9
Variance and variance estimation for calibration estimators

NR causes both

- a problem due to bias

  and

- a problem with variance estimation

  (which we now discuss)

The accuracy measured by MSE has two parts :

Due to samling: A variance

Due to NR: A variance and $(\text{bias})^2$

$$MSE_{pq}(\hat{Y}_W) \approx \underbrace{V_p(\hat{Y})}_{\text{due to sampling}} + \underbrace{E_p V_q(\hat{Y}_W|s) + E_p(B^2_{W|s})}_{\text{due to } NR}$$

$\hat{Y}$ 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 $$B_{W|s} = E_q\left((\hat{Y}_W - \hat{Y}) \middle| s \right) = 0$$

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 ***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 $$\hat{Y}_W \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{Y}_W)}$$

with $$z_{\alpha/2} = 1.96$$

gives ≈ 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

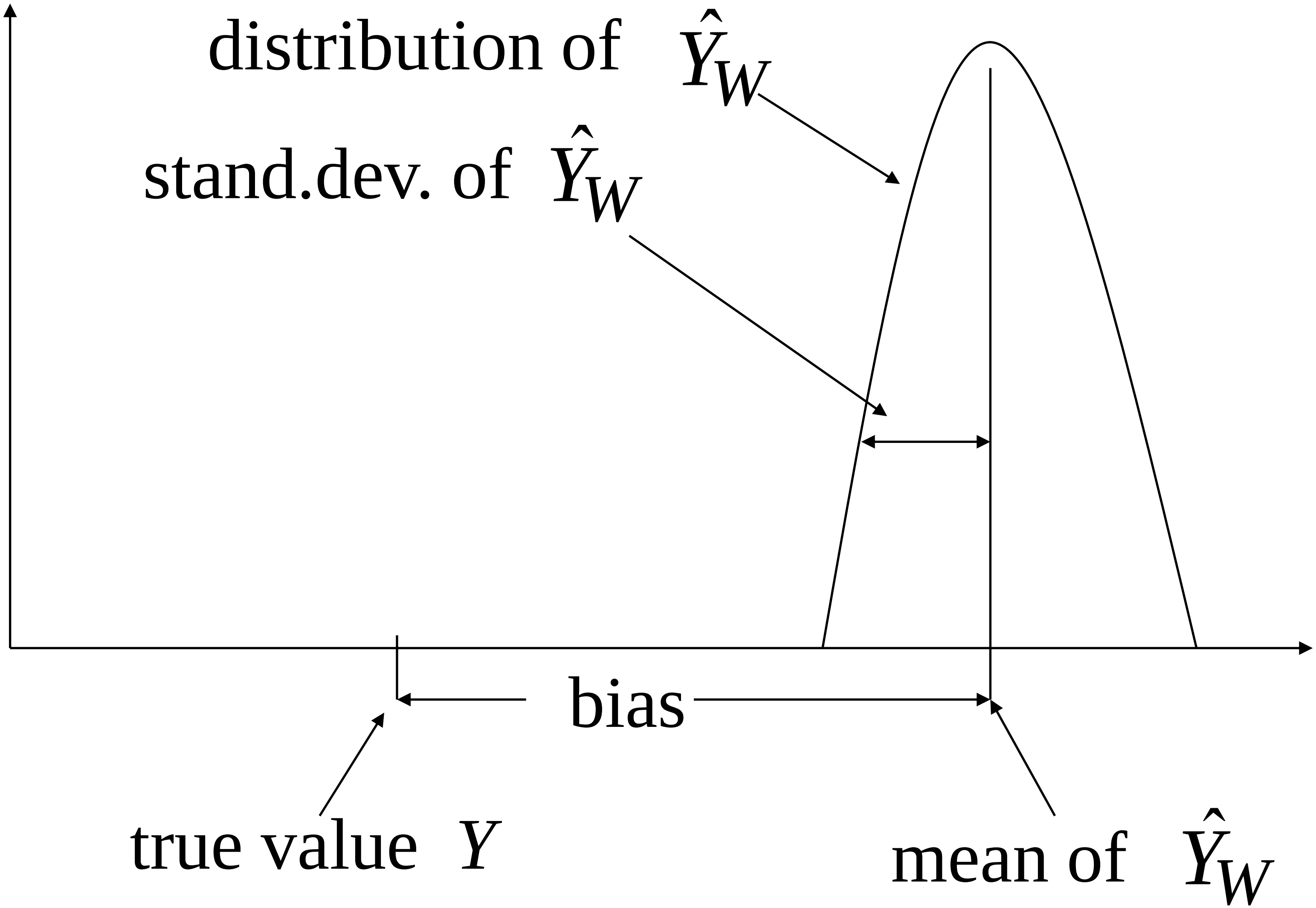
It is obvious that

$$\hat{Y}_W \pm 1.96\sqrt{\hat{V}(\hat{Y}_W)}$$

can give ≈ 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 : bias > stand. dev.



In this case, coverage of conf.int. ≈ 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 :

$$V_{SAM} + V_{NR} = V_p(\hat{Y}) + E_p V_q(\hat{Y}_W \mid s)$$

In the book, we create an estimator of each component ,

$$\hat{V}_{SAM} \text{ and } \hat{V}_{NR}$$

then add them to get an

estimator of total variance :

$$\hat{V}(\hat{Y}_W) = \hat{V}_{SAM} + \hat{V}_{NR}$$

This is done under general conditions :

- any sampling design
- any auxiliary vector $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^{\text{o}} \end{pmatrix}$

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 .

Recall : Auxiliary information statement

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

To illustrate the general procedure (given in the book)

$$\hat{V}(\hat{Y}_W) = \hat{V}_{SAM} + \hat{V}_{NR}$$

let us examine the expressions obtained in a familiar situation :

- STSRS sampling
- each stratum is a group for NR adjustment.

This is the procedure
"simple expansion by stratum" .

STSRS; each stratum an adjustment group.

$$\mathbf{x}_k = \mathbf{x}_k^* = \boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{hk}, \ldots, \gamma_{Hk})'$$

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

The "1" indicates the stratum to which k belongs

STSRS; each stratum an adjustment group.

In stratum h ,

n_h are sampled from N_h by SRS

m_h out of n_h respond

The general formulas give the weights

$$d_k = \frac{N_h}{n_h} \quad ; \quad v_k = \frac{n_h}{m_h} \quad ; \quad w_k = d_k v_k = \frac{N_h}{m_h}$$

STSRS; each stratum an adjustment group.

A important justification: The general formulas for the estimated variance components give *easily understood expressions* :

Estimated ***sampling variance*** :

$$\hat{V}_{SAM} \approx \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{yr_h}^2$$

$S_{yr_h}^2$ = *y*-variance computed in r_h

(the response set in stratum *h*)

STSRS; each stratum an adjustment group.

Estimated ***NR variance*** :

$$\hat{V}_{NR} \approx \sum_{h=1}^{H} N_h^2 \left(\frac{1}{m_h} - \frac{1}{n_h}\right) S_{yr_h}^2$$

Makes good sense. It is like "taking m_h from n_h"

Factors: $\left(\frac{1}{n_h} - \frac{1}{N_h}\right) + \left(\frac{1}{m_h} - \frac{1}{n_h}\right) = \frac{1}{m_h} - \frac{1}{N_h}$

$\Rightarrow$ Estimated ***total variance*** :

$$\hat{V}(\hat{Y}_W) \approx \sum_{h=1}^{H} N_h^2 \left(\frac{1}{m_h} - \frac{1}{N_h}\right) S_{yr_h}^2$$

The general formulas for estimated variance components

The formulas are lenghtly and complex.

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.: CLAN/ETOS .

This variance estimation procedure, although not perfect, has been shown to work well (see simulations in the book) .

Caution: Variance estimates are occasionally unstable, can be sensitive to ‘large weights’.