correlated through the origin

$Var(\hat{\tau}_r)$ is less than $Var(N\bar{y})$

that is the ratio estimator is more efficient then the HT estimator

High correlation through the origin



Ratio estimator

2 - The case when N is unknow, then it provides a way to estimate the total

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

Since when N is unknown, one cannot use the HT estimator

$$N\bar{y}.$$

HT estimator under srs

$N_d \bar{y}_d$ is the HT estimator for the domain d

infact if s.r.s is used in the domain

$\pi_i = \frac{n_d}{N_d}$ and $\hat{t}_{HT} = \sum y_i / \pi_i$

$$= \sum y_i / N_d \cdot (n_d)^{-1}$$

$$= \sum \frac{y_i}{n_d} \cdot N_d =$$

$$= N_d \cdot \frac{\sum y_i}{n_d} =$$

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

Note that for simplicity we used N_d for N_d.

Properties of the Ratio estimator

- This estimator is not unbiased, but it is unbiased for large samples when the sampling design is a simple random sampling

- The ratio estimator of a mean is

$$\hat{\mu}_r = \frac{\bar{y}}{\bar{x}} \cdot \mu_x$$

Variability of the Ratio estimator

Variance of the estimator

$$Var(\hat{\mu}_r) \approx \left(\frac{N-n}{N}\right) \cdot \frac{\sigma_r^2}{n}$$

where

$$\sigma_r^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \frac{\tau_y}{\tau_x} \cdot x_i\right)^2$$

is estimated by

$$s_r^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \frac{\bar{y}}{\bar{x}} \cdot x_i\right)^2$$

Confidence Interval for $\hat{\mu}_r$

Approximate 100(1-a)% confidence interval

$$\hat{\mu}_r \pm t_{n-1,\alpha/2}\sqrt{\hat{V}ar(\hat{\mu}_r)}$$

where $t_{n-1,a/2}$ is the percentile read on the **t Student** distribution table

Confidence Interval for $\hat{\tau}_r$

Approximate 100(1-a)% confidence interval based on t student tables for

$$\hat{\tau}_r = N\hat{\mu}_r = \frac{\bar{y}}{\bar{x}} \cdot \tau_x$$

is obtained using

$$\boxed{\hat{V}ar(\hat{\tau}_r) = N \cdot (N - n)\frac{s_r^2}{n}}$$

Ratio estimator *r*

- In some cases we are interested in estimating

$$R = \frac{\tau_y}{\tau_x} \left(\text{also, } \frac{\mu_y}{\mu_x} \right)$$

That is for example ratio such as the monthly food budget compared to the monthly income per family

Ratio estimator *r*

- The sample ratio is the estimate for R

$$r = \frac{\bar{y}}{\bar{x}}$$

- where *r* is the ratio between the sample mean of y and x variables

Variance of *r*

Variance of the estimator

$$Var(r) \approx \left(\frac{N-n}{N\mu_x^2}\right)\frac{\sigma_r^2}{n}$$

Estimated variance of the estimator

$$\hat{V}ar(r) \approx \left(\frac{N-n}{N\mu_x^2}\right)\frac{s_r^2}{n}$$

Confidence Interval for *r*

Approximate 100(1-a)% confidence interval based on t student tables for

$$r = \frac{\bar{y}}{\bar{x}}$$

is obtained using

$$\boxed{\hat{V}ar(r) \approx \left(\frac{N-n}{N\mu_x^2} \right) \frac{s_r^2}{n}}$$

- Examples of application of the ratio estimator during the R lab