

.

# Baltic-Nordic-Ukrainian Workshop

Valmiera, Latvia

24-28 August 2012

Lectures by  
Carl-Erik Särndal

Lectures

by

Carl-Erik Särndal  
Örebro University  
Statistics Sweden

Baltic-Nordic-Ukrainian Workshop

Valmiera, Latvia

24-28 August 2012

As announced in the program:

Lecture 1 : Interplay between survey theory and the demands of official statistics production, a theory of science perspective

Lecture 2 : The data collection stage: Responsive design and balancing the set of respondents

Lecture 3 : The estimation stage: Calibrated weighting for nonresponse bias reduction and preferably without increased variance

.

Lecture 1:

Interplay between survey theory and  
official statistics production

Subtitle: Comments on

Survey science in the last 200 years  
with an emphasis on the last 100 years  
in particular on the last 60 years

My lecture received its structure  
from 4 recent publications in our area

three Ph.D. theses  
one important review paper

.

Important research  
and methodological progress  
is realized in our area

Area = the geographical area

Here : Baltic-Nordic-Ukrainian

Area = our discipline :

Survey science, broadly defined

I selected

Three recent Ph.D. theses in our area

*On unequal probability sampling designs*

Anton Grafström, Umeå University 2010

*Estimation of domains under restrictions built upon  
generalized regression and synthetic estimators*

Natalja Lepik, Tartu University 2011

*Paradigms in statistical inference for finite populations  
up to the 1950's*

Vesa Kuusela, University of Helsinki 2011

.

Congratulations to Anton, to Natalja , to Vesa !

And to all other successful recent Ph.D's in our area !

And I wish I were a Ph.D. student again !



An important review article

Groves, R. M. and Lyberg, L. (2010)

*Total survey error, past, present, future.*

Public Opinion Quarterly **24**, 849-879

## Four themes

*Inference for finite populations*

*Unequal probability sampling*

*Estimation for domains by GREG*

*Total survey error*

## Those four references

are very different

but what unites them is “our area” :

finite populations, survey statistics

The differences reflect

the vitality of our discipline

the different motivations of researchers in our area

The differences lend structure to my lecture

## Outline of my lecture

1. The historical perspective up until 1950: Kuusela
2. The journey from the 1950's : Modern progress with classical roots (examples : Grafström, Lepik)
3. The reality in practice today: Groves and Lyberg
4. Discussion and prospects for the future, in particular for math/stat work

# 1. A historical perspective

An intellectual adventure beginning over 200 years ago

In the interest of the national authority (the king, the state, the decision makers) or the local authority

We need to know, to find out,

about our population,

too large to collect data on everybody  
(persons, or farms, or firms, or hospitals)

## A historical perspective

Louis XIV , King of France, wants to know :

How many subjects do I have  
in New France (Québec) ?

Jean Talon's 1666 census of New France :  
there were 3215 people  
and 538 separate families

That was not so hard - they were not many

.

150 years later,

it was already much more difficult :

the population of France  $> 30 \times 10^6$

## Kuusela's thesis

The method of Laplace 1783;  $n = 30$  départements

Ratio estimator :

Pop. births  $\times$  sample population/sample births

1802 estimate :  $28.4 \times 10^6$

Compare: Central Statistical Office of France:

Year 1801 :  $27.3 \times 10^6$

Year 1806 :  $29.1 \times 10^6$



.

The method of Laplace (1783) marks the beginning of  
mathematical statistical inference  
for finite populations

A significant step forward !

Kuusela discusses

Early 20th century key figures :

Kiær in Norway, Bowley in England

1934 Neyman's foundation for design-based inference

1936 Gallup poll = quota sampling; called representative

## Classical theory

Kuusela:

“The classical theory of survey sampling was more or less completed in 1952 when Horvitz and Thompson published a paper on a general theory for constructing unbiased estimates”

“The random sampling approach was almost universally accepted”

60 years ago already !

The classical period  
extensions in 1950's , 1960's

- The teaching flourished
- The research continued, somewhat hesitatingly
- Gave a “hard core” for the future

## Teaching the classical theory

Thousands of students became familiar with  
Cochran (editions 1953, 1963, 1976)

Des Raj & Murthy & others.

## Research

Theory was in a way finished, complete, in 1952.

Some said : There is nothing more to be done.

Survey sampling was seen by many  
in 1960's as “a dead field”

These observers did not see very far.

We have come a long way since 1950's  
with a “neo-classical” perspective

## The hard core

One of the lasting contributions  
of the classical period :

gave a **hard core** for survey theory,  
making it a mature math/stat science

.

Kuusela's thesis

- is devoted only to the sampling error
- the many other “errors in surveys” play no role  
(not a criticism)
- has an undertone of nostalgia for the classical period,  
when surveys were simple



## 2. The journey from the 1950's

After 1952, espcecially since 1970's :

Neoclassical theory flourished

with “ a hard core”

born out of the classical theory

The *hard core* of a research programme, what is it ?

We owe the term to Imre Lakatos 1922-1974

*The hard core* consists of theoretical assumptions that cannot be abandoned or altered without abandoning the programme altogether.

More modest theories, formulated in order to explain evidence that threatens the hard core, are called auxiliary hypotheses.

.

The *hard core* of survey science :

Postulates a finite collection of objects (units)  
from which some are sampled  
and a subsample is observed  
with more or less measurement error

Model based methods (from the 1970's and on) still  
within the hard core

## Neoclassical research traditions since 1970's

- Unequal probability sampling designs
- Forms of inference: design-based, model based etc
- GREG and calibration
- Small area estimation
- Nonresponse treatment
- Longitudinal surveys
- Confidentiality
- Editing

## Grafström's thesis

The research tradition: Unequal probability sampling

The mathematical base: Probability theory, probability distributions applied to finite universes

The survey background: Large units should be selected with high probability

## Grafström's thesis

“Wonderful opportunity to learn more about mathematical statistics and sampling”

The roots: Hájek (1964), posthumous book (1981)

M.R. Sampford (1967)

K.R.W. Brewer 1970's ; Brewer & Hanif (1983)

The thesis contains many references  
from last 20 years:

A modern, lively discipline, with classical roots !

## Lepik's thesis

“Our study method is mathematical”

The research tradition: Auxiliary information as in GREG

The mathematical base: Advanced matrix algebra;  
multivariate theory;

The survey context: Estimation, design-based, in additive  
manner for sub-populations (domains) (design-based)

Oldest reference: 1976

Modern : Only 3 of 36 references are older than 1990;  
only 13 older than 2000 .

A modern, lively discipline, with classical roots !

.

Without skilled mathematical work,  
these theses would not have been produced

They are manifestations of  
a mature survey sampling science



## Aftermath of the classical period

Three inference theories for finite populations,  
Ray Chambers (2012) Pak. J. Statist.

- The design-based
- The model assisted design-based
- The model based

“All 3 are in use in major statistical agencies”

### 3. Practice today

The development of survey science  
(to serve the interests of national statistical agencies)  
is driven by

- The (increasing) needs for statistics in society
- The costs of production

(It was so in the past, is so today, and always)

Article by Groves and Lyberg (2010) :  
Total Survey Error: Past, Present, Future

- expresses “the reality of surveys”
- expresses very important concerns

but is in striking contrast with the math/stat work  
(the statistical inference aspect)

that I have reviewed

.

Groves and Lyberg (2010)  
on Total Survey Error

- called a paradigm
- cannot be measured
- but provides a framework for our thinking
- no reference to the math/stat work in survey science from the 1950's

.

Unlike the 3 theses mentioned,  
Groves and Lyberg (2010) do not discuss  
statistical inference for finite populations

The focus instead: The quality of survey results

Groves and Lyberg try to explain what they see as a (misguided ?) overemphasis on the math/stat work

Deming *Some Theory of Sampling* (1950) “... focuses entirely on sampling error properties ... not surprising ... sampling was not universally accepted and had to be vigorously promoted at the time”

Hansen, Hurwitz & Madow (1953) devote 9 pages out of 638 in their book to “response and other non-sampling errors in surveys”.

Bring attention instead to issues such as :

- measurement and questionnaire
- types of nonresponse: refusal vs. non-contact
- mode effects vs. respondent effect;
- self-administered data delivery vs. interview
- coverage error vs. nonresponse error

Groves & Lyberg say :

“The isolation of survey statisticians and methodology from the mainstream of social statistics has ... retarded the importation of model based approaches to many of the error components in the total survey error format.”

Examples of such models:

structural equation model building, hierarchical linear models, latent class models



.

Article by Robert Groves (1987) titled:  
Survey research is a methodology without a unifying  
theory

“A theory of surveys would unite social science  
concepts with the statistical properties of survey  
estimates” (i.e., accuracy, bias and variance)

We do not have such a theory (of inference)

## 4. Discussion

Themes :

Survey science in a broad perspective

The future of excellent math/stat work in survey science ?

Is the mathematical orientation “misguided” ?

Is survey science “mature” ?

## Discussion

Do we as mathematical statisticians  
not see the forest for only trees ?

There is an old saying about the forest and the trees :

“We cannot see the forest for only trees”

## The forest and the trees

John Polanyi , distinguished Canadian scientist ;  
Nobel laureate chemistry 1986;  
eminent philosopher-chemist

In his address to the 1998 M.D. graduates,  
Fac. of Medicine, Univ. of Toronto,  
Polanyi says:

## Polanyi :

“Nature deals in forests, scientists seldom even in trees. We decompose what we see, to the level of atoms and molecules ... But in the process of delving for hidden patterns, the large pattern called a forest can be lost to view. Then the strength of science, which lies in its sharp but narrow focus, becomes its weakness.”

Polanyi :

“Since there is no right solution to this problem of balance between viewing the whole and the part, one finds different “styles” in science ... Though there is no right style, there is a wrong one which is to abandon the problem of balance and neglect the whole in favour of its details”

The underlining is mine.

## Questions arising

The problem of balance between the whole and its parts

Do math/stat survey scientists fail to strike the balance  
the total survey picture versus its minute details ?

Do we devote too much myopic attention to minor  
details ?

My answer is both a YES and a NO



## Survey science

I need here to make a distinction

- Math/stat survey scientists
- Other stakeholders and contributors to survey science  
(sociologists, economists, political scientists)

At this workshop, perhaps in majority the first category.

## Survey science

in the sense of “inference for finite populations”  
(as for example in Kuusela, Grafström, Lepik)

- is mathematical
- the best of it has (over the years) had tremendous impact on practice

We should be proud of that.

Illustration: IASS jubilee commemorative volume 2001  
(Landmark papers in Survey Statistics):  
19 papers, almost all mathematical

Historically, some math/stat contributions have been extremely important in the advancement of survey science.

Beginning with Neyman & Hansen & others,  
1930's to 1950's

Continuing with the impact of models (from 1970's) :  
Model assisted, Model based, and so on

.

As a result of his/her training,  
a natural instinct of math/stat survey scientist :

Set (probability) bounds on the error in statistics for  
finite populations

For this, there is a toolbox: The methods of statistical  
inference

Now today, given the enormous complexity of modern  
surveys and the multitude of errors,  
what is the future of math/stat work in survey  
science?

Yet the message of Groves & Lyberg is very important

They minimize the recent math/stat contributions.

They do not say so explicitly, but it is implied that

much work of math/stat character is just

“little trees or bushes in the big forest”

My impression : they feel that the quest for balance

between the whole (of survey science) and its parts

is not well served by a focus on math/stat “details”.

Yes, I believe there is, in survey science,  
a certain conflict

between the view of the whole and the view on its  
parts.

How strike the balance ?

I have no satisfactory answer – perhaps you have

## A personal view (and hope)

I would like survey theory to progress,  
to make decisive leaps forward,  
with math/stat means.

Because the math/stat resolution  
of an important practical problem has  
a tremendous “convincing power”  
This is in the nature of mathematical language.

But this is not easy, in view of the complexity of modern  
surveys.

## Conclusion : Questions we need to ask

- What is *the value* of math/stat work in survey science ?
- What is *the future* of math/stat work in survey science ?



.

What is *the value* of math/stat work in survey science ?

Many important math/stat contributions to survey science have been realized

- In the classical period (before 1952)
- In the neo-classical period (after 1952)

.

What is *the future* of math/stat work in survey science ?

- In defining our research, we should strive for the desirable balance between the whole picture of survey science and its parts.
- We must ask ourselves : Is the direction of our work sufficiently “in balance” ?

.

## Choosing directions (themes, problems) for one's research

- A young PhD student relies heavily on the advice (the preferences) of the thesis director (professor)
- Established senior researchers (professors)  
can always justify :  
    “I am continuing my research”  
    (because it was well received in the past)

.

Revolutionary progress by math/stat (in the manner of  
Neyman and others) still not excluded  
(but it was easier then)

I wish all of us good luck for the next 60 years !  
In 2072, almost 300 years since Laplace !

This ends my Lecture 1.

Thank you for your attention !

## References

- Bohm, D. and Peat, F.D. (2000). Science, Order, and Creativity, 2<sup>nd</sup> edn. London: Routledge.
- Franchet, Y. and Nanopoulos, P. (1997). Statistical science and the European statistical system: Expectations and perspectives. In Proc. Conference in honour of S. Franscini. Basel: Birkhäuser.
- Groves, R. (1987). Survey research is a methodology without a unifying theory. Public Opinion Quarterly, 51, 156-172.
- Groves, R. M. and Lyberg, L. (2010) Total survey error, past, present, future. Public Opinion Quarterly 24, 849-879
- Kuusela, V. (2011) Paradigms in Statistical Inference for Finite Populations. Thèse de doctorat (en statistiques), Université d'Helsinki.
- Lakatos, I. (1970). A chapter in: Criticism and the Growth of Knowledge. Cambridge Univ. Press.
- Laudan, L. (1977). Progress and its Problems. Toward a theory of scientific growth. LA: Univ. of California Press.
- Platek, R. and Särndal, C.E. (2001). Can a statistician deliver? J. Official Statistics, 17, 1 – 127 (with 16 discussions)

.

# Baltic-Nordic-Ukrainian Workshop

Valmiera, Latvia

24-28 August 2012

Lectures by  
Carl-Erik Särndal

Lectures

by

Carl-Erik Särndal  
Örebro University  
Statistics Sweden

Baltic-Nordic-Ukrainian Workshop

Valmiera, Latvia

24-28 August 2012



.

Survey nonresponse :

statisticians have a role to play

(a) at the data collection stage

(b) at the estimation stage.

The tasks (a) and (b) interact.

In these two lectures we examine these tasks  
and their interaction.

Lecture 2 : The data collection stage:  
Responsive design; balancing the set of respondents

Lecture 3 : The estimation stage: Calibrated  
weighting for nonresponse bias reduction and  
preferably without increased variance

## Lecture 2 :

The data collection stage : Responsive design, and balancing the set of respondents

Most surveys today experience high nonresponse –  
in some cases  $> 50\%$  –  
an impediment to survey quality,  
in establishment surveys  
as in surveys on individuals and households.

Let's face it: Nonresponse is here to stay.

While keeping nonresponse rates high,  
we should integrate nonresponse  
– and other non-sampling errors for that matter -  
into theory; not view it as a “disease”,  
treated by “repairs” of various kinds.

Starting points:

The *study variables* (y-variables) are affected by non-random nonresponse (even conditional on x-vector).

Estimates more or less biased.

Bias can never be eliminated completely.

Ignorable nonresponse (MAR) does not exist.

*Auxiliary variables* (x-variables) crucially important.

These are variables known at least for all units in the selected probability sample, respondents as well as non-respondents.

“Multivariate auxiliary” is a starting point

## Background of this lecture

The ideas of *Responsive design*

European developments:

Statistics Netherlands (RISQ project)

Statistics Sweden

survey environments

rich in auxiliary information

.

Aspects of *Responsive design* : Data collection monitored with the aid of indicators of *balance* and/or *representativity* (of the respondents), and *distance* (between respondents and non-respondents).

These are computed on selected auxiliary variables.

- Data collection: Evolving over time; the original data collection may be modified by interventions; focus on *balanced response* at the end
- Estimation stage : Focus on *adjusting for bias* still affecting the estimates, despite balancing at the data collection stage.

Both activities require auxiliary variables

Challenge: make the most of such variables



## Outline of Lecture 2

1. Probability sampling background
2. Balanced response; measuring imbalance
3. Monitoring the data collection; responsive design

# 1. Probability sampling background

Probability sample  $s$  from  $U = \{1, \dots, k, \dots, N\}$

*Inclusion prob. of  $k$  :*  $\pi_k = \Pr(k \in s)$

known for all units  $k \in U$

$\Rightarrow$  design weight  $d_k = 1/\pi_k$

Target of estimation :  $Y = \sum_U y_k$

The (unrealistic) case of

Sampling and full response

Full response :

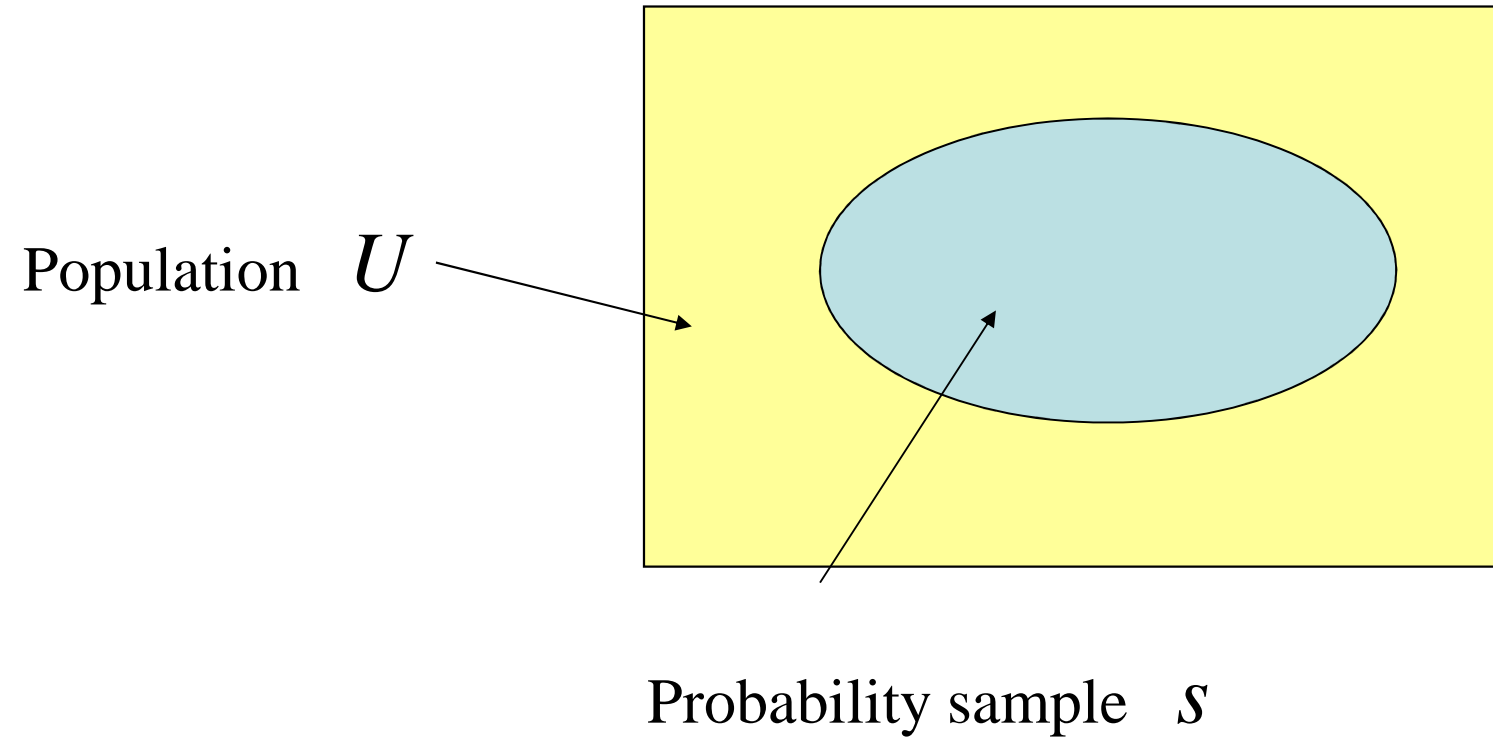
Study variable  $y_k$  observed all  $k \in s$

sample  $s \subset$  population  $U$

Unbiased estimation possible, e.g. by HT estimator

$$\hat{Y}_{HT} = \sum_s d_k y_k \quad \text{with } d_k = 1 / \pi_k$$

# Sampling and full response



.

.

A well known concept : *balanced sample*

sample mean = population mean

for measurable variables

.

Full response

and known population auxiliary total

$$\sum_U \mathbf{x}_k = N \bar{\mathbf{x}}_U$$

Brings a regression adjustment to HT estimator :

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + N(\bar{\mathbf{x}}_U - \bar{\mathbf{x}}_s)' \mathbf{b}$$

$$\text{where } \bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k / \sum_s d_k$$

If  $\bar{\mathbf{x}}_s = \bar{\mathbf{x}}_U$  : balanced sample

and the adjustment term is ZERO

An extensive literature exists on balanced sampling for full response

R. Royall and collaborators in the 1970's  
(model based approaches)

The cube method (Deville and Tillé, 2006)  
combines probability sampling with balancing

At the sampling stage, realize with the cube method  
what one would otherwise have realized later  
at the estimation stage by GREG estimation

For full response, balanced sampling brings the regression adjustment term in the GREG to zero.

Similarly, we shall see for nonresponse:

balancing the response set

brings zero nonresponse adjustment



More realistically here : *Nonresponse*

$$U \supset s \supset r$$

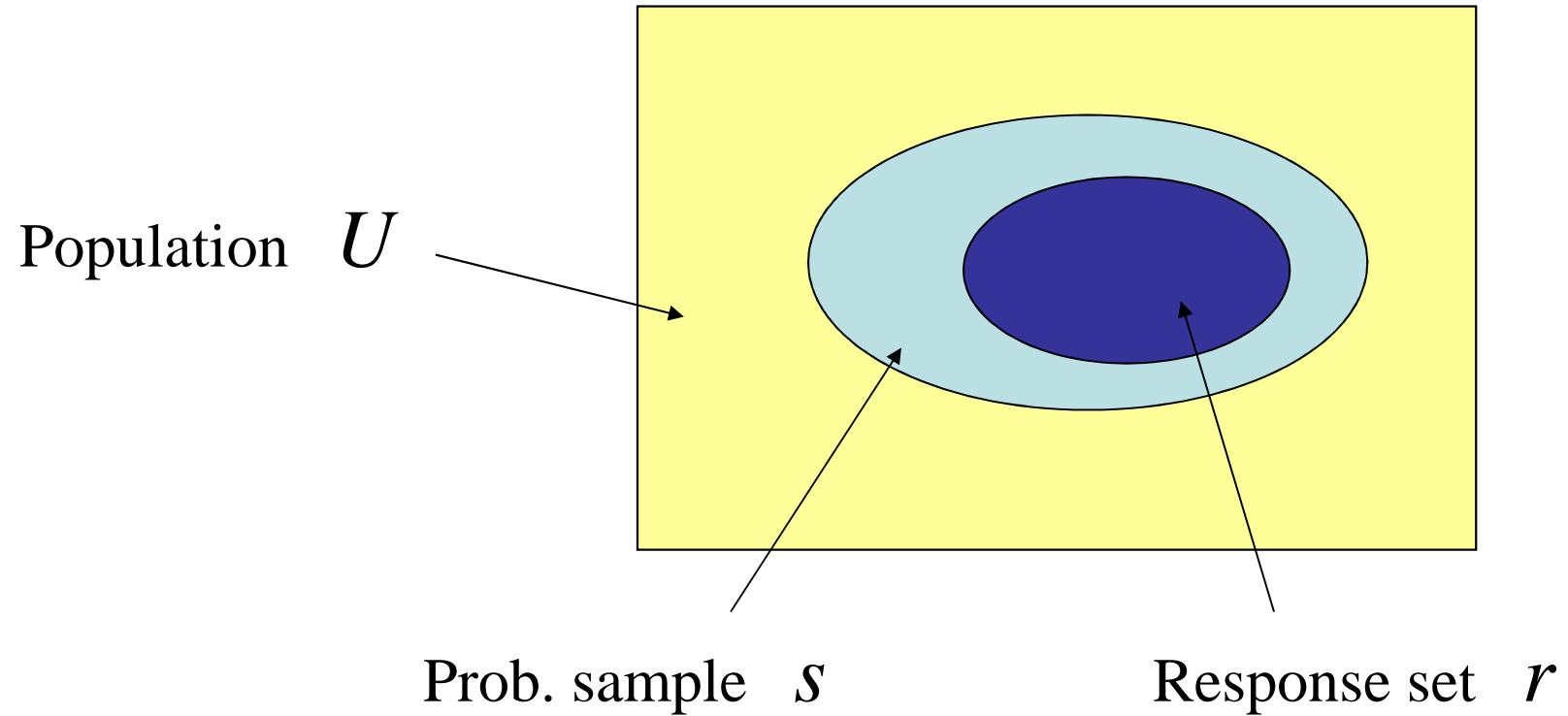
population  $\supset$  sample  $\supset$  response set

The response set  $r$  is the subset of sample  $s$   
for which  $y_k$  is observed  
Response rate (weighted) :

$$P = \sum_r d_k / \sum_s d_k$$

(Some call  $r$  the sample; not so here)

Our situation :



$r$  = the set where  $y_k$  observed

$$U \supset s \supset r$$

population  $\supset$  sample  $\supset$  response set

Desirable, unrealized estimator (would be unbiased) :

$$\hat{Y} = \sum_r d_k \frac{1}{\theta_k} y_k \quad \text{with } d_k = 1 / \pi_k$$

$\theta_k$  is the unknown response prob. of unit  $k$

Had  $\theta_k$  been known, we could stop here

Said differently : Under nonresponse,  
the theory for unbiased estimation fails because  
observation probability is unknown  
and  $<$  inclusion probability

$$\Pr(\text{observe } y_k) = \pi_k \times \theta_k$$

$$\pi_k = \Pr(k \in s) \quad \text{known inclusion prob.}$$

$$\theta_k = \Pr(k \in r | s) \quad \text{unknown response prob.}$$

But  $\theta_k$  unknown : We work without it

Desirable, unrealized estimator (would be unbiased) :

$$\hat{Y} = \sum_r d_k \frac{1}{\theta_k} y_k \quad \text{with } d_k = 1 / \pi_k$$

The two-phase weighting that would make this estimator unbiased does not work because the weight  $1/\theta_k$  is unknown

We must work without knowing  $\theta_k$

The response rate

$$P = \sum_r d_k / \sum_s d_k$$

is an often computed survey characteristic

But gets more attention than it deserves ;  
in itself, insufficient to portray the harm  
done by non-response

Given today's high nonresponse,

the *quality* of the response data set is what counts ,  
much more so than its relative size (the response rate).

Quality features:

the composition, the balance, the representativity

those are the aspects of the response  
that we must measure

## 2. Balanced response ; measuring imbalance

Objectives in this section: To discuss

- the concept of *balanced response set*
- a measure of *imbalance*
- the *distance*

(between respondents and non-respondents)



.

Well known concept : *balanced sample*

The concept of *balanced response*, less well known,  
is essential here

.

Why examine *balance*, and *imbalance*  
of the response ?

Comparative perspectives:

In a repeated survey: Is this year's response  
better balanced than last year's ?

Multinational survey: Do participating countries differ  
in the degree of balance they get ?

.

.

Dynamic perspective in one & the same survey:

During data collection,

can we influence the balance,  
improve it by interventions ?

.

In the dynamic perspective: Response set  
grows larger as more and more units respond

There is an *ultimate response set*  $r$  ,  
one that we have in a sense created  
through judicious intervention

We want it to be well balanced

For this we need tools & concepts

.

One such concept :

***Balance***

refers to ***equality of means***

for important measured variables

Imbalance , Balance , Distance

are concepts built on auxiliary variables

Auxiliary vector value  $\mathbf{X}_k$

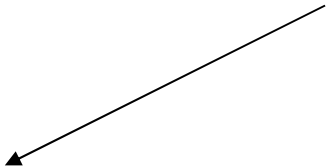
known for  $k \in s$ , perhaps for  $k \in U$

(But study variable  $y_k$  recorded for  $k \in r$  only)

$$r \subset s \subset U$$

Think multivariate !

$j$ :th aux. variable  
continuous or categorical


$$\mathbf{X}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})'$$

recorded at least for  $k \in s$ , maybe for all  $k \in U$

Dimension  $J$  can be considerable  
maybe 50 or more

## Auxiliary vector

$$\mathbf{x}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})' \quad \text{available for } k \in s$$

One of the simplest examples is multivariate :  
Classification vector

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

where the only “1” indicates class membership, as for ex.

$$J = 4 \times 6 \quad \text{size-by-industry classes}$$

Alternatively, “side by side”: size + industry + turnover

$$J = 4+6+5-2$$



Auxiliary vector     $\mathbf{X}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})'$

Is often composed from several classifications  
arranged “side-by-side”

(rather than crossed, to avoid small or zero cells)

.

## Auxiliary variables and vectors

$$\mathbf{x}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})'$$

assumed available for  $k \in s$

Often, several classifications “side-by-side” :

No classification ( $\mathbf{x}_k = 1$ ):  $J = 1$

One classification :  $J = 1 + (J_1 - 1)$

Two classifications :  $J = 1 + (J_1 - 1) + (J_2 - 1)$

Three classifications :  $J = 1 + (J_1 - 1) + (J_2 - 1) + (J_3 - 1)$

$J$  may be 50 or more .

Auxiliary vector     $\mathbf{X}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})'$

One category excluded in each classification  
in order to preserve matrix invertibility.

## Criterion $H_3$ ; order of selection

Step	Variable entering	$H_3 \times 10^3$	<i>RDF</i>
0	(trivial)	0	10.6
1	EDUCATION LEVEL (3)	186	6.0
2	POSTAL CODE CLUSTER (6)	250	5.6
3	COUNTRY OF BIRTH (2)	281	5.5
4	INCOME CLASS (3)	298	2.4
5	AGE CLASS (4)	354	3.1
6	SEX (2)	364	2.8
7	URBAN DWELLER (2)	374	2.6
8	INDEBTEDNESS (3)	381	2.3

***RDF* = relative deviation from unbiased est.**

## Criterion $H_1$ ; order of selection

Step	Variable entering	$H_1 \times 10^3$	<i>RDF</i>
0	(trivial)	0	10.6
1	INCOME CLASS (3)	76	4.5
2	EDUCATION LEVEL (3)	107	2.0
3	HAVE CHILDREN (2)	114	1.4
4	URBAN DWELLER (2)	118	1.1
5	SEX (2)	123	0.7
6	MARITAL STATUS (2)	125	0.5
7	DAYS UNEMPLOYED (3)	121	0.9
8	MONTHS SICKNESS (3)	120	1.0

***RDF* = relative deviation from unbiased est.**

## Order of selection, an example

Step	Variable
0	(none)
1	EDUCATION LEVEL (3)
2	POSTAL CODE CLUSTER (6)
3	COUNTRY OF BIRTH (2)
4	INCOME CLASS (3)
5	AGE CLASS (4)
6	SEX (2)
7	URBAN DWELLER (2)
8	INDEBTEDNESS (3)

## Now we confront respondents with full sample

For  $j^{\text{th}}$  variable  $x_j$ , compute

$$D_j = \bar{x}_{jr} - \bar{x}_{js}$$

respondent  
mean

full sample  
mean

As a vector :  $\mathbf{D} = (D_1, \dots, D_j, \dots, D_J)'$

## Confronting respondents with full sample

Mean difference vector, dimension  $J \geq 1$

$$\mathbf{D} = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$$

;

$$\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k / \sum_r d_k \quad ; \quad \bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k / \sum_s d_k$$

Sample design weighted :  $d_k = 1/\pi_k$



Balanced response set  $r$  :

Respondents *on average* equal to  
full sample, with respect to the chosen  $\mathbf{x}_k$

$$\mathbf{D} = \underbrace{\bar{\mathbf{x}}_r}_{\text{responding}} - \underbrace{\bar{\mathbf{x}}_s}_{\text{all sampled}} = \mathbf{0}$$

Intuitively desirable, but hard to realize completely

Goal for *data collection* : try to get high balance

*Estimation stage*: adjustment still needed,  
but part of the job done

Why seek balance ?

Because balance on an  $\mathbf{x}$ -vector strongly related  
to the study variable  $y$

$\Rightarrow$  even the simple expansion estimator  
is close to unbiased

$$\hat{Y}_{EXP} = N \bar{y}_r = N \times \frac{\sum_r d_k y_k}{\sum_r d_k}$$

Show it as an exercise !

Normally,  $\mathbf{D} = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s \neq \mathbf{0}$  : Response is *unbalanced*

$\mathbf{D}$  is multivariate;  
we need a *univariate* measure of  
*imbalance*.

To this end, use  $\mathbf{D}'\Sigma_s^{-1}\mathbf{D}$   
a quadratic form in  $\mathbf{D}$

$J \times J$  weighting matrix :  $\Sigma_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k$

assumed non-singular

Notation :  $IMB = \mathbf{D}'\Sigma_s^{-1}\mathbf{D}$

Note:  $IMB$  (for Imbalance) depends on

- (i) the composition of the auxiliary vector  $\mathbf{x}_k$
- (ii) the composition of  $r$  , given  $s$

$IMB(r, \mathbf{x}_k | s)$  would be more informative notation

But let us use simply  $IMB$

Properties of the imbalance statistic

$$IMB = \mathbf{D}'\Sigma_s^{-1}\mathbf{D}$$

- balanced response  $\Leftrightarrow \mathbf{D} = \mathbf{0} \Rightarrow IMB = 0$
- $IMB \geq 0$  any outcome  $(s, r)$  and vector  $\mathbf{x}_k$

The imbalance statistic

$$IMB = \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \boldsymbol{\Sigma}_s^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)$$

is an extension

to multivariate auxiliaries

of the univariate

$$\frac{\text{mean difference}}{\text{stand.dev.}} = \frac{\bar{x}_r - \bar{x}_s}{S_x}$$

Interposing the inverse of  $\Sigma_s$  “standardizes”

and permits a simple upper bound

to be stated on  $IMB = \mathbf{D}'\Sigma_s^{-1}\mathbf{D}$

For any outcome  $(s, r)$  and vector  $\mathbf{X}_k$ , we have

$$0 \leq IMB \leq \frac{1}{P} - 1 \quad P = \text{response rate}$$

20% nonresponse :  $0 \leq IMB \leq 0.25$

50% nonresponse :  $0 \leq IMB \leq 1$

*IMB* is “not a big number”

But  $IMB = 0.20$  can mean large imbalance  
compared with  $IMB = 0$  which is perfect balance



The ratio  $\frac{IMB}{1/P - 1}$

(between 0 and 1) measures the degree  
to which  $\mathbf{X}_k$  explains the response

(Exercise: Show it!)

.

Experience with survey data shows

*IMB* usually not close to its upper bound  $\frac{1}{P} - 1$

Usually  $IMB < 0.3$

but depending greatly on the choice of  $\mathbf{x}$ -vector :

- the number of  $x$ -variables
- how well they “explain” nonresponse

For fixed response  $r$  and given sample  $S$  ,  
adding more variables to  $\mathbf{X}_k$  increases  $IMB$   
(proof not given here)

A bigger  $\mathbf{X}$ -vector has more imbalance, naturally,  
because more variables on which means have to agree

The trivial  $\mathbf{x}$ -vector  $\mathbf{X}_k = 1$  has  $IMB = 0$   
yet is a totally unattractive vector

## Achieving well balanced response

is a challenge we impose on the data collection

The task is tougher the more  $x$ -variables  
we decide to balance on  
(but rewards may be greater)

.

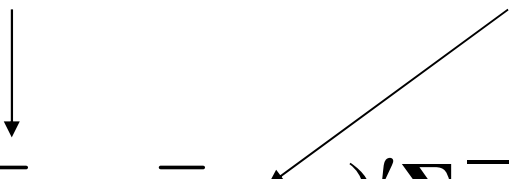
Some simple functions of the *IMB* statistic  
are very useful :

- The idea of distance (between respondents and nonrespondents)
- The notion of Balance (imbalance with opposite sign)
- Related is the R-indicator (*R* for representativity; the RISQ project)

.

## The notion of distance

between respondents  $r$  and nonrespondents  $nr = s - r$


$$dist_{r|nr} = \{ (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{s-r})' \Sigma_s^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{s-r}) \}^{1/2}$$

Its simple relation to imbalance :

$$dist_{r|nr} = \frac{1}{1-P} \sqrt{IMB}$$

$$P = \sum_r d_k / \sum_s d_k = \text{response rate}$$

The distance  $dist_{r|nr} = \frac{1}{1-P} \sqrt{IMB}$

For example, 40% nonresponse, and  $IMB = 0.16$

$$\Rightarrow dist_{r|nr} = 1$$

Property :  $dist_{r|nr} \leq \frac{1}{\sqrt{P(1-P)}}$

For ex., nonresponse 50%  $\Rightarrow$  distance  $\leq 2$

even for the most ill-structured response  $r$

Experience shows : It is seldom  $> 0.5$

again depending greatly on our choice of **X**-vector



The notion of balance  
of the response set

$$BI = 1 - \sqrt{\frac{IMB}{1/P - 1}}$$

*BI* for Balance Indicator ;  
between 0 and 1

A legitimate objective :

Achieve small distance

so as to have “respondents like nonrespondents”  
when data collection ends

## The Swedish Living Conditions Survey 2009

Telephone interview survey.

*WinDATI events* (contact attempts) are registered

**Ordinary** data collection: 3 weeks;  
for some units,  $> 30$  contact attempts;  
at the end, resp. rate = 60.4 %

**Follow-up**, 3 weeks, final resp. rate = 67.4%

.

Ordinary data collection

(with  $> 20$  call attempts for many units)

Follow-up data collection

(with  $> 10$  call attempts for many units)

All these attempts - is it worth it ?

.

## Monitoring the data collection

In a dynamic perspective : A series of expanding response sets, viewed as a function of the time point  $a$

$$r^{(1)} \subset r^{(2)} \subset \dots \subset r^{(a)} \subset \dots$$

For simplicity, let  $r$  denote any one of the growing sample subsets

For the Swedish LCS 2009, we compute  
the imbalance statistic

$$IMB = \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \mathbf{\Sigma}_s^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)$$

and the distance respondents-to-nonrespondents

$$\begin{aligned} dist_{r|nr} &= \{(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{s-r})' \mathbf{\Sigma}_s^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{s-r})\}^{1/2} \\ &= \frac{1}{1-P} \sqrt{IMB} \end{aligned}$$

More specifically, we compute *IMB* and the distance repeatedly during data collection (for a series of growing response sets  $r$ ) and for the vector

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

of dimension  $2^3 = 8$   
defined by crossing of three dichotomous x-variables,

educ  $\times$  owner  $\times$  origin

.

## The actual LCS 2009 data collection file

Attempt #	$100 \times P$	$100 \times IMB$	$\text{dist}_{r/nr}$
1 ordin	12.8	4.13	0.233
5 ordin	44.3	2.99	0.310
12 ordin	57.7	2.78	0.394
End ordin	60.4	2.72	0.417
1 fol-up	61.4	2.61	0.418
4 fol-up	64.6	2.37	0.435
Final	67.4	2.36	0.471

Note: The distance increases the whole time



## The actual LCS 2009 data collection

The distance between respondents and nonrespondents increases the whole time

From 0.310 at attempt # 5

To 0.471 at the end of data collection

Nonrespondents become more and more unlike respondents.

This is disturbing, even unacceptable

# The actual LCS 2009 data collection

Distance increases the whole time

Alternative interpretation :

Respondents are becoming less and less representative.

Signs of an inefficient data collection.

Why continue data collection

according to an unchanged format,

and just get “more of the same” ?

Mathematical note : We are assuming  $\mathbf{x}$ -vectors of the type:

There exists a constant vector  $\boldsymbol{\mu}$  such that

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

Most vectors of interest are of this type, for ex.

$$\text{If } \mathbf{x}_k = (1, x_k)', \text{ take } \mu = (1, 0)'$$

$$\text{If } \mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, \dots, \gamma_{jk}, \dots, \gamma_{Jk})' = (0, \dots, 1, \dots, 0)'$$

$$\text{take } \mu = (1, 1, \dots, 1)'$$

### 3. Monitoring the data collection (a form of Responsive Design)

A dynamic perspective : Data collection extends over a period of time (days, weeks)

We can perhaps make suitable interventions or changes underway

to obtain in the end a well balanced response set.

## Monitoring the data collection

Dynamic perspective : Monitor the data collection, seen as  
function of the contact attempts (attempt 1, attempt 2 ...)  
or of the data collection days, (day 1, day 2 ...)

and perhaps make suitable interventions or changes

Using tools that we now have : *IMB* , and functions of *IMB*

## Monitoring the data collection

A series of expanding response sets, viewed as a function of the time point  $a$

$$r^{(1)} \subset r^{(2)} \subset \dots \subset r^{(a)} \subset \dots$$

For simplicity,  $r$  denotes any one of the growing sample subsets

.

During data collection, how can we reduce  $IMB = \mathbf{D}'\Sigma_s^{-1}\mathbf{D}$   
bring  $\mathbf{D}$  closer to the zero vector ?

What interventions in the data collection ?

What modifications of an original plan ?

## Monitoring the data collection

Differentiate the sample units,  
via their observable characteristics,

intervene and halt the contact attempts in sample subgroups  
where “realistic expectations” on the response  
have already been met -

it does not pay to pursue those any more,  
it just gives “more of the same”



Collecting “more of the same”

is often unproductive,

does not reduce imbalance  $IMB = \mathbf{D}'\Sigma_s^{-1}\mathbf{D}$

does not reduce  $dist_{r|nr} = \frac{1}{1-P} \sqrt{IMB}$

## Considerations

- a good data collection should show decreasing distance  $dist_{r/nr}$  as  $r$  expands, should be small at the end
- $dist_{r/nr}$  a more suitable tool for monitoring than  $IMB$ , which tends to decrease as  $r$  grows towards full response  $S$  (because  $\mathbf{D} \rightarrow \mathbf{0}$ )

The  $\mathbf{x}$ -vector used to define  $IMB$  is general.

Only one case discussed here:

The particularly transparent case of  $J$   
mutually exclusive and exhaustive groups

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

*e.g.* size-by-industry

$IMB$  takes a simple expression :

a sum of non-negative terms, one for each group

## The case of $J$ sample subgroups

$IMB$  has a simple expression :

$$IMB = \sum_{j=1}^J C_j \quad \text{with} \quad C_j = W_{js} \times \left( \frac{P_j}{P} - 1 \right)^2$$

$W_{js}$  = class  $j$  proportion out of the whole sample  $s$

$P_j$  = response rate, group  $j$

$P$  = overall response rate

$J$  sample subgroups

$$IMB = \sum_{j=1}^J C_j \quad \text{with} \quad C_j = W_{js} \times \left(\frac{P_j}{P} - 1\right)^2$$

A procedure: Compute  $IMB$  and the  $C_j$  repeatedly during data collection.

Response proportion  $P$  increases.

We observe continuously

$P_j$  = response rate, group  $j$   
and we can follow which groups  
contribute most to imbalance

those overrepresented :  $P_j > P$

those underrepresented :  $P_j < P$

$J$  sample subgroups

$$IMB = \sum_{j=1}^J C_j \quad \text{with} \quad C_j = W_{js} \times \left( \frac{P_j}{P} - 1 \right)^2$$

Those overrepresented :  $P_j > P$   
those we do not need any more of,  
although they are “an easy way out”  
for the interviewers  
because they are “easy cases”

$J$  sample subgroups

$$IMB = \sum_{j=1}^J C_j \quad \text{with} \quad C_j = W_{js} \times \left( \frac{P_j}{P} - 1 \right)^2$$

Desirable goal :

Make all  $P_j$  equal  $\Rightarrow IMB = 0$  : Perfect balance  
for the groups *that we decided to monitor*

But seldom will we realize it completely in practice

## We return to the Swedish Living Conditions Survey 2009:

Telephone interview survey.

WinDATI events are registered

We have seen signs that the current data collection is inefficient : Distance between respondents and nonrespondents increases as the data collection proceeds.



## Swedish Living Conditions Survey 2009

Experiments were carried out “in retrospect” :

$2^3 = 8$  sample subgroups identified by

$$\mathbf{x} = (\text{educ} \times \text{owner} \times \text{origin})$$

called *monitoring vector*

Data collection considered stopped in a group  
when its response rate had reached 60%

Consequence: We disregard some already collected  
y-data (to get better balance)

## Experiments with Swedish Living Conditions Survey 2009

Data collection was stopped in a group  
when its response rate had reached 60%

Some groups stop sooner than others;  
in the end, all groups tend to have more equal  
response rates

The terms  $C_j = W_{js} \times \left(\frac{P_j}{P} - 1\right)^2$  less variable

$$IMB = \sum_{j=1}^J C_j \quad \text{is reduced}$$

## Experiment with the LCS 2009 data

Attempt #	$100 \times P$	$100 \times IMB$	distance
7 ordin	50.9	3.07	0.357
8 ordin	52.5	2.81	0.353
9 ordin	53.8	2.49	0.341
15 ordin	56.0	1.59	0.287
3 fol-up	58.6	1.09	0.252
Final	58.9	0.82	0.220

Now the distance is decreasing,  
thanks to interventions  
(data collection stopped in groups with  $P > 60\%$ )

## Experiment with the LCS 2009 data

	$100 \times P$	$100 \times IMB$	distance
Final	58.9	0.82	0.220

Compare actual LCS 2009 data collection :

Final	67.4	2.36	0.417
-------	------	------	-------

Despite much smaller response rate (58.9 vs. 67.4)  
get much smaller distance (0.22 vs 0.42)

## Scenario for data collection stage (Responsive Design)

an example :

- Decide on a monitoring vector  $\mathbf{X}_k$
- During data collection, compute group response rate  $P_j$   
for  $j = 1, \dots, J$
- If  $P_j$  has reached “reasonable expectations”,  
cease data collection in that group  $j$
- Focus data collection on other groups, until the end
- Proceed to estimation stage and nonresponse adjustment  
of estimates

General procedure based on  $\mathbf{x}$ -vector of arbitrary type  
(with continuous and/or categorical variables)  
based on response propensity  $\hat{P}_k$

- At several points in the data collection, compute  $\hat{P}_k$   
for all  $k \in s$
- At point 1, stop data coll. for those units  $k$  having  
attained “high  $\hat{P}_k$ ” (e.g. the 20% highest), set those aside
- At point 2, stop data coll. for next 20%
- And so on until the end
- Proceed to estimation stage and  
nonresponse adjustment of estimates

Lectures

by

Carl-Erik Särndal  
Örebro University  
Statistics Sweden

Baltic-Nordic-Ukrainian Workshop

Valmiera, Latvia

24-28 August 2012

Lecture 3 : The estimation stage:

Calibrated weighting for nonresponse bias reduction  
and preferably without increased variance

The scenario is now changed: response  $r$  is fixed  
cannot be improved any more ;  
we have to live with it in the estimation



Estimation stage : adjusting for nonresponse

$$r \subset s \subset U$$

response  $\subset$  sample  $\subset$  population

$y_k$  recorded  $k \in r$  only

Response set  $r$  is fixed

cannot be improved any more ;

Objective: Construct an efficient **x**-vector

Available : a supply of aux. variables, perhaps many

Objective: construct an efficient  $\mathbf{x}$ -vector,  
used to compute calibrated weights

to reduce as much as possible  
the bias still affecting the estimates

despite (incomplete) balancing at the data collection stage

How do we select, in a stepwise or other fashion,  
the  $x$ -variables that adjust the most ?

“Pick best ones first” is one option

Numerous  $y$ -variables complicates the question

Effective adjustment for one is maybe not so for others

For sake of theory, must look at one of them

## Estimators

of the population total  $\sum_U y_k$

Unbiased, but not available under nonresponse

Horvitz-Thompson, for full response

$$\hat{Y}_{FUL} = \sum_s d_k y_k$$

## Estimators under nonresponse

- Basic, but poor choice, considerably biased :  
the crude expansion estimator

$$\hat{Y}_{EXP} = N \bar{y}_r \qquad \bar{y}_r = \sum_r d_k y_k / \sum_r d_k$$

- Adjusted, less biased,  
by calibration on a potent **x**-vector :

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

$m_k$  = adjustment factor computed on chosen **x**-vector

Calibration estimator of  $Y = \sum_U y_k$

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

uses adjustment factor

$$m_k = \underbrace{\left( \sum_s d_k \mathbf{x}_k \right)' \sum_r d_k \mathbf{x}_k \mathbf{x}_k'}_{\text{row vector}}^{-1} \underbrace{\mathbf{x}_k}_{\text{column}}$$

Weights  $d_k m_k$  calibrated to  $\sum_s d_k \mathbf{x}_k$

**Note :**  $\mathbf{x}_k$  here may be different from the  $\mathbf{x}_k$  used to monitor the data collection

Calibration estimator       $\hat{Y}_{CAL} = \sum_r d_k m_k y_k$

with       $m_k = \underbrace{\left( \sum_s d_k \mathbf{x}_k \right)' \sum_r d_k \mathbf{x}_k \mathbf{x}_k'}_{\text{row vector}}^{-1} \underbrace{\mathbf{x}_k}_{\text{column}}$

For some  $x$ -variables, information  
all the way up to the population level  
(« star variables »)

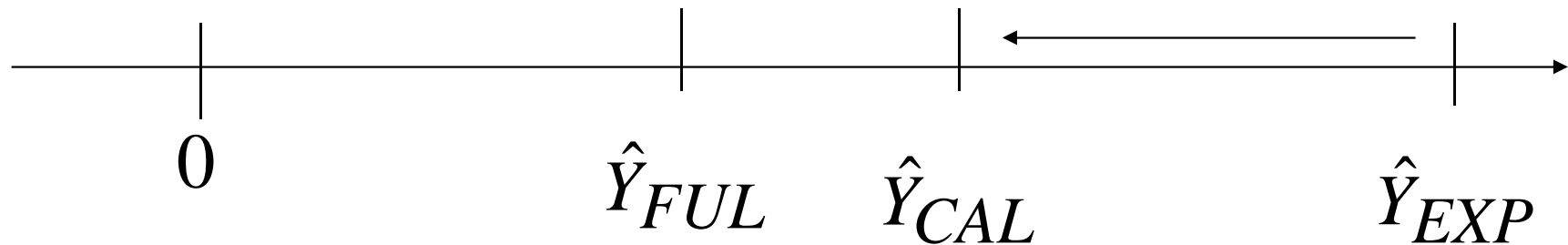
$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$$

Weights  $d_k m_k$  calibrated to       $\begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$

Building  $\mathbf{x}$ -vector from scratch

more & more variables added to  $\mathbf{x}$  :

fixed  $r$  and  $s$

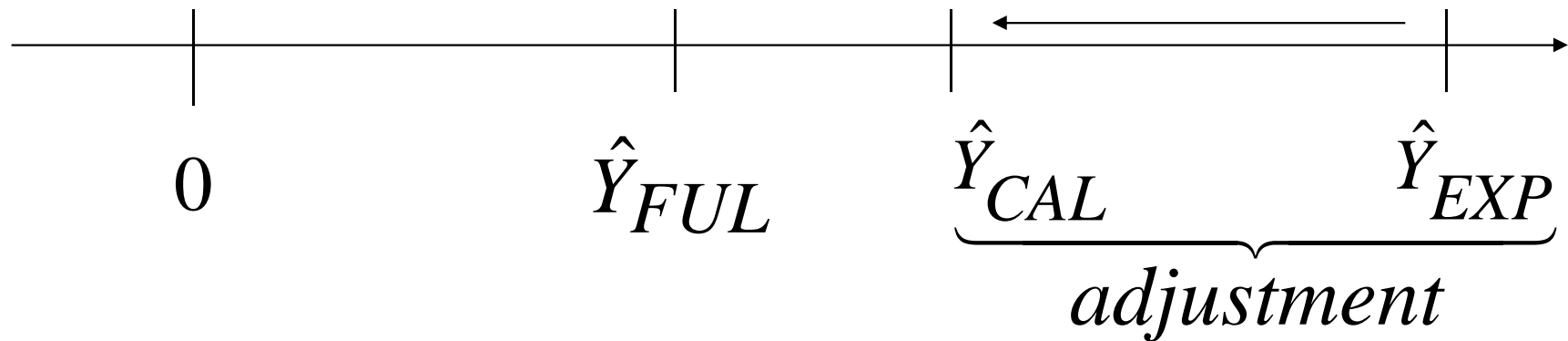


$\hat{Y}_{CAL}$  moves away from  $\hat{Y}_{EXP}$  (very biased)

and approaches  $\hat{Y}_{FUL}$  (without bias)



When  $\mathbf{X}_k$  improves, for fixed  $r$  and  $s$



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

goes diminishing  
but probably not to zero

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

$\hat{Y}_{EXP} - \hat{Y}_{CAL}$       computable adjustment  
 changes with the choice of **X**-vector,  
 If large, suggests a considerable bias  
 has become adjusted for

$\hat{Y}_{EXP} - \hat{Y}_{FUL}$       not computable, unchanging

Let us examine the computable *standardized adjustment*

$$StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y}$$

$S_y$  = stand. dev. of  $y$

computed on the response  $r$

## Interpretation

Consider

$$StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y} = 0.10 \quad (\text{fairly typical})$$

Then we have

moved away 0.10 stand.dev.

from the primitive mean estimate  $\hat{Y}_{EXP} / \hat{N} = \bar{y}_r$

to obtain adjusted estimate :

$$\frac{\hat{Y}_{CAL}}{\hat{N}} = \frac{\hat{Y}_{EXP}}{\hat{N}} - 0.10 \times S_y$$

$$StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y} = 0.10 \quad (\text{fairly typical})$$

Seemingly small, it can mean  
a very large move, compared with  $\frac{S_y}{\sqrt{n}}$

$$\text{for ex. } n = 10,000 \Rightarrow \frac{S_y}{\sqrt{n}} = \frac{S_y}{100} = 0.01 \times S_y$$

$$\text{adjustment } \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N}} = 0.10 S_y \quad 10 \text{ times greater}$$

Experience with data shows :

$$StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y} \quad \text{seldom} > 0.3$$

In practice, we can always compute  $StAdj$

But for our understanding we should ask :

What factors determine the  $StAdj$  ?

Traditional wisdom holds :

$\mathbf{x}$ -vector should (must) explain study variable  $y$

$\mathbf{x}$ -vector should (must) explain the response

At best, it does so to a degree only

These are two factors we expect to find in *StAdj*

We do, but there is a third important factor

Some work shows: 3 factors determine

$$StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y} = \sqrt{IMB} \times R_{y,\mathbf{x}} \times R_{D,C}$$

where  $IMB$  is the imbalance (still remaining)

$R_{yx}$  and  $R_{DC}$  are correlation coefficients



DE

$$StAdj = \sqrt{IMB} \times R_{y,\mathbf{x}} \times R_{D,C}$$

$$\text{We have } R_{y,\mathbf{x}} \leq 1 \quad ; \quad |R_{D,C}| \leq 1$$

$$\text{and typically } 0 < IMB < 0.3$$

$$\text{For ex. } StAdj = 0.5 \times 0.8 \times 0.2 = 8\%$$

$$\text{Adjustment} = 0.08 \text{ stand.dev.}$$

$$\Rightarrow \text{adjusted est.} = \frac{\hat{Y}_{CAL}}{\hat{N}} = \frac{\hat{Y}_{EXP}}{\hat{N}} - 0.08 \times S_y$$

## The three factors

$$StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y} = \sqrt{IMB} \times R_{y,\mathbf{x}} \times R_{D,C}$$

## The first factor

$$\sqrt{IMB} = \sqrt{\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}} \quad \text{sqrt. imbalance}$$

This factor depends on  $\mathbf{x}$  only, not on any of the (many)  $y$ 's  
It measures the degree to which  $\mathbf{x}$  explains the response

Perfect balance :  $IMB = 0$  : No adjustment occurs

weighting matrix :  $\boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' / \sum_r d_k$

.

Note : over  $r$  not  $S$

## The second factor

$R_{y,\mathbf{X}}$  = coef. multiple corr. between  $y$  and  $\mathbf{X}$

based on data  $(y_k, \mathbf{x}_k)$  ,  $k \in r$ ,  $d$ -weighted

### The third factor

$R_{D,C}$  = coeff. of corr. between  $D_j$  et  $C_j$

Viewed as  $J$  data points,  $(D_j, C_j)$ ,  $j = 1, \dots, J$  ;

$D_j = \bar{x}_{jr} - \bar{x}_{js}$  deviation,  $x$ -variable  $j$

$C_j = \text{covariance}(x_j, y)$

The third factor  $R_{D,C}$  is high

if the large deviations  $D_j = \bar{x}_{jr} - \bar{x}_{js}$

go together with the large correlations  $x_j$ -to- $y$

$j = 1, 2, \dots, J = \#$  variables in  $\mathbf{X}$ -vector

.

.

*Large adjustment*      $StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y}$      occurs if

- Large imbalance still needing to be compensated for
- High relationship  $y$ -to- $\mathbf{X}$
- High relationship between  
    deviations  $D_j$  and covariances  $C_j$ 
  - large deviations matched with  
    high correlations

DE

Properties of  $StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y}$

when we add more  $x$ -variables to the vector  $\mathbf{X}_k$ :

- first factor  $IMB$  increases
- second factor  $R_{y,\mathbf{x}}$  increases
- third factor  $R_{DC}$  may not increase  
in abs.value but may be fairly constant

$StAdj$  does not necessarily increase



DE

Criteria for stepwise selection  
of variables for the  $\mathbf{x}$ -vector

$$\cdot H_3 = \sqrt{IMB}$$

$$\cdot H_2 = \sqrt{IMB} \times R_{y,\mathbf{x}}$$

$$\cdot H_1 = \sqrt{IMB} \times R_{y,\mathbf{x}} \times |R_{D,C}|$$

Advantage of  $H_3$ : computed only from the values  $\mathbf{X}_k$  ;  
does not involve the  $y$ -variable .

$H_1$  and  $H_2$  depend on both  $y$  and  $\mathbf{x}$

## Criterion $H_3$ ; order of selection

Step	Variable entering	$H_3 \times 10^3$	<i>RDF</i>
0	(trivial)	0	10.6
1	EDUCATION LEVEL (3)	186	6.0
2	POSTAL CODE CLUSTER (6)	250	5.6
3	COUNTRY OF BIRTH (2)	281	5.5
4	INCOME CLASS (3)	298	2.4
5	AGE CLASS (4)	354	3.1
6	SEX (2)	364	2.8
7	URBAN DWELLER (2)	374	2.6
8	INDEBTEDNESS (3)	381	2.3

***RDF* = relative deviation from unbiased est.**

## Discussion and conclusion

How should the nonresponse problem be treated

At the data collection stage ?

At the estimation stage ?

.

## Discussion

The important difference from the theory point of view :

Data collection stage: The response set is  
“tailored”, to some degree constructed

Estimation stage : The response set is fixed;  
Estimation theory is the basis for  
nonresponse adjustment

.

## Discussion

The data collection:

Responsive design is a prominent topic in the survey literature today.

It can give us ideas and tools to obtain a high quality set of respondents

.

I have discussed important *measurable quality features* of the response set, relative to a stated auxiliary vector **X**

They refer to the composition of the response set :

balance , distance

.

## Discussion

Responsive design and “creative data collection” should not be approached as a topic separate from the estimation .

A combined look at the data collection phase and the estimation phase is recommended .

Much work remains to do here .

.

## Discussion

The estimation stage:

The remaining bias still needs to be adjusted for.

Estimation theory is important.

One must not believe the task is finished after balancing the data collection on a chosen  $\mathbf{x}$ -vector.

.



Thank you for your attention

.

.

### Selected references

- Groves, R.M. and Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A*, **169**.
- Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. New York: Wiley.
- Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, **35**, 101-113.
- Schouten, B., Shlomo, N. and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, **27**, 231-253.
- Särndal, C.E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: Wiley.
- Särndal, C.E. (2011a). Dealing with Survey Nonresponse in Data Collection, in Estimation (Morris Hansen lecture). *Journal of Official Statistics*, **27**, 1-21.
- Särndal, C.E. (2011b). Three factors to signal nonresponse bias, with applications to categorical auxiliary variables. *International Statistical Review*, **79**, 233-254.
- Lundquist, P. and Särndal, C.E. (2012). Aspects of responsive design with applications to the Swedish Living Conditions Survey. Report 2011:1, Statistics Sweden