Propensity Score Weighting and Calibrated Weighting How do they compare ?

Carl-Erik Särndal

Workshop
Kuressaare 2008

My presentation has three parts :

- Personal remarks on Survey Sampling Theory (in the presence of nonresponse)
- Identifying auxiliary vectors for calibration
- Propensity Score method(s)

Part 1: Remarks on Sampling Theory for surveys with nonresponse (NR)

- NR unavoidable today
- Not only unavoidable; it is alarmingly high
- 50% NR not unusual nowadays
- Statistics continue to be produced by "trusted agencies" from such "infected data sources"
- Today, Survey Sampling Theory is, necessarily, "Statistical Theory for surveys with NR"

Remarks on Survey Sampling Theory (SST)

How does SST respond to “the plague of NR” ?

- Classical (design-based) theory does not make room for NR.
- But SST ought to recognize NR from the outset : Incorporate NR in “the ground rules”.

What do I mean by this ?

Remarks on Survey Sampling Theory

Objective: Estimate pop. total(s) of y-variable(s)

Classical ground rules :

There is a prob. sampling design ;

a sample s is drawn from pop. U ; $s \subset U$,

known inclusion probabilities π

There exists information about aux. vector $\mathbf{x}_k$

Researcher's aim : Invent

new sampling designs, new uses of aux. info.

to ***minimize variance***

Remarks on Survey Sampling Theory

Realistic ground rules (still design-based) :

There is a prob. sampling design;
sample s drawn from pop. U , $s \subset U$,
known inclusion probabilities π_k

NR occurs : y is observed, not for s,
only for the response set r ; $r \subset s$.
unknown response probabilities

There exists info. about aux. vector $\mathbf{x}_k$

...

Researcher's aim: Use of aux. info. to reduce bias and variance.

Faking design-based ("cheating")

Often practiced; not recommended .

Manipulate the sampling weight d_k :

multiply it by "ad hoc factor" a_k

then pretend $d_k \, a_k$

is the inclusion probability of k

Alternative :

Abandon design-based theory;

believe instead in a theory that is more accommodating (and pays less attention to NR bias).

Make assumptions, formulate models, and so on

Remarks on Survey Sampling Theory

Much research devoted to
"fixing the NR predicament"

Broad methodologies:

Imputation   Adjustment weighting

Both important, both requiring powerful aux. info.

Tend to be treated as "issues in their own right",
rather than "integrated into SST".

Under design-based ground rules,

what is possible, what is not ?

Impossible : Complete removal of bias;
quantification (estimation) of the bias

Possible : Compare and rank aux. vectors
in regard to their potential for bias reduction;
a partial removal of bias.

Reducing NR bias

Bias is reduced by efficient weighting, based on a powerful auxiliary vector.

We need tools for ranking alternative auxiliary vectors in regard to their potential for bias reduction.

Reducing NR bias

What info. is available?

What admin. registers & other sources ?

Statistics Sweden has access to many potential aux. variables, esp. for individuals and households. They form a ***vast supply of aux. info.***

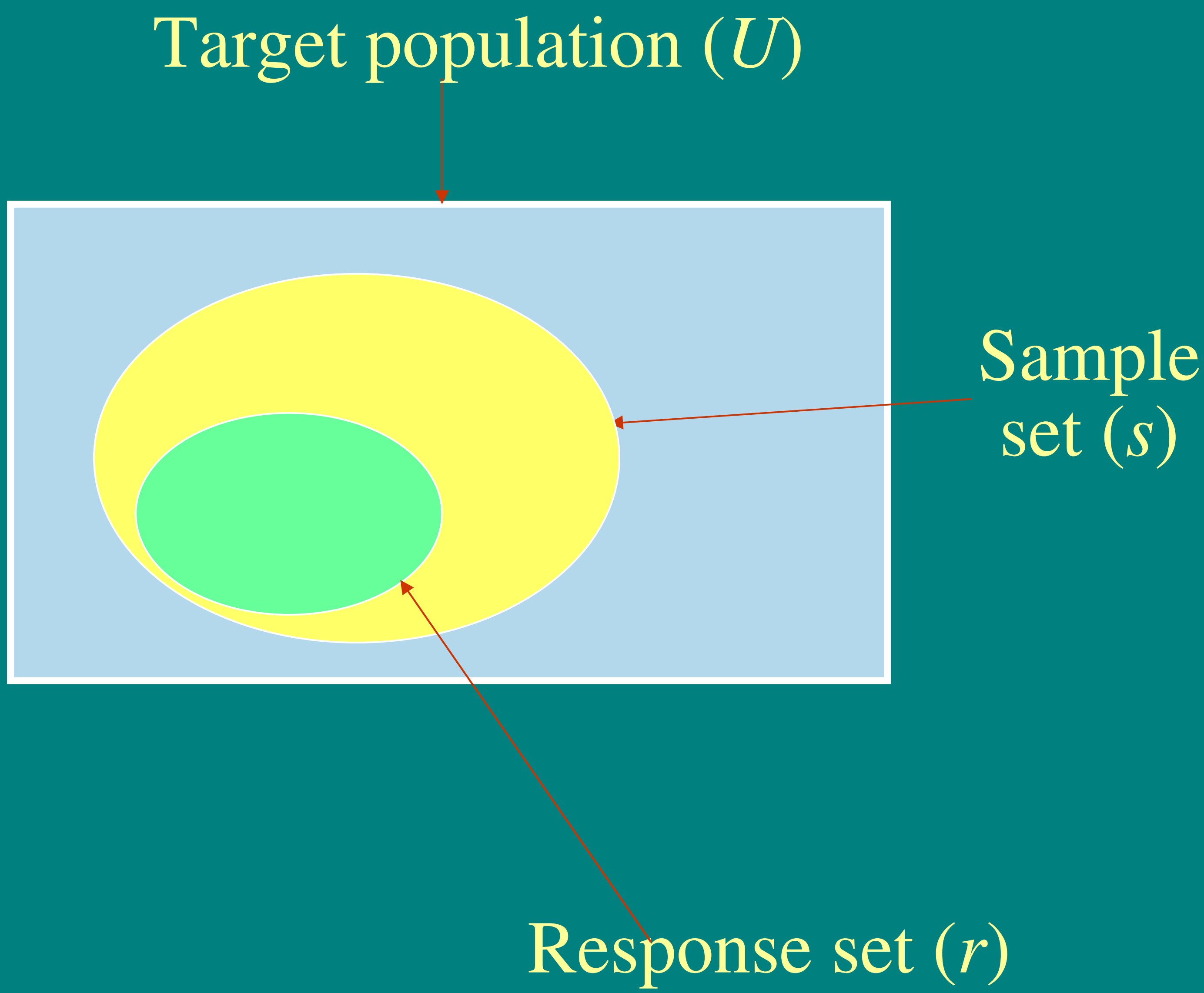
In practice, the question is one of selection :

Which aux. var. should be selected for the aux. vector ?

Reducing NR bias

- In recent years, Statistics Sweden has gained considerable experience in ***calibration for NR***.
- Clients demand “calibrated weighting”.
- Relies on a vast recent literature on calibration theory

Part 2 : Identifying suitable aux. variables



Objective :

estimate population y-total $\quad Y = \sum_U y_k$

y continuous or categorical

In practice, many totals and/or functions of totals need to be estimated. We focus on one total.

Ground rules (design-based)

Population U

of units $k = 1, 2, \ldots, N$

Sample s (subset of U)

Non-sampled : $U - s$

Response set r (subset of s)

Sampled but non-responding : $s - r$

Ground rules (design-based)

The response set r

is the set for which we observe y_k

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

Missing y-data : y_k for $k \in s - r$

Ground rules (design-based)

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

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

Known ***design weight*** of k : $d_k = 1/\pi_k$

Ground rules (design-based)

Phase two: ***Response selection***

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

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

Ground rules (design-based)

The auxiliary information

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

When both types of info. are present :

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

known total

estimated total
(random var.)

aux. vector

information

When both types of info present :

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

Example :

$$\mathbf{x}_k = (\underbrace{0,\ldots,1,\ldots0}_{\text{identifies age/sex group for } k \in U} \qquad \underbrace{0,\ldots,1,\ldots0}_{\text{identifies interviewer for } k \in s})'$$

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

identifies interviewer for $k \in s$

Are these ground rules design-based ?

Yes : They preserve the concept of
a finite population $\{1,\ldots, k,\ldots, N\}$;

To unit k belongs :

- A probability to observe k :

$$\Pr(k \in s)\Pr(k \in r|s) = \pi_k \theta_k \quad \text{although } \theta_k \text{ unknown}$$

- An auxiliary vector value $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$

- y-value, known if k responds

Objective

Not to claim that "under these conditions (models, etc.), our estimation is unbiased"

Unbiased estimation is impossible;
all situations are non-ignorable.

Instead, the objective is :

Rank the available **x**-vectors;
identify one likely to give a low bias.

When the search ends, we still do not know how much bias remains.

Steps in

<u>the calibration approach</u>

- State the ***information*** you wish to rely on.
- Formulate the corresponding ***aux. vector***
- State the ***calibration equation***
- Specify the ***starting weights*** (usually the sampling weights)
- Compute adjusted weights - the ***calibrated weights*** - that respect the calibration equation
- Use the adjusted weights to compute ***calibration estimators***

A category of auxiliary vectors

Consider vectors with the following property :

There exists a constant vector μ such that

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \quad \text{for all } k \in U$$

This "in-line property" is present in most aux. vectors of interest in practice.

Example 1 : Continuous x-variable

$$\mathbf{x}_k = (1, x_k)'$$

Take $\boldsymbol{\mu} = (1, 0)'$

Then , as required :

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \times 1 + 0 \times x_k = 1 \quad \text{for all } k$$

Example 2 : The classification vector

$$\mathbf{x}_k = \underbrace{(0,\ldots,1,\ldots,0)}'$$

identifies the category of k

Take $\boldsymbol{\mu} = (1,\ldots,1,\ldots,1)'$

Then, as required :

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \quad \text{for all } k$$

Calibration estimator

$$\hat{Y}_{CAL} = \sum_r w_k y_k$$

with w_k calibrated so that

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

that is,

$$\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^* \quad ; \quad \sum_r w_k \mathbf{x}_k^\circ = \sum_s d_k \mathbf{x}_k^\circ$$

population info

sample info.

”Bias-equivalent” calibration estimator

$$\tilde{Y}_{CAL} = \sum\nolimits_r w_k y_k$$

$$w_k = d_k \times m_k \quad = \quad \text{design weight} \times \text{adjustment}$$

$$m_k = \mathbf{f}_r' \mathbf{x}_k \quad ; \quad \mathbf{f}_r' = \underbrace{(\sum\nolimits_s d_k \mathbf{x}_k)'}_{vector} \underbrace{(\sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}}_{inverted\ matrix}$$

Calibrated “only” to the sample level:

$$\sum\nolimits_r d_k m_k \mathbf{x}_k = \sum\nolimits_s d_k \mathbf{x}_k \quad ; \quad \mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^{\mathrm{o}} \end{pmatrix}$$

unbiased “control”

The adjustment factor m_k

is a derived (univariate) random variable

$$m_k = \underbrace{(\textstyle\sum_s d_k \mathbf{x}_k)'}_{vector} \underbrace{(\textstyle\sum_r d_k \mathbf{x}_k \mathbf{x}'_k)^{-1}}_{inverted\ \ matrix} \times \mathbf{x}_k$$

- computable for $k \in s$;
- used for $k \in r$ in

  computing $\tilde{Y}_{CAL} = \sum_r d_k m_k y_k$

The adjustment factor m_k

When is it effective for bias reduction ?

Särndal & Lundström *J.Off.Stat.* 2008

If $m_k \approx \theta_k^{-1} = (\text{response prob.})^{-1}$,

$$\Rightarrow E(\tilde{Y}_{CAL}) \approx \text{ unbiased for } Y$$

The adjustment factor

$$m_k = \left(\sum\nolimits_s d_k \mathbf{x}_k\right)' \left(\sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \mathbf{x}_k$$

has interesting statistical properties

The mean of m_k

is the same for every aux. vector : $\bar{m}_{r;d} = \dfrac{\sum_r d_k m_k}{\sum_r d_k} = \dfrac{1}{P}$

where $P = \dfrac{\sum_r d_k}{\sum_s d_k} =$ survey response rate

Interpretation: On average, the adjustment factor in

$\tilde{Y}_{CAL} = \sum_r d_k m_k y_k$ is equal to (response rate)$^{-1}$

regardless of the auxiliary vector used

The variance of m_k

$$S_m^2 = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \overline{m}_{r;d})^2$$

depends on the aux. vector .

Development gives $cv_m^2 = S_m^2 / \overline{m}_{r;d}^2 = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}$

$$\mathbf{D} = \overline{\mathbf{x}}_{s;d} - \overline{\mathbf{x}}_{r;d} \qquad ; \qquad \mathbf{\Sigma} = \frac{1}{\sum_r d_k} \sum_r d_k \mathbf{x}_k \mathbf{x}'_k$$

"contrast vector"

The value of cv_m^2 increases as $\mathbf{x}_k$ expands (same property as R^2 in regression analysis.)

Simplest possible **x**-vector

$$\mathbf{x}_k = 1 \quad \text{for all } k$$

The calibration estimator is then the **Expansion estimator**

$$\tilde{Y}_{EXP} = \frac{\sum_s d_k}{\sum_r d_k} \times \sum_r d_k y_k$$

1/(response rate)

and $cv_m^2 = 0$

- Adding further x-variables

to the **x**-vector increases the value of cv_m^2

- One can show that this is

likely to decrease the bias in $\tilde{Y}_{CAL}$

$\Rightarrow$ Stepwise (forward or backward) selection of x-variables

Stepwise (forward or backward) selection
of x-variables

By successive increments of cv_m^2 (or of S_m^2)

A procedure independent of the y-variable(s)

Currently practiced at Statistics Sweden

Using s_m^2 to select *x*-variables

Example:

The 2006 Swedish National Crime Victim and Security Study (BRÅ)
(Data collection and calibration by Statistics Sweden)

Särndal & Lundström
J.Off.Stat. 2008
Estimation in Surveys with NR. Wiley 2006

Step	Auxiliary variable entering	Number of groups	$s_m^2 \times 1000$
0	------	-----	0
1	Country of birth	2	20.0
2	Income group	3	27.6
3	Age group	6	31.3
4	Gender	2	35.1
5	Martial status	2	38.6
6	Region	21	40.7
7	Family size group	5	41.4
8	Days unemployed	6	41.9
9	Urban centre dweller	2	42.3
10	Occupation	10	42.7

Searching the most suitable aux. vector

extensions

currently explored at Statistics Sweden

(results tentative)

Objective

Two factors influence the bias :

Relation y-to-$\mathbf{x}$

Relation y-to-response propensity

Rank the many available **x**-vectors;

identify one likely to give lowest possible bias.

When the search stops, we must still accept :

unknown remaining bias (but reduced)

Searching an effective aux. vector

Consider three estimators,
the first two computable, the third hypothetical

- $\tilde{Y}_{CAL} = \sum_r d_k m_k y_k$ moderate bias

  with $m_k = \mathbf{f}_r' \mathbf{x}_k$; $\mathbf{f}_r' = (\sum_s d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}$

- $\tilde{Y}_{EXP} = (1/P) \sum_r d_k y_k = \hat{N} \, \bar{y}_{r;d}$ large bias

- $\tilde{Y}_{FUL} = \sum_s d_k y_k$ ideal: unbiased, but requiring full response

The ideal

$$\tilde{Y}_{FUL} = \sum\nolimits_s d_k y_k$$

- unbiased but not computable due to NR
- ”bias-equivalent” with the GREG

calibrated according to $\sum_s w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$

Three differences of interest :

$$T_1 = \tilde{Y}_{EXP} - \tilde{Y}_{CAL}$$ computable

$$T_2 = \tilde{Y}_{CAL} - \tilde{Y}_{FUL}$$ not computable

$$T = T_1 + T_2 = \tilde{Y}_{EXP} - \tilde{Y}_{FUL}$$ not computable

We want a near-zero value of

$$\text{bias ratio} = \frac{\tilde{Y}_{CAL} - \tilde{Y}_{FUL}}{\tilde{Y}_{EXP} - \tilde{Y}_{FUL}} = 1 - \frac{T_1}{T}$$ not computable

But T_1 is computable

$\Rightarrow$ Find $\mathbf{x}_k$ to make T_1 large

$$\text{bias ratio} = \frac{\tilde{Y}_{CAL} - \tilde{Y}_{FUL}}{\tilde{Y}_{EXP} - \tilde{Y}_{FUL}} = 1 - \frac{T_1}{T}$$

- is $=$ 1 for the trivial **x**-vector $\mathbf{x}_k = 1$
- is near 0 for a highly efficient **x**-vector

Objective : maximize $T_1 = (\bar{\mathbf{x}}_{s;d} - \bar{\mathbf{x}}_{r;d})'\mathbf{B}_{\mathbf{x}}$

where $\mathbf{B}_{\mathbf{x}} = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \sum_r d_k \mathbf{x}_k y_k$

regression y on **x**

y-variance $$S_y^2 = \frac{1}{\sum_r d_k} \sum_r d_k (y_k - \bar{y}_{r;d})^2$$

proportion explained by **x** : $$R_{y,\mathbf{x}}^2 = \frac{\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}}{S_y^2}$$

proportion explained by *m* : $$R_{y,m}^2 = \frac{(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2}{(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}) \times S_y^2} = \frac{(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2}{cv_m^2 \times S_y^2}$$

where

$$\mathbf{C} = \left(\sum_r d_k (\mathbf{x}_k - \bar{\mathbf{x}}_{r;d}) y_k\right) / \left(\sum_r d_k\right)$$ covariance vector

$$\mathbf{D} = \bar{\mathbf{x}}_{s;d} - \bar{\mathbf{x}}_{r;d}$$ contrast vector

$$\boldsymbol{\Sigma} = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' / \sum_r d_k$$ cross-prod. matrix

Maximize $\left[\frac{T_1}{\hat{N} \times S_y}\right]^2 = R^2_{y,m} \times cv^2_m = \Lambda^2_{\mathbf{CD}} \times R^2_{y,\mathbf{x}} \times cv^2_m$

$cv^2_m = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}$ coeff. of var. of m

$R^2_{y,m} = \frac{(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C})^2}{(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}) \times S^2_y}$ explained by m

$R^2_{y,\mathbf{x}} = \frac{\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}}{S^2_y}$ explained by $\mathbf{x}$

$\Lambda^2_{\mathbf{CD}} = \frac{(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C})^2}{(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})}$ (cosine)^2

betw. vectors **C** and **D**

Stepwise (forward or backward) selection
of x-variables

while paying attention to important y-variables

Based on successive increments of

- $R^2_{y,m} \times cv^2_m = [\frac{T_1}{\hat{N} \times S_y}]^2$
- $R^2_{y,\mathbf{x}} \times cv^2_m$

Currently explored at Statistics Sweden

Part 3: Propensity score method(s) for NR

Main idea:

Response propensities are estimated,

then grouped into subintervals of (0,1),

then used for weighting, by the inverse of response rate, by sub-interval

Origins of propensity score method :

observational studies for causal effects;
treatments assigned to experimental units
but without the benefits of randomization

Rosenbaum and Rubin :

The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983

Reducing bias in observational studies using subclassification on the propensity score. *JASA* 1984

These authors consider :

A nonrandomized design ;
compare two treatments,
$z = 0$ or $z = 1$

A central concept is
the propensity score

$$e(\mathbf{x}) = \Pr(z = 1 | \mathbf{x})$$

where $\mathbf{x}$ is a vector of observed covariates

Formulation not in terms of finite populations

Translated into the framework for
finite population theory :

Treatment 0 or 1 $\Leftrightarrow$ response/nonresponse

An assumption we may hesitate to make :

The auxiliary vector $\mathbf{x}$ is such that

R (the response indicator) and

y (the study variable whose total is to be estimated)

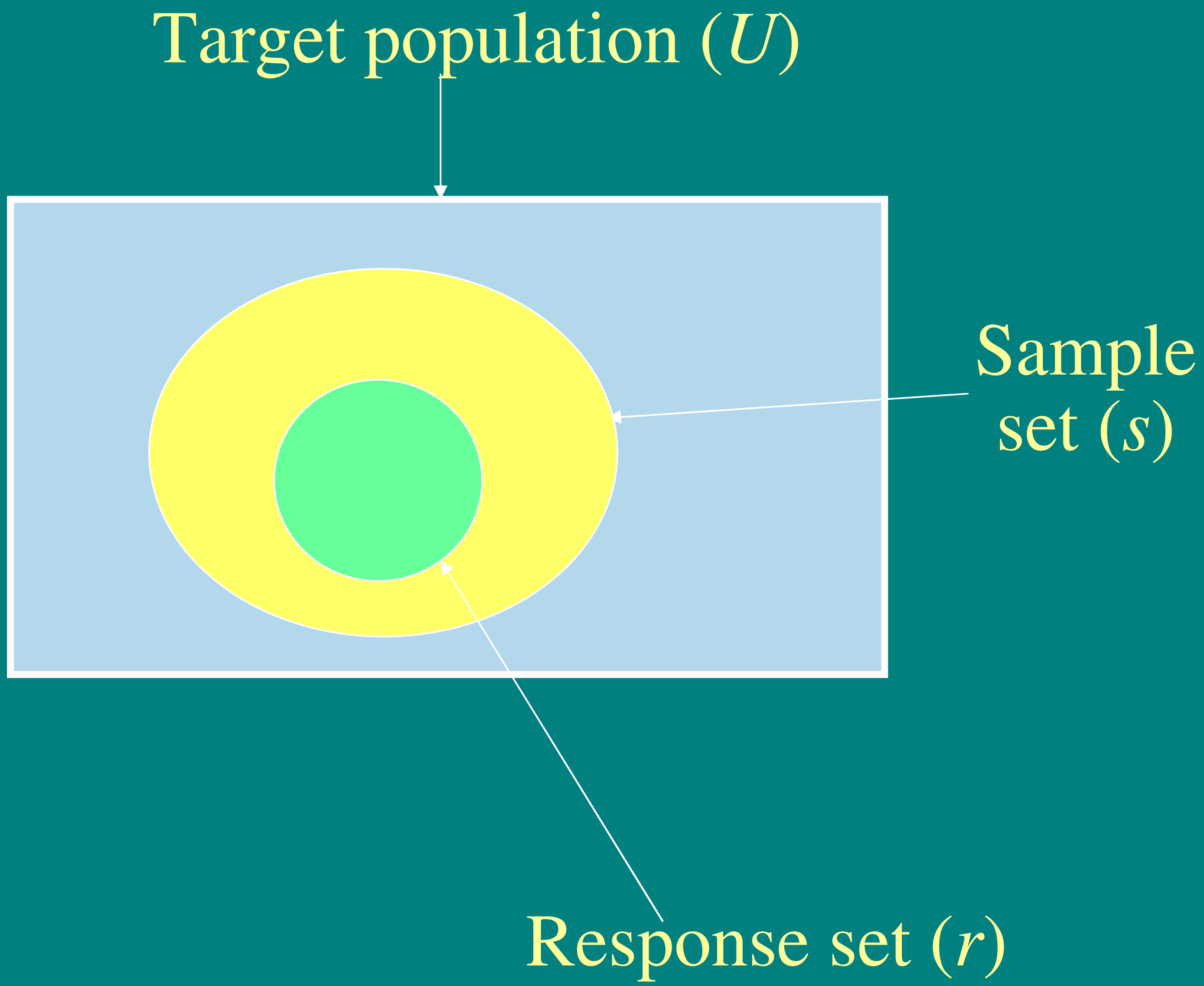
are *conditionally independent* (or almost so) .

Propensity score method

Applications of the method :

- the single-sample situation
- the two-sample situation

Single-sample application



Propensity score method; single-sample application

Prototype : $\hat{Y} = \sum_r d_k \frac{1}{\theta_k} y_k$ (unbiased if θ_k known)

- estimate θ_k by $\hat{\theta}_k$; $k \in s$

- sort the values $\hat{\theta}_k$ into J sub-intervals of (0,1)

$\tilde{P}_k = m_j / n_j$, all $k \in$ group j ; $j = 1, \ldots, J$

- compute $\hat{Y} = \sum_r d_k \frac{1}{\tilde{P}_k} y_k$

sampling weight NR adjustment

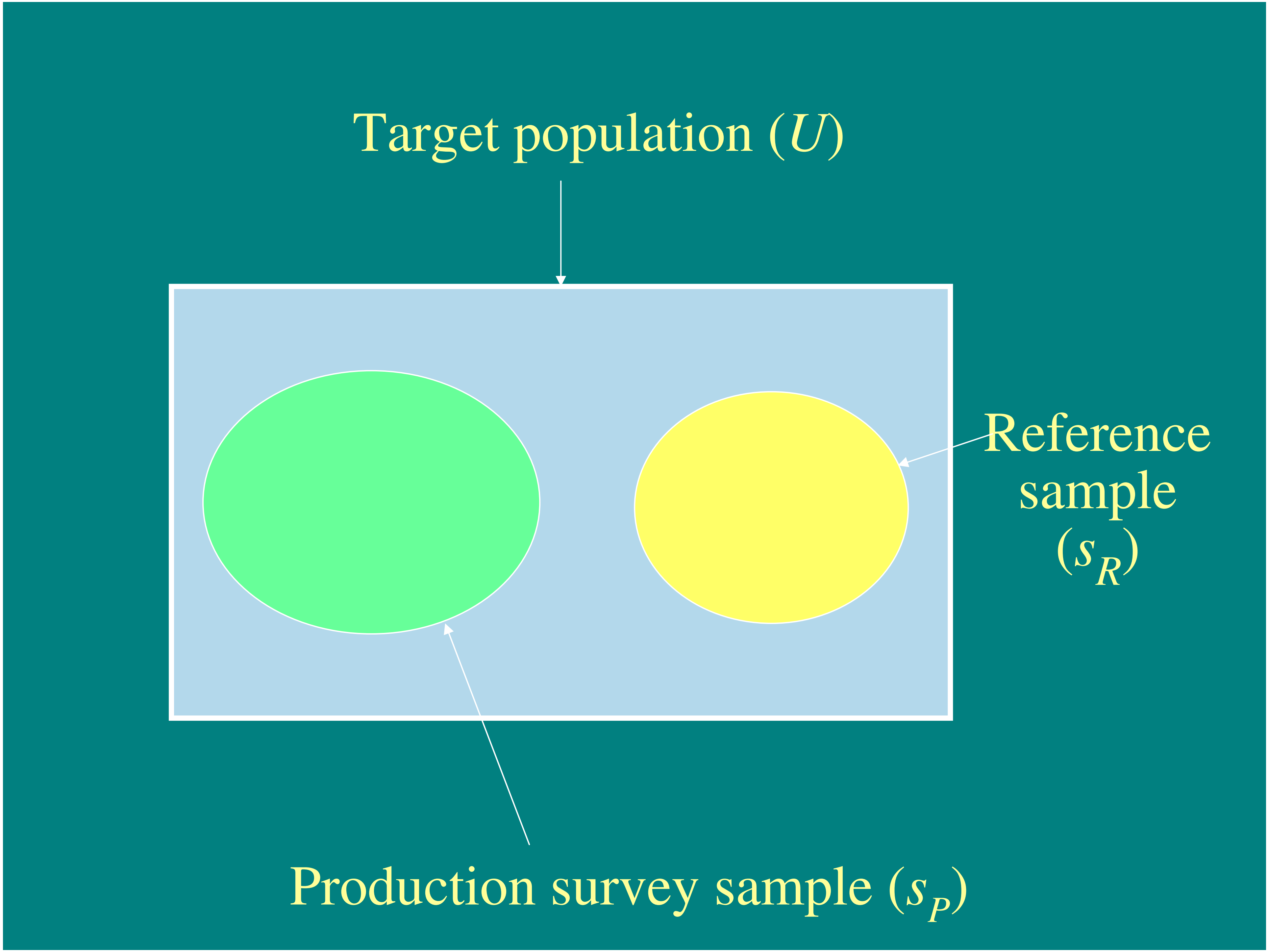
Propensity score method; two-sample application

- A reference survey, done with probability sampling, used to derive estimated response propensities.

  This is "a proper survey",
  in the eyes of traditional survey theory

- The production survey (non-probability sampling; e.g., web survey), in which the variable(s) of interest y are observed, then used to produce estimates.

  It is "an improper survey";
  data collection uncontrolled, hap-hazard.


Target population (U)
Reference sample (s_R)
Production survey sample (s_P)

Propensity score method; two-sample application

- How can this work ?

  The key :

  Some auxiliary variables are observed in both surveys

Propensity score method; two-sample application

- Reference survey serves to derive response propensities, by interval : Set

$$\tilde{P}_k = (\text{response rate})^{-1} \quad \text{for all} \quad k \in \text{group } j\,; \quad j = 1,\ldots, J$$

- These are used as adjustment weights in obtaining y-estimates from the production survey

Propensity score method; two-sample application

Attractive features: Cost advantage:

Although the reference survey may be expensive, the production survey may be much less expensive e.g., no expensive follow-up .

Less attractive features :

The reference survey will have some NR, so reliance on its results contributes further to bias.

Crucial question : Can the production survey (although improper) produce estimates of sufficient quality (limited bias) ?

A look at
propensity score method (two-sample)
from the perspective of calibration theory

Common variables, measured in both surveys, form a vector $\mathbf{x}_C$ of auxiliary variables for calibration

y measured only in the production survey.

Reference survey $s_R \subset U$

Design weights : $d_k = 1/\pi_k$

Data : $\mathbf{x}_{Ck}$ for $k \in s_R$

$\Rightarrow \sum_{s_R} d_k \mathbf{x}_{Ck}$ (design unbiased for $\mathbf{x}_C -$ total)

Production survey $s_P \subset U$

Absence of design weights

Data : $(y_k, \mathbf{x}_{Ck})$ for $k \in s_P$

Seek weights w_k calibrated so that

$$\sum_{s_P} w_k \mathbf{x}_{Ck} = \sum_{s_R} d_k \mathbf{x}_{Ck}$$

random but unbiased
control quantity

Then compute calibration estimator
from the production survey y-data :

$$\hat{Y}_{CAL} = \sum_{s_P} w_k y_k$$

Question arising for the calibration :

What should be the starting weights ?

- Constant (equal to 1), to express ignorance ?
- Other (more or less arbitrary) choice ?
- Is the choice really important ?

Which is the overriding consideration:

- proper (design-based) starting weights ? ***or***
- the power of the aux. vector for the calibration ?

Proposition :

More important :
create a powerful aux. vector;

The choice of starting weights an issue of
secondary importance.

Future examination needed.

If we accept this reasoning,
do we abandon
Classical Survey Sampling Theory ?

We will see …

Concluding remarks

The broader question for the NR problem is not
Do we use this or that imputation technique ?
This or that weighting method ?

Instead:

Do we statisticians really believe that trustworthy information can come from surveys with less than 50% response ?

Some say NO

Some say, apparently, YES : We know how to impute; we know how to use weighting, and so on

When NR is as high as 50%
Is the output from the survey worthless?
Or does it still have some value, as information for our society ?

The community of statisticians (that includes you and me) has not (yet) succeeded to develop a concerted stand,

including clear criteria (in mathematical statistical or other terms) for assessing the information value of output from surveys with large NR.