Calibration of Weights in Surveys with Nonresponse and Frame Imperfections

A course presented at Eustat
Bilbao, Basque Country
January 26-27, 2009
by
Sixten Lundström and Carl-Erik Särndal
Statistics Sweden
http://www.scb.se/statistik/publications/UY9999_2003032_IB_X97%2535%2536%2531.pdf

1_1
Introduction
Welcome to this course

with the title:

*Calibration of Weights in Surveys with Nonresponse and Frame Imperfections*

---

The title of the course suggests two objectives:

- To study *calibration* as a general method for estimation in surveys; this approach has attracted considerable attention in recent years.

- A focus on problems caused by *nonresponse*: bias in the estimates, and how to reduce it.
Key concepts

Finite population $U$:

$N$ objects (elements) : persons, 
or farms, or business firms, or ... 

Sample $s$:
A subset of the elements in $U$ : $s \subseteq U$

Sampling design:
How to select a sample $s$ from $U$
or, more precisely, from the list
of the elements in $U$ (the frame population)

Key concepts

Probability sampling:
Every element in the population has
a non-zero probability
of being selected for the sample

In this course we assume that
probability sampling is used.
There is a well-defined survey objective. For ex., information needed about employment: How many unemployed persons are there in the population?

Study variable: $y$

- with value $y_k = 1$ if $k$ unemployed
- $y_k = 0$ if $k$ not unemployed

‘Unemployed’ is a well-defined concept (ILO)

Number of unemployed to be estimated:

$$\sum_{k=1}^{N} y_k = \sum_{k \in U} y_k$$

Key concepts

A survey often has many study variables ($y$-variables).

- **Categorical** study variables:
  
  Frequently in surveys of individuals and households (number of persons by category)

- **Continuous** study variables:
  
  Frequently in business surveys (monetary amounts)
Key concepts

There may exist *other variables* whose values are known and can be used to improve the estimation. They are called *auxiliary variables*.

*Calibration* is a systematic approach to the use of auxiliary information.

---

Key concepts

Auxiliary variables play an important role
- in the sampling design (e.g., stratification)
- in the estimation (by calibration)

In this course we discuss only how aux. information is used in the estimation.
Key concepts

*Ideal survey conditions*:
- The only error is sampling error.
- All units selected for the sample provide the desired information (no *nonresponse*).
- They respond correctly and truthfully (no *measurement error*).
- The frame population agrees with the target population (no *frame imperfections*).

This course

*Ideal conditions*:
They do not exist in the real world. But they are a starting point for theory.

Session 1_4 of this course discuss uses of aux. information under ideal conditions.
Objective: Unbiased estimation; small variance.
This course

Nonresponse (abbreviated NR):
All of those selected for the sample
do not response, or they respond to
part of the questionnaire only

A troubling feature of surveys today:
NR rates are very high.
‘Classical survey theory’ did not need
to pay much attention to NR.

This course

Most of this course - Sessions 1_5 to 2_6 -
is devoted to the situation:

\textit{sampling error and NR error}

Objective:
Describe approaches to estimation;
Reduce as much as possible
both bias (due to NR) and variance
This course

In the concluding Session 2.7 we add another complication:

*Frame imperfections*: The target population is not identical to the frame population

Not discussed in the course:

*Measurement error*: Some of the answers provided are wrong

---

Research on NR in recent years

Two directions:

- **Preventing** NR from occurring (methods from behavioural sciences) - We do not discuss this

- **Dealing with** (‘adjusting for’) NR once it has occurred (mathematical and statistical sciences); the subject of this course.
Categories of NR

- *Item NR*: The selected element responds to some but not all questions on the questionnaire
- *Unit NR*: The selected element does not respond at all; among the reasons:
  - refusal, not-at-home, and others

Basic considerations for this course

- NR is a *normal, but undesirable feature* of essentially all sample surveys today
- NR causes *bias* in the estimates
- We must still make the best possible estimates
- Bias is never completely eliminated, but we strive to reduce it as far as possible
- Small variance no consolation, because $(bias)^2$ can be the dominating part of MSE
Why is NR such a serious problem?

The intuitive understanding: Those who happen to respond are often not ‘representative’ for the population for which we wish to make inferences (estimates).

The result is bias: Data on the study variable(s) available only for those who respond. The estimates computed on these data are often systematically wrong (biased), but we cannot (completely) eliminate that bias.

Consequences of NR

- $(\text{bias})^2$ can be the larger part of MSE
- NR increases survey cost; follow-up is expensive
- NR will increase the variance, because fewer than desired will respond. But this can be compensated by anticipating the NR rate and allowing ‘extra sample size’
- Increased variance often a minor problem, compared with the bias.
**Treatment of NR**

- NR may be treated by *imputation* primarily the *item NR*; not discussed in this course.

- NR may be treated by (adjustment) *weighting* primarily the *unit NR*; it is the main topic in this course.

Neither type of treatment will resolve the real problem, which is bias.

**Starting points**

- Adjustment methods never completely eliminate the NR bias for a given study variable. This holds for the methods in this course, and for any other method.

- NR bias may be *small for some* of the usually many study variables, but *large for others*; unfortunately, we have no way of knowing.
Comments, questions

• The course is theoretical, but has a very practical background
• Different countries have very different conditions for sampling design and estimation. The Scandinavian countries have access to many kinds of registers, providing extensive sources of auxiliary data.
• We are curious: What are the survey conditions in your country?
• What do you consider to be ‘high NR’ in your country?

Literature on nonresponse

• little was said in early books on survey sampling (Cochran and other books from the 1950’s)
• in recent years, a large body of literature, many conferences
• several statistical agencies have paid considerable attention to the problem
Our background and experience for work on NR methodology

  http://www.scb.se/statistik/publikationer/OV9999_2000I02_BR_X97%e3%96P0103.pdf


Our background


Särndal & Lundström (2009): Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. Submitted for publication
Important earlier works

Olkin, Madow and Rubin (editors):  
*Incomplete data in sample surveys.*  

Groves, Dillman, Eltinge and Little (editors):  
*Survey Nonresponse.*  
New York:  Wiley (2001)

These books examine NR from many different perspectives.

---

A comment

The nature of NR is sometimes described by terms such as

*ignoreable, MAR, MCAR, non-ignoreable*

These distinctions not needed in this course
1.2
Introductory aspects of the course material

Planning a survey

*The process usually starts with a general, sometimes rather vague description of a problem (a need for information)*

The statistician must determine the survey objective as clearly as possible:

- What exactly is the problem?
- Exactly what information is wanted?
Types of fact finding

Options:

• An experiment?

• A survey?

• Other?

The statistician’s formulation

must specify:

• the finite population and the subpopulations (domains) for which information is required

• the variables to be measured and the parameters to be estimated
The target population ($U$)

$$\sum_{k} y_k$$

Domain ($U_q$)

Parameters:

$$Y = \sum_{k} y_k$$

$$Y_q = \sum_{k} y_k \quad \text{where} \quad q = 1, \ldots, Q$$

$$\psi = f(Y_1, \ldots, Y_m, \ldots, Y_M)$$

Aspects of the survey design that need to be considered:

- Data collection method
- Questionnaire design and pretesting
- Procedures for minimizing response errors
- Selection and training of interviewers
- Techniques for handling nonresponse
- Procedures for tabulation and analysis
No survey is perfect in all regards!

**Sampling errors (examined)**

**Nonsampling errors**

• Errors due to non-observation
  
  Undercoverage (examined)
  Nonresponse (examined)

• Errors in observations
  
  Measurement
  Data processing
Sampling error and nonresponse error

Target population ($U$)

Sample set ($s$)

Response set ($r$)

A simple experiment to illustrate sampling error and nonresponse error

Parameter to estimate: The proportion, in %, of elements with a given property:

$$P = \frac{100}{N} \sum_{U} y_k$$

where

$$y_k = \begin{cases} 1 & \text{if element $k$ has the property} \\ 0 & \text{otherwise} \end{cases}$$

Let us assume $P = 50$
Sampling design: SI, \( n \) from \( N \)

Assume no auxiliary information available

Estimator of \( P \) if full response:

\[
\hat{P} = \frac{100}{n} \sum_s y_k
\]

Estimator of \( P \) if \( m \) out of \( n \) respond:

\[
\hat{P}_{NR} = \frac{100}{m} \sum_r y_k
\]

Let us study what happens if the response distribution is as follows, where \( \theta_k = \Pr(k \text{ responds}) \):

\[
\theta_k = \begin{cases} 
0.5 & \text{if element } k \text{ has the property} \\
0.9 & \text{otherwise}
\end{cases}
\]

Note: The response is directly related to the property under estimation.

100 repeated realizations \((s, r)\)
\( n = 30 \)

**Full-response**

**Nonresponse**

\( n = 300 \)

**Full-response**

**Nonresponse**
Comments

- In practice, we never know the response probabilities. To be able to study the effect of nonresponse, assumptions about response probabilities are necessary.

- Increasing the sample size will not reduce the nonresponse bias. As a matter of fact, the proportion of MSE due to the bias will increase with increasing sample size, as we now shall show.

We consider response distributions of the type:

\[ \theta_k = \begin{cases} 
\theta^\ast & \text{if element } k \text{ has the property} \\
0.9 & \text{otherwise}
\end{cases} \]

Consider four such response distributions:

(1) \( \theta^\ast = 0.5 \);  (2) \( \theta^\ast = 0.85 \);
(3) \( \theta^\ast = 0.88 \);  (4) \( \theta^\ast = 0.89 \);
100 repeated realizations \((s, r)\); for each of these, we compute

\[ \hat{P}_{NR} = \frac{100}{m} \sum_r y_k \]

then compute

the proportion of MSE due to squared bias:

\[ \text{RelB}^2 = 100 \times \frac{\text{Bias}^2}{\text{MSE}} \]

where

\[ \text{MSE} = \text{Var} + \text{Bias}^2 \]

<table>
<thead>
<tr>
<th>(\theta^*)</th>
<th>(n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>30</td>
</tr>
<tr>
<td>0.85</td>
<td>300</td>
</tr>
<tr>
<td>0.88</td>
<td>1000</td>
</tr>
<tr>
<td>0.89</td>
<td>2000</td>
</tr>
</tbody>
</table>

\(\text{RelB}^2\) for different sample sizes and resp. distrib.
The proportion of MSE due to squared bias…

(i) increases with increasing sample size

(ii) is rather high for large sample sizes even when the difference between the response probabilities for elements with the property and elements without the property is small.

The high proportion will cause the confidence interval to be invalid, as we now show.

The usual 95% confidence interval would be computed as

\[ \hat{P}_{NR} \pm 1.96 \sqrt{\frac{\hat{P}_{NR}(100 - \hat{P}_{NR})}{m}} \]

Problem: The coverage rate does not reach 95% when there is NR.
Coverage rate (%) for different sample sizes for the response distribution with:

$$\theta_k = \begin{cases} 
0.85 & \text{if element } k \text{ has the property} \\
0.9 & \text{otherwise}
\end{cases}$$

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>30</th>
<th>300</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>93.2</td>
<td>92.6</td>
<td>87.1</td>
<td>77.9</td>
<td></td>
</tr>
</tbody>
</table>

Sampling, nonresponse and undercoverage error

Frame population

Target population

"Persisters"

Overcoverage

Undercoverage
Different sets

$R$: Target population elements with complete or partiell response

$NR$: Target population elements with no or inadequate response

$O$: Elements in the sample which we do not know if they belong to the target population or the overcoverage

$\phi$: Elements in the sample which belong to the overcoverage

Different sets (contin.)

$C$: Target population elements with complete response

$NC$: Target population elements with partiell response
Breakdown of the sample size $n$

\[ n = n_R + n_{NR} + n_O + n_{\phi} \]

Unweighted response rate =

\[ \text{Unweighted response rate} = \frac{n_R}{n_R + n_{NR} + u \times n_O} \]

where $u$ is the rate of $O$ that belongs to the nonresponse.
Weighted response rate =

\[ \frac{\sum_R d_k}{\sum_R d_k + \sum_{NR} d_k + u \sum_O d_k} \]

NR is an increasingly serious problem. It must always be taken into account in the estimation.

We illustrate this by some evidence.
The Swedish Labour Force Survey - Time series of the nonresponse rate

Nonresponse analysis in the Survey on Life and Health

<table>
<thead>
<tr>
<th>Age group</th>
<th>18-34</th>
<th>35-49</th>
<th>50-64</th>
<th>65-79</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate(%)</td>
<td>54.9</td>
<td>61.0</td>
<td>72.5</td>
<td>78.2</td>
</tr>
<tr>
<td>Country of birth</td>
<td>Nordic countries</td>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>------------------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response rate (%)</td>
<td>66.7</td>
<td>50.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Income class (in thousands of SEK)</th>
<th>0-149</th>
<th>150-299</th>
<th>300-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>60.8</td>
<td>70.0</td>
<td>70.2</td>
</tr>
<tr>
<td>Marital status</td>
<td>Married</td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Response rate (%)</td>
<td>72.7</td>
<td>58.7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education level</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>63.7</td>
<td>65.4</td>
<td>75.6</td>
</tr>
</tbody>
</table>
**International experience**

*Lower response rate for:*
- Metropolitan residents
- Single people
- Members of childless households
- Young people
- Divorced / widowed people
- People with lower educational attainment
- Self-employed people
- Persons of foreign origin

---

**This course will show:**

*Use of (the best possible) auxiliary information* will reduce

- the nonresponse bias
- the variance
- the coverage errors
Survey response in your organization

Trends in survey response rates? Increasing?

What are some typical response rates? In the Labour Force Survey for ex.? Reason for concern?

Have measures been introduced to increase survey response?

Have measures been introduced to improve estimation? By more efficient use of auxiliary information, or by other means?
Some response rates

The Swedish Household Budget Survey

1958 86 %
2005 52 %

The Swedish Labour Force Survey

1970 97 %
2005 81 %
The use of auxiliary information under ideal survey conditions

Review: Basic theory for complete response

Important concepts in *design-based estimation* for finite populations:

- Horvitz-Thompson (HT) estimator
- Generalized Regression (GREG) estimator
- Calibration estimator
The progression of ideas

Unbiased estimators for common designs (1930’s and on). Cochran (1953) and other important books of the 1950’s:
• stratified simple random sampling (STSI)
• cluster & two-stage sampling

Horvitz-Thompson (HT) estimator (1952):
arbitrary sampling design; the idea of individual inclusion prob’s

The progression of ideas

**GREG estimator** (1970’s):
arbitrary *auxiliary vector* for model assisted estimation

**Calibration estimator** (1990’s):
identify powerful *information* ; use it to compute weights for estimation (with or without NR)
Concurrently, development of *computerized tools* : CLAN97, Bascula, Calmar, others
Basic theory for complete response

Population $U$ of elements $k = 1, 2, \ldots, N$

Sample $s$ (subset of $U$)

Non-sampled (non-observed): $U - s$

Complete response: all those sampled are also observed (their $y$-values recorded)

<table>
<thead>
<tr>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finite population $U = {1, 2, \ldots, k, \ldots, N}$</td>
</tr>
<tr>
<td>Sample from $U$</td>
</tr>
<tr>
<td>Sampling design $p(s)$</td>
</tr>
<tr>
<td>Inclusion prob. of $k$</td>
</tr>
<tr>
<td>Design weight of $k$</td>
</tr>
<tr>
<td>Joint incl. prob. of $k$ and $\ell$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$\pi_k$</td>
</tr>
<tr>
<td>$d_k = 1/\pi_k$</td>
</tr>
<tr>
<td>$\pi_k\ell$</td>
</tr>
</tbody>
</table>
Notation

Study variable \( y \)

Its value for element \( k \) \( y_k \)

We want to estimate \( \sum_U y_k \)

Usually, a survey has many \( y \) – variables
Can be categorical or continuous

Notation

**Domain** = Sub-population

A typical domain : \( U_q \)

It is a subset of \( U : U_q \subseteq U \)

Domain total to estimate : \( \sum_{U_q} y_k \)
**Notation**

Domain-specific $y$-variable $y_q$

Its value for element $k$ $y_{qk}$

$y_{qk} = y_k$ in domain, $y_{qk} = 0$ outside

Domain total to estimate: $\sum_{U_q} y_k = \sum_{U} y_{qk}$

for ex.: total of disposable income (the variable) in single-member households (the domain)

---

**The approach to estimation**

must handle a variety of practical circumstances

A typical survey has many $y$-variables:
One for every socio-economic concept
One for every domain of interest (every new domain adds a new $y$-variable)
A $y$-variable is often both categorical (“zero-one”) and domain-specific (= 0 outside domain).
For ex.: Unemployed (variable) among persons living alone (domain).
Even though the survey has many $y$-variables, we can focus on one of them and on the estimation of its unknown population total

$$Y = \sum_U y_k$$

**HT estimator**  
for complete response:

$$\hat{Y}_{HT} = \sum_S d_k y_k$$

Design weight of $k$: $d_k = 1/\pi_k$

Auxiliary information not used at the estimation stage
HT estimator
for complete response:

Variance
\[ V(\hat{Y}_{HT}) = \sum\sum U F_{k\ell} y_k y_\ell \]
\[ F_{k\ell} = \frac{d_k d_\ell}{d_{k\ell}} - 1 \] for \( \ell \neq k \)
\[ d_{k\ell} = \frac{1}{\pi_{k\ell}} ; \]
\[ F_{kk} = d_k - 1 \]

For ex., for SI sampling, we have \( \hat{Y}_{HT} = N\bar{y}_s \)
and
\[ V(N\bar{y}_s) = N^2 \left( \frac{1}{n} - \frac{1}{N} \right) S_y^2 U \]

HT estimation
for complete response:

The variance estimator
\[ \hat{V}(\hat{Y}_{HT}) = \sum\sum S d_{k\ell} F_{k\ell} y_k y_\ell \]
It has familiar expressions for ‘the usual designs’.

For STSI, with \( n_h \) from \( N_h \) in stratum \( h \)
\[ \hat{Y}_{HT} = \sum_{h=1}^{H} N_h \bar{y}_{sh} \]
with estimated variance
\[ \sum_{h=1}^{H} N_h^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) S_{ysh}^2 \]
**Auxiliary vector**

denoted \( \mathbf{x} \); its dimension may be large

Its value for element \( k \): \( \mathbf{x}_k \)

To qualify as auxiliary vector,
*must know more* than just \( \mathbf{x}_k \) for \( k \in s \)

For example, know \( \mathbf{x}_k \) for \( k \in U \)

Or know the total \( \sum_U \mathbf{x}_k \)

---

**GREG estimator of** \( Y = \sum_U y_k \) (1980’s)

\[
\hat{Y}_{GREG} = \sum_s d_k y_k + (\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k)' \mathbf{B}_{s;d}
\]

HT est. of \( Y \) + regression adjustment;
an estimator of 0

\( \mathbf{B}_{s;d} \) is a regression vector,
computed on the sample data
**GREG estimator; alternative expression**

\[ \hat{Y}_{GREG} = \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k) \]

Population sum of predicted values
Sample sum of weighted residuals

\[ \hat{y}_k = \mathbf{x}_k' \mathbf{B}_{s; \ell} \]

computable for \( k \in U \)

---

The **auxiliary information** for GREG is:

\[ \sum_U \mathbf{x}_k = \text{pop. total of aux. vector} \]

Examples:

- A continuous \( x \)-variable
  \[ \mathbf{x}_k = (1, x_k)' \Rightarrow \sum_U \mathbf{x}_k = (N, \sum_U x_k)' \]

- A classification of the elements
  \[ \mathbf{x}_k = (0, \ldots, 1, \ldots, 0)' \Rightarrow \sum_U \mathbf{x}_k = (N_1, \ldots, N_j, \ldots, N_J)' \]
\( \hat{Y}_{GREG} \) contains
the estimated regression vector

\[
\mathbf{B}_{s;d} = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k \mathbf{x}_k \mathbf{y}_k)
\]

matrix to invert \( \times \) column vector
is a (nearly unbiased) estimator of its population counterpart:

\[
\mathbf{B}_U = (\sum_U \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U \mathbf{x}_k \mathbf{y}_k)
\]

**System of notation**
for means, regression coefficients, etc.

First index: *the set of elements* that defines the quantity ("the computation set")
then *semi-colon*, then
Second index: *the weighting* used in the quantity.

Examples:

\[
\bar{y}_{s;d} = \frac{\sum_s d_k y_k}{\sum_s d_k}
\]
weighted sample mean

\[
\mathbf{B}_{s;d} = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k \mathbf{x}_k \mathbf{y}_k)
\]
If the need arises to be even more explicit:

\[ 
B_{(y:x)S;d} = \left( \sum_S d_k x_k' x_k \right)^{-1} \left( \sum_S d_k x_k y_k \right) 
\]

Regression of \( y \) on \( x \), computed over the sample \( S \) with the weighting \( d_k = 1/\pi_k \)

---

**System of notation**

Absence of the second index means:
the weighting is uniform ("unweighted").

Examples:

\[ 
\bar{y}_U = \frac{1}{N} \sum_U y_k 
\]

unweighted population mean

\[ 
B_U = \left( \sum_U x_k' x_k \right)^{-1} \left( \sum_U x_k y_k \right) 
\]

(unweighted regr. vector)
**Estimators as weighted sums**

**HT estimator:**

\[ \hat{Y}_{HT} = \sum_s d_k y_k \]

The weight of \( k \) is \( d_k = 1 / \pi_k \)

---

**Estimators as weighted sums**

**GREG estimator as a weighted sum:**

\[ \hat{Y}_{GREG} = \sum_s d_k g_k y_k \]

The weight of element \( k \) is

\[ d_k g_k = \text{design weight} \times \text{adjustment factor based on the auxiliary info.} \]
The GREG estimator

gives element $k$ the weight $d_k g_k$

where

$$d_k = 1/ \pi_k$$

$$g_k = 1 + \lambda'_s x_k$$

$$\lambda'_s = (\sum_U x_k - \sum_s d_k x_k)'(\sum_s d_k x_k x'_k)^{-1}$$

GREG estimator; computation

$$\hat{Y}_{GREG} = \sum_s d_k g_k y_k$$

1. Matrix inversion $(\sum_s d_k x_k x'_k)^{-1}$

2. Compute

$$\lambda'_s = (\sum_U x_k - \sum_s d_k x_k)'(\sum_s d_k x_k x'_k)^{-1}$$

3. Compute $g_k = 1 + \lambda'_s x_k$

4. Finally compute $d_k g_k$

Several software exists for this.
Comment

Matrix inversion is part of the weight computation

\[ \lambda' = (\sum_U x_k - \sum_S d_k x_k)'(\sum_S d_k x_k x_k')^{-1} \]

row vector matrix inversion

GREG estimator

\[ \hat{Y}_{GREG} = \sum_S d_k g_k y_k \]

Property of the weights :

\[ \sum_S d_k g_k x_k = \sum_U x_k \, (\text{known total}) \]

They are calibrated to the known information
Bias of GREG: is very small, already for modest sample sizes

Bias/stand. dev. is of order \( n^{-1/2} \)

Bias decreases faster than the stand.dev. For practical purposes we can forget the bias (assuming full response).

Variance estimation for GREG:
Well known since the 1980's

Comment
Weights of the form \( d_k (1 + \lambda' x_k) \)
will be seen often in the following:
the design weight multiplied by an adjustment factor of the form

\[ 1 + \lambda' x_k \]
Note:
When we examine estimation for NR, (Sessions 1_5 and following), the weights will again have the form design weight $\times$ adjustment factor

but then the estimators will be biased, more or less, depending on the strength of the auxiliary vector

Auxiliary information: An example

For every $k$ in $U$, suppose known:

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

- The value $x_k$ of a continuous variable $x$
  e.g., $x_k =$ income of $k$

Many aux. vectors can be formulated to transmit some or all of this total information. Let us consider 5 of these vectors.
<table>
<thead>
<tr>
<th>Vector $x_k$</th>
<th>Info $\sum U x_k$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_k$</td>
<td>$\sum U x_k$</td>
<td>total population income</td>
</tr>
<tr>
<td>$(1, x_k)'$</td>
<td>$(N, \sum U x_k)'$</td>
<td>population size and total population income</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vector</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(0, x_k, 0, 0, 0, 0)'$</td>
<td>$(\sum_{U_{11}} x_k, ..., \sum_{U_{23}} x_k)'$</td>
</tr>
<tr>
<td>$(0, 0, 0, 0, 0, x_k, 0, 0, 0, 0)'$</td>
<td>$(N_{11}, ..., N_{23}, \sum_{U_{11}} x_k, ..., \sum_{U_{23}} x_k)'$</td>
</tr>
<tr>
<td>$(1, 0, 0, x_k, 0)'$</td>
<td>$(N_1, N_2, \sum_{U_1} x_k, \sum_{U_2} x_k, \sum_{U_3} x_k)'$</td>
</tr>
</tbody>
</table>
For each of the five formulated vectors, 

\[ \hat{Y}_{GREG} = \sum_s d_k g_k y_k \]

will have a certain mathematical form:

Five different expressions, but all of them are special cases of the general formula for \( g_k \).

(No need to give them individual names - they are just special cases of \textit{one estimator} namely GREG)

For example, with the aux. vector \( x_k = (1, x_k)' \)

\[ \hat{Y}_{GREG} = \sum_s d_k g_k y_k \]

takes the form that \textit{the old literature} calls the \textit{(simple) regression estimator},

\[ \hat{Y}_{GREG} = N \left\{ \bar{y}_s; d + (\bar{x}_U - \bar{x}_s; d) B_s; d \right\} \]

In modern language: It is \textit{the GREG estimator for the aux vector} \( x_k = (1, x_k)' \)
1.5
Introduction to estimation in surveys with nonresponse

Target population ($U$)

Sample set ($s$)

Response set ($r$)
Our objective: To estimate \( Y = \sum U y_k \) with an estimator denoted \( \hat{Y}_{NR} \) representing either

\[
\hat{Y}_W = \sum_r w_k y_k \quad \text{(weighting only)}
\]

\[
\hat{Y}_{IW} = \sum_r w_k y_{ik} \quad \text{(imputation followed by weighting)}
\]

Imputation followed by weighting

A typical survey has many \( y \)-variables, indexed \( i = 1, \ldots, I \).

Response set for variable \( i \) : \( r_i \)

Response set for the survey: The set of elements having responded to \textit{at least one item} : \( r \)
Imputation followed by weighting

\[ \hat{Y}_{IW} = \sum_r w_k y_{\bullet k} \]

where

\[ y_{\bullet k} = \begin{cases} y_k & \text{for } k \in r_i \\ \hat{y}_k & \text{for } k \in r - r_i \end{cases} \]

Imputation for item NR: The imputed value \( \hat{y}_k \) takes the place of the missing value \( y_k \).

Components of error

\[ \hat{Y}_{NR} - Y = (\hat{Y} - Y) + (\hat{Y}_{NR} - \hat{Y}) \]

Total error = Sampling error + NR error

\( \hat{Y} \) is the estimator of \( Y \) that would be used under complete response (\( r = s \))

\( \hat{Y}_{NR} \) is the “NR-estimator” for \( Y \)
Two phases of selection

1. $s$ is selected from $U$.

2. Given $s$, $r$ is realised as a subset from $s$.

The two probability distributions are

- $p(s)$ (known)
- $q(r|s)$ (unknown)

Both are taken into account in evaluating bias and variance.

We use the conditional argument:

For expected value:

$$E_{pq}(\cdot) = E_p[E_q(\cdot|s)]$$

For variance:

$$V_{pq}(\cdot) = V_p[E_q(\cdot|s)] + E_p[V_q(\cdot|s)]$$
The basic statistical properties of $\hat{Y}_{NR}$

The bias:

$$B_{pq}(\hat{Y}_{NR}) = E_{pq}(\hat{Y}_{NR}) - Y$$

The accuracy, measured by MSE:

$$MSE_{pq}(\hat{Y}_{NR}) = V_{pq}(\hat{Y}_{NR}) + \left(B_{pq}(\hat{Y}_{NR})\right)^2$$

The bias will be carefully studied in this course. It has two components

$$B_{pq}(\hat{Y}_{NR}) = E_{pq}(\hat{Y}_{NR}) - Y$$

$$= [E_p(\hat{Y}) - Y] + [E_{pq}(\hat{Y}_{NR} - \hat{Y})]$$

$$= B_{SAM} + B_{NR}$$

sampling bias + NR bias

$B_{SAM}$ is zero (for HT) or negligible (for GREG)
The variance

By definition

\[ V_{pq}(\hat{Y}_{NR}) = E_{pq}(\hat{Y}_{NR} - E_{pq}(\hat{Y}_{NR}))^2 \]

It can be decomposed into two components

\[ V_{pq}(\hat{Y}_{NR}) = V_{SAM} + V_{NR} \]

sampling variance + NR variance

The sampling variance component:

\[ V_{SAM} = V_p(\hat{Y}) = E_p[(\hat{Y} - E_p(\hat{Y}))^2] \]

depends only on the sampling design \( p(s) \)

For ex., under SRS, if the full response estimator is \( \hat{Y} = N \bar{y}_S \) then the well-known expression

\[ V_{SAM} = N^2 \left( \frac{1}{n} - \frac{1}{N} \right) s^2_y U \]
The NR variance component is more complex:

\[ V_{NR} = E_p V_q (\hat{Y}_{NR} | s) + V_p (B_{NR} | s) + 2Cov_p (\hat{Y}, B_{NR} | s) \]

where

\[ B_{NR} | s = E_q (\hat{Y}_{NR} - \hat{Y}) | s \] (conditional NR bias)

Add the squared bias to arrive at the

the measure of accuracy:

\[ MSE_{pq}(\hat{Y}_{NR}) = V_p (\hat{Y}) + E_p V_q (\hat{Y}_{NR} | s) + E_p (B_{NR}^2 | s) + 2Cov_p (\hat{Y}, B_{NR} | s) + 2B_{SAM} B_{NR} + (B_{SAM})^2 \]

\( B_{SAM} \) is negligible, and if \( Cov \) term small, then

\[ MSE_{pq}(\hat{Y}_{NR}) \approx V_p (\hat{Y}) + E_p V_q (\hat{Y}_{NR} | s) + E_p (B_{NR}^2 | s) \]
The accuracy has two parts:

\[
MSE_{pq}(\hat{Y}_{NR}) \approx V_p(\hat{Y}) + E_p V_q(\hat{Y}_{NR}|s) + E_p(B_{NR|s}^2) \\
\text{due to sampling} \quad \text{due to NR}
\]

The main problem with NR:

The term involving the bias, \( E_p(B_{NR|s}^2) \), can be a very large component of MSE.
1.6
Weighting of data. Types of auxiliary information. The calibration approach.

Structure
Target population $U$

Sample $s$

Response set $r$
### Notation and terminology

**Population** $U$
- of elements $k = 1, 2, \ldots, N$

**Sample** $s$ (subset of $U$)
- Non-sampled: $U - s$

**Response set** $r$ (subset of $s$)
- Sampled but non-responding: $s - r$
- $U \supseteq s \supseteq r$

### The objective

- remains to estimate the total $Y = \sum_{U} y_k$

In practice, many $y$-totals and functions of $y$-totals.
- But we can focus here on one total.
- No need at this point to distinguish *item* $NR$ and *unit* $NR$.
- *Perfect coverage* assumed.
The response set $r$ is the set for which we observe $y_k$

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

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

where $r \subseteq s \subseteq U$

Nonresponse means that $r \subseteq s$

Full response means that $r = s$
with probability one

Two phases of selection

Phase one: Sample selection
with known sampling design

Phase two: Response selection
with unknown response mechanism
Phase one: *Sample selection*

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

**Known inclusion prob. of** \( k \): \( \pi_k \)

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

---

Phase two: *Response selection*

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

**Unknown response prob. of** \( k \): \( \theta_k \)

**Unknown response influence of** \( k \): \( \phi_k = 1/\theta_k \)
A note on terminology

\[ d_k = \frac{1}{\pi_k} \quad \text{computable weight} \]

\[ \phi_k = \frac{1}{\theta_k} \quad \text{unknown; not a weight, called influence} \]

Sample weighting combined with response weighting

Desired (but impossible) combined weighting:

\[ d_k \times \phi_k = \frac{1}{\pi_k} \times \frac{1}{\theta_k} \]

known unknown
Desirable nonresponse weighting

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

Cannot be computed, because unknown influences \( \phi_k = 1/\theta_k \)

We present the calibration approach.

But first we look at a more traditional approach.

Most estimators in the traditional approach are special cases of the calibration approach.
Traditional approach: The principal idea is to derive estimates $\hat{\theta}_k$ of the unknown response prob. $\theta_k$.

Then use these estimates in constructing the estimator of the total $\hat{Y}$.

An often used form of this approach:

Starting from $\hat{Y} = \sum_r d_k \frac{1}{\theta_k} y_k$

replace $1/\theta_k$ by $1/\hat{\theta}_k$

We get $\hat{Y} = \sum_r d_k \frac{1}{\hat{\theta}_k} y_k$

sampling weight  NR adjustment weight
A large literature exists about this type of estimator:

$$\hat{Y} = \sum_r d_k \frac{1}{\hat{\theta}_k} y_k$$

Estimation of $\theta_k$ done with the aid of a response model:

- response homogeneity group (RHG:s)
- logistic

The term response propensity is sometimes used.

The idea behind response homogeneity groups (RHG:s)

The elements in the sample (and in the response set) can be divided into groups. Everyone in the same group responds with the same probability, but these probabilities can vary considerably between the groups.
Example: STSI sampling
RHG's coinciding with strata
(each stratum assumed to be an RHG)

\[ d_k \frac{1}{\hat{\theta}_k} = \frac{N_h}{n_h} \frac{m_h}{m_h} = \frac{N_h}{m_h} \]

\[ \hat{Y} = \sum_{h=1}^{H} \frac{N_h}{n_h} \frac{m_h}{m_h} \sum_{r_h} y_{hk} = \sum_{h=1}^{H} N_h \bar{y}_{rh} \]

The procedure is convenient but oversimplifies the problem. It is a special case of the calibration approach.

A variation of the traditional approach
Start with 2-phase GREG estimator

\[ \hat{Y} = \sum_r d_k \frac{1}{\hat{\theta}_k} g_{0k} y_{rk} \]

After estimation of the response prob, we get

\[ \hat{Y} = \sum_r d_k \frac{1}{\hat{\theta}_k} g_{0k} y_{rk} \]
A general method for estimation in the presence of NR should

• be easy to understand
• cover many survey situations
• offer a systematic way to incorporate auxiliary information
• be computationally easy
• be suitable for statistics production (in NSI:s)

One can maintain that the calibration approach satisfies these requirements. There is an extensive literature since 1990.
Steps in the calibration approach

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

**Pedagogical note**

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

Generality is at the price of a certain level of abstraction.

The formulation uses linear algebra. Knowledge of regression theory is helpful.
Why can we not use the design weights $d_k = 1/\pi_k$ without any further adjustment?

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

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

We must expand the design weights.

**Information**

may exist at the population level

at the sample level
Levels of information

Distinguish:

• **Information at the population level.** Such info, taken from population registers, is particularly prevalent and important in Scandinavia, The Netherlands, and increasingly elsewhere in Europe.

• **Information at the sample level.** Such info may be present in any sample survey.
Levels of information

Notation: Two types of auxiliary vector

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

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

Auxiliary vector, population level

Two common situations:

- \( \mathbf{x}_k^* \) *known value* for every \( k \) in \( U \)
  
  (given in the frame, or coming from admin.reg.

- the total \( \mathbf{X}^* = \sum_{U} \mathbf{x}_k^* \) is *imported*
  
  from accurate outside source

\( \mathbf{x}_k^* \) need not be known for every \( k \)
Sources of variables for the star vector $X_k^*$

- the existing frame
- by matching with other registers

Examples of variables for the star vector:
For persons: age, sex, address, income
To related persons: Example, in survey of school children, get (by matching) variables for parents

**Auxiliary vector, sample level**

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

Hence we can compute and use

$$\hat{X}^{\circ} = \sum_s d_k X_k^{\circ}$$

It is *unbiased information*, not damaged by NR
Examples of variables for the moon vector $X_k^\circ$

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

The information statement

- Specifies the information at hand; totals or estimated totals
- May refer to either level: Population level, sample level
- It is not a model statement
Information is something we know; it provides input for the calibration approach.

(By contrast, a model is something you do not know, but venture to assume.)

Statement of auxiliary information
sampling, then nonresponse

<table>
<thead>
<tr>
<th>Set of units</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population $U$</td>
<td>$\sum_U x_k^*$ known</td>
</tr>
<tr>
<td>Sample $S$</td>
<td>$x_k^\circ$ known, $k \in s$</td>
</tr>
<tr>
<td>Response set $r$</td>
<td>$x_k^*$ and $x_k^\circ$ known, $k \in r$</td>
</tr>
</tbody>
</table>
• The auxiliary vector
  General notation : $x_k$

• The information available about that vector
  General notation : $X$

Three special cases :
  • population info only
  • sample info only
  • both types of info
• population info only

\[ x_k = x_k^* ; \quad X = \sum_U x_k^* \] (known total)

• sample info only

\[ x_k^\circ = x_k^\circ ; \quad X = \sum_s d_k x_k^\circ \] (unbiasedly estimated total)

• both types of info

\[
x_k = \begin{pmatrix} x_k^* \\ x_k^\circ \end{pmatrix} ; \quad X = \begin{pmatrix} \sum_U x_k^* \\ \sum_s d_k x_k^\circ \end{pmatrix}
\]

Example:

\[ x_k = (0, \ldots, 1, \ldots, 0, 0, \ldots, 1, \ldots, 0)' \]

identifies age/sex group for \( k \in U \)

identifies interviewer for \( k \in s \)
For **the study variable** \( y \)

we know (we have observed):

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

Missing values:

\[ y_k \quad \text{for} \quad k \in s - r \]

The **calibration estimator** is of the form

\[ \hat{Y}_W = \sum_r w_k y_k \]

with \( w_k = d_k v_k \)

where \( d_k = 1/\pi_k \), and the factor \( v_k \) serves to

- expand the design weight \( d_k \) for unit \( k \)
- incorporate the auxiliary information
- reduce as far as possible bias due to NR
- reduce the variance
Note: We want $v_k > 1$ for all (or nearly all) $k \in r$, in order to compensate for the elements lost by NR.

Primary interest:
Examine the (remaining) bias in $\hat{Y}_W = \sum_r w_k y_k$
attempt to reduce it further.
Recepie: Seek better and better auxiliary vectors for the calibration!
(Sessions 2_3, 2_4, 2_5)

Secondary interest (but also important):
Examine the variance of $\hat{Y}_W$
find methods to estimate it.
Mathematically, the adjustment factor \( v_k \) can be determined by different criteria, for example

- \( v_k = 1 + \lambda' x_k \) linear in the aux. vector
- \( v_k = \exp(\lambda' x_k) \) exponential

Determine first \( \lambda \)

(implicitly or by numeric methods)

---

**Linear adjustment factor**

\( v_k \) is determined to satisfy:

(i) \( v_k = 1 + \lambda' x_k \) linearity

and

(ii) \( \sum_r d_k w_k x_k = X \) calibration to the given information \( X \)

Now determine \( \lambda \)
From (i) and (ii) follow
\[
\lambda' = \lambda'_r = \left( X - \sum_r d_k x_k \right) \left( \sum_r d_k x_k x_k' \right)^{-1}
\]
assuming the matrix non-singular.
Then the desired calibrated weights are
\[
w_k = d_k v_k = d_k \left( 1 + \lambda'_r x_k \right)
\]

Computational note:
Possibility of *negative weights*:
\[
d_k v_k = d_k \left( 1 + \lambda' x_k \right)
\]
with
\[
\lambda' = \left( X - \sum_r d_k x_k \right) \left( \sum_r d_k x_k x_k' \right)^{-1}
\]
can be negative. It does happen, but rarely.
Computational note:

The vector

\[
\left( \mathbf{x} - \sum_{r} d_k \mathbf{x}_k \right) \left( \sum_{r} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1}
\]

is not near zero, as it was for the GREG estimator (in the absence of NR).

**Properties of the calibrated weights**

\[
w_k = d_k \left( 1 + \lambda'_r \mathbf{x}_k \right)
\]

1. They **expand**: 
   \[ w_k > d_k \quad \text{all } k, \text{ or almost all} \]

2. \[ \sum_{r} w_k = N = \text{population size} \]
   
   under a simple condition
Note: if both types of information, then

\[ x_k = \left( \begin{array}{c} x_k^* \\ \hat{x}_k \\ x_k \end{array} \right) \]

and the information input is

\[ X = \left( \begin{array}{c} \sum_U x_k^* \\ \sum_x d_k x_k^* \end{array} \right) \]

When both types of information present, it is also possible to \textbf{calibrate in two steps}:

First on the sample information; gives intermediate weights.

Then in step two, the intermediate weights are calibrated, using also the population information, to obtain the final weights \( w_k \).
Consistency

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

If \( x_k \) is known for \( k \in s \), the statistical agency can sum over \( s \) and publish the unbiased estimate
\[
\hat{X} = \sum_s d_k x_k
\]

Users often require that this estimate coincide with the estimate obtained by summing over \( r \) using the calibrated weights:
\[
\hat{X}_W = \sum_r w_k x_k
\]

Calibration makes this consistency possible

Almost all of our aux. vectors are of the form:

There exists a constant vector \( \mu \) such that
\[
\mu' x_k = 1 \quad \text{for all } k
\]

For example, if \( x_k = (1, x_k)' \), then \( \mu = (1,0)' \).
When $x_k$ is such that $\mu'x_k = 1$ for all $k$

then the weights simplify:

$$w_k = d_k v_k = d_k \{X' \left( \sum_r d_k x_k x_k' \right)^{-1} x_k \}$$

where

$$X = \begin{pmatrix} \sum_U x_k^* \\ \sum_S d_k x_k^o \end{pmatrix}$$

is the information input

---

A summary of this session: We have

- discussed two types of **auxiliary information**

- introduced the idea of a weighting (of responding elements) that is **calibrated** to the given information

- hinted that calibrated weighting gives **consistency**, and that it often leads to both reduced NR bias and reduced variance. More about this later.
1_7
Comments on the calibration approach

The calibration approach

Some features:

- Generality (any sampling design, and auxiliary vector)
- ”Conventional techniques” are special cases
- Computational feasibility (software exists)
The calibration approach brings generality

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

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

Another feature of the calibration estimator:
Perfect estimates under certain condition

Consider the case where
\[ x_k = x_k^* \quad \text{and} \quad X = X^* = \sum U x_k^* \]

Assume that \[ y_k = (x_k^*)' \beta^* \] holds for every \( k \in U \) (perfect linear regression), then
\[ \hat{Y}_W = \sum U y_k = Y \]

No sampling error, no NR-bias!
Recall: We have specified the weights as

\[ w_k = d_k \nu_k \quad ; \quad \nu_k = 1 + \lambda'_r x_k \]

where

\[ \lambda'_r = \left( \mathbf{X} - \sum_r d_k x_k \right) \left( \sum_r d_k x_k x_k' \right)^{-1} \]

They satisfy the calibration equation

\[ \sum_r w_k x_k = \mathbf{X} \]

But they are not unique: They are not the only ones that satisfy the calibration equation.

In fact, for a given \( x_k \)-vector with given information input \( \mathbf{X} \), there exist many sets of weights that satisfy the calibration equation

\[ \sum_r w_k x_k = \mathbf{X} \]

In other words, “calibrated weights” is not a unique concept.

Let us examine this.
The calibration procedure takes certain initial weights and transforms them into (final) calibrated weights.

The initial weights can be specified in more than one way.

Consider the weights \( w_k = d_{\alpha k} v_k \)

where \( v_k = 1 + \lambda'_r z_k \)

\[ \lambda'_r = \left( X - \sum_r d_{\alpha k} x_k \right) \left( \sum_r d_{\alpha k} z_k x'_k \right)^{-1} \]

\( d_{\alpha k} \) is an initial weight

\( z_k \) is an instrument vector (may be \( \neq x_k \))

These \( w_k \) satisfy the calibration equation

\[ \sum_r w_k x_k = \sum U x_k \]

for any choice of \( d_{\alpha k} \) and \( z_k \)

(as long as the matrix can be inverted)
The ”natural choices”

\[ d_{\alpha k} = d_k = 1 / \pi_k \quad \text{and} \quad z_k = x_k \]

are used most of the time and will be called the standard specifications.

---

**An important type of z-vector**

There exists a constant vector \( \mu \) not dependent on \( k \) such that

\[ \mu'z_k = 1 \quad \text{for all} \quad k \in U \]

When \( z_k = x_k \), this condition reads:

\[ \mu'x_k = 1 \quad \text{for all} \quad k \in U \]

Almost all of our x-vectors are of this type.
Different initial weights may produce the same calibrated weights

When the \( z \)-vector satisfies \( \mu'z_k = 1 \) for all \( k \)
then
\[
d_{\alpha k} = d_k
\]
and
\[
d_{\alpha k} = C d_k
\]
give the same calibrated weights

Example

- SI sampling; \( n \) from \( N \)
- \( z_k = x_k = x_k^* = 1 \)

Then the initial weights
\[
d_{\alpha k} = d_k = \frac{N}{n}
\]
and
\[
d_{\alpha k} = d_k \frac{n}{m} = \frac{N}{m}
\]
give the same calibrated weights, namely,
\[
w_k = \frac{N}{m}
\]
Invariant calibrated weights are also obtained in the following situation:

- STSI with strata $U_p$; $n_p$ from $N_p$; $p = 1, \ldots, P$
- $z_k = x_k = x_k^*$ = stratum identifier

Then the initial weights

$$d_{\alpha k} = d_k = \frac{N_p}{n_p}$$

and

$$d_{\alpha k} = d_k \times (\frac{n_p}{m_p}) = \frac{N_p}{m_p}$$

give the same calibrated weights, namely

$$w_k = \frac{N_p}{m_p}$$

Usually the components of $z_k$ are functions of the $x$-variables.

For example, if $x_k = (x_{1k}, x_{2k})'$

we get calibrated weights by taking

$$z_k = (\sqrt{x_{1k}}, \sqrt{x_{2k}})'$$
The well-known **Ratio (RA) estimator** is obtained by the specifications

\[ x_k = x_k^* = z_k = 1 \]

**Note**: Non-standard specifications!

They give

\[ \hat{Y}_W = \sum_U x_k \times \frac{\sum_r d_k y_k}{\sum_r d_k x_k} \]

---

**A perspective on the weights**: We can write the calibrated weight as the sum of two components

\[ w_k = w_{Mk} + w_{Rk} \]

\[ = \text{“Main term” + “Remainder”} \]

with

\[ w_{Mk} = d_{ak} \left\{ X' \left( \sum_r d_{ak} z_k x'_k \right)^{-1} z_k \right\} \]

\[ w_{Rk} = d_{ak} \left\{ 1 - \left( \sum_r d_{ak} x_k \right)^{\prime} \left( \sum_r d_{ak} z_k x'_k \right)^{-1} z_k \right\} \]
We can specify a constant vector $\mu$ not dependent on $k$ such that $\mu'z_k = 1$ for all $k$.

Then $w_{Rk} = 0$ and $w_k = w_{Mk}$

(An example: $z_k = x_k = (1, x_k)'$ and $\mu = (1, 0)'$)

When $w_{Rk} = 0$, the calibrated weights have simplified form

$$w_k = w_{Mk} = d_{ck} \left\{ X' \left( \sum_r d_{rk} z_k x_k' \right)^{-1} z_k \right\}$$

Under the standard specifications:

$$w_k = w_{Mk} = d_k \left\{ X' \left( \sum_r d_k x_k x_k' \right)^{-1} x_k \right\}$$
Agreement with the GREG estimator

If \( r = s \) (complete response), and if
\[
x_k = x_k^* \quad \text{and} \quad z_k = c \ x_k^*
\]
for any positive constant \( c \), then

the calibration estimator and
the GREG estimator
can be shown to be identical.
Traditional estimators as special cases of the calibration approach

The family of calibration estimators includes many ‘traditional estimator formulas’

Let us look at some examples.

The standard specification

\[ d_{\alpha k} = d_k \quad \text{and} \quad Z_k = X_k \]

is used (unless otherwise stated).
An advantage of the calibration approach:

We need not any more think in terms of ‘traditional estimators’ with specific names.

All of the following examples are special cases of the calibration approach, corresponding to simple formulations of the auxiliary vector \( \mathbf{x}_k \)

The simplest auxiliary vector

\[
\mathbf{x}_k = \mathbf{x}_k^* = 1 \quad \text{for all} \quad k
\]

The corresponding information is weak:

\[
\sum_U \mathbf{x}_k = \sum_U 1 = N
\]

Calibrated weights (by the general formula):

\[
w_k = d_k \times \frac{N}{\sum_r d_k}
\]

\[
\hat{Y}_W = N \bar{y}_r; d = N \frac{\sum_r d_k y_k}{\sum_r d_k} = \hat{Y}_{\text{EXP}}
\]

known as the **Expansion estimator**
The simplest auxiliary vector

\[ x_k = x_k^* = 1 \]

In particular, for SI (\( n \) sampled from \( N \);
\( m \) respondents):

\[ w_k = \frac{N}{n} \frac{n}{m} = \frac{N}{m} \]

- weakness the aux. vector \( x_k = 1 \) recognizes no differences among elements
- bias usually large
One can show, for any sampling design,

\[
\text{bias}(\hat{Y}_{\text{EXP}}) / N \approx \bar{y}_{U;\theta} - \bar{y}_U
\]

Note the difference between two means:

The **theta-weighted mean**

\[
\bar{y}_{U;\theta} = \frac{\sum U \theta_k y_k}{\sum U \theta_k}
\]

The **unweighted mean**

\[
\bar{y}_U = \frac{\sum U y_k}{N}
\]

When \( y \) and \( \theta \) are highly correlated, that difference can be very large (more about this later).

---

Comment on the Expansion Estimator

Despite an often large nonresponse bias, the **expansion estimator** is (surprisingly enough) often used by practitioners and researchers in social science.

This practice, which has developed in some disciplines, cannot be recommended.
The classification vector ("gamma vector")

Elements classified into $P$ dummy-coded groups

$$\gamma_k = (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{Pk})'$$

$$= (0, \ldots, 1, \ldots, 0)'$$

The only entry ‘1’ identifies the group (out of $P$ possible ones) to which element $k$ belongs.

The classification vector

Typical examples:

- Age groups
- Age groups by sex (complete crossing)
- Complete crossing of $>2$ groupings
- Groups formed by intervals of a continuous $x$-variable
The classification vector
as a star vector

\[ x_k = x_k^* = \gamma_k = (0, \ldots, 1, \ldots, 0)' \]

The associated information:

The vector of population class frequencies

\[ \sum_U x_k^* = (N_1, \ldots, N_p, \ldots, N_P) \]

Calibrated weights (by the general formula):

\[ w_k = d_k \times \frac{N_p}{\sum_{r_p} d_k} \text{ for all } k \text{ in group } p \]

---

The classification vector
as a star vector:

\[ x_k = x_k^* = \gamma_k \]

The calibration estimator takes the form

\[ \hat{y}_W = \sum_{p=1}^{P} N_p \bar{y}_{r_p} d = \hat{y}_{PWA} \]

known as the

Population Weighting Adjustment estimator
**Population Weighting Adjustment estimator**

A closer look:

\[
\hat{Y}_{PWA} = \sum_{p=1}^{P} N_p \bar{y}_{r_p,d}
\]

with

\[
\bar{y}_{r_p,d} = \frac{\sum_{r_p} d_k y_k}{\sum_{r_p} d_k} = \text{weighted group } y\text{-mean for respondents}
\]

\(N_p\) = known group count in the population

---

**The classification vector**

as a moon vector

\(x_k = x_k^\circ = \gamma_k = (0,\ldots,1,\ldots,0)'

Information for calibration:

the unbiasedly *estimated* class counts

\(\hat{N}_p = \sum_{s_p} d_k, \quad p = 1,2,\ldots,P\)

The general formula gives the weights

\[
w_k = d_k \times \frac{\sum_{s_p} d_k}{\sum_{r_p} d_k} \quad \text{for all } k \text{ in group } p
\]
The classification vector

as a moon vector: \( x_k = x^\circ_k = \gamma_k \)

In particular for SI sampling:

\[
w_k = \frac{N}{n} \frac{n_p}{m_p} \quad \text{for all } k \text{ in group } p.
\]

Sampling weight

NR adjustment

by inverse of

group response rate

The classification vector

as a moon vector

\[ x_k = x^\circ_k = \gamma_k = (0,\ldots,1,\ldots,0)' \]

\[
\hat{y}_W = \sum_{p=1}^{P} \hat{N}_p \bar{y}_{p;d} = \hat{Y}_{WC}
\]

known as

Weighting Class estimator
Weighting Class estimator

\[ \hat{Y}_{WC} = \sum_{p=1}^{P} \hat{N}_p \tilde{y}_{r,p} ; d \]

Class sizes not known but estimated: \( \hat{N}_p = \sum_{s} d_k \)

\[ \bar{y}_{r,p} ; d = \frac{\sum_{r_p} d_k y_k}{\sum_{r_p} d_k} = \text{weighted group } y\text{-mean for respondents} \]

A continuous \( x\)-variable

for example, \( x_k = \text{income} \); \( y_k = \text{expenditure} \)

Two vector formulations are of interest:

- \( x_k = x_k^* = x_k \) \quad \text{Info: } \sum_U x_k = \sum_U x_k \)
- \( x_k = x_k^* = (1, x_k)' \) \quad \text{Info: } \sum_U x_k = (N, \sum_U x_k)' \)
The Ratio Estimator

is obtained by formulating

\[ x_k = x_k^* = x_k \quad \text{and} \quad z_k = 1 \quad \text{(non-standard!)} \]

weights \( w_k = d_k \times \frac{\sum_U x_k}{\sum_r d_k x_k} \)

calibration estimator \( \hat{y}_W = (\sum_U x_k) \frac{\sum_r d_k y_k}{\sum_r d_k x_k} = \hat{y}_{RA} \)

**Not very efficient** for controlling bias.
A better use of the \( x \)-variable:
create size groups or “include an intercept”

---

The (simple) Regression Estimator

A better use of the \( x \)-variable:

\[ x_k = x_k^* = (1, x_k)' = z_k \]

calibrated weights given by:

\[ d_{kv_k} = d_k \times N \left( \frac{1}{\sum_r d_k} + \frac{\bar{x}_U - \bar{x}_{r; d}}{\sum_r d_k (x_k - \bar{x}_{r; d})^2} (x_k - \bar{x}_{r; d}) \right) \]

The calibration estimator takes the form

\[ \hat{y}_W = N \{ \bar{y}_{r; d} + (\bar{x}_U - \bar{x}_{r; d}) B_{r; d} \} = \hat{y}_{REG} \]

regression coefficient
The (simple) Regression Estimator

A closer look:

\[ \hat{Y}_{REG} = N \left[ \bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d}) B_{r;d} \right] \]

with

\[ \bar{x}_{r;d} = \frac{\sum_r d_k x_k}{\sum_r d_k} \]

\[ \bar{y}_{r;d} \text{ analogous } y\text{-mean} \]

\[ B_{r;d} = \frac{\sum_r d_k (x_k - \bar{x}_{r;d})(y_k - \bar{y}_{r;d})}{\sum_r d_k (x_k - \bar{x}_{r;d})^2} \]

regression of \( y \) on \( x \)

Combining

a classification and a continuous \( x \)-variable

Information about \( \textit{both} \)

(i) the \( \text{classification vector} \)

\[ \gamma_k = (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{pk})' \]

and

(ii) a \( \text{continuous variable} \) with value \( \chi_k \)
Known group totals for a continuous variable

The vector formulation:
\[ x_k = x_k^* = (\gamma_{1k} x_k, \ldots, \gamma_{pk} x_k, \ldots, \gamma_{pk} x_k)' = x_k \gamma_k \]

Information for \( p = 1, \ldots, P \):
\[ \sum_{U_p} x_k \]
\[ z_k = \gamma_k = (0, \ldots, 1, \ldots, 0)' \quad \text{(not standard)} \]

gives the **SEPRAG** (separate regression) estimator

---

Known group counts and group totals for a continuous variable

The vector formulation:
\[ x_k = x_k^* = (\gamma_k', x_k \gamma_k)' = z_k \]
\[ (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{pk}, x_k \gamma_{1k}, \ldots, x_k \gamma_{pk}, \ldots, x_k \gamma_{pk})' \]

Information for \( p = 1, \ldots, P \):
\[ N_p \quad \text{and} \quad \sum_{U_p} x_k \]

gives the **SEPREG** (separate regression) estimator
The Separate Regression Estimator

\[ \hat{Y}_W = \sum_{p=1}^{P} N_p \left( \bar{y}_{r_p; d} + \left( \bar{x}_{U_p} - \bar{x}_{r_p; d} \right) \bar{B}_{r_p; d} \right) = \hat{Y}_{SEPREG} \]

Marginal counts for a two-way classification

- \( P \) groups for classification 1 (say, age by sex)
- \( H \) groups for classification 2 (say, profession)

\[ x_k = x_k^* = \]
\[ = (\gamma_1 k, \ldots, \gamma_P k, \delta_1 k, \ldots, \delta_{H-1} k, \delta_H -1, k)' \]
\[ = (0, \ldots, 1, \ldots, 0, 0, \ldots, 1, \ldots, 0)' \]

Calibration on the \( P + H - 1 \) marginal counts.

Note: \( H - 1 \)

Gives the two-way classification estimator
List of ‘traditional estimators’

(We shall refer to them later.)

<table>
<thead>
<tr>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansion (EXP)</td>
</tr>
<tr>
<td>Weighting Class (WC)</td>
</tr>
<tr>
<td>Population Weighting Adjustment (PWA)</td>
</tr>
<tr>
<td>Ratio (RA)</td>
</tr>
<tr>
<td>Regression (REG)</td>
</tr>
<tr>
<td>Separate Ratio (SEPRA)</td>
</tr>
<tr>
<td>Separate Regression (SEPREG)</td>
</tr>
<tr>
<td>Two-Way Classification (TWOWAY)</td>
</tr>
</tbody>
</table>

Comment: No need to give individual names to the traditional estimators. All are calibration estimators.

For example, although known earlier as ‘regression estimator’,

\[
\hat{Y}_{REG} = N\left(\bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d}) B_{r;d}\right)
\]

is now completely described as the calibration estimator for the vector \( x_k = x_k^* = (1, x_k)' \)
Your set of course materials contains an appendix with a number of exercises.

Some of these ask you to formulate (verbally) your response to a given practical situation, others require an algebraic derivation.
You are encouraged to consider these exercises, now during the course, or after the course.

Exercises 1 and 2 reflect practical situations that survey statisticians are likely to encounter in their work. Think about the (verbal) answers you would give.
2.1 Calibration with combined use of sample information and population information

Different levels of auxiliary information

Population level $X_k^*$
Sample level $X_k^o$
Recall the traditional approach

Find estimates $\hat{\theta}_k$ of the unknown response prob. $\theta_k$

Then form

$$\hat{Y} = \sum_r d_k \frac{1}{\hat{\theta}_k} y_k$$

If population totals are available, there may be a second step: Use $d_k/\hat{\theta}_k$ as starting weights; get final weights by calibrating to the known population totals

Alternative traditional approach

Start from 2-phase GREG estimator

$$\hat{Y} = \sum_r d_k \frac{1}{\bar{\theta}_k} g_{\theta_k} y_k$$

After estimation of the response prob, we get

$$\hat{Y} = \sum_r d_k \frac{1}{\hat{\theta}_k} g_{\hat{\theta}_k} y_k$$
The first step in traditional approaches:

The idea: Adjust for nonresponse by model fitting

An explicit model is formulated, with the $\theta_k$ as unknown parameters.

The model is fitted, $\hat{\theta}_k$ is obtained as an estimate of $\theta_k$, and $1/\hat{\theta}_k$

is used as a weight adjustment to $d_k$

Ex. Logistic regression fitting

Frequently used: Subgrouping

The sample $s$ is split into a number of subgroups (Response homogeneity groups)

The inverse of the response fraction within a group is used as a weight adjustment to $d_k$
The traditional approach often gives
the same result as the calibration approach

We return to the calibration estimator

\[ \hat{Y}_W = \sum_r w_k y_k \]

Let us consider alternatives
for computing the \( w_k \)

Single-step or two-step may be used.

We recommend single-step, as follows:

Initial weights: \( d_{\alpha k} = d_k \)

Auxiliary vector: \( x_k = \begin{pmatrix} x_k^* \\ \circ \ x_k \end{pmatrix} \)

Calibration equation: \( \sum_r w_k x_k = \begin{pmatrix} \sum U x_k^* \\ \sum_s d_k x_k^\circ \end{pmatrix} \)

Then compute the \( w_k \)
Two variations of two-step:

Two-step A

and

Two-step B

**Two-step A**

Step 1:

Initial weights: \( d_k \)

Auxiliary vector: \( x_k = x_k^\circ \)

Calibration equation: \( \sum_r w_k^\circ x_k^\circ = \sum_s d_k x_k^\circ \)
**Two-step A (cont.)**

**Step 2:**

Initial weights: \( W_k^o \)

Auxiliary vector: \( x_k = \begin{pmatrix} x_k^* \\ x_k^o \end{pmatrix} \)

Calibration equation: \( \sum_r w_{2,Ak} x_k = \begin{pmatrix} \sum_U x_k^* \\ \sum_s d_k x_k^o \end{pmatrix} \)

---

**Two-step B**

**Step 1:**

Initial weights: \( d_k \)

Auxiliary vector: \( x_k = x_k^o \)

Calibration equation: \( \sum_r w_k^o x_k^o = \sum_s d_k x_k^o \)
Two-step B (cont.)

Step 2:

Initial weights: \( w_k^o \)

Auxiliary vector: \( \mathbf{x}_k = \mathbf{x}_k^* \)

Calibration equation: \( \sum_r w_{2k} \mathbf{x}_k^* = \sum U \mathbf{x}_k^* \)

Here no calibration to the sample information
\( \sum_s d_k \mathbf{x}_k^o \)

An example of calibration with information at both levels

Sample level: \( \mathbf{x}_k^o = \mathbf{\gamma}_k = (\gamma_{1k}, \ldots, P_k, \ldots, \gamma P_k)' \)

(classification for \( k \in s \))

Population level: \( \mathbf{x}_k^* = (1, \mathbf{x}_k)' \)

\( (x_k \text{ a continuous variable with known population total}) \)
Single-step

Initial weights: \( d_{\alpha k} = d_k \)

Auxiliary vector: \( x_k = \begin{pmatrix} x_k \\ \gamma_k \end{pmatrix} \)

Calibration equation: \( \sum_r w_k x_k = \left( \frac{\sum_U x_k}{\sum_s d_k \gamma_k} \right) \)

Two-step A

Step 1:

Initial weights: \( d_k \)

Auxiliary vector: \( x_k^\circ = \gamma_k \)

Calibration equation: \( \sum_r w_k x_k^\circ = \sum_s d_k \gamma_k \)
Two-step A (cont.)

Step 2:
Initial weights: \( w_k^0 \)
Auxiliary vector: \( x_k = \begin{pmatrix} x_k \\ \gamma_k \end{pmatrix} \)
Calibration equation: \( \sum_r w_{2Ak} x_k = \left( \frac{\sum_U x_k}{\sum_S d_k \gamma_k} \right) \)

Two-step B

Step 1:
Initial weights: \( d_k \)
Auxiliary vector: \( x_k^0 = \gamma_k \)
Calibration equation: \( \sum_r w_k^0 x_k^0 = \sum_S d_k \gamma_k \)
Two-step B (cont.)

Step 2:

Initial weights: \( w_k^\circ \)

Auxiliary vector: \( \mathbf{x}_k^* = (1, x_k)' \)

Calibration equation: \[ \sum_r w_{2Bk} \mathbf{x}_k^* = \left( \begin{array}{c} N \\ \sum_U x_k \end{array} \right) \]

Comments:

In general, Single-step, Two-step A and Two-step B give different weight systems. But we expect the estimators to have minor differences only.

There is no disadvantage in mixing the population information with the sample information. It is important that both sources are allowed to contribute.
The Two-step B procedure resembles the traditional approach, and has been much used in practice.

**Step 1:** Adjust for nonresponse

**Step 2:** Achieve consistency of the weight system and reduce the variance somewhat

But we recommend the Single-step procedure.

---

**Monte Carlo simulation**

10,000 SI samples each of size $n = 300$ drawn from experimental population of size $N = 832$, constructed from actual survey data: Statistics Sweden’s **KYBOK** survey

Elements classified into four administrative groups; sizes: 348, 234, 161, 89
Monte Carlo simulation

**Information**: For every $k \in U$, we know

- membership in one of 4 admin. groups
- the value $x_k$ of a continuous variable
  
  $x = \text{sq.root revenues}$

We can use all or some of the info.

**Study variable**: $y = \text{expenditures}$

---

Monte Carlo simulation measures computed

\[
\text{RelBias} = 100 \left[ \text{Ave}(\hat{Y}_W) - Y \right] / Y
\]

\[
\text{Ave}(\hat{Y}_W) = \frac{1}{10,000} \sum_{j=1}^{10,000} \hat{Y}_W(j) / 10,000
\]

\[
\text{Variance} = \frac{1}{9,999} \sum_{j=1}^{10,000} \left[ \hat{Y}_W(j) - \text{Ave}(\hat{Y}_W) \right]^2 \times 10^{-8}
\]
### Monte Carlo simulation; logit response

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RelBias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP</td>
<td>5.0</td>
<td>69.6</td>
</tr>
<tr>
<td>Single-step</td>
<td>-0.6</td>
<td>9.7</td>
</tr>
<tr>
<td>Two-step A</td>
<td>-0.6</td>
<td>9.8</td>
</tr>
<tr>
<td>Two-step B</td>
<td>-0.8</td>
<td>9.5</td>
</tr>
</tbody>
</table>

### Monte Carlo simulation; increasing exp response

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RelBias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP</td>
<td>9.3</td>
<td>70.1</td>
</tr>
<tr>
<td>Single-Step</td>
<td>-2.4</td>
<td>8.2</td>
</tr>
<tr>
<td>Two-step A</td>
<td>-2.3</td>
<td>8.3</td>
</tr>
<tr>
<td>Two-step B</td>
<td>-3.0</td>
<td>8.0</td>
</tr>
</tbody>
</table>
Our conclusion

In practice there are no rational grounds for selecting another method than the Single-step procedure.
2.2 Analysing the bias remaining in the calibration estimator

Important to try to reduce the bias? Most of us would say YES, OF COURSE.

A (pessimistic) argument for a NO:
There is no satisfactory theoretical solution;
the bias cannot be estimated.
It is always unknown
(because the response probabilities unknown)

The approach that we present not pessimistic.
Important to try to reduce the bias?

Yes. It is true that the bias due to NR cannot be known or estimated.

But we must strive to reduce the bias. We describe methods for this.

Calibration is not a panacea.

No matter how we choose the aux. vector, the calibration estimator (or any other estimator) will always have a remaining bias.

The question becomes: How do we reduce the remaining bias?

Answer: Seek ever better $X^*_k$

We need procedures for this search (Sessions 2_3, 2_4, 2_5)
Improved auxiliary vector
will (usually) lead to
reduced bias , reduced variance

**Interesting quantities are:**

(a) the *mean squared error*

\[ \text{MSE} = (\text{Bias})^2 + \text{Variance} \]

and

(b) *proportion of MSE due to squared bias*

\[ (\text{Bias})^2 / \{(\text{Bias})^2 + \text{Variance}\} \]
A bad situation: bias > stand. dev.

\[ \frac{\text{bias}}{\text{stand. dev. of } \hat{Y}} \]

true value \( Y \)  
mean of \( \hat{Y} \)

Bad situation: squared bias represents a large portion of the MSE

\[ \hat{Y} \pm 1.96 \times \sqrt{\hat{V}(\hat{Y})}, \]

estimated stand.dev.

will almost certainly \textbf{not contain} the unknown value \( Y \) for which we want to state valid 95% confidence limits
We know:

**Variance**
- is often small (and tends to 0)
  - compared to
- **squared bias** (does not tend to 0)

Both **bias** and **variance** are theoretical quantities (expectations), stated in terms of values for the whole finite population

Variance can be estimated, but not the bias.

---

**The bias of the calibration estimator**

- The calibration estimator is not without bias. (Same holds for any other type of estimator.)
- The bias comes (almost entirely) from the NR, *not* from the probability sampling.
- If 100% response, the calibration estimator becomes the (almost) unbiased GREG estimator.
- Both bias and variance of the calibration estimator depend on the strength of the auxiliary vector. Important: Seek powerful auxiliary vector.
The bias of the calibration estimator

Recall the general definition:

\[ \text{bias} = \text{expected value of estimator} - \text{value of parameter under estimation} \]

What is ‘expected value’ in our case?

We assess expected value, bias and variance \textit{jointly} under:

- the known sample selection \( p(s) \)
- the unknown response mechanism \( q(r|s) \)

\[ \text{bias} (\hat{Y}_W) = E_{pq} (\hat{Y}_W) - Y \]

Our assumptions on the unknown \( q(r|s) \) are ‘almost none at all’.
The bias of the calibration estimator

Derivation of the bias is an evaluation in two phases:

\[ \text{bias}(\hat{Y}_W) = E_p\left(E_q(\hat{Y}_W \mid s)\right) - Y \]

Let us evaluate it!

Approximate expression is obtainable for any auxiliary vector and any sampling design.

Before evaluating the bias in a general way (arbitrary sampling design, arbitrary aux. vector), let us consider a simple example.
Example: The simplest auxiliary vector

\[ x_k = x^*_k = 1 \quad \text{for all } k \]

\[ \hat{Y}_{\text{EXP}} = N \bar{y}_{r;d} = N \frac{\sum_r d_k y_k}{\sum_r d_k} \]

Weighted respondent mean, expanded by \( N \)

Recommended exercise:
Use first principles to derive its bias!

We find

\[ \text{bias}(\hat{Y}_{\text{EXP}} / N) \approx \bar{y}_{\theta} - \bar{y}_U \]

\[ \bar{y}_{\theta} = \frac{\sum_U \theta_k y_k}{\sum_U \theta_k} \quad \text{theta-weighted mean} \]

\[ \bar{y}_U = \frac{1}{N} \sum_U y_k \quad \text{simple unweighted mean} \]

Why approximation?
Answer: Exact expression hard to obtain.
It is a close approx.? Yes.
The bias of the expansion estimator

The \textit{theta-weighted population mean} can differ considerably from the \textit{unweighted population mean}, (both of them unknown), so bias can be very large. These means differ considerably when $y$ and $\theta$ have high correlation.

Suppose the correlation between $y$ and $\theta$ is $0.6$. Then simple analysis shows that

\[
\text{bias}(\hat{Y}_{\text{EXP}} / N) \approx 0.6 \times cv(\theta) \times S_{yU}
\]

where

\[
cv(\theta) = S_{\theta U} / \bar{\theta}_U \quad \text{the coeff. of variation of } \theta
\]

\[
S_{yU} \quad \text{the stand. dev. of } y \text{ in } U
\]
If the response probabilities $\theta$ do not vary at all, then

$$cv(\theta) = S_{\theta U} / \bar{\theta}_U = 0$$

and

$$\text{bias}(\hat{Y}_{\text{EXP}} / N) \approx 0$$

As long as all elements have the same response prob. (perhaps considerably < 1), there is no bias.

But suppose

$$cv(\theta) = S_{\theta U} / \bar{\theta}_U = 0.1$$

Then

$$\text{bias}(\hat{Y}_{\text{EXP}} / N) \approx 0.6 \times 0.1 \times S_{yU} = 0.06 S_{yU}$$

This bias may not seem large, but the crucial question is: How serious is it compared with

$$\text{stand.dev}(\hat{Y}_{\text{EXP}} / N)$$
\[ \text{Var}(\hat{Y}_{\text{EXP}} / N) \approx \frac{1}{m} S_{yU}^2 \]

(a crude approximation; SI sampling assumed)

Suppose \( m = 900 \) responding elements

\[ \text{stand.dev}(\hat{Y}_{\text{EXP}} / N) \approx 0.033 S_{yU} \]

compared with :

\[ \text{bias}(\hat{Y}_{\text{EXP}} / N) \approx 0.06 S_{yU} \]

Then

\[ (\text{Bias})^2 / [(\text{Bias})^2 + \text{Variance}] = \]

\[ (0.06)^2 / [(0.06)^2 + (1/900)] = \]

\[ 0.0036 / (0.0036 + 0.0011) = 77 \% \]

Impossible then to make valid inference by confidence interval!
We return to the General calibration estimator

For a specified auxiliary vector $x_k$ with corresponding information $X$, let us evaluate its bias.

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

with

$$w_k = d_k y_k = d_k (1 + \lambda'_r x_k)$$

$$\lambda'_r = (X - \sum_r d_k x_k) (\sum_r d_k x_k x_k')^{-1}$$

matrix inversion

**The Calibration Estimator : Its bias**
Deriving the bias of the calibration estimator requires an evaluation of

\[ \text{bias}(\hat{Y}_W) = E_p(E_q(\hat{Y}_W|s)) - Y \]

This exact bias expression does not tell us much. But it is \textit{closely approximated} by a much more informative quantity called nearbias \( \hat{Y}_w \)

Comments on approximation:
All ‘modern advanced estimators’, GREG and others, are complex (non-linear). We cannot assess the exact variance of GREG, but there is an excellent approximation.

Likewise, for the calibration estimator, we work \textit{not} with the exact expression for bias and variance, but with close approximations.
Derivation of the bias:

Technique: Taylor linearization.
Keep the leading term of the development; for this term, we can evaluate the expected values in question.

\[
\text{Calibration estimator close approximation to its bias}
\]
\[
bias(\hat{Y}_w) \approx \text{nearbias}(\hat{Y}_w)
\]
where
\[
\text{nearbias}(\hat{Y}_w) = - \sum_U (1 - \theta_k) e_{\theta k}
\]
with
\[
e_{\theta k} = y_k - x_k' B_{U;\theta}
\]
\[
B_{U;\theta} = \left(\sum_U \theta_k x_k x_k'\right)^{-1} \sum_U \theta_k x_k y_k
\]
nearbias($\hat{Y}_w$) = $- \sum_U (1 - \theta_k) e_{\theta k}$

is important in the following

It is a general formula, valid for:

• any sampling design
• any aux. vector
• it is a close approximation (verified in simulations)

Comments

• Detailed derivation of nearbias, see the book
• For given auxiliary vector, nearbias is the same for any sampling design, but depends on the (unknown) response prob’s
• nearbias is a function of certain regression residuals (not the usual regression residuals)
• The variance does depend on sampling design
Comments

- The nearbias formula makes no distinction between “star variables” and “moon variables”
- In other words, for bias reduction, an $x$-variable is equally important when it carries info to the pop. level (included in $x_k^*$) as when it carries info only to the sample level (included in $x_k^\circ$)

Surprising conclusion, perhaps.

But for variance, the distinction can be important.

Example: Let $x_k$ be a continuous aux. variable
- Info at population level: $x_k = x_k^* = (1, x_k)'$
  \[ N \text{ and } \sum U x_k \text{ known} \]
  \[ \Rightarrow \hat{Y}_W = \hat{Y}_{REG} = N \{ \bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d})B_{r;d} \} \]
- Info at sample level only: $x_k = x_k^\circ = (1, x_k)'$
  \[ \Rightarrow \hat{N} = \sum_s d_k \text{ and } \sum_s d_k x_k \text{ computable} \]
  \[ \Rightarrow \hat{Y}_W = \hat{N} \{ \bar{y}_{r;d} + (\bar{x}_s;d - \bar{x}_{r;d})B_{r;d} \} \]
  \[ \text{where } \bar{x}_{s;d} = \frac{\sum_s d_k x_k}{\hat{N}} \]

The two estimators differ, but same nearbias.
• Can nearbias be zero? (Would mean that the calibration estimator is almost unbiased.)

   \textit{Answer}: Yes.

• Under what condition(s)?

   \textit{Answer}: There are 2 conditions, each sufficient to give \textbf{nearbias} = 0.

• Can we expect to satisfy these conditions in practice?

   \textit{Answer}: Not completely. We can reduce the bias.

\begin{center}
\textbf{Conditions for nearbias = 0}
\end{center}

In words: \text{nearbias}(\hat{Y}_W) = 0

under either of the following conditions:

Condition 1: The influence \( \phi \) has \textbf{perfect} linear relation to the aux. vector

Condition 2: The study variable \( y \) has \textbf{perfect} linear relation to the aux. vector
Condition 1

nearbias = 0 if the influence $\phi$ has perfect linear relation to the auxiliary vector:

$$\text{nearbias}(\hat{Y}_W) = 0 \quad \text{if, for all } k \text{ in } U,$$

$$\phi_k = \frac{1}{\theta_k} = 1 + \lambda' x_k$$

for some constant vector $\lambda$

Exercise: Show this!

Comments:

1. The requirement $\phi_k = 1 + \lambda' x_k$ must hold for all $k \in U$.
2. It is not a model. (A model is something you assume as a basis for a statistical procedure.) It is a population property.
3. It requires the influence to be linear in $x_k$
4. If it holds, nearbias = 0
Condition 2

nearbias = 0 if the study variable \( y \) has perfect linear relation to the aux. vector.

$$
\text{nearbias} = 0 \quad \text{if, for all } k \in U, \\
y_k = \beta' x_k \\
\quad \text{for some constant vector } \beta
$$

Exercise: Show this!

Note:

$$
y_k = \beta' x_k \quad \text{for all } k \in U
$$

is not a model.

It is a population property saying that nearbias = 0 if the \( y \)-variable has perfect linear relation to the aux. vector.
Example: auxiliary vector $x_k = x_k^* = (1, x_k)'$
gives regression estimator:

$$
\hat{Y}_W = \hat{Y}_{REG} = N\{\bar{y}_r;d + (\bar{x}_U - \bar{x}_r;d)B_r;d\}
$$

nearbias = 0 if:

$$
\phi_k = a + bx_k , \text{ all } k \in U \quad \text{Condition 1}
$$
or if

$$
y_k = \alpha + \beta x_k , \text{ all } k \in U \quad \text{Condition 2}
$$

Comment

We have found that

nearbias ($\hat{Y}_W$) = 0

1. if the influence $\phi$ has perfect linear relation to the aux. vector

2. if the $y$-variable has perfect linear relation to the aux. vector.
Comment

There are many $y$-variables in a survey:

- One for every socio-economic concept measured in the survey
- One for every domain (sub-population) of interest

To have $\text{nearbias} = 0$ for the whole survey requires that every one of the many $y$-variables must have perfect linear relation to the auxiliary vector.

Not easy (or impossible) to fulfill.

Comment

Therefore, the first condition is the more important one

If satisfied, then $\text{nearbias}(\hat{Y}_w) = 0$

for every one of the many $y$-variables
Can the statistician *control* the remaining bias? make nearbias smaller?

Can the bias be controlled?

We would like to *come close* to *one or both* of:

1. the influence \( \phi \) has *perfect* linear relation to the aux. vector
2. every \( y \)-variable of interest has *perfect* linear relation to the aux. vector

We propose *diagnostic tools* (Sessions 2_3, 2_4).
Questions that we shall consider in the following sessions:

What aux. vector should we use?
How do we evaluate different choices of aux. vector?

A comment on auxiliary vectors

Almost all vectors we are interested are of the following type:

It is possible to specify a constant vector $\mu$ such that $\mu' x_k = 1$ for all $k$
Example 1: A continuous $x$-variable

$$x_k = (1, x_k)'$$

Take $\mu = (1, 0)'$

The property is present:

$$\mu' x_k = 1 \times 1 + 0 \times x_k = 1 \text{ for all } k$$

Example 2: The classification vector

$$x_k = \gamma_k = (0, \ldots, 1, \ldots, 0)'$$

Take $\mu = (1, \ldots, 1, \ldots, 1)'$

The property is present:

$$\mu' x_k = 1 \text{ for all } k$$
Equivalent expressions for \textit{nearbias}

for the \textit{x}-vector type $\mathbf{\mu}'x_k = 1$ for all $k$

\[ \text{nearbias}(\hat{Y}_W) = \]

(i) $-\sum_U e_{\theta k}$

(ii) $(\sum_U x_k)'(B_{U;\theta} - B_U)$

(iii) $\sum_U (\theta_k M_k - 1)y_k$

We now comment on (ii); we need (iii) later.

Expression (ii):

\[ \text{nearbias}(\hat{Y}_W) = (\sum_U x_k)'(B_{U;\theta} - B_U) \]

$B_{U;\theta} = (\sum_U \theta_k x_k x_k')^{-1} \sum_U \theta_k x_k y_k$ \text{ weighted}

$B_U = (\sum_U x_k x_k')^{-1} \sum_U x_k y_k$ \text{ unweighted}
This shows nearbias as a function of the difference between two regression coefficients.

Interpretation: NR causes systematic error in the estimated regression relationship (reason: ‘non-random selection’). We would like to estimate the ordinary regression coefficient $B_U$, but because of NR we obtain an estimate of $B_U;\theta$.

nearbias($\hat{Y}_w$) = $(\sum_U x_k)'(B_U;\theta - B_U)$

What is the nearbias under conditions 1 and 2?

Condition 1: $y_k = \beta'x_k$ for all $k \in U$

$\Rightarrow$ $B_U;\theta = B_U$ and nearbias = 0

Condition 2: $\phi_k = \lambda'x_k$ for all $k \in U$

$\Rightarrow$ $(\sum_U x_k)'(B_U;\theta - B_U) = 0$ (show this!) and nearbias = 0
Comment on terminology

We do not need concepts such as

MAR, MCAR, ignorable NR,
non-ignorable NR

In our view: All situations non-ignorable.
Selecting the most relevant auxiliary information

Auxiliary information can be used both at the design stage and at the estimation stage.
Commonly used sampling designs

- Simple random sampling (SI)
- Stratified simple random sampling (STSI)
- Cluster sampling
- Two-stage sampling
- Probability-proportional-to-size

The design stage

Two important steps in building the auxiliary vector:

(i) making an inventory of potential auxiliary variables

(ii) selecting the most suitable of these variables and preparing them for entry into the auxiliary vector

The estimation stage
Inventory of potential auxiliary variables

Example of an extensive data source:
Sweden’s Total population register (TPR):
A complete listing of the population of individuals (around 9 million)

Some of the variables in TPR:
Unique personal identity number, name and address, date of birth, sex, marital status, country of birth and taxable income.

Recall:
If the nonresponse is considerable and not counteracted by effective adjustment then

(i) the squared bias term is likely to dominate the MSE

(ii) the possibilities for valid statistical inference are reduced; valid confidence intervals cannot be computed
Guidelines for the construction of an auxiliary vector

**Principle 1:** The auxiliary vector (or the instrument vector) should explain the inverse response probability, called the response influence

**Principle 2:** The auxiliary vector should explain the main study variables

**Principle 3:** The auxiliary vector should identify the most important domains

**Principle 1 fulfilled:**
The bias of the calibration estimates reduced for *all* study variables

**Principle 2 fulfilled:**
The bias is reduced in the estimates for the main study variables, and the variance is also reduced
Principle 3 fulfilled:

For the main domains, both bias and variance will be reduced

The general formula for the nearbias (Session 2-2) can guide our search for a powerful auxiliary vector. It also answers the question:

When is the nearbias = 0, for a given estimator?

Let us look at some traditional estimators.

Standard specifications assumed, unless otherwise stated.

The $x$-vector is a ’star vector’ in most of these examples
Prospects for zero nearbias with traditional estimators

Expansion estimator:  \( \hat{Y}_{EXP} = N \bar{y}_{r,d} \)

Auxiliary vector: \( x_k = 1 \)

Zero nearbias if

(i) \( \phi_k = a \) for all \( k \in U \)  

or if

(ii) \( y_k = \alpha \) for all \( k \in U \)

Weighting class estimator:  \( \hat{y}_{WC} = \sum_{p=1}^{P} \bar{N}_p \bar{y}_{r,p,d} \)

Population weighting adjustment estimator:

\( \hat{y}_{PWA} = \sum_{p=1}^{P} N_p \bar{y}_{r,p,d} \)

Aux. vector \( x_k = \gamma_k = \) class indicator vector

Moon vector for \( \hat{y}_{WC} \), star vector for \( \hat{y}_{PWA} \)

Zero nearbias if

(i) \( \phi_k = \alpha_p \) for all \( k \in U_p \)  

or if

(ii) \( y_k = \beta_p \) for all \( k \in U_p \)
Ratio estimator: \( \hat{Y}_{RA} = (\sum_{U} x_k) \frac{y_{r;d}}{x_{r;d}} \)

Auxiliary vector: \( x_k = x_k \)

Instrument vector: \( z_k = 1 \)

Zero nearbias if

(i) \( \phi_k = a \) for all \( k \in U \) or if

(ii) \( y_k = \alpha x_k \) for all \( k \in U \)

Separate ratio estimator:

\( \hat{Y}_{SEPRA} = \sum_{p=1}^{P} (\sum_{U_p} x_k) \frac{y_{r_p:d}}{x_{r_p:d}} \)

Auxiliary vector: \( x_k = x_k \gamma_k \)

Instrument vector: \( z_k = \gamma_k \)

Zero nearbias if

(i) \( \phi_k = a_p \) for all \( k \in U_p \) or if

(ii) \( y_k = \alpha_p x_k \) for all \( k \in U_p \)
Regression estimator:

$$\hat{Y}_{\text{REG}} = N\{y_r;d + (x_U - x_r;d)B_r;d\}$$

Auxiliary vector: $$x_k = (1, x_k)'$$

Zero nearbias if

(i) $$\phi_k = a + bx_k$$ or if

(ii) $$y_k = \alpha + \beta x_k$$

Separate regression estimator:

$$\hat{Y}_{\text{SEPREG}} =$$

$$= \sum_{p=1}^{P} N_p \left\{ y_{r_p};d + (x_{U_p} - x_{r_p};d)B_{r_p};d \right\}$$

Auxiliary vector: $$x_k = (\gamma_k', x_k \gamma_k')'$$

Zero nearbias if

(i) $$\phi_k = a_p + b_p x_k$$ or if

(ii) $$y_k = \alpha_p + \beta_p x_k$$ for all $$k \in U_p$$
Two-way estimator:

\( \hat{Y}_{TWOWAY} \) (expression somewhat complicated)

Auxiliary vector: \( \mathbf{x}_k = (\gamma'_k, \delta'_k)' \)

\( \gamma \) indicates classes \( p=1, \ldots, P; \)
\( \delta \) indicates classes \( h=1, \ldots, H \)

Zero nearbias if

(i) \( \phi_k = a_p + b_h \) \quad \text{or if}

(ii) \( y_k = \alpha_p + \beta_h \)

Conclusion:

Best suited for fulfilling Principle 1: SEPREG or TWOWAY

Best suited for fulfilling Principle 2: The same two vectors

Worst: For Principle 1, EXP and RA. But RA is better than EXP for Principle 2.
Monte Carlo simulation

10,000 SI samples
each of size $n = 300$ drawn from
experimental population of size $N = 832$,
constructed from actual survey data:
Statistics Sweden’s KYBOK survey

Elements classified into four administrative
groups; sizes: 348, 234, 161, 89

---

Monte Carlo simulation

**Information:** For every $k \in U$, we know
- membership in one of 4 admin. groups
- the value $x_k$ of a continuous variable
  $x = \text{sq.root revenues}$

  We can use some or all of that info.

**Study variable:** $y = \text{expenditures}$
Monte Carlo simulation

We used two response distributions, called:
(1) Logit
(2) Increasing exponential

Average response prob.: 86% (for both)

Response probability $\theta$ increases with $x$ and with $y$

Corr. between $y$ and $\theta$:
$\approx 0.70$ (logit) ; $\approx 0.55$ (incr. exp.)

Monte Carlo simulation
measures computed

$$\text{RelBias} = 100 \left[ \frac{\text{Ave}(\hat{Y}_W) - Y}{Y} \right]$$

$$\text{Ave}(\hat{Y}_W) = \frac{1}{10,000} \sum_{j=1}^{10,000} \hat{Y}_W(j) / 10,000$$

$$\text{Variance} = \frac{1}{9,999} \sum_{j=1}^{10,000} \left[ \hat{Y}_W(j) - \text{Ave}(\hat{Y}_W) \right]^2 \times 10^{-8}$$
### Monte Carlo simulation; logit response

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RelBias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansion (EXP)</td>
<td>5.0</td>
<td>69.6</td>
</tr>
<tr>
<td>Weighting Class (WC)</td>
<td>2.2</td>
<td>59.4</td>
</tr>
<tr>
<td>Population Weighting Adjustment (PWA)</td>
<td>2.2</td>
<td>37.1</td>
</tr>
<tr>
<td>Ratio (RA)</td>
<td>2.5</td>
<td>27.5</td>
</tr>
<tr>
<td>Regression (REG)</td>
<td>-0.6</td>
<td>9.5</td>
</tr>
<tr>
<td>Separate Ratio (SEPRA)</td>
<td>0.7</td>
<td>11.8</td>
</tr>
<tr>
<td>Separate Regression (SEPREG)</td>
<td>-0.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Two-Way Classification (TWOWAY)</td>
<td>0.5</td>
<td>21.7</td>
</tr>
</tbody>
</table>

### Monte Carlo simulation; increasing exp. response

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RelBias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansion (EXP)</td>
<td>9.3</td>
<td>70.1</td>
</tr>
<tr>
<td>Weighting Class (WC)</td>
<td>5.7</td>
<td>57.7</td>
</tr>
<tr>
<td>Population Weighting Adjustment (PWA)</td>
<td>5.7</td>
<td>36.3</td>
</tr>
<tr>
<td>Ratio (RA)</td>
<td>3.9</td>
<td>26.1</td>
</tr>
<tr>
<td>Regression (REG)</td>
<td>-2.7</td>
<td>8.1</td>
</tr>
<tr>
<td>Separate Ratio (SEPRA)</td>
<td>2.0</td>
<td>11.3</td>
</tr>
<tr>
<td>Separate Regression (SEPREG)</td>
<td>-0.8</td>
<td>7.4</td>
</tr>
<tr>
<td>Two-Way Classification (TWOWAY)</td>
<td>0.5</td>
<td>20.3</td>
</tr>
</tbody>
</table>
What do we learn from the simulations?

Bias ↓ when the auxiliary vector ‘gets better’ (more informative)

Variance also ↓, as expected

For ex., SEPREG clearly uses much more information than EXP or RA

We want to be more precise about ‘informative’. This will follow.

The search for a powerful auxiliary vector

Principle 1

Tool 1.1: Nonresponse analysis

Tool 1.2: Bias indicator $q^2$

Principle 2

Tool 2.1: Analysis of important target variables

Tool 2.2: Indicator $IND2$
**A new indicator** (not yet published)

We have developed a new indicator, denoted $H_1$, which takes into consideration both Principle 1 and Principle 2. $H_1$ is a product of $q^2$ and a factor depending on the relation between the target variable $y$ and the auxiliary vector.

That is, $H_1 = q^2 	imes f(y, x)$

**Some further tools**

- Transforming the auxiliary variables
- Choosing a powerful instrument vector
- Analysing the distribution of the weights (for ex.: any extreme weights?)
### Tool 1.1: Nonresponse analysis

**Example 1:** The Survey on Life and Health  
(postal survey; Statistics Sweden)

<table>
<thead>
<tr>
<th>Age group</th>
<th>18-34</th>
<th>35-49</th>
<th>50-64</th>
<th>65-79</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>54.9</td>
<td>61.0</td>
<td>72.5</td>
<td>78.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country of birth</th>
<th>Nordic countries</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>66.7</td>
<td>50.8</td>
</tr>
<tr>
<td>Income class (in thousands of SEK)</td>
<td>0-149</td>
<td>150-299</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>Response rate (%)</td>
<td>60.8</td>
<td>70.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marital status</th>
<th>Married</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>72.7</td>
<td>58.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education level</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>63.7</td>
<td>65.4</td>
<td>75.6</td>
</tr>
</tbody>
</table>
Conclusions from this nonresponse analysis:

- The response propensities vary quite a lot between groups
- Without any weighting, one expects a disturbingly large nonresponse bias
- Some of the presumptive auxiliary variables are related, for example, income and education level. What is the simultaneous effect? Should both be used or just one?

<table>
<thead>
<tr>
<th>Sex</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>73.1</td>
<td>78.1</td>
</tr>
<tr>
<td>Age group</td>
<td>16-29</td>
<td>30-40</td>
</tr>
<tr>
<td>---------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Response rate (%)</td>
<td>76.8</td>
<td>74.6</td>
</tr>
<tr>
<td>51-65</td>
<td>76.2</td>
<td>76.1</td>
</tr>
<tr>
<td>66-74</td>
<td>76.2</td>
<td>76.1</td>
</tr>
<tr>
<td>75-79</td>
<td>76.2</td>
<td>76.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country of birth</th>
<th>Nordic countries</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>77.7</td>
<td>57.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marital status</th>
<th>Married</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>78.3</td>
<td>73.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big cities/others</th>
<th>Big cities</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>72.1</td>
<td>77.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Income (in thousands of SEK)</th>
<th>0-149</th>
<th>150-299</th>
<th>300-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate (%)</td>
<td>69.9</td>
<td>78.1</td>
<td>82.2</td>
</tr>
</tbody>
</table>
Conclusions from the nonresponse analyses:

The two surveys show a very similar response propensity structure.

This agrees with a general conclusion (seen also in other surveys). But sometimes the survey topic (respondent’s interest in the topic, for ex.) can affect the nature of the response propensity.

We seek an indicator for Principle 1 that gives us information on the simultaneous effect of the auxiliary variables.

\[
q^2
\]

(Described in Session 2_4)
Tools for Principle 2

Tool 2.1: Analysis of important target variables

Example: The Survey on Life and Health

Four important dichotomous study variables (attributes) are:

(a) Poor health
(b) Avoiding staying outdoors after dark
(c) Difficulties in regard to housing
(d) Poor personal finances
### Auxiliary variable: Sex

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>7.5</td>
<td>8.9</td>
</tr>
<tr>
<td>(b)</td>
<td>7.8</td>
<td>21.1</td>
</tr>
<tr>
<td>(c)</td>
<td>2.6</td>
<td>2.4</td>
</tr>
<tr>
<td>(d)</td>
<td>19.6</td>
<td>19.8</td>
</tr>
</tbody>
</table>

### Auxiliary variable: Age class

<table>
<thead>
<tr>
<th>Attribute</th>
<th>18-34</th>
<th>35-49</th>
<th>50-64</th>
<th>65-79</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>4.3</td>
<td>6.6</td>
<td>10.6</td>
<td>10.9</td>
</tr>
<tr>
<td>(b)</td>
<td>11.8</td>
<td>11.4</td>
<td>14.3</td>
<td>23.4</td>
</tr>
<tr>
<td>(c)</td>
<td>5.9</td>
<td>2.8</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>(d)</td>
<td>31.0</td>
<td>26.6</td>
<td>12.5</td>
<td>9.6</td>
</tr>
</tbody>
</table>
### Auxiliary variable: Country of birth

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Nordic countries</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>8.0</td>
<td>11.7</td>
</tr>
<tr>
<td>(b)</td>
<td>14.7</td>
<td>18.3</td>
</tr>
<tr>
<td>(c)</td>
<td>2.4</td>
<td>4.2</td>
</tr>
<tr>
<td>(d)</td>
<td>19.2</td>
<td>28.5</td>
</tr>
</tbody>
</table>

### Auxiliary variable: Income group (in thousands of SEK)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>0-149</th>
<th>150-299</th>
<th>300-</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>10.0</td>
<td>7.2</td>
<td>4.0</td>
</tr>
<tr>
<td>(b)</td>
<td>18.6</td>
<td>12.6</td>
<td>8.1</td>
</tr>
<tr>
<td>(c)</td>
<td>3.8</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>(d)</td>
<td>25.3</td>
<td>16.5</td>
<td>6.9</td>
</tr>
</tbody>
</table>
## Auxiliary variable: Marital status

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Married</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>8.2</td>
<td>8.2</td>
</tr>
<tr>
<td>(b)</td>
<td>13.8</td>
<td>16.3</td>
</tr>
<tr>
<td>(c)</td>
<td>1.1</td>
<td>4.3</td>
</tr>
<tr>
<td>(d)</td>
<td>14.1</td>
<td>26.5</td>
</tr>
</tbody>
</table>

## Auxiliary variable: Education level

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>10.5</td>
<td>7.3</td>
<td>4.6</td>
</tr>
<tr>
<td>(b)</td>
<td>19.1</td>
<td>12.6</td>
<td>12.9</td>
</tr>
<tr>
<td>(c)</td>
<td>1.7</td>
<td>3.2</td>
<td>1.8</td>
</tr>
<tr>
<td>(d)</td>
<td>17.5</td>
<td>21.6</td>
<td>16.8</td>
</tr>
</tbody>
</table>
**Conclusions** from the analysis of important target variables:

- Sex important for explaining variable (b)
- Marital status important for variable (d)
- Age class and country of birth important for most of the four variables
- Income group and education level are both important, but seem to give almost the same information
- Question arising: What is the simultaneous effect of these aux. variables?

Thus, we seek an indicator for Principle 2 that can inform us about the simultaneous effect of the auxiliary variables.

**Recall**: The NR-bias of $\hat{Y}_W$ will be small if the residuals from the regression of $y$ on $x$ are small.
Tool 2.2: Indicator \textit{IND2}

\textit{IND2} measures how close the residuals are to zero:

\[ \text{IND2} = 1 - \frac{\sum_r d_k y_{sk} (y_k - \hat{y}_k)^2}{\sum_r d_k y_{sk} (y_k - \bar{y}_r; \bar{d}_v)^2} \]

where

\[ v_{sk} = 1 + (\sum_s d_k x_k - \sum_r d_k x_k)'(\sum_r d_k z_k x_k)^{-1} z_k \]

and

\[ \hat{y}_k = x_k' (\sum_r d_k v_{sk} z_k x_k')^{-1} \sum_r d_k v_{sk} z_k y_k \]

Some empirical evidence follows in Session 2_5.

Further tools

Transforming a continuous auxiliary variable

- Forming size groups based on the variable values (often a very useful practice)

- Transforming the value of \( x_k \). We may prefer \( \sqrt{x_k} \) or \( \ln x_k \)
Further tools
Choose a "powerful" instrument vector

We know that the near-bias is zero if

$$\phi_k = 1 + \lambda' z_k$$

holds for $$k \in U$$ and some constant, non-random vector $$\lambda$$.

Thus, we should try to find "the best instrument vector"!

Example:

Suppose $$x$$ is a continuous aux. variable. Consider the auxiliary vector $$x_k = x_k$$

and an instrument vector of the form

$$z_k = x_k^{1-v}$$

where the value of $$v$$ is to be suitably determined
The nearbias is zero if \( \phi_k = 1 + a x_k^{1-\nu} \)

where \( a \) is a constant

If we believe that the response probabilities are constant throughout the population then \( \nu = 1 \) is an appropriate choice.

RA-estimator

If we believe that \( \phi_k \) increases with \( x_k \) we should use a value \( \nu < 1 \).

**Further tools**

**Analysing the weights**

Some weights too large?

- Could make the estimate for some domains too large
- The variance estimator may deteriorate

Some weights negative?

- Most users dislike negative weights
Our recommendations

(i) Make an inventory of potential aux. variables
(ii) Categorize the continuous aux. variables
(iii) Calculate $q^2$ and IND2 for different aux. vectors
(iv) Calculate the weights $\nu_k$ for the ”best” aux. vector
(v) If some of the $\nu_k$ are negative or ”too large”, drop the aux. variable that has the smallest effect on $q^2$ (or on IND2).

Sample-based selection of auxiliary variables may affect important properties of the estimator

”The choice of stratification variables cannot be made solely on the basis of the available observations. Over or under-representation of some groups can mislead us about the relationship between the target and the stratification variable. There has to be additional information about the homogeneity of the target variable.”

(Bethlehem, 1988)
Examples of sample-based selection of auxiliary variables

• collapsing of groups
• restricting or ”trimming” the weights
• avoiding near-collinearity by excluding unnecessary auxiliary variables
Intuitively, the better the aux. vector $x_k$, the better the calibration estimator:

Smaller bias, smaller variance.

- How can we analyze this more precisely?
- How do we construct the aux. vector?
- We may have access to *many* aux. variables; how do we choose?
- Primary objective here: reduce bias!
This session and the next are based on the article:

C.E. Särndal and S. Lundström (2008):
Assessing auxiliary vectors
for control of nonresponse bias
in the calibration estimator.

We consider aux. vectors of the form:

\[ \mu' x_k = 1 \quad \text{for all } k \]
Recall: We have seen three expressions for nearbias

Expression (ii):

\[
\text{nearbias}(\hat{Y}_W) = (\sum_U x_k)'(B_{U,0} - B_U)
\]

Recall: \( x_k = \begin{pmatrix} x_k^* \\ x_k^\circ \end{pmatrix} \)

Now consider expression (iii)

\[
\text{nearbias}(\hat{Y}_W) = \sum_U (\theta_k M_k - 1)y_k
\]

where

\[
M_k = \frac{(\sum_U x_k)'(\sum_U \theta_k x_k x_k^\prime)^{-1} x_k}{\text{vector defined over } U}
\]

\( M_k \) is a scalar value, unknown, linear in \( x_k \)
The value $M_k$ depends
on the aux. vector $x_k = \begin{pmatrix} x_k^* \\ x_k^o \end{pmatrix}$
on on the response prob. $\theta_k$
but not on the $y$-variable

Examination of $M_k$, $k \in U$, helps
understanding the bias

Recall: nearbias = 0 if
influence $\phi_k = 1/\theta_k = \lambda'x_k$, all $k \in U$

For this ideal (non-existent) aux. vector, we have
$M_k = \phi_k$ for all $k$ (exercise: show this !)
$\Rightarrow \theta_kM_k = 1 \Rightarrow$ nearbias = 0

For a less than ideal aux. vector,
$M_k$ is an optimal predictor of $\phi_k$, as we now show.
Properties of $M_k$

Property 1. $M_k$ is an optimal predictor (estimate) of the unknown influence $\phi_k$

Proof: We want to predict (estimate) the influences, because this would give

$$\hat{Y} = \sum_r d_k \hat{\phi}_k y_k$$

as a good substitute for the unbiased (but unrealizable) estimator

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

---

Weighted LSQ prediction:

Let $x_k$ be a fixed aux. vector. Determine $\phi_k$ as a linear function of $x_k$, so as to minimize

$$WSS = \sum_U \theta_k (\phi_k - \lambda' x_k)^2$$

Minimize $WSS$: find best $\lambda$, say, $\lambda = \hat{\lambda}$

$\implies$ Predicted influence: $\hat{\phi}_k = \hat{\lambda}' x_k = M_k$

Recommended exercise: verify the details!
We have concluded:

\( M_k \) is the best predictor (for the given aux. vector) of the influence \( \phi_k \).

For the trivial aux. vector, \( x_k = 1 \) for all \( k \)

\[
\hat{Y}_W = N \bar{y}_{r;d} = \hat{Y}_{EXP} \quad \text{(Expansion estimator)}
\]

and \( M_k = 1/\bar{\theta}_U \) for all \( k \)

\[
\Rightarrow \quad \text{nearbias}(N\bar{y}_{r;d}) = \\
\sum_U (\theta_k M_k - 1) y_k = N(\bar{y}_U;\theta - \bar{y}_U)
\]

\( \bar{y}_U;\theta - \bar{y}_U \) = weighted minus unweighted mean

---

**Recall notation**

weighted mean \( \bar{y}_U;\theta = \frac{\sum_U \theta_k y_k}{\sum_U \theta_k} \)

unweighted mean \( \bar{y}_U = \frac{\sum_U y_k}{N} \)
Properties of $M_k$

Property 2. Mean and variance of $M_k$

Weighted mean:

$$\bar{M}_{U;\theta} = \frac{\sum U \theta_k M_k}{\sum U \theta_k} = \frac{N}{\sum U \theta_k} = \frac{1}{\bar{\theta}_U}$$

Weighted variance:

$$S^2_{M|U;\theta} = \frac{\sum U \theta_k (M_k - \bar{M}_{U;\theta})^2}{\sum U \theta_k} = Q^2$$

($Q^2$ is simpler notation)

We have $Q^2 = \bar{M}_{U;\theta} (\bar{M}_U - \bar{M}_{U;\theta})$

---

Properties of $M_k$

Property 3. The variance $Q^2$ of the $M_k$ is approx. linearly related to the nearbias:

Suppose we compare $\hat{Y}_W$ (with any $x_k$) with its simplest form $N \bar{y}_{r;d} = \hat{Y}_{EXP}$ ($x_k = 1$)

Consider the nearbias ratio:

$$\frac{\text{nearbias} (\hat{Y}_W)}{\text{nearbias} (N \bar{y}_{r;d})} = \frac{\sum U (\theta_k M_k - 1)y_k}{N (\bar{y}_{U;\theta} - \bar{y}_U)}$$

Objective: Choose $x_k$ to make it small!
Properties of $M_k$

One can show (details not given here):

$$\frac{\text{nearbias}(\hat{y}_W)}{\text{nearbias}(\hat{y}_{r,d})} \approx 1 - \frac{Q^2}{Q^2_{\text{sup}}}$$

where

$$Q^2_{\text{sup}} = (1/\bar{\theta}_U)(\bar{\phi}_U - 1/\bar{\theta}_U)$$

is the value of $Q^2$ for the ideal (unattainable) case

$$\phi_k = 1/\theta_k = \lambda'x_k \quad \text{all} \; k \in U$$

Note: $0 \leq 1 - \frac{Q^2}{Q^2_{\text{sup}}} \leq 1$

Conclusion: In the choice between different aux. vectors, we should select the one that maximizes the variance $Q^2$ of the $M_k$

But $Q^2$ cannot be computed; the values $M_k$ involve sums over the whole population $U$, and contain unknown $\theta$

We replace the $M_k$ by computable analogues
Sample-based analogue of $M_k$

Replace unknown population sums in $M_k$ by corresponding computable estimates

$$m_k = (\sum_s d_k x_k)' (\sum_r d_k x_k x_k')^{-1} x_k$$

This scalar value, defined for $k \in s$, depends

- on the sampling design
- on the outcome of the response phase
- on the choice of aux. vector $x_k$

Sample-based analogue of $M_k$

For $k \in s$, we can compute

$$m_k = (\sum_s d_k x_k)' (\sum_r d_k x_k x_k')^{-1} x_k$$

We can compute the (weighted) mean and variance over $r$:

$$\bar{m}_{r:d} = \frac{1}{\sum_r d_k} \sum_r d_k m_k$$

$$S^2_{m|r:d} = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r:d})^2$$
An analysis shows

$$\bar{m}_{r,d} = \frac{\sum_r d_km_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} = \frac{1}{(\text{weighted}) \text{response rate}}$$

Hence the mean of $m_k$ is the same for every aux. vector $x_k$. But the variance depends on the aux. vector (short notation $q^2$):

$$S^2_{m|r;d} = \frac{1}{\sum_r d_k} \sum_d \sum_r d_k (m_k - \bar{m}_{r,d})^2 = q^2$$

Some properties of $q^2$

1. $q^2$ is a variance, hence non-negative
2. Alternative expression:
   $$q^2 = \bar{m}_{r;d} (\bar{m}_s;d - \bar{m}_{r;d})$$
3. The simple aux. vector $x_k = 1$ gives $q^2 = 0$
4. When new variables are added to the aux. vector, the effect is an increase in the value of $q^2$ (compare $R^2$ in regression analysis).
Practical use of $q^2$

For low bias, choose $x_k$ to make $q^2$ large.
The reason: The bias ratio is

$$\frac{\text{nearbias}(\hat{y}_W)}{\text{nearbias}(N \bar{y}_{r;d})} \approx 1 - \frac{Q^2}{Q_{\text{sup}}}$$

where $Q^2$ is the (unknown) variance of $M_k$.
Ideally: choose $x_k$ to make $Q^2$ large.

Now $q^2$ is an estimator of $Q^2$.

$\Rightarrow$ Choose $x_k$ so as to make the computable ‘indicator’ $q^2$ large.

Thus $q^2$ is a useful tool for comparing $x$-vectors, to find “the best one” (the one giving lowest bias)

We can regard $m_k$ as a “proxy value” for the unknown influence.

The more the $m_k$ vary (within limits), the better the prospects for small bias in the calibration estimator.

We call $q^2$ a “bias indicator”

Empirical illustrations
in the continuation of this session.
Comparing different aux. vectors

Suppose a supply of $x$-variables is available for the survey. Our objective: Build a good aux. vector from this supply.

• **Stepwise forward**
  
  Start with the simple vector $x_k = 1$;
  
  add one $x$-variable at a time

• **Stepwise backward**

  Start with all available $x$-variables;
  
  eliminate one at a time

---

Procedure for comparing different aux. vectors

**Stepwise forward**

Start with the simple vector $x_k = 1$;

add one $x$-variable at a time

Step 1. Compute $q^2$ for all vectors of the form $(1, x_k)$, where $x_k$ is one of the available $x$-variables. If there are $J$ available $x$-variables, we get $J$ values of $q^2$. Keep the $x$-variable that gives the largest of these values.
Procedure for comparing different aux. vectors

Stepwise forward

Step 2. Add a second $x$-variable, namely, the one that gives the largest increment among the $J - 1$ computed new values of $q^2$.

And so on, in steps 3, 4, …

A note on the case where the weights are computed with an instrument vector.

Then $\hat{y}_W = \sum_r w_k y_k$

with $w_k = d_k v_k = d_k (1 + \lambda_r' z_k)$

where $\lambda_r' = (X - \sum_r d_k x_k) (\sum_r d_k z_k x_k')^{-1}$

$x_k = \begin{pmatrix} x_k^* \\ x_k^\circ \end{pmatrix}$ ; $X = \begin{pmatrix} \sum_U x_k^* \\ \sum_s d_k x_k^\circ \end{pmatrix}$
Then we define instead $m_k$ as

$$m_k = 1 + \left( X_s - \sum_r d_k x_k \right) \left( \sum_r d_k z_k x'_k \right)^{-1} z_k$$

with

$$X_s = \begin{pmatrix} \sum_s d_k x^*_k \\ \sum_s d_k x^o_k \end{pmatrix}$$

Then compute $q^2$ as the variance of these values $m_k$; then proceed as before, with stepwise construction of the aux. vector.

---

A note on the approximation of the bias ratio

$$\frac{\text{nearbias}(\hat{Y}_W)}{\text{nearbias}(N \bar{y}_r; d)}$$

More precisely, we have

$$\text{nearbias}(\hat{Y}_W) = \text{nearbias}(N \bar{y}_r; d) \times \left( 1 - \frac{Q^2}{Q_{sup}^2} \right) + \Delta$$

What is the size of $\Delta$?

We have

$$\Delta = \sum_U \theta_k M_k E_k$$

with

$$E_k = y_k - \bar{y}_U - (\phi_k - \bar{\phi}_U) \frac{\bar{y}_U - \bar{y}_{U,0}}{\bar{\phi}_U - 1/\theta_U}$$
Recall from Session 2.4:

$$\frac{\text{nearbias}(\hat{Y}_W)}{\text{nearbias}(N y_r)} = (1 - \frac{Q^2}{Q_{\text{sup}}^2}) + \Delta$$

where

- $Q_{\text{sup}}^2$ is a constant,
- $\Delta$ is a residual term
- $Q^2$ is a function of $x_k$ and $\theta_k$
- $q^2$ is an estimator of $Q^2$
The indicator $q^2$ is computed from the values $x_k; k \in s$. It does not depend on the $y$-variable.

**Comparing aux. vectors**: We have reason to believe that the vector with the largest $q^2$ gives the smallest bias.

Important practical questions:

Does $q^2$ order the aux. vectors in a "correct" way
- on average?
- for every possible sample?

No, not always because…

$q^2$ is subject to sampling variability and

$\Delta$ is not always small (depends on $y$)

Let us look at some simulations.
Population of size $N = 832$, derived from Statistics Sweden’s KYBOK survey (see Session 2_3).

**Information:** For every $k \in U$, we know
- membership in one of 4 admin. groups
- the value of a continuous variable $x = \text{sq.root revenues}$

**Study variable:** $y = \text{expenditures}$

---

**Monte Carlo simulation**

We used two response distributions, called:
(1) Logit
(2) Increasing exponential

Average response prob.: 86% (for both)

Response probability $\theta$ increases with $x$ and with $y$

Corr. between $y$ and $\theta$:
$\approx 0.70$ (logit) ; $\approx 0.55$ (incr. exp.)
I. Monte Carlo simulation

Measures computed as averages over 10,000 repetitions (s, r); size of every s: n = 300

\[ Aveq^2 = \text{Average of } q^2 \times 10^3 \]

\[ AveIND2 = \text{Average of } IND2 \times 10^2 \]

\[ RelBias = 100 \left[ Ave(\hat{Y}_W) - Y \right] / Y \]

\[ Ave(\hat{Y}_W) = \frac{10,000}{\sum_{j=1}^{10,000} \hat{Y}_W(j)} / 10,000 \]

Response distribution: Logit

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Aveq^2</th>
<th>AveIND2</th>
<th>RelBias</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP</td>
<td>0.0</td>
<td>0.0</td>
<td>5.0</td>
</tr>
<tr>
<td>WC</td>
<td>2.7</td>
<td>43.3</td>
<td>2.2</td>
</tr>
<tr>
<td>PWA</td>
<td>2.7</td>
<td>43.3</td>
<td>2.2</td>
</tr>
<tr>
<td>REG</td>
<td>2.2</td>
<td>83.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>SEPREG</td>
<td>6.0</td>
<td>88.1</td>
<td>-0.2</td>
</tr>
<tr>
<td>TWOWAY</td>
<td>5.7</td>
<td>67.4</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The estimators are described in Session 1_8
**Response distribution**: Increasing exponential

<table>
<thead>
<tr>
<th>Estimator</th>
<th>( Aveq^2 )</th>
<th>( AveIND2 )</th>
<th>RelBias</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP</td>
<td>0.0</td>
<td>0.0</td>
<td>9.4</td>
</tr>
<tr>
<td>WC</td>
<td>3.4</td>
<td>42.3</td>
<td>5.7</td>
</tr>
<tr>
<td>PWA</td>
<td>3.4</td>
<td>42.3</td>
<td>5.7</td>
</tr>
<tr>
<td>REG</td>
<td>9.4</td>
<td>81.7</td>
<td>-2.7</td>
</tr>
<tr>
<td>SEPREG</td>
<td>18.3</td>
<td>88.1</td>
<td>-0.8</td>
</tr>
<tr>
<td>TWOWAY</td>
<td>18.0</td>
<td>67.1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The estimators are described in Session 1_8

---

**This simulation shows**:  

- a clear tendency (although not a perfect relationship) that larger values of \( Aveq^2 \) accompany the estimators with small bias  
- that the relationship between \( y \) and \( x \) has an effect on the bias.

*Example*: \( Aveq^2 \) is larger for WC (and PWA) than for REG, but the RelBias is smaller. This is explained by the fact that \( AveIND2 \) is smaller for WC (and PWA) than for REG.
II. Monte Carlo simulation

For every possible sample, does $q^2$ correctly order the auxiliary vectors?

We examine four of the six estimators: SEPREG, REG, WC and EXP.

$q^2$ is random; it depends on the outcome $(s,r)$.

For every outcome, we can rank the four estimators by their value of $q^2$. The perfect ordering would be

$$q^2(\text{SEPREG}) \geq q^2(\text{REG}) \geq q^2(\text{WC}) \geq q^2(\text{EXP})$$

because this is the ordering based on the absolute value of $\text{RelBias}$.

---

Reasons for using only 4 of the 6 estimators in the study:

(i) WC and PWA have the same nearbias

(ii) SEPREG and TWOWAY have almost the same nearbias
For each repetition \((s,r)\), we rank order the estimators by the size of \(q^2\), and assign rank values: 1 (to the estimator with the largest \(q^2\)), 2, 3 and 4 (to the estimator with the smallest \(q^2\)).

We then compute the average rank ordering \((AveOrd)\) over the 10,000 repetitions. The results are shown in the following pictures.

**Response distribution**: Logit

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Aveq^2</th>
<th>AveIND2</th>
<th>RelBias</th>
<th>AveOrd</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP</td>
<td>0.0</td>
<td>0.0</td>
<td>5.0</td>
<td>4.00</td>
</tr>
<tr>
<td>WC</td>
<td>2.7</td>
<td>43.3</td>
<td>2.2</td>
<td>2.40</td>
</tr>
<tr>
<td>REG</td>
<td>2.2</td>
<td>83.4</td>
<td>-0.6</td>
<td>2.60</td>
</tr>
<tr>
<td>SEPREG</td>
<td>6.0</td>
<td>88.1</td>
<td>-0.2</td>
<td>1.00</td>
</tr>
</tbody>
</table>
**Response distribution**: Increasing exponential

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$Aveq^2$</th>
<th>$AveIND2$</th>
<th>RelBias</th>
<th>AveOrd</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP</td>
<td>0.0</td>
<td>0.0</td>
<td>9.4</td>
<td>4.00</td>
</tr>
<tr>
<td>WC</td>
<td>3.4</td>
<td>42.3</td>
<td>5.7</td>
<td>2.97</td>
</tr>
<tr>
<td>REG</td>
<td>9.4</td>
<td>81.7</td>
<td>-2.7</td>
<td>2.03</td>
</tr>
<tr>
<td>SEPREG</td>
<td>18.3</td>
<td>88.1</td>
<td>-0.8</td>
<td>1.00</td>
</tr>
</tbody>
</table>

This simulation experiment shows:

- SEPREG always (in every sample) receives rank 1 (agreeing with the fact that its bias is the smallest)
- EXP always receives rank 4 (and it has the highest bias)
- Between WC and REG, the pattern is not clear-cut. One important reason is that the relationship between $y$ and $x$ has an effect.
Use of the bias indicator $q^2$ in the Swedish National Crime Victim and Security Study (a telephone interview survey)

Survey objective: Measure trends in certain types of crimes, in particular crimes against the person.

Sampling design: STSI of 10,000 persons (strata: 21 regions (“län”) × 3 age groups)

Overall response rate: 77.8%

Statistics Sweden’s data base LISA contains many potential auxiliary variables.

For example:

Type of family, number of children in different age groups, education level, profession, branch of industry, number of days with illness, number of days of unemployment, number of days in early retirement pension, income of capital, and so on

How do we select?
Preparation:

(i) An initial set of potential auxiliary variables was selected by a subjective procedure

(ii) Aux. variables were used at the sample level (moon variables)

(iii) Continuous variables are used as grouped; all variables used are then grouped.

The use of $q^2$ as a tool for stepwise forward selection of variables:

- In each step, the auxiliary vector expands by adding the (grouped) variable causing the largest increase in $q^2$

- Variables enter in the ”side-by-side” manner (or ”+”)
## Results

<table>
<thead>
<tr>
<th>Step</th>
<th>Auxiliary variable entering</th>
<th>Number of groups</th>
<th>Value of $1000 \times q^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-----</td>
<td>-----</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>Country of birth</td>
<td>2</td>
<td>20.0</td>
</tr>
<tr>
<td>2</td>
<td>Income group</td>
<td>3</td>
<td>27.6</td>
</tr>
<tr>
<td>3</td>
<td>Age group</td>
<td>6</td>
<td>31.3</td>
</tr>
<tr>
<td>4</td>
<td>Gender</td>
<td>2</td>
<td>35.1</td>
</tr>
<tr>
<td>5</td>
<td>Marital status</td>
<td>2</td>
<td>38.6</td>
</tr>
<tr>
<td>6</td>
<td>Region</td>
<td>21</td>
<td>40.7</td>
</tr>
<tr>
<td>7</td>
<td>Family size group</td>
<td>5</td>
<td>41.4</td>
</tr>
<tr>
<td>8</td>
<td>Days unemployed</td>
<td>6</td>
<td>41.9</td>
</tr>
<tr>
<td>9</td>
<td>Urban centre dweller</td>
<td>2</td>
<td>42.3</td>
</tr>
<tr>
<td>10</td>
<td>Occupation</td>
<td>10</td>
<td>42.7</td>
</tr>
</tbody>
</table>

## Observations:

- Successive increases in $q^2$ taper off (as expected).
- It seems hardly motivated to go beyond the sixth variable (region)
The final choice of auxiliary vector was:
Region+gender+age group+country of birth+
+ income group+urban centre dweller

Principles that also played a role:
(i) The auxiliary vector should be robust. The survey will be conducted yearly; the client prefers having the same vector over time.

(ii) The auxiliary vector should contain region and age group, because they identify the most important domains.

(iii) An auxiliary vector should well explain the (main) study variables
NR causes both
• a problem due to bias
  and
• a problem with variance estimation
  (which we now discuss)
Recall from Session 1.5:

The accuracy has two parts:

\[ \text{MSE}_{pq}(\hat{Y}_W) \approx \underbrace{V_p(\hat{Y})}_{\text{due to sampling}} + \underbrace{E_p V_q(\hat{Y}_W|s) + E_p(B^2_{W|s})}_{\text{due to NR}} \]

\( \hat{Y} \) is the full response estimator

A serious problem: the bias component \( E_p(B^2_{W|s}) \) may be large

The variance of the calibration estimator \( \hat{Y}_W \)

Assuming that \( B_{W|s} = E_q(\hat{Y}_W - \hat{Y})|s| = 0 \)

the variance is the sum of two components:

- **Sampling variance** \( V_{SAM} = V_p(\hat{Y}) \)
  
  \( \hat{Y} \) is the full response estimator

- **Nonresponse variance** \( V_{NR} = E_p V_q(\hat{Y}_W|s) \)
The variance of the calibration estimator

\[ V_{NR} \] is the \textit{additional variance} incurred by getting fewer observations than desired.

NR increases variance.

We can always ‘oversample’ to counterbalance the increased variance.

The more serious consequence of NR is the systematic error (the bias).

---

Objective:
Obtain valid confidence statements

so that

\[ \hat{Y}_W \pm \frac{z_{\alpha}}{2} \sqrt{V(\hat{Y}_W)} \]

with \( z_{\alpha} / 2 = 1.96 \)

gives \( \approx 95\% \) confidence.

We can count on approx. normal distribution, but a non-negligible bias would distort the confidence. The interval may become invalid.
Objective: Obtain valid confidence statements

It is obvious that

\[ \hat{Y}_W \pm 1.96\sqrt{\hat{V}(\hat{Y}_W)} \]

can give \(\approx 95\%\) confidence

only if \(\text{bias}(\hat{Y}_W)\) fairly small compared with the estimated stand.dev \(\sqrt{\hat{V}(\hat{Y}_W)}\)

A bad situation: \(\text{bias} > \text{stand. dev.}\)

In this case, coverage of conf.int. \(\approx 0\)
We proceed under the assumption that we have succeeded in reducing the NR bias to modest levels (by the methods seen in earlier sessions). We shall construct an estimator of the variance \( \hat{V}(\hat{Y}_W) \) by estimating each of the two components:

\[
V_{SAM} + V_{NR} = V_P(\hat{Y}) + E_p V_q(\hat{Y}_W|s)
\]

We create an estimator of each component, \( \hat{V}_{SAM} \) and \( \hat{V}_{NR} \), then add them to get an estimator of total variance:

\[
\hat{V}(\hat{Y}_W) = \hat{V}_{SAM} + \hat{V}_{NR}
\]

We do this under very general conditions:

- any sampling design
- any auxiliary vector \( \mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix} \)
A dilemma
for the variance estimation

Estimating the variance components runs into the same problem as the point estimation:
The $y$-data available only for the response set are ‘not representative’, because of non-random NR.

Unknown influences $\phi_k = 1/\theta_k$

Comment

Variance estimation is a more sensitive issue than point estimation.

Variance implies *squared numbers*; more sensitive to weighting.
An approach to variance estimation

Had the influences $\phi_k = 1/\theta_k$ been known, we could have used the two-phase GREG estimator

$$\hat{Y}_{GREG\ 2\ ph} = \sum_r d_k \frac{1}{\theta_k} g\theta_k y_k$$

Given that the $\theta_k$ are known, we know the expression for the variance, and how to estimate it.

We note now that the two-phase GREG estimator

$$\hat{Y}_{GREG\ 2\ ph} = \sum_r d_k \frac{1}{\theta_k} g\theta_k y_k$$

is equal to the calibration estimator

$$\hat{Y}_W = \sum_r d_k v_k y_k$$

if $\phi_k = 1/\theta_k = v_k$
The proposed variance estimator for $\hat{Y}_W$ builds on this identity with the two-phase GREG estimator $\hat{Y}_{GREG\ 2\ ph}$.

The known formula for $V(\hat{Y}_{GREG\ 2\ ph})$ has two components. In those components, we replace $1 / \theta_k$ by the adjustment factor $v_k$ (already computed for the point estimator).

We thus obtain an ‘ad hoc’ estimator of each component.
Recall: \( v_k = 1 + \lambda'_r x_k \)

where \( \lambda'_r = (X - \sum_r d_k x_k) (\sum_r d_k x_k x'_k)^{-1} \)

and \( x_k = \begin{pmatrix} x^*_k \\ x^o_k \end{pmatrix} ; \quad X = \begin{pmatrix} \sum_U x^*_k \\ \sum_S d_k x^o_k \end{pmatrix} \)

The procedure gives \( \hat{V}_{SAM} \) and \( \hat{V}_{NR} \)

Adding them: \( \hat{V}(\hat{Y}_w) = \hat{V}_{SAM} + \hat{V}_{NR} \)

The components will contain two types of residual (but no regression is ever fitted).
One residual for each component.
The residuals reflect the available information.
Recall: Auxiliary information statement

<table>
<thead>
<tr>
<th>Set of units</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population $U$</td>
<td>$\sum_U x_k^*$ known</td>
</tr>
<tr>
<td>Sample $S$</td>
<td>$x_k^*$ known, $k \in s$</td>
</tr>
<tr>
<td>Response set $R$</td>
<td>$x_k^*$ and $x_k^\circ$ known, $k \in r$</td>
</tr>
</tbody>
</table>

The residuals for *NR variance component* are adjusted for *both* kinds of aux. info

$$\hat{e}_k = y_k - x_k^*/B_{r;dv}^* - x_k^\circ/B_{r;dv}^\circ$$

Residuals for *Sampling variance component* are adjusted only for the “population info”:

$$\hat{e}_k^* = y_k - x_k^*/B_{r;dv}^*$$

For details, see the book.
The regression coefficient is computed as

\[
B_{r;dv} = \begin{pmatrix}
B^*_{r;dv} \\
B^0_{r;dv}
\end{pmatrix} = \left( \sum_r d_k v_k x_k x_k' \right)^{-1} \left( \sum_r d_k v_k x_k y_k \right)
\]

Note the weighting: \( d_k v_k \)

\( v_k \) a proxy for the unknown \( \phi_k = 1 / \theta_k \)

To illustrate the general formula

\[
\hat{V}(\hat{Y}_W) = \hat{V}_{SAM} + \hat{V}_{NR}
\]

it is a good idea to note what the expressions look like in a familiar situation:

- STSI sampling
- each stratum used as a group for NR adjustment.

Procedure “simple expansion by stratum”
STSI; each stratum an adjustment group.

\[ x_k = x_k^* = \gamma_k = (\gamma_{1k}, \ldots, \gamma_{hk}, \ldots, \gamma_{Hk})' \]

\[ = (0, \ldots, 1, \ldots, 0)' \]

The “1” indicates the stratum to which \( k \) belongs.

---

STSI; each stratum an adjustment group.

In stratum \( h \),

\( n_h \) are sampled from \( N_h \) by SI sampling

\( m_h \) out of \( n_h \) are found to respond

The general formulas give the weights

\[ d_k = \frac{N_h}{n_h} \; ; \; v_k = \frac{n_h}{m_h} \; ; \; w_k = d_k v_k = \frac{N_h}{m_h} \]

Recommended exercise: Derive \( v_k \) in this case!
STSI; each stratum an adjustment group.

The general formulas for the estimated variance components give easily understood expressions:

Estimated **sampling variance**:

\[ \hat{V}_{SAM} \approx \sum_{h=1}^{H} N_h^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) S_{yrh}^2 \]

\[ S_{yrh}^2 = y\text{-variance computed in } r_h \]

(the response set in stratum \( h \))

STSI; each stratum an adjustment group.

Estimated **NR variance**:

\[ \hat{V}_{NR} \approx \sum_{h=1}^{H} N_h^2 \left( \frac{1}{m_h} - \frac{1}{n_h} \right) S_{yrh}^2 \]

Makes good sense. It is like “taking \( m_h \) from \( n_h \)”

Factors: \( \left( \frac{1}{n_h} - \frac{1}{N_h} \right) + \left( \frac{1}{m_h} - \frac{1}{n_h} \right) = \frac{1}{m_h} - \frac{1}{N_h} \)

Estimated **total variance**:

\[ \hat{V}(\hat{Y}_w) \approx \sum_{h=1}^{H} N_h^2 \left( \frac{1}{m_h} - \frac{1}{N_h} \right) S_{yrh}^2 \]
General formulas for estimated variance components

The following pictures show abstract and lengthy general formulas.

They are of particular interest for the specialist in variance estimation.

The practitioner wants to know ‘if it works’.

The answer is ‘yes’. Software is available, for ex.: CLAN97.

*Estimator of sampling variance*

\[
\hat{V}_{SAM} = \\
\sum_{k} \sum_{\ell} (d_k d_{\ell} - d_k \ell) (v_k \hat{e}_k^\ast) (v_\ell \hat{e}_\ell^\ast) \\
- \sum_{k} d_k (d_k - 1) v_k (v_k - 1) (\hat{e}_k^\ast)^2
\]

with

\[
\hat{e}_k^\ast = y_k - x_k^\ast B_{r;dv}^\ast
\]
Estimator of nonresponse variance

\[ \hat{V}_{NR} = \sum_r v_k (v_k - 1) (d_k \hat{e}_k)^2 \]

with

\[ \hat{e}_k = y_k - x'_k B_{r;dv} = y_k - x^*_k B_{r;dv} - x^*_k B_{r;dv} \]

The special case \( x_k = x_k^* \)

(only population info)

\[ \hat{e}^*_k = \hat{e}_k = y_k - x^*_k \left( \sum_r d_k v_k x^*_k x^*_k \right)^{-1} \left( \sum_r d_k v_k x^*_k y_k \right) \]
This variance estimation, although not perfect in all respects, has been shown to work well (see simulations in the book).

Caution: Variance estimates are occasionally unstable, can be sensitive to ‘large weights’.
2.7 Estimation in the presence of both nonresponse and frame imperfections

Frame population: $U_F$  
Target population: $U$

"Persisters"  
($U_P = U \cap U_F$)

Overcoverage ($U_F - U_P$)  
Undercoverage ($U - U_P$)
The estimation procedure needs to deal simultaneously with sampling error, nonresponse error and coverage error.

Not a trivial step to accommodate the third kind of error and derive a firmly established methodology!

Few "conventional methods" to compare with.
Problems:
• the absence of observed $y$-data from the undercoverage set
• the absence of correct auxiliary vector total for $U$
• difficulties of decomposing the nonresponse set $OF$ into its subsets $OP$ and $O \backslash P$, for example, identifying the elements that need imputation

Two procedures for estimating $Y_U$

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

(ii) by direct estimation of the target population total $Y_U$
i) a. Estimation of the persister total \( Y_{UP} \)

The persister set \( UP \) is a domain of \( UF \) and the corresponding response sets are \( r_P \) and \( r_F = r_P \cup r_{\neg P} \).

Let us define

\[
y_{Pk} = \begin{cases} y_k & \text{if } k \in UP = U \cap UF \\ 0 & \text{otherwise} \end{cases}
\]

\[
\hat{Y}_{UPW} = \sum_{r_F} w_k y_{Pk} = \sum_{r_P} w_k y_k
\]

where \( w_k = d_k \nu_k \) and

\[
\nu_k = 1 + \left( \sum_{UF} x^*_k - \sum_{r_F} d_k x^*_k \right) \left( \sum_{r_F} d_k x_k^* (x_k^*)' \right)^{-1} x_k^*
\]
**Ex.** A commonly used estimator of the persister total

\[
\hat{Y}_{UPW} = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{Ph} + m_{\setminus Ph}} \sum r_{Ph} y_k =
\]

\[
= \sum_{h=1}^{H} \frac{N_{Fh}}{m_{Fh}} \sum r_{Fh} y Pk
\]

If \( U_F \) is divided into strata, \( U_{Fh}, h = 1, \ldots, H \)

STSI: \( n_{Fh} \) from \( N_{Fh} \); \( m_{Fh} \) respond

Aux. vector: \( x_k = x_k^* = \gamma_k \)

---

i) b. Estimation of the undercoverage total

\( Y_{U-U_P} \)

In the book we do not suggest any particular method for estimating the undercoverage total.
ii) Direct estimation of the target population total $Y_U$

Let $\hat{X}$ denote an approximation of $\sum_U x_k^*$

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

where

$$w_k = d_k v_k$$

and

$$v_k =$$

$$= 1 + \left( \hat{X} - \sum r_P d_k x_k^* \right) \left( \sum r_P d_k x_k^* (x_k^*)' \right)^{-1} x_k^*$$

Ex. A commonly used estimator of the target population total

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

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

STSI: $n_{Fh}$ from $N_{Fh}; m_{Fh}$ respond

Aux. vector: $x_k = x_k^* = \gamma_k$
Variance estimators

are derived with the aid of proxies for

\[ \phi_k = 1 / \theta_k \]

Let us look at the two cases

(i) Estimation of the persister total

and

(ii) Direct estimation of the target population total

Variance estimation

Case i) Estimation of the persister total

\[ \hat{\phi}_k = \nu_k \]

where

\[ \nu_k = 1 + \left( \sum_{U_F} x_k^* - \sum_{r_F} d_k^* x_k^* \right) \left( \sum_{r_F} d_k^* x_k^* (x_k^*)' \right)^{-1} x_k^* \]
Ideal: Calibrate from \( r_P \) to \( r_P \cup o_P \)

But impossible if \( o_P \) is not identified

Surrogate procedure: Calibrate from \( r_F \) to \( U_F \)

Variance estimation

Case ii) Direct estimation of the target population total (two alternatives)

\[
(1) \quad \hat{\phi}_k = v_k
\]

where

\[
v_k = 1 + (\sum_{U_F} x_k^* - \sum_{r_F} d_k x_k^*)' \left( \sum_{r_F} d_k x_k^* (x_k^*)' \right)^{-1} x_k^*
\]
(2) \( \hat{\phi}_k = \nu P_k \)

where

\[
\nu P_k = 1 + (\sum_{\rho \cup \rho} d_k x_k^* - \sum_{\rho} d_k x_k^*)'(\sum_{\rho} d_k x_k(x_k^*)')^{-1} x_k^*
\]

**A case study**

The survey on ”Transition from upper secondary school to higher education”

We call it the School Survey.

Important study variables:

(a) The intentions to pursuing studies at university

(b) The university programmes viewed as the most interesting
The estimator used before the redesign

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

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

(i) The overcoverage is considerable greater than the undercoverage

(ii) The response propensity is very low among nonpersisters
The solution:

We discovered a good approximation $\tilde{X}$ of $\sum_U x_k^*$ and estimated the target population total by the direct estimation method.

Aux. variables:
- "final mark" at the end of grade 9
- parental variables: level of education, income and civil status

Some results
- The estimates of totals undergo considerable change
- Estimates of proportions undergo little change
- The estimated variances for proportions were not much reduced
The course has presented ‘a general way of thinking’ about estimation in sample surveys with NR and frame imperfections:

*Estimation by calibration*

As a result, instead of a few specific (‘traditional’) estimator formulas, we have seen a general way to produce estimators; we have focused on the question: how do we choose an appropriate *auxiliary vector*, with the corresponding *auxiliary information*. 
The approach is simple to explain to users. The approach relies on important statistical concepts, but a fairly limited number of concepts.

Computationally, the approach is not highly complex or demanding.

We do believe that survey methodologists (in particular) need to have a solid understanding of the theory behind the approach.

As a result, this course has examined the theory in some detail; a number of theoretical expressions have been presented.

The course has emphasized that the key to “conclusions of acceptable quality” in a survey (with a perhaps considerable NR) is to identify powerful auxiliary information for the calibration.

We have specified some tools that are useful in this search.
We hope you enjoyed the course!

Thank you for listening!
Exercise 1

A scenario: Someone in your organization is seeking your opinion on a survey with NR. He or she says: “With a sample size of 1,500, we got 1,000 responses, so we still have a lot of data to base our statistics and our conclusions on. I do not think the NR is a problem.”

Formulate your response to the person making this statement.
**Exercise 2**

*A scenario*: As a methodologist, you are called upon to discuss survey NR treatment with a user in your organization. More specifically, you need to:

- Convince the user about the need for NR bias adjustment
- Explain to the user (a) the favourable effects of calibration, and (b) the nature and the properties of the calibrated weights

*Formulate your responses* to the user.

---

**Exercise 3**

The simulation experiment in Session 1_2 ends with a table titled “Coverage rate (%) for different samples sizes …” Explain (with the aid of basic statistical concepts) why, as a result of the NR, the coverage rate drops when the sample size increases, other things being equal.
Exercise 4

The simplest auxiliary vector

\[ x_k = x_k^* = 1 \]

Show that the calibrated weights are

\[ w_k = d_k \frac{N}{\sum_r d_k} \]

Consequence for SI sampling:

\[ w_k = \frac{N}{n} \frac{n}{m} = \frac{N}{m} \]

\[ m = \text{number of respondents} \]

See Session 1_8

Exercise 5

Start from the general formula for the calibrated weights. Take

\[ x_k = x_k^* = \gamma_k \]

Show that the weights are

\[ w_k = d_k N_p / \sum_r d_k \]

for \( k \) in group \( p \), so that the estimator becomes

\[ \hat{Y}_{PWA} = \sum_{p=1}^P N_p \gamma_r p \cdot d \]

See Session 1_8
**Exercise 6**

Start from the general formula for the calibrated weights. Take
\[ x_k = x_k = \gamma_k \]
Show that the weights are
\[ w_k = d_k \left( \sum_s d_k \right) / \left( \sum_r d_k \right) \]
for \( k \) in group \( p \).
For SI sampling: \[ w_k = \frac{N n_p}{n m_p} \]
See Session 1_8

---

**Exercise 7**

Consider the weights \[ w_k = d_k v_k \]
where \[ v_k = 1 + \lambda'_r z_k \]
\[ \lambda'_r = \left( X - \sum_r d_k x_k \right) \left( \sum_r d_k z_k x'_k \right)^{-1} \]
where \( z_k \) is an instrument vector
Show that, for any \( z_k \), these weights satisfy the calibration equation
\[ \sum_r w_k x_k = \sum U x_k \]
See Session 1_7


**Exercise 8**

Invariant calibrated weights are obtained in the following situation:

- STSI with strata $U_p$; $n_p$ from $N_p$; $p = 1, \ldots, P$
- $z_k = x_k = x_k^* = \text{stratum identifier}$

Then the initial weights

$$d_{\alpha k} = d_k = N_p / n_p$$

and

$$d_{\alpha k} = d_k \times (n_p / m_p) = N_p / m_p$$

give the same calibrated weights, namely $w_k = N_p / m_p$

Show this! See Session 1_7

**Exercise 9**

Suppose the correlation between $y$ and $\theta$ is 0.6. Then show that

$$\text{bias}(\hat{Y}_{\text{EXP}} / N) \approx 0.6 \times cv(\theta) \times S_{yU}$$

where

$$cv(\theta) = S_{\theta U} / \bar{\theta}_U$$

the coeff. of variation of $\theta$

$$S_{yU}$$

the stand. dev. of $y$ in $U$

See Session 2_2
Show that  
\[ \text{nearbias}(\hat{Y}_W) = -\sum_{U} (1-\theta_k) e_{\theta_k} \]
becomes 0 if  
\[ \phi_k = 1 + \lambda' x_k \]
holds for all \( k \) in \( U \)
and some constant vector \( \lambda \)

See Session 2.2