Adaptive and Responsive Design : thoughts from a time related perspective

Carl-Erik Särndal

Stockholm University & Statistics Sweden

Keynote address, 4th workshop,

Advances in Adaptive & Responsive Survey Design

Manchester, UK

November 9 – 10, 2015

2015-11-09

Thank you for the invitation

and for an opportunity to review

and to question some ideas, mine and those of others;

and to be self-critical

Acknowledgement : Co-operation with

Statistics Sweden

Stockholm University,

University of Tartu, Estonia

(although not for the opinions & thoughts in this talk)

…

Origins of Responsive (or Adaptive) design

R/A design for short

20 or so years ago we did not have this term, but is now an every-day word, among survey statisticians at least.

What circumstances or trends, in statistics production or in society, created the movement - the concept - R/A design ?

Survey theory & practice has always evolved in response to important trends in “the real world”, in society.

What drives the R/A design movement ?

…

R/A design ; I make three introductory points :

- R/A design: A sign of our times.
- R/A design: A “return to the sources”.
- R/A design: A welcome and positive development.

R/A design : A sign of our times

Data collection today faced with high nonresponse and high cost.

Thus two aspects seem to drive R/A design:

(a) Nonresponse and (b) Cost of data collection.

Difficulty to make contact with selected units and receive solicited data from those.

Cost: Why spend money on getting observations of marginal value at best, observations “that we already have enough of”?

R/A design : A return to the sources.

Classical texts in survey statistics –
the first that promoted “scientific sampling”
- as probability sampling was then called -
were those of the 1950’s - such as Cochran (1953):
Simple random sampling ; Stratified random sampling ;
Sampling in 2 or more stages, etc, etc.

These are ways to *select*
(a small number of) population units
so that the sample is *random* - (and representative ?)

R/A design : A return to the sources.

Classical texts in survey statistics – Cochran (1953) & others -
showed designs for *randomized selection* of sample,
and thereby the result of the data collection,
because in those days, nonresponse very low

“Selected unit” was essentially synonymous with "observed unit”

R/A design : A return to the sources.

The selection designs of the 1930's to 50's,

stratified, two-stage, etc.,

still govern the way statistical agencies think and do

their surveys today – remarkable, some 70 years later

A reason: these were simple and convincing concepts.

But high nonresponse ruined – in part - their validity.

"Selected unit" is no longer "observed unit"

R/A design : A return to the sources.

Nonresponse has caused much concern, even crises,

noticed in particular this year in Sweden

Media attention : “Statistics Sweden signaling high response”

as if they are unable to handle the situation

R/A design : A return to the sources.

By “return to the sources” I mean:

R/A design is a return to emphasis on “good set of observed units”

from the finite population.

“Good” at least in a sense of “representative” or “balanced”

or some similar notion,

even if falling short of “known observation probability”

…

R/A design : A return to the sources.

R/A design is a return to emphasis on “good set of observed units”
from the finite population.

I find that encouraging.
It fills a void that I have been aware of for some time
in the theoretical discussion on survey statistics. - I’ll explain.

R/A design : A return to the sources.

The term “representative”

Neyman (1934): Representative if

the variance can be estimated from the sample itself.

R/A design: a positive development

I was intrigued when first exposed to the ideas of R/A design

Enthusiastic even: I thought
this is how, in times of high nonresponse,
we return to an emphasis on
an “impeccable”, “irreproachable” set of observed units
from the finite population

R/A design: a positive development

And profit from it: The ideas of R/A design
may bring benefits :
improved accuracy (reduced bias) in the estimates
we make in presence of (high) nonresponse

My expectations were fulfilled,
but to a degree only, as I will explain

…

R/A design: a positive development

Estimation theory in survey statistics: prominent theme since 1970.

Different modes of statistical inference from finite population emerged and were hotly debated :

Design based, *model based*, *design-based model assisted*, *calibration*, etc.

Small area estimation

R/A design: a positive development

Small area estimation :

a prime demonstration, from 70's and on, of estimation theory "pushed far"

estimation with almost no observations -

but hardly any attention paid to the selection of units ,

to whether they respond or not

R/A design: a positive development

I co-authored, in the 80's, the book *Model Assisted Survey Sampling*

A friend, by far my senior, said at the time (some 25 years ago) :

Look Carl, your book is not about *survey sampling*,

it is about (design-based) *estimation in surveys*.

He saw a lack of emphasis on the data collection

The book should perhaps have been called

Estimation in probability sampling surveys.

R/A design: a positive development

I now see the R/A design movement as a desirable direction,

because it brings back an emphasis on the source of the data,

the set of *observed* units on which we will

set *estimation theory* in motion.

Three points I have just made of the R/A design movement

- R/A design: A sign of our times.
- R/A design: A “return to the sources”.
- R/A design: A welcome, a positive development.

Randomness of selected and observed units

The importance of “random sample” is an insight of the 1920’s – 1930’s.

But simple random sampling can give high variance in the estimates.

How resolve that? Select an “informed sample”, but at the same time

a probability sample

$\Rightarrow$ *Stratified* simple random sampling

Randomness of observed units

What created *stratified* SRS, some 80 years ago?

It was the idea to have an *informed sample* that was at the same time

a *probability sample*.

It is *enough* if it is random within identifiable population subgroups

called *strata*.

Stratified SRS is

"a little less random" (less entropy) than SRS

but a little more "information receptive"

Randomness of selected and observed units

Strata are identifiable subgroups of the finite population.

How identify membership ?

Requires *information* about the population,

to see group membership of all units

That is called *auxiliary information*.

Randomness of observed units

Stratified SRS combines two ideas

Representativeness (of population groups called strata) and
Randomness (imposed by probability sampling).

Stratified SRS requires the *sample s from U to be representative*
or *balanced* with respect to

$\mathbf{x}_k = (0,...,1,...,0)'$ stratum identifier, known all population units

hence satisfying the balancing equation $\sum_s d_k \mathbf{x}_k = \sum_U \mathbf{x}_k$

$d_k = (\text{inclusion prob})^{-1} = (\text{stratum sampling rate})^{-1} = N_h/n_h$

Randomness of units under nonresponse

Similarly in in nonresponse treatment.

(Auxiliary) Information is present; use it!

to get *representativeness* or *balance* in the ultimate set of respondents

- R/A designs can help -

and then of course, use it in the estimation phase also

Randomness of selected and observed units

Trouble is, in nonresponse treatment, despite all the auxiliary information,

we cannot get the full *randomness*

to completely meet the design-based estimation theory

With today's high nonresponse, an issue is :

Randomness vis-à-vis Representativeness

Representativeness or *balancing* helps, but does not manage to

"fill the vast open space" of the *randomness* ,
the protection that "scientific sampling" granted

But that's the best we can do under nonresponse

Private survey institutes (in Sweden, Canada and elsewhere) typically say :

We took a *representative sample* of 1000 persons.

They may add : Representative *with respect to* (list a few variables)

- Some institutes say: A "Mini-Sweden"

Randomness vis-à-vis representativeness

More explicitly:

Representative with respect to
sex×*age group*×*region*

is not as good - in design based thinking - as

Stratified simple random sample, within strata defined by
sex×*age group*×*region*

Representativeness does not manage to “fill the vast open space” of randomness. “Mini-Sweden” is not sufficient.

...

R/A designs : Auxiliary information essential

What do R/A designs rely on, what makes them work ?

The availability of auxiliary information,

the material on which to structure the adaptive data collection.

including both *Register information* (plentiful in Sweden)

and *Paradata* (features of the data collection)

Who's got auxiliary information?

Statistics Sweden has got it, lots of it,

for surveys on individuals & households in particular

As have a few others, notably in northern Europe.

Many others have not got it.

Differences (between countries)

in access to auxiliary information

restricts or hampers the dialogue

My approach (in the past few years) has been :

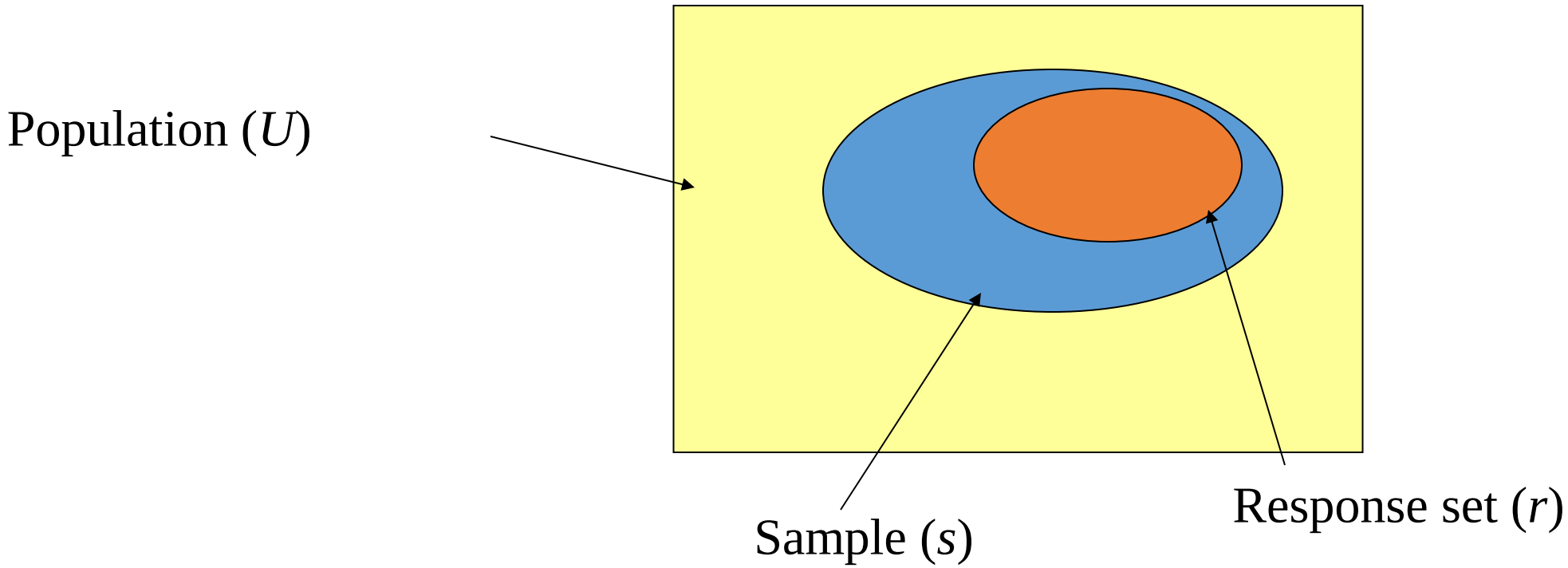
R/A design followed by an estimation phase poses

a problem of *statistical inference* :

“selection, observing some of the selected, estimation based on those observed”

R/A design extends my interest in statistical inference for surveys

This is the situation :



s is the probability sample

r is the response set, i.e., where survey variable *y* is observed

(*r* is not ”the sample”. *s* is the sample)

Data collection is followed by *estimation*

There is a decisive date, a fatal date, a date of no return :
The day data collection is declared ended ;
after that, no more y-values collected; r is fixed
and estimation has to take place

R/A design is (to a considerable degree) about

"*Before* the fatal date"

Estimation theory - adjustment weighting, and so on - is about

"*After* the fatal date"

Researchers tend to be divided into

"the before-people" and "the after-people"

The before-people concentrate on action in data collection

e.g., with R/A designs

Anticipate the problem, ahead of the fatal date:

influence what responding units *k* we shall end up with;

change the composition of the set *r* before the fatal date.

"The anticipators"

The *after-people* investigate what can be done in in estimation
with the response that was ultimately obtained

Attenuate the problem after the fatal date:
estimation theory, including weighting methods

“The attenuators”

“Anticipators” versus “Attenuators”

Nothing new : A similar dichotomy,
20 or so years ago, was the distinction was between

“Reducers” and “Adjusters”

…

The R/A design movement killed
the dream of the Reducers
i.e., to strive for (and get) nearly 100% response

To me, the R/A design is “healthier”
because it seems to accept
the inevitability of (perhaps considerable) nonresponse

There again, differences in opinion, perhaps,
between North America and (northern) Europe

…

I used to belong in the after-people crowd ;

spent much effort on for ex. selecting best x-variables
for non-response adjustment in the estimation

The R/A design movement re-oriented me

To combine “before” and “after” is important,
for the inference perspective

Given my theory background, I continue to look at

the nonresponse dilemma as a problem of statistical inference ,

but one to which to which R/A design has added

an interesting new dimension

…

What results do the R/A movement need to present in the next few years ?

One important question is

Is there *Increased accuracy* in the survey estimates
if adaptive data collection is practiced ?

Answers depend on the environment

- information rich environment
- little information is present

…

Increased accuracy (less bias) in the survey estimates
if *adaptive data collection* is practiced ?

This *question of statistical inference*
can be addressed in one of the various modes :
quasi-design-based, model based and so on

May give contradicting conclusions, not surprisingly
because those modes have never (completely) agreed

more evidence needed

…

We are concerned with sets of units indexed by k

population $U = \{1, 2, ..., k, ..., N\}$ size N

sample $s \subseteq U$ size n

response $r \subseteq s \subseteq U$ size m

$d_k = (\text{known inclusion prob of unit } k)^{-1}$

but response prob unknown

...

Variables entering the discussion :

- **Target variable** y_k observed $k \in r$
- **Auxiliary vector** $\mathbf{x}_k$ known all $k \in s$ *(*or all $k \in U$*)*
- **Response indicator** I_k observed $k \in s$

$$I_k = 1 \text{ if } k \in r \text{ , otherwise } = 0$$

target parameter $Y = \sum_U y_k = N\,\bar{y}_U$

…

Basics: There is bias

For the survey variable y

the response mean $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ (known; biased)

risks to differ (a lot) from

the sample mean $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ (unknown; unbiased)

The difference, in expectation, is ***bias***, in statistical jargon

response rate $P = \sum_s d_k I_k / \sum_s d_k$

Balance : “agreement of *means* of variables” - an old concept

More specifically, “balanced” if
means in a *smaller set* of units
are equal to
corresponding means in a *larger set* that contains the smaller set.

When means differ : ***Imbalance***, more or less severe

…

The imbalance in the study variable y : $\bar{y}_r - \bar{y}_s$

Cannot measure it, but can analyze it

Sample mean : $\bar{y}_S = \sum_s d_k y_k / \sum_s d_k$ (unknown)

Response mean : $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ (computable)

. .

…

Unknown imbalance for the study variable y : $\bar{y}_r - \bar{y}_s$

$\bar{y}_s$ is the benchmark with which $\bar{y}_r$ is compared

because $\hat{N}\bar{y}_s$ unbiased for the total $Y = \sum_U y_k$

We know $\bar{y}_r - \bar{y}_s \neq 0$ but not by how much

Can we assess the deviation ? Influence it ?

In the data collection ? In the estimation ?

- How is imbalance in *y* related to the (observable) imbalance in the **x**-vector

- How does it depend on the relationship between **x** and *y* ?

...

For the auxiliary vector $\mathbf{x}$: $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$

Balanced if $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, otherwise response imbalanced w.r.t. $\mathbf{x}$

The difference is a vector
so make it
scalar by forming a quadratic form in $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$

Vector $\mathbf{x}$, when composed of categorical variables (paradata or register), may have several hundred possible values

Make-up of the auxiliary vector is important for the inference

There is

a monitoring vector $\mathbf{x}_{MV}$ used in adaptive data collection

a calibration vector $\mathbf{x}_{CAL}$ used in weight computation

They can contain *x*-variables known

- for the sample *s* (paradata and/or register variables)
- for the population *U* (usually register variables)

A number of cases and choices arise for the statistical inference

…

Make-up of the auxiliary vector

Consider here

one realistic case : $\mathbf{x}_{MV} = \mathbf{x}_{CAL} = \mathbf{x}$

where **x** consists entirely of x-variables known at the sample level

(paradata and/or register)

…

Then decomposition of the imbalance in survey variable y :

$$\bar{y}_r - \bar{y}_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r + (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$$

Why is this of interest ? Answer :

It highlights two undesirable differences :

Due to imbalance in the auxiliary : $\bar{\mathbf{x}}_r \neq \bar{\mathbf{x}}_s$

and

Due to inconsistent regression : $\mathbf{b}_r \neq \mathbf{b}_s$

Decomposition of the imbalance in *y* :

$$\bar{y}_r - \bar{y}_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r + (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$$

The first term on right hand side
we can clearly reduce by adaptive design
by making the two **x**-means come closer

The second term is the critical one;
does adaptive design reduce it ?

The critical term : $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s = (\hat{\bar{Y}}_{CAL} - \hat{\bar{Y}}_{FUL}) / \hat{N}$

Problem: The regression in the sample does not hold for the response .

Inconsistent regression

$\mathbf{b}_r \neq \mathbf{b}_s$ is the always the case

Lin. regr. coeff. vectors, y on $\mathbf{x}$:

$\mathbf{b}_s = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k \mathbf{x}_k y_k)$ for the sample s; unknown

$\mathbf{b}_r = (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_r d_k \mathbf{x}_k y_k)$ computable for response r

I expected R/A design to be beneficial

- improved accuracy, less bias -

Found this was fulfilled, but to a degree only

The critical term: $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s = (\hat{Y}_{CAL} - \hat{Y}_{FUL}) / \hat{N}$

is reduced, but to a modest extent, by reduced imbalance

(Theoretical and empirical results available; not shown here)

…

Imbalance of response r with respect to specified auxiliary vector $\mathbf{x}$:

$$IMB(r, \mathbf{x}|s) = P^2(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)$$

Simple descriptive measure contrasting r with s

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

$$\boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k \quad \text{weighting matrix}$$

Compact notation: IMB

.

IMB is a small number : For any *r*, *s* and **x**-vector,

$$0 \leq IMB \leq P(1-P) \leq 0.25$$

To reduce *IMB* as much as possible

- a desirable goal in data collection -

is equivalent to reducing the explanation of the response indicator

as much as possible, to a small (near-zero) value

Concluding remarks

I welcome the R/A movement: In the face of high nonresponse

it is an attempt to put survey statistics “back into the fold”

of “a good set of observed units”

not necessarily to “get” the whole designated set of units

(the whole selected probability sample)

That may be out of reach

Concluding remarks

but use R/A design to get at least a “good” (representative, balanced) picture of the probability sample or of the population

This may be considerably better than if no efforts had been made in data collection to direct & adapt it

As I see it, the pursuit of R/A designs entails:

a trust in ideas that have been

- and apparently still are -

proven and central to survey statistics :

A “good” set of responding units

(those who deliver the survey variable *y*)

Those were objectives in sampling theory

already in the 1930’s and 1940’s

…

What is the future, the prospects of R/A design

What results does the R/A movement need to present
in the next few years
to keep the movement viable and interesting?

…

A challenge for the R/A design movement

How can we succeed in making the ideas of R/A design “penetrate”,
convince the grass roots of a survey organization -
- conservative, as national statistical agencies tend to be -
and to make the agencies implement and use these ideas,

as if they were as natural as “stratification”
and other “self-evident” concepts.

A challenge for the R/A design movement

A beginning: Make the survey organization

"start thinking in these terms"

Not even that modest ambition is simple . A lot of effort remains.

Big easy data

Some ask : Why worry about realizing, at considerable cost perhaps,
a set of respondents - out of a fixed (probability) sample -

that has representativeness, or balance,

by the ideas of the R/A movement

when in the future "representativeness" and "balance"

may be replaced by other notions, take other expressions.

A major (and well respected) national statistical agency : Statistics Sweden.

My views coloured by long affiliation in the past

with Statistics Canada and Statistics Sweden.

Features:

- Statistics Sweden is a national statistical agency, a component in a decentralized Swedish statistical system.
- Methodology know-how dispersed among different agencies, each of them responsible for their own statistics production.
- Lots of auxiliary variable available by national register of population and the other principal registers that can be matched, using Sweden's unique personal identifier (10 digit number) as key
- Reliance on this extensive source of auxiliary information

Theoretical results on the deviation of CAL

$$\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\,\Delta_r \quad \text{where} \quad \boldsymbol{\Delta}_r = (\mathbf{b}_r - \mathbf{b}_s)'\mathbf{x}_s$$

The approx. variance is the sum of

a constant term $\left(\frac{1}{m} - \frac{1}{n}\right) S^2_{y,gr}$

and a penalty term $\frac{IMB}{p^2}\,\frac{S^2_{y,gr}}{m}$

Theoretical results on the deviation of CAL

$$\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\ \Delta_r \quad \text{where} \quad \Delta_r = (\mathbf{b}_r - \mathbf{b}_s)'\mathbf{x}_s$$

Mean and variance of Δ_r over sets r with fixed mean $\overline{\mathbf{x}}_r$ (fixed ***IMB***)

under conditions:

- s self-weighting sample from U ; n from N,
- $\mathbf{x}$ a mutually exclusive group indicator

Response rate $p = m/n$

$$\overline{\Delta} = \text{mean}(\ \Delta_r \,|\overline{\mathbf{x}}_r, m, s\ ) = 0$$

$$S_\Delta^2 = \text{var}(\ \Delta_r \,|\overline{\mathbf{x}}_r, m, s\ ) \approx (\frac{1}{m} - \frac{1}{n})S_{y,gr}^2 + \frac{IMB}{p^2}\frac{S_{y,gr}^2}{m}$$

Thank you for listening

…