Statistiska centralbyrån Statistics Sweden

10
Estimation in the presence of both nonresponse and frame imperfections

The estimation procedure needs to deal simultaneously with :

sampling error
nonresponse error
coverage error

It is not a trivial step to accomodate the third kind of error and derive a firmly established methodology!

There are few ”conventional methods” to compare with. Further development is needed !

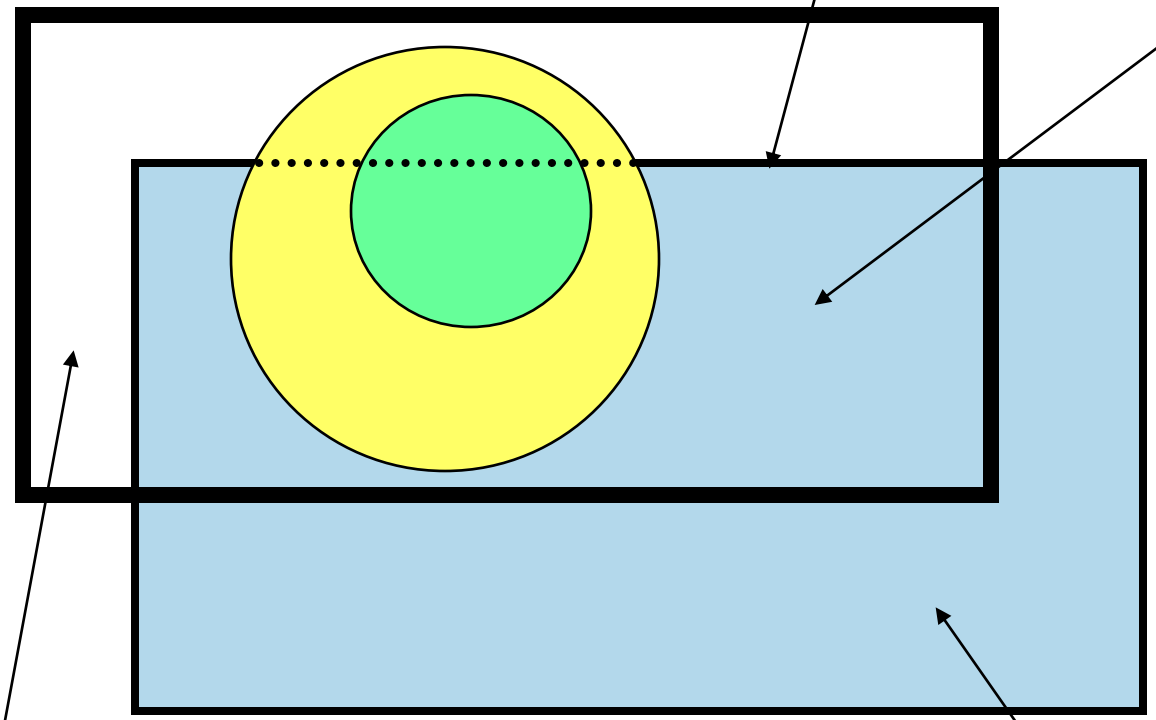
Here we show how ”calibration thinking” may be applied to deal with this complex situation.

Frame population: U_F

Target population: U

"Persisters"

$(U_P = U \cap U_F)$



Overcoverage $(U_F - U_P)$

Undercoverage $(U - U_P)$

A more detailed system of notation is needed :

s_F = sample from the frame U_F
r_F = response subset of the sample s_F
o_F = nonresponse subset of s_F

$s_F = r_F \cup o_F$

subscript P = ”in the target population”
subscript \P = ”not in the target population”

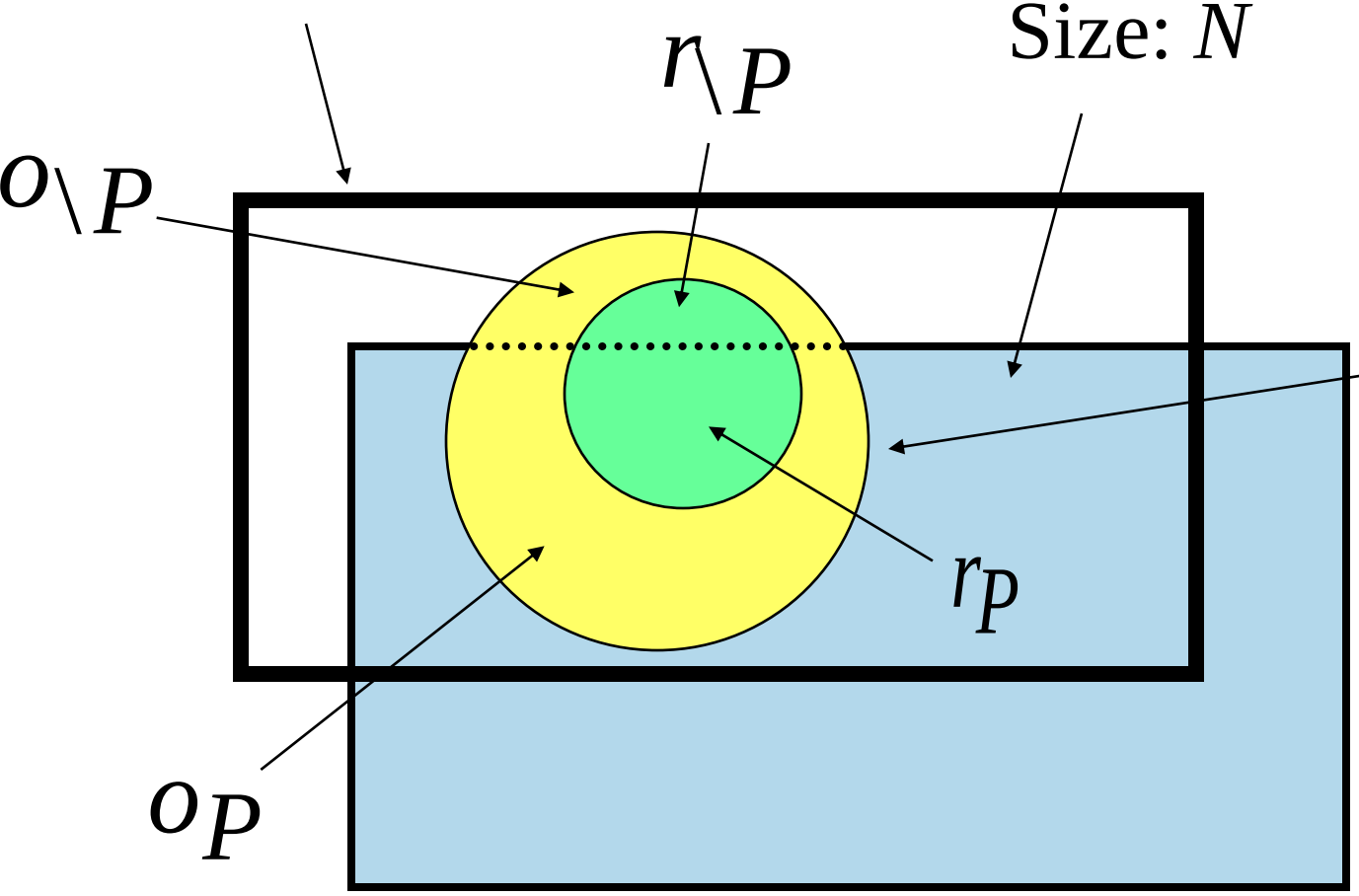
$r_F = r_P \cup r_{\backslash P}$
$o_F = o_P \cup o_{\backslash P}$

Frame population: U_F

Size: N_F

Target population: U

Size: N



Sample: s_F

Size: n_F

$$o_F = o_P \cup o \setminus P$$

Problems:

- the absence of observed *y*-data from the undercoverage set

- the absence of a correct auxiliary vector total for the target population $\boldsymbol{U}$

- difficulties of decomposing the nonresponse set o_F into its subsets o_P and $o \setminus P$. (This would be needed for example to identify the elements that need imputation.)

Two procedures for estimating Y_U

(i) by the sum of (a) an estimate of the persister total Y_{UP} and (b) an estimate of the undercoverage total Y_{U-UP}

(ii) by direct estimation of the target population total Y_U

i) a. Estimation of the persister total Y_{U_P}

The persister set U_P is a domain of U_F and the corresponding response sets are r_P and $r_F = r_P \cup r_{\setminus P}$

Let us define

$$y_{Pk} = \begin{cases} y_k & \text{if } k \in U_P = U \cap U_F \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{Y}_{U_P W} = \sum_{r_F} w_k y_{Pk} = \sum_{r_P} w_k y_k$$

where $w_k = d_k v_k$ and

$$v_k =$$

$$= 1 + (\sum_{U_F} \mathbf{x}_k^* - \sum_{r_F} d_k \mathbf{x}_k^*)' \left(\sum_{r_F} d_k \mathbf{x}_k^* (\mathbf{x}_k^*)' \right)^{-1} \mathbf{x}_k^*$$

Example

U_F is divided into strata, U_{Fh}, $h = 1,..., H$

STSRS: n_{Fh} from N_{Fh}; m_{Fh} respond

Aux. vector: $\mathbf{x}_k = \mathbf{x}_k^* = \boldsymbol{\gamma}_k =$ stratum identifier

The general formula gives a commonly used estimator of the persister total:

$$\hat{Y}_{U_P W} = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{Ph} + m_{\setminus Ph}} \sum_{r_{Ph}} y_k =$$

$$= \sum_{h=1}^{H} \frac{N_{Fh}}{m_{Fh}} \sum_{r_{Fh}} y_{Pk}$$

i) b. Estimation of the undercoverage total

$$Y_{U-U_P}$$

An estimator is produced using a judgemental or model-dependent procedure.

ii) Direct estimation of the target population total Y_U

Let $\tilde{\mathbf{x}}$ denote an approximation of $\sum_U \mathbf{x}_k^*$

$\hat{Y}_{UW} = \sum_{r_P} w_k y_k$ where

$w_k = d_k v_k$ and

$$v_k = 1 + \left(\tilde{\mathbf{x}} - \sum_{r_P} d_k \mathbf{x}_k^* \right)' \left(\sum_{r_P} d_k \mathbf{x}_k^* (\mathbf{x}_k^*)' \right)^{-1} \mathbf{x}_k^*$$

Example

U_F is divided into strata, U_{Fh}, $h = 1,..., H$

STSRS: n_{Fh} from N_{Fh} ; m_{Fh} respond

Aux. vector: $\mathbf{x}_k = \mathbf{x}_k^* = \boldsymbol{\gamma}_k$ = stratum indicator

The resulting estimator of the target population total is

$$\hat{Y}_{UW} = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{Ph}} \sum r_{Ph} y_k$$

It is a commonly used estimator (at Statistics Sweden).

A case study : The survey

Transition from upper secondary school to higher education carried out by Statistics Sweden, redesigned in 2006.

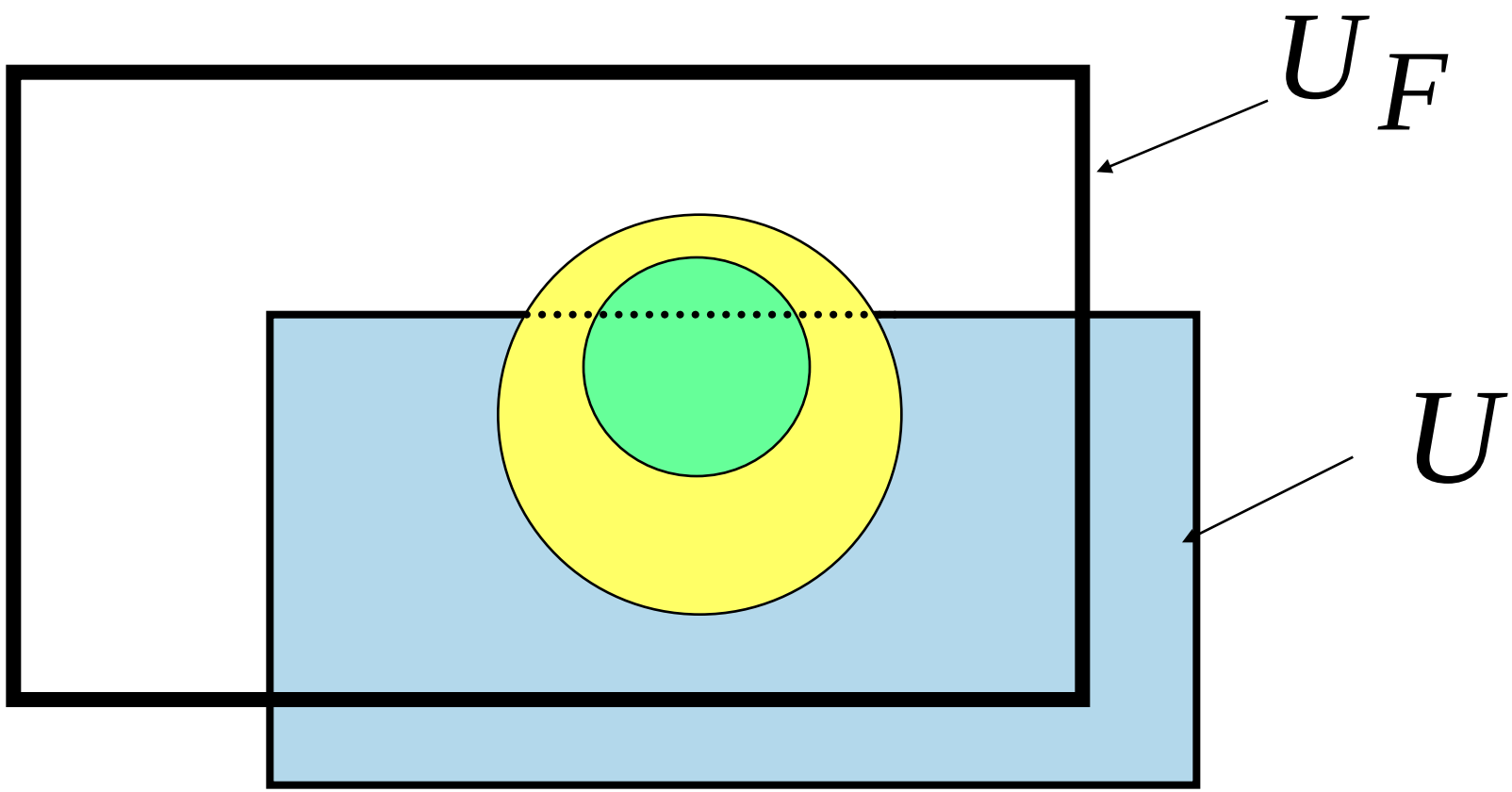
Single stage sampling.

Elements: Students in third year upper secondary school.

Important study variables:

(a) Intentions of pursuing studies at university

(b) The university programmes viewed as the most interesting ones



U : Third-year upper secondary students

U_F: Second-year upper secondary students, preceding year

The estimator used before the redesign

$$\hat{Y}_{U_P W} = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{Ph} + m_{\backslash Ph}} \sum r_{Ph}\, y_k = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{Fh}} \sum r_{Fh}\, y_{Pk}$$

At first look one would believe that it is an underestimation, but it turns out to be an overestimation for the following reasons:

(i) The overcoverage is considerable greater than the undercoverage $(N_F > N)$

(ii) The response propensity is very low among nonpersisters $(m_{\backslash P_h}$ near zero$)$

The solution:

We discovered a good approximation $\tilde{\mathbf{X}}$

of $\sum_U \mathbf{x}_k^*$ and estimated the target

population total by the direct estimation method

Aux. variables:

-”final mark” at the end of grade 9

- parental variables: level of education, income and civil status

Improvements compared with the old design:

- The estimates of totals undergo considerable change
- Estimates of proportions undergo little change
- The estimated variances for proportions were not much reduced