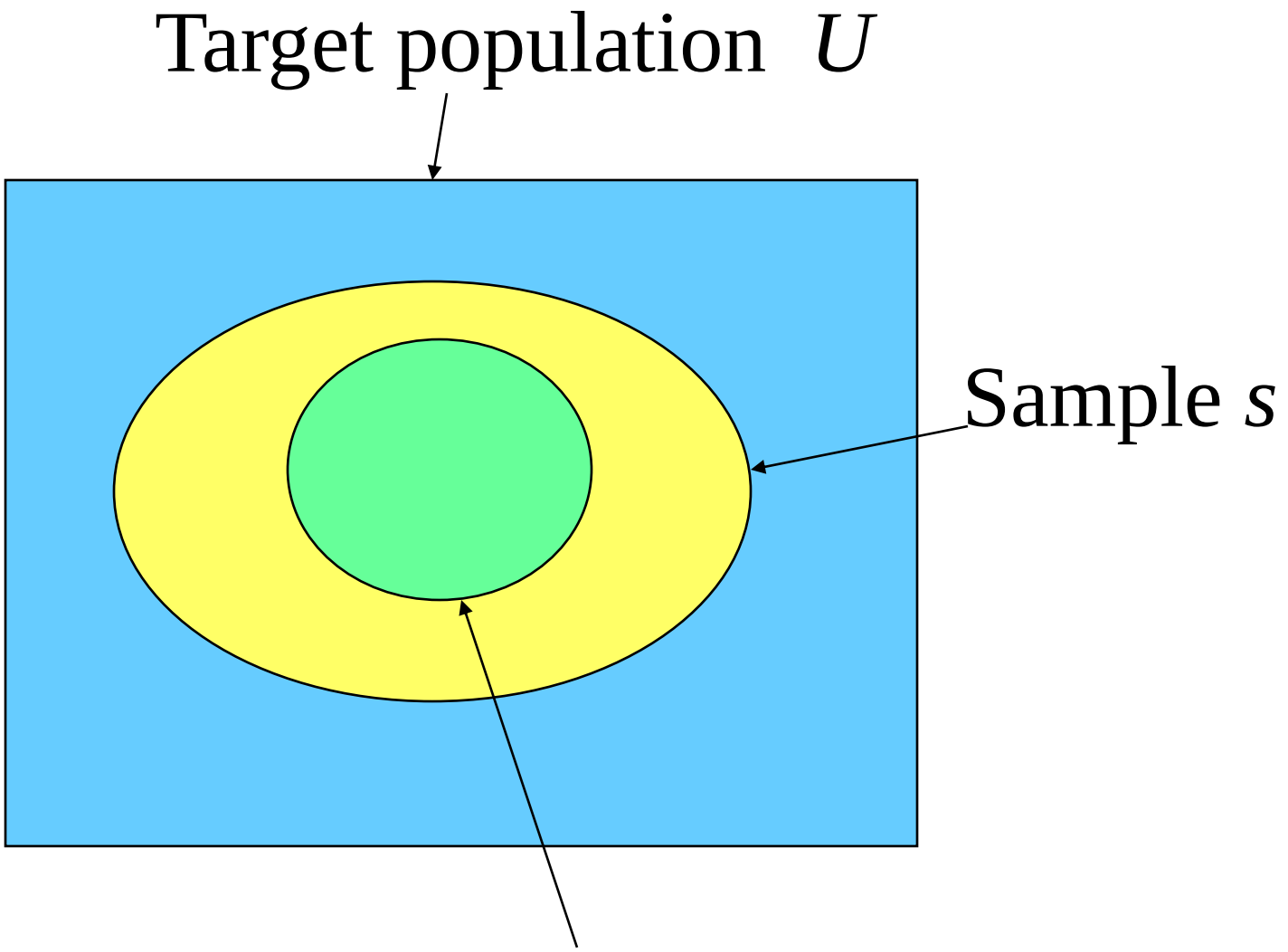
Statistiska centralbyrån Statistics Sweden

3
Weighting of data. Types of auxiliary information. The calibration approach.

Structure



Notation and terminology

Population U

of elements $k = 1, 2, ..., N$

Sample s (subset of U)

Non-sampled : $U - s$

Response set r (subset of s)

Sampled but non-responding : $s - r$

$$U \supseteq s \supseteq r$$

The objective

is to estimate the total $Y = \sum_U y_k$

In practice, many *y*-totals and functions of *y*-totals.
But we can focus here on one total.
No need at this point to distinguish
item NR and **unit NR**.
Perfect frame coverage assumed.

The response set r

is the set for which we observe y_k

Available y-data : y_k for $k \in r$

Missing values : y_k for $k \in s - r$

where $r \subseteq s \subseteq U$

Nonresponse means that $r \subset s$

Full response means that $r = s$

with probability one

Two phases of selection

Phase one : ***Sample selection***
with known ***sampling design***

Phase two : ***Response selection***
with unknown ***response distribution***

Phase one: ***Sample selection***

Known ***sampling design*** : $p(s)$

Known ***inclusion prob.*** of k : π_k

Known ***design weight*** of k :

$$d_k = 1/\pi_k$$

Phase two: ***Response selection***

Unknown ***response distribution*** : $q(r|s)$

Unknown ***response prob.*** of k: θ_k

Unknown ***response influence*** of k:

$$\phi_k = 1/\theta_k$$

A note on terminology

$d_k = 1/\pi_k$ computable ***weight***

$\phi_k = 1/\theta_k$ unknown; not a weight, called ***influence***

Sample weighting combined with response weighting

Desired (but impossible) combined weighting :

$$d_k \times \phi_k = \frac{1}{\pi_k} \times \frac{1}{\theta_k}$$

known unknown

Why can we not use the design weights $d_k = 1/\pi_k$ without any further adjustment ?

Answer: They are not large enough when there is NR.

$$\hat{Y} = \sum_r d_k y_k \quad \Rightarrow \text{underestimation}$$

We must expand the design weights.

Desirable nonresponse weighting

$$\hat{Y} = \sum_r \frac{d_k}{\theta_k} y_k = \sum_r d_k \phi_k y_k$$

Cannot be computed,

because unknown influences $\phi_k = 1/\theta_k$

The calibration approach

Calibration estimation is a highly general approach. It covers many situations arising in practice.

Calibration uses auxiliary information.

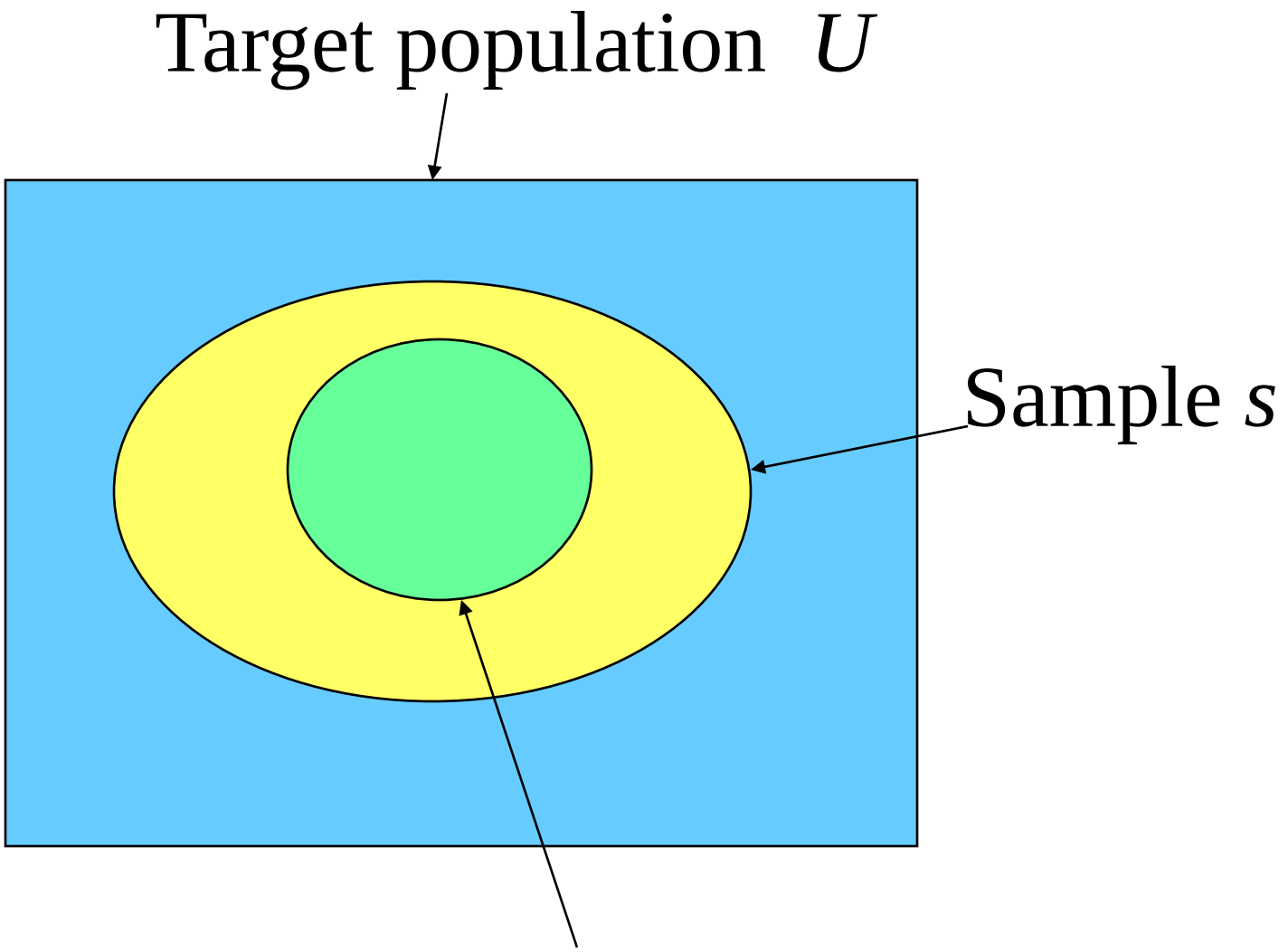
Auxiliary information

may exist

at the ***population level***

at the ***sample level***

Structure



Levels of information

Distinguish :

- ***Information at the population level***. Such info, taken from ***population registers***, is particularly prevalent and important in Scandinavia, The Netherlands, and increasingly elsewhere in Europe
- ***Information at the sample level.*** Such info may be present in any sample survey

Levels of information

Notation : Two types of auxiliary vector

$\mathbf{x}_k^*$ transmits information
at the ***population level***

$\mathbf{x}_k^\circ$ transmits information
at the ***sample level***

Auxiliary vector, population level

Distinguish two situations :

- $\mathbf{x}_k^*$ ***known value*** for every k in U

  given in the frame, or coming from admin. registers

- the total $\mathbf{X}^* = \sum_U \mathbf{x}_k^*$ is ***imported*** from accurate outside source

  $\mathbf{x}_k^*$ need not be known for every k

Sources of variables for the star vector $\mathbf{x}_k^*$

- the existing frame
- by matching with other registers

Examples of variables for the star vector :
For persons : age, sex, address, income

To related persons: Example, in survey of school children, get (by matching) variables for parents.

Auxiliary information: Examples

For every k in U, suppose known :

- Membership in one out of 2 3 = 6 possible groups, e.g., ***sex*** by ***age group***

- The value x_k of a continuous variable x

  e.g., x_k = income of k

Many aux. vectors can be formulated
to transmit ***some or all of this total information*** .

Let us consider **5** of these vectors .

Vector $\mathbf{x}_k^*$	Info $\sum_U \mathbf{x}_k^*$	Description
x_k	$\sum_U x_k$	total population income
$(1, x_k)'$	$(N, \sum_U x_k)'$	population size and total population income

Vector	Info
$(0, x_k, 0, 0, 0, 0)'$	$(\sum_{U_{11}} x_k, \ldots, \sum_{U_{23}} x_k)'$
	population income by age/sex group
$(0, 1, 0, 0, 0, 0, 0, x_k, 0, 0, 0, 0)'$	$(N_{11}, \ldots, N_{23}, \sum_{U_{11}} x_k, \ldots, \sum_{U_{23}} x_k)'$
	size of age/sex groups, and population income by groups
$(1, 0, 0, 0, x_k, 0)'$	$(N_{1.}, N_{2.}, \sum_{U_{.1}} x_k, \sum_{U_{.2}} x_k, \sum_{U_{.3}} x_k)'$
	size of sex groups, and income by age groups

Auxiliary vector, sample level

$\mathbf{x}_k^{\circ}$ is a ***known value*** for every k in s
(observed for the sample units)

$\sum_U \mathbf{x}_k^{\circ}$ unknown

Hence we can compute and use

$$\sum_s d_k \mathbf{x}_k^{\circ}$$

It is ***unbiased information*** ,
not damaged by NR

Examples of variables for the moon vector $\mathbf{x}_k^{\circ}$

- Identity of the interviewer
- Ease of establishing contact with selected sample element
- Other survey process characteristics
- Basic question method ("easily observed features" of sampled elements)
- Register info transmitted *only* to the sample data file, for convenience

The information statement

- Specifies the ***information*** at hand ;
  known or estimated totals
- May refer to either ***level***:
  Population level, sample level
- It is *not* a model statement

Statement of auxiliary information

sampling, then nonresponse

Set of units	Information
Population U	$\sum_U \mathbf{x}_k^*$ known
Sample s	$\mathbf{x}_k^\circ$ known, $k \in s$
Response set r	$\mathbf{x}_k^*$ and $\mathbf{x}_k^\circ$ known, $k \in r$

- The auxiliary vector

  General notation : $\mathbf{x}_k$

- The information available about that vector

  General notation : $\mathbf{X}$

Three special cases :

- population info only
- sample info only
- both types of info

- population info only

$$\mathbf{x}_k = \mathbf{x}_k^* \;; \quad \mathbf{X} = \sum_U \mathbf{x}_k^* \quad \text{(known total)}$$

- sample info only

$$\mathbf{x}_k = \mathbf{x}_k^\circ \quad ; \quad \mathbf{X} = \sum_s d_k \mathbf{x}_k^\circ$$

(unbiasedly estimated total)

- both types of info

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^{*} \\ \mathbf{x}_k^{o} \end{pmatrix} \quad ; \quad \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^{*} \\ \sum_s d_k \mathbf{x}_k^{\circ} \end{pmatrix}$$

Example :

$$\mathbf{x}_k = (\ 0,...,1,...,0 \quad 0,...,1,...0\ )'$$

identifies age/sex group for $k \in U$

identifies interviewer for $k \in s$

For **the study variable** y

we know (we have observed) :

$$y_k \text{ for } k \in r ; \quad r \subset s \subset U$$

Missing values :

$$y_k \text{ for } k \in s - r$$

The ***calibration estimator*** is of the form

$$\hat{Y}_W = \sum_r w_k y_k$$

with $w_k = d_k v_k$

where $d_k = 1/\pi_k$, and the factor v_k

serves to

- expand the design weight d_k for unit k
- incorporate the auxiliary information
- reduce as far as possible bias due to NR
- reduce the variance

Note: We want $v_k > 1$ for all (or nearly all) $k \in r$, in order to compensate for the elements lost by NR.

Primary interest :

Examine the (remaining) bias in $\hat{Y}_W = \sum_r w_k y_k$
attempt to reduce it further.
Recepie: Seek better and better auxiliary vectors for the calibration!
(Sessions 6, 7 and 8)

Secundary interest (but also important):

Examine the variance of $\hat{Y}_W$
find methods to estimate it .

The adjustment factor

v_k is determined to satisfy :

(i) The calibration equation $\sum_r d_k v_k \mathbf{x}_k = \mathbf{X}$.

Consistency with the given information.

and

(ii) $v_k = 1 + \boldsymbol{\lambda}' \mathbf{x}_k$ linearly related to $\mathbf{x}_k$

Now determine $\boldsymbol{\lambda}$

From (*i*) and (*ii*) follow

$$\boldsymbol{\lambda}' = \boldsymbol{\lambda}'_r = \left(\mathbf{x} - \sum_r d_k \mathbf{x}_k\right)\left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}$$

assuming the matrix non-singular.
Then the desired calibrated weights are

$$w_k = d_k v_k = d_k (1 + \boldsymbol{\lambda}'_r \mathbf{x}_k)$$

Easily computed (even for large dimension of $\mathbf{x}_k$)

Properties of the calibrated weights

$$w_k = d_k(1+\boldsymbol{\lambda}_r'\mathbf{x}_k)$$

1. They transform the design weights d_k and makes them larger:

$$w_k > d_k \quad \text{all } k \text{, or almost all}$$

2. $\sum_r w_k = N$ = population size

under a simple condition

Note : if both types of information, then

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$$

and the information input is

$$\mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$$

Consistency

is also an important motivation for calibration (in addition to bias reduction and variance reduction)

If $\mathbf{x}_k$ is known for $k \in s$, the statistical agency can sum over s and publish the unbiased estimate

$$\hat{\mathbf{X}} = \sum_s d_k \mathbf{x}_k$$

Users may require that this estimate coincide with the estimate obtained by summing over r using the calibrated weights : $\hat{\mathbf{X}}_W = \sum_r w_k \mathbf{x}_k$

Calibration makes this ***consistency*** possible

Calibration in two steps

An alternative procedure (with slightly different results):

In step 1 ”from *r* to *s*”, using the $\mathbf{x}_k^{\circ}$ information and get ”intermediate weights”

In step 2 use those weights and calibrate on both

$\mathbf{x}_k^{*}$ and $\mathbf{x}_k^{\circ}$ information.

The calibration approach

Some features:

- Is a new approach
- Generality (any sampling design, any auxiliary vector)
- ”Conventional techniques” are special cases
- Computational feasibility

The calibration approach brings generality

Earlier : Specific estimators were used for surveys with NR. They had names, such as Ratio estimator, Weighting Class estimator and so on.

Now : Most of these ‘conventional techniques’ are simple special cases of the calibration approach. Specific names no longer needed.

All are calibration estimators.

Another feature of the calibration estimator:

Perfect estimates under certain condition

Consider the case where

$$\mathbf{x}_k = \mathbf{x}_k^* \quad \text{and} \quad \mathbf{X} = \mathbf{X}^* = \sum_U \mathbf{x}_k^*$$

Assume that $y_k = (\mathbf{x}_k^*)'\boldsymbol{\beta}^*$ holds for every

$k \in U$ (perfect linear regression), then

$$\hat{Y}_W = \sum_U y_k = Y$$

No sampling error, no NR-bias!

A summary of this session: We have

- discussed two types of auxiliary information
- introduced the idea of a weighting (of responding elements) that is calibrated to the given information
- pointed out that calibrated weighting gives consistency, and that it often leads to both reduced NR bias and reduced variance . More about this later.