Statistiska centralbyrån Statistics Sweden

2
Introductory aspects of the course material

Survey errors

Sampling errors **(considered in the course)**

Nonsampling errors

- Errors due to non-observation

  Undercoverage **(considered in the course)**

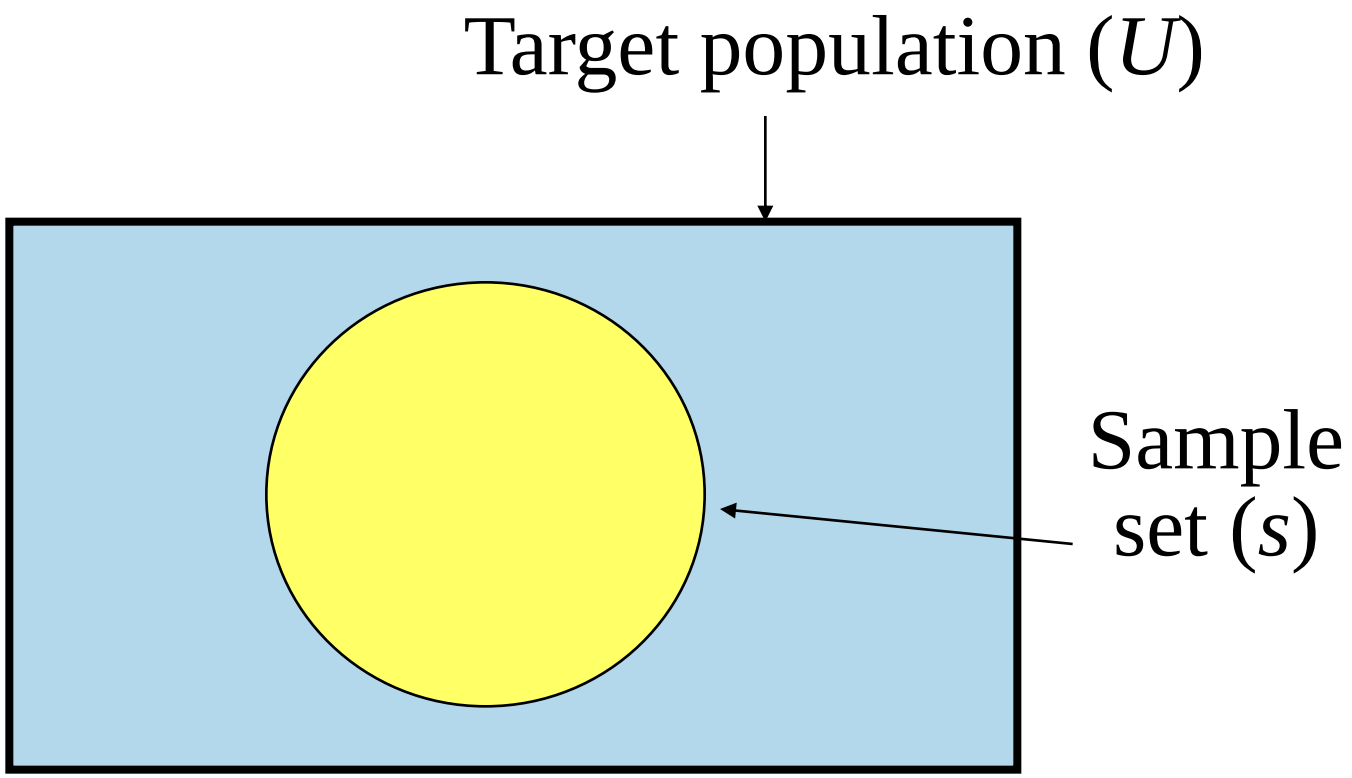
  Nonresponse **(considered in the course)**

- Errors in observations

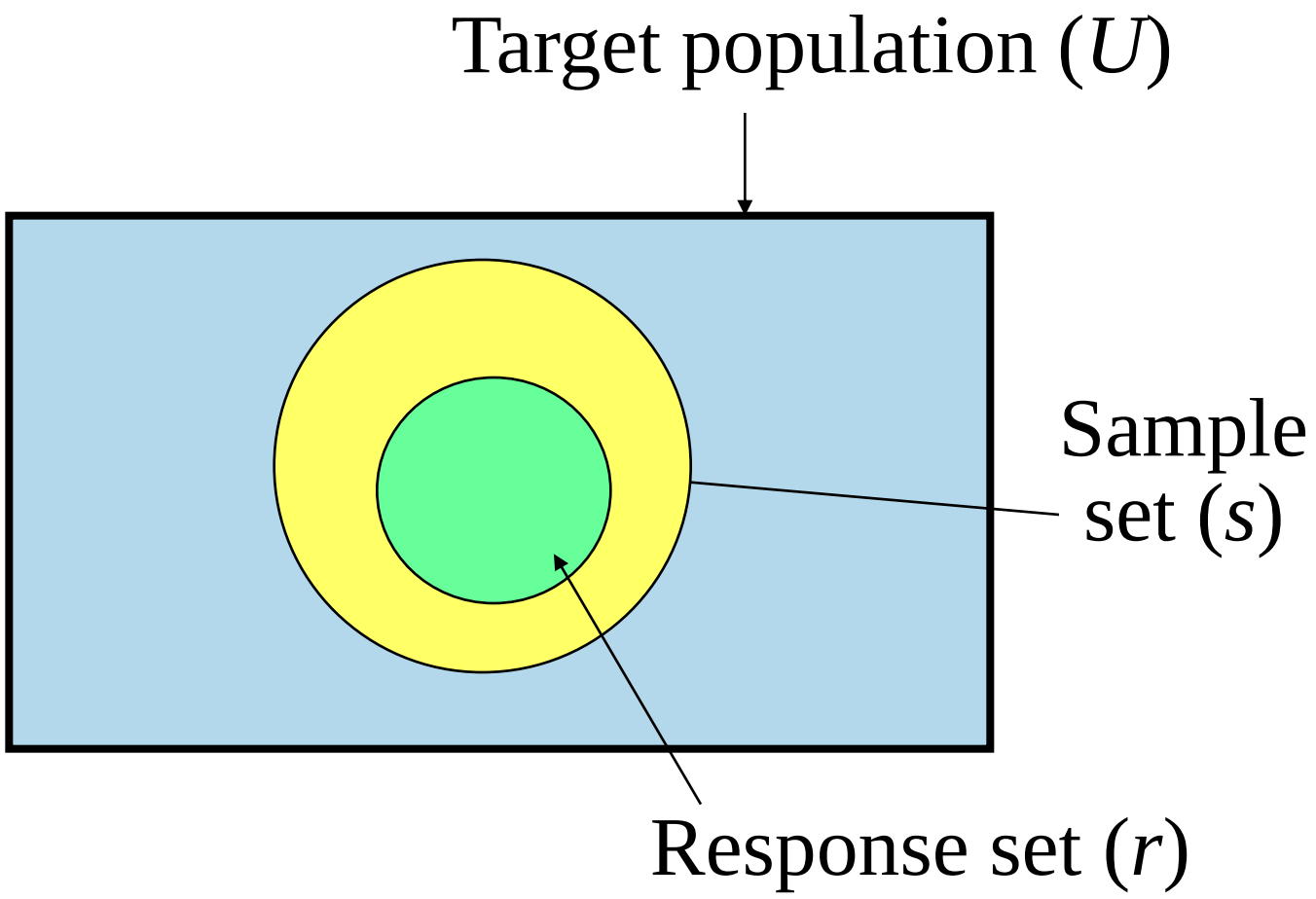
  Measurement

  Data processing

Sampling error



Sampling error and nonresponse error



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

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

Let us assume $P = 50$

Sampling design: SRS , ***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 $\theta_k = \Pr(k \text{ responds})$:

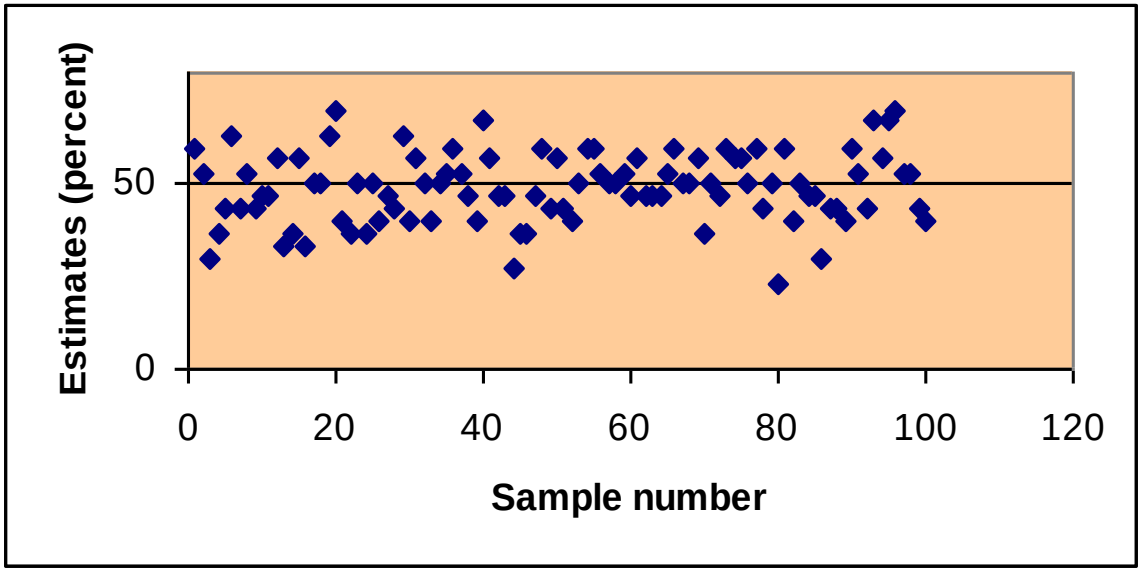
$$\theta_k = \begin{cases} 0.5 & \text{if element } k \text{ has the property} \\ 0.9 & \text{otherwise} \end{cases}$$

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

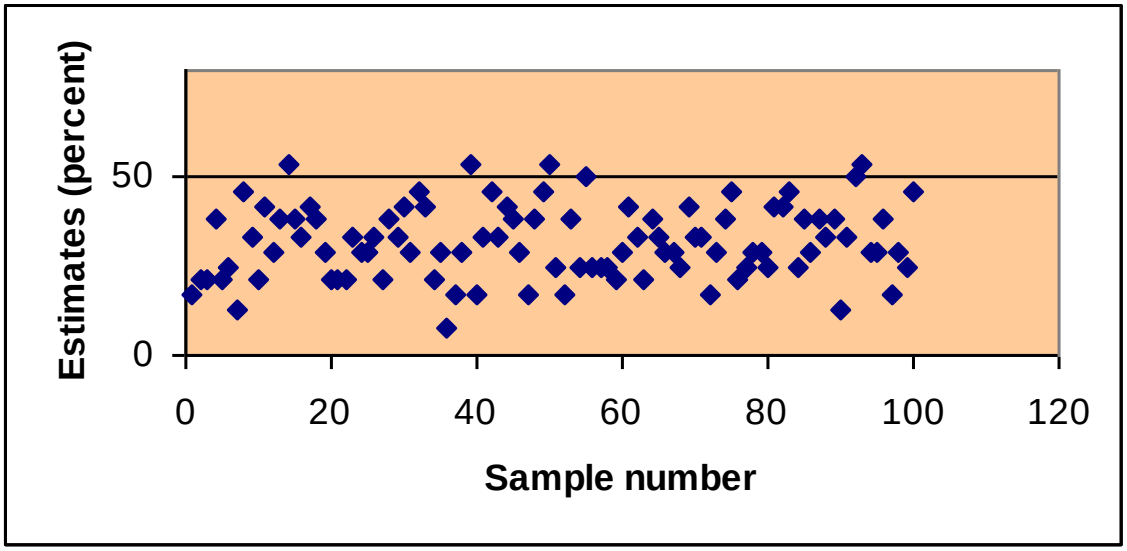
100 repeated realizations (s, r); N large

***n*=30**

Full-response

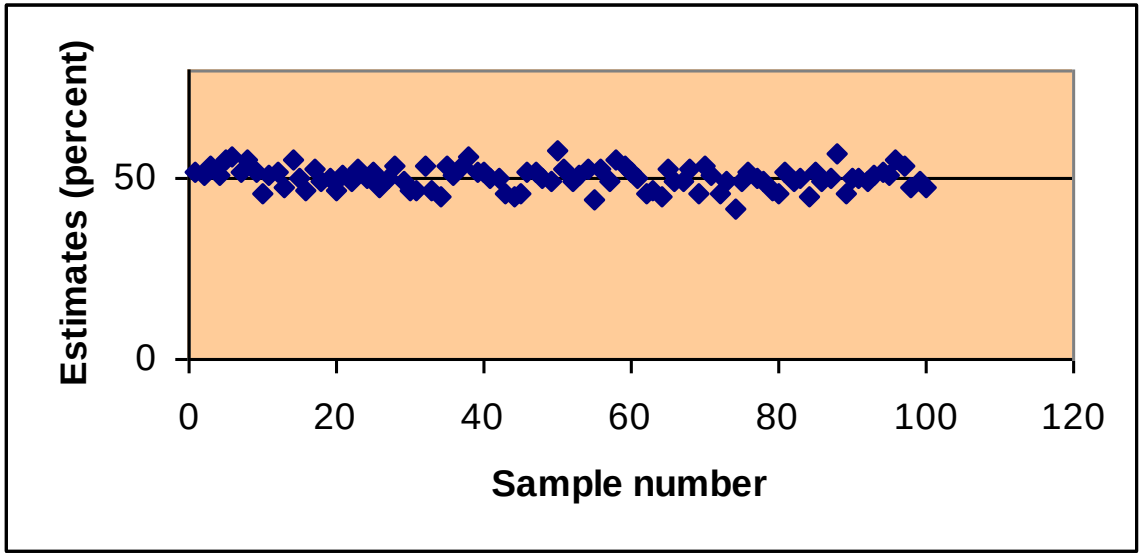


Nonresponse

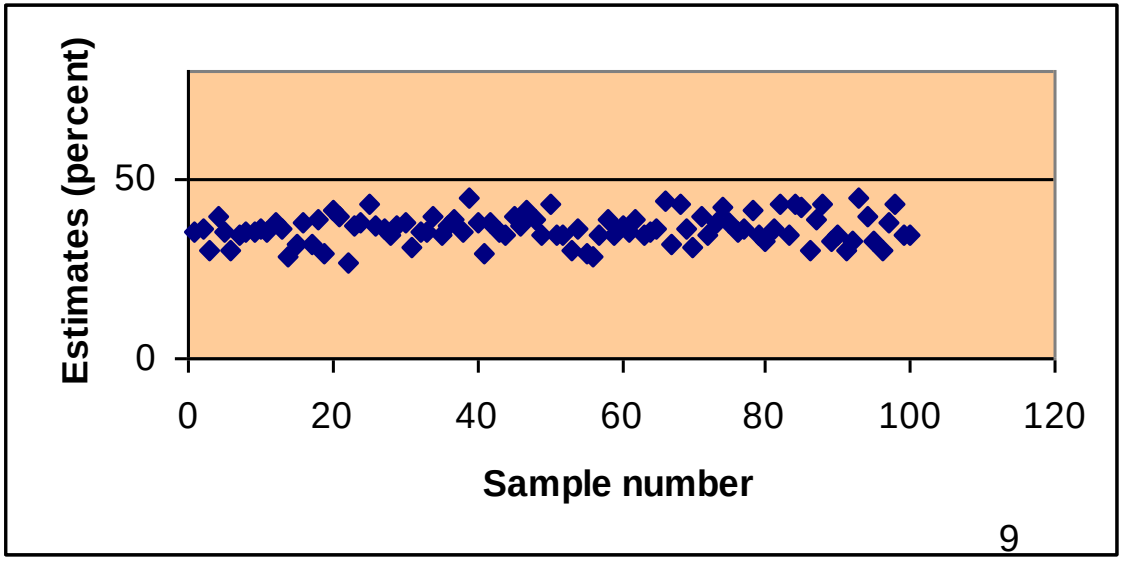


n=300

Full-response



Nonresponse



Important comments

In practice, we never know the response distribution and its response probabilities. The methods that we propose in his course for NR adjustment do not require any assumptions on the response distribution.

But when we do experiments (such as this one), or when we make evaluations of theoretical properties (expected value and variance), then the response distribution enters into consideration.

Comments

- The reasoning (here and in the whole course) is based on “long run behaviour.

- The graphs show that increased sample size will not reduce the nonresponse bias.

- But because the variance decreases, the proportion of MSE due to the bias will increase with increasing sample size, as we shall now see.

We experiment with several response distributions of the type :

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

Consider four such response distributions :

$$(1)\,\theta^* = 0.5; \quad (2)\,\theta^* = 0.85;$$

$$(3)\,\theta^* = 0.88; \quad (4)\,\theta^* = 0.89;$$

100 repeated realizations (s, r); for each of these, we compute the estimate

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

then compute the Monte Carlo estimate of

the proportion of MSE due to squared bias :

$$RelB^2 = 100 \times \frac{Bias^2}{MSE}$$

where

$$MSE = Var + Bias^2$$

$RelB^2$ for different sample sizes and resp. distrib.

θ^*	n			
	30	300	1000	2000
0.50	65.1	94.9	98.4	99.2
0.85	2.6	17.2	42.2	59.1
0.88	0.4	3.2	10.1	19.4
0.89	0.1	0.8	2.6	5.9

The proportion of MSE due to squared bias,

(i) increases with increasing sample size

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

The high proportion of MSE due to bias 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 caused by the bias:
The coverage rate does not reach
the desired 95%. (The interval is invalid.)

Coverage rate (%) as a function of sample size

for the response distribution with

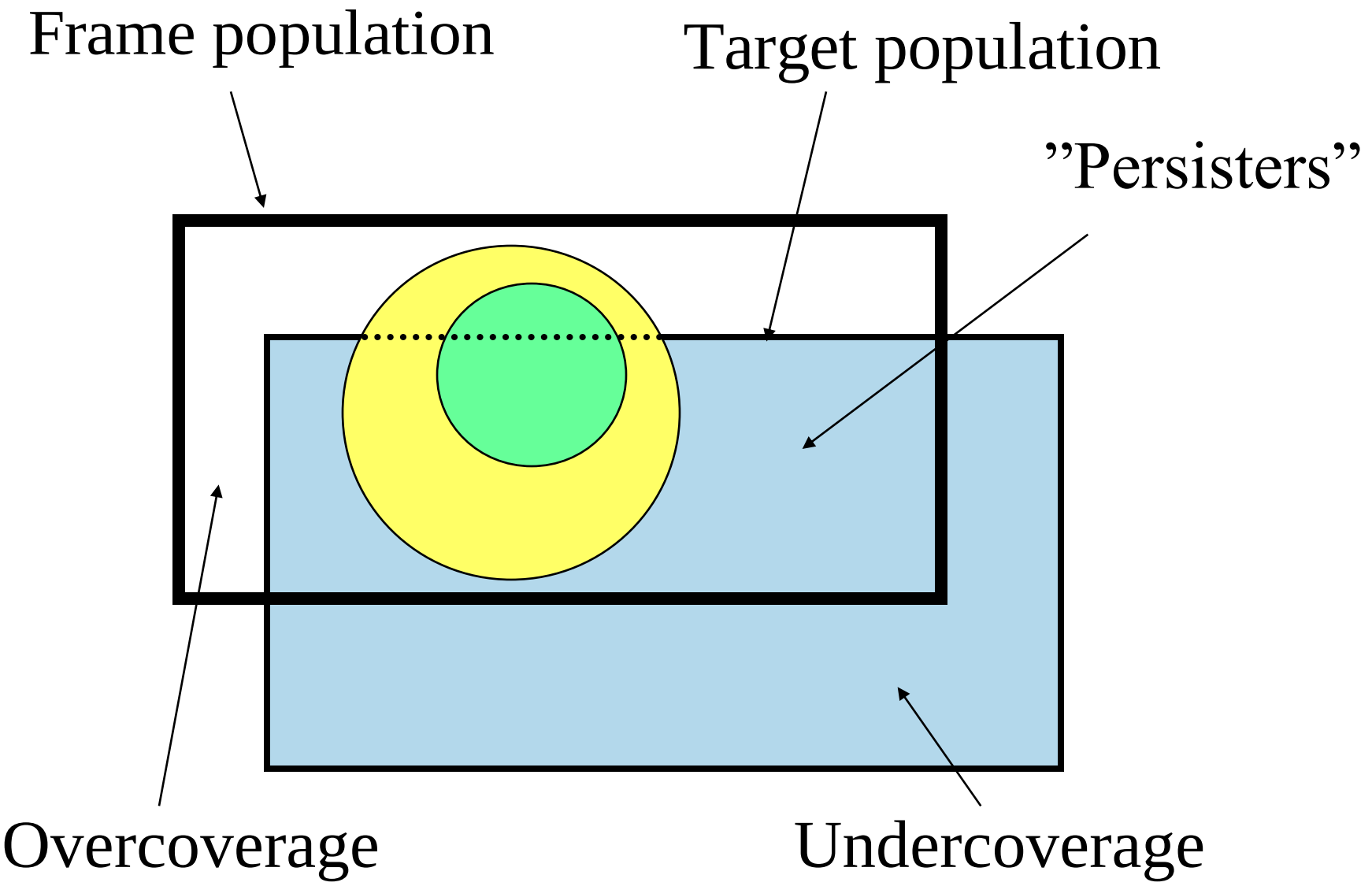
$$\theta_k = \begin{cases} 0.85 \text{ if element } k \text{ has the property} \\ 0.9 \text{ otherwise} \end{cases}$$

Sample size (n)			
30	300	1000	2000
93.2	92.6	87.1	77.9

As n increases:

- Variance is reduced
- Bias approximately the same

Sampling, nonresponse and undercoverage error



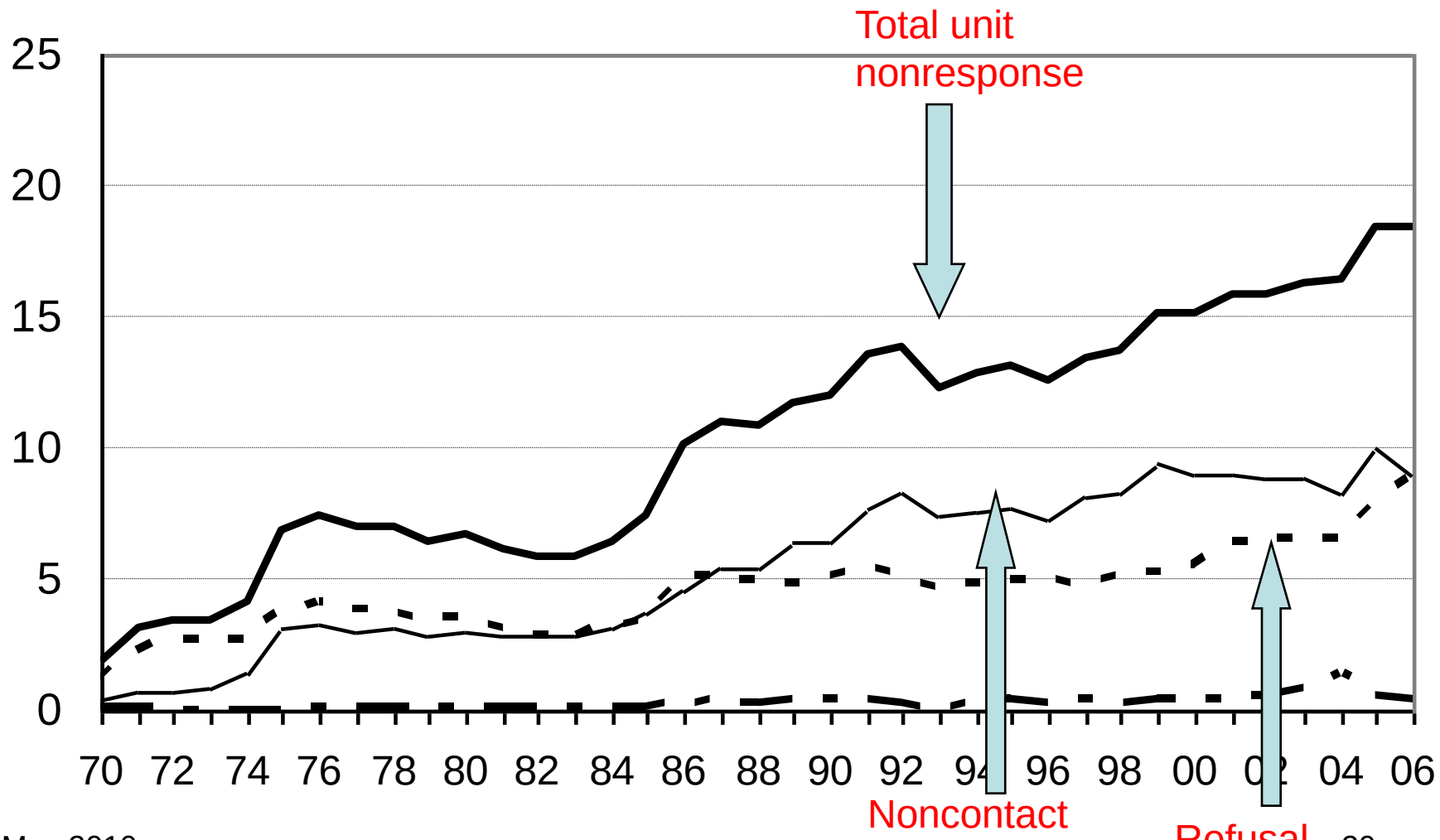
Increasing NR rates

A disturbing fact: In the past few decades, NR rates have increased dramatically in many countries.

We illustrate this by some Swedish evidence.

Consequently, NR must now always be taken into account in the estimation.

The Swedish Labour Force Survey - Time series of the nonresponse rate



International experiences

Categories that tend to have high NR :

- 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

Traditional NR analysis

The Swedish National Crime Victim and Security Study

(telephone interview survey; Statistics Sweden)

Sex	Male	Female
Response rate (%)	73.1	78.1

Age group	16-29	30-40	41-50
Response rate (%)	76.8	74.6	75.0

51-65	66-74	75-79
76.2	76.1	71.0

Country of birth	Nordic countries	Others
Response rate (%)	77.7	57.8

Marital status	Married	Others
Response rate (%)	78.3	73.6

Big cities/others	Big cities	Others
Response rate (%)	72.1	77.6

Income (in thousands of SEK)	0-149	150-299	300-
Response rate (%)	69.9	78.1	82.2

Conclusions from this nonresponse analysis:

- The response propensities vary quite a lot between groups

- Without adjustment weighting, one can expect a disturbingly large nonresponse bias

Given the results of such *nonresponse analyses*, the readers of survey reports will ask:

Can we draw *valid conclusions* from surveys with such high rates of NR and such variability in the response propensities?

In this course we try to show:

The use of (the best possible) ***auxiliary information*** (via the calibration approach) will reduce

- the nonresponse bias
- the variance
- the coverage errors

Although NR bias is reduced, it may still be questionable whether valid conclusions (confidence intervals) can be obtained.