Simplified variance estimation for multistage sampling - Application to longitudinal surveys

Guillaume Chauvet

Ensai-IRMAR

26/10/2018

1. Context of this work
2. Simplified variance estimation for two-stage sampling
3. Bootstrap variance estimation
4. Cross-sectional Estimation at Initial time
5. Cross-sectional estimation at time $t+1$: Panel and refreshment sample
6. Related/further work

Context of this work

First work : Panel Politique de la Ville

The survey "Panel Politique de la Ville" (PPV) has been set up to study the life conditions for the inhabitants in the districts of the sensitive urban areas (ZUS). The four waves of survey from 2011 to 2014 aim at focusing on :

- residential mobility inside districts,
- the perception of public policies,
- the impact of public policies on recipients.

The original sample is selected through two-stage sampling (Couvert et al., 2016) :

- drawing of a stratified sample of districts,
- inside districts, drawing of a sample of households by using a sampling frame built from the Census,
- inside households, all the individuals are surveyed.

Second work : Household Finance and Consumption Survey (HFCS)

The HCFS is conducted in a decentralised manner, though some common guidelines are provided by the HCF Network (HCFN) + coordination by the European Central bank. The methodology depends on the country : from 1 to 3 stages of sampling.

The HFCS provides detailed household-level data on different aspects of household balance sheets and associated economic and demographic variables. This includes income, employment and measures of consumption.

The priority of the HCFS is on cross-sectional estimations (i.e., at a given point in time).
The HFC Network (HFCN) also recommends the introduction of a panel component to measure individual changes over time.

Some common challenges

At initial time t, the sample is selected by means of a multistage sampling design. At further times, a split-panel strategy is used :

- tracking of a part of the units (individuals or households) surveyed at former time,
- selection of a refreshment sample of households and individuals to represent the newborns.

Purpose of my work :

- weighting for cross-sectional estimations,
- weighting for longitudinal estimations,
- assorted variance estimation :
  - linearization variance estimation for PPV,
  - bootstrap variance estimation for HCFS.

In this talk : focus on cross-sectional estimation + bootstrap

Simplified variance estimation for two-stage sampling

Two-stage sampling design

We are interested in a finite population U of units, say households. These households are grouped into larger Primary Sampling Units (PSUs), say municipalities. We note U_I the population of PSUs.

We are interested in the total

$$\begin{aligned} Y &= \sum_{k \in U} y_k \text{ for some variable of interest } y_k, \\ &= \sum_{u_i \in U_I} Y_i \text{ with } Y_i \text{ the total of } y_k \text{ on } u_i. \end{aligned}$$

A sample of households is selected by two-stage sampling :

- Sample S_I of municipalities with inclusion probabilities π_{Ii},
- Inside any $u_i \in S_I$, subsample S_i of households.

To simplify the presentation, no stratification at the first stage.

Horvitz-Thompson estimator

The Horvitz-Thompson (HT) estimator for the total Y is

$$\hat{Y}_\pi = \sum_{u_i \in S_I} \frac{\hat{Y}_i}{\pi_{Ii}} \quad \text{with} \quad \hat{Y}_i \text{ the estimator in the PSU } u_i.$$

For first-stage sampling designs with very large entropy, $\hat{Y}_\pi$ is consistent + HT variance estimator consistent (Chauvet and Vallée, 2018).

Simple variance estimators : Hajek-Rosen (HR) or with-replacement (WR)

$$\hat{V}_{HR} = \frac{n_I}{n_I - 1} \sum_{u_i \in S_I} (1 - \pi_{Ii}) \left(\frac{\hat{Y}_i}{\pi_{Ii}} - \frac{\sum_{u_j \in S_I} (1 - \pi_{Ij}) \frac{\hat{Y}_j}{\pi_{Ij}}}{\sum_{u_j \in S_I} (1 - \pi_{Ij})} \right)^2,$$

$$\hat{V}_{WR} = \frac{n_I}{n_I - 1} \sum_{u_i \in S_I} \left(\frac{\hat{Y}_i}{\pi_{Ii}} - \frac{\sum_{u_j \in S_I} \frac{\hat{Y}_j}{\pi_{Ij}}}{n_I} \right)^2.$$

HR : underestimates the variance. WR : overestimates the variance.

Advantage of these simplified variance estimators

They require only total estimators $\hat{Y}_i$ inside PSUs :

- no need to produce unbiased variance estimators $\hat{V}_i$ inside PSUs,
- can be applied with any type of second-stage sampling designs (e.g., systematic sampling),
- consistent when $n_I/N_I \to 0$ (Chauvet and Vallée, 2018).

They can be applied if we may (approximatly) write the estimator as a sum of independent random variables :

$$\hat{Y}_\pi = \sum_{u_i \in S_I} \frac{\hat{Y}_i}{\pi_{Ii}}.$$

Important to catch the treatment of unit non-response.

The second variance estimator may be easily bootstrapped.

Bootstrap variance estimation

Computation of bootstrap weights

In case of the Household Finance and Consumption Survey (HFCS), a bootstrap variance estimator is needed.

Bootstrap = computational technique which enables to produce variance estimations by repeatedly mimicking the sampling process + estimation steps.

Proposed bootstrap = bootstrap of Primary Sampling Units (PSUs) :

- does not need to bootstrap the sample selection inside the PSUs $\Rightarrow$ fairly simple to implement,
- overestimation of the 1st-stage sampling variance $\Rightarrow$ conservative,
- variance of further stages (e.g., non-response) correctly accounted for.

Computation of bootstrap weights

The bootstrap weights are obtained as follows :

1. Selection of a with-replacement sample of $n_I - 1$ PSUs in the original sample, with equal probabilities (**Bootstrap of PSUs**).
   The bootstrap weight of the PSU u_i is :

$$W_{Ii} = \frac{n_I}{n_I - 1} \times \text{Number of selections of } u_i.$$

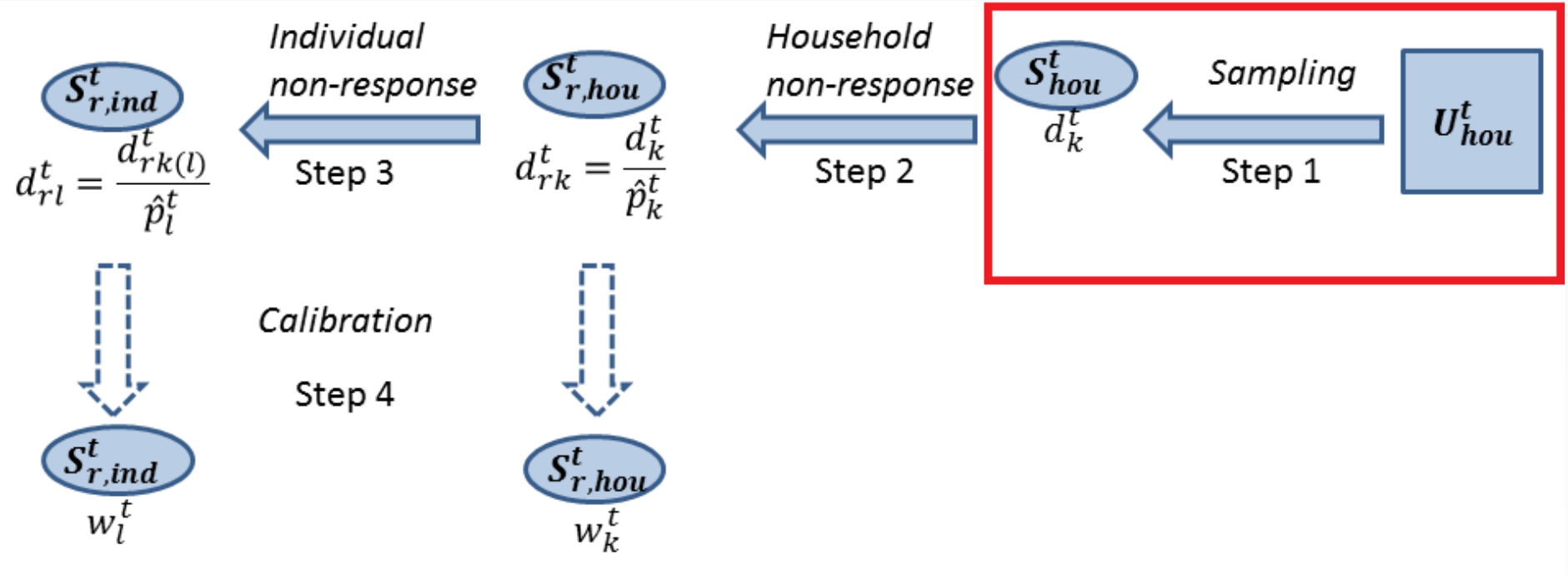
2. The bootstrap sampling weight of some household $k \in u_i$ is :

$$d_{k*} = d_k \times W_{Ii}.$$

3. These bootstrap sampling weights are adjusted for all the following estimation steps :
   - Household non-response,
   - Individual non-response,
   - Calibration.
4. Steps 1-3 are repeated a large number of times (say $B = 500$)

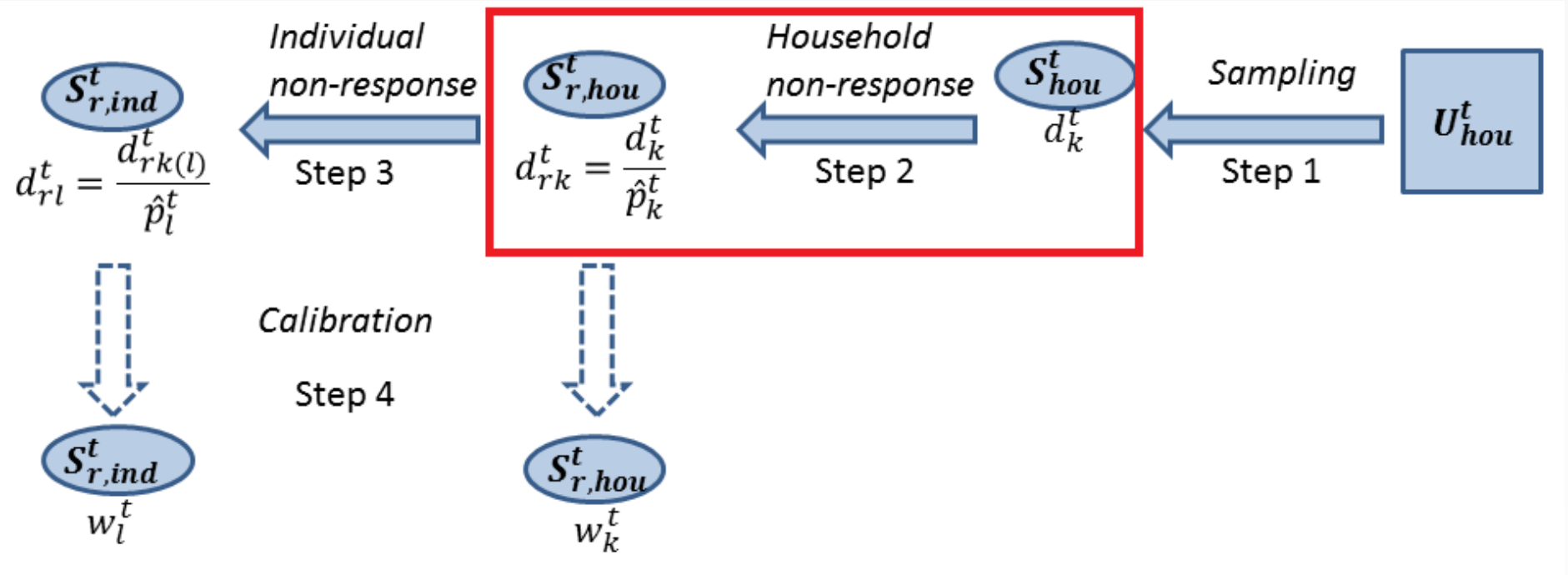
Cross-sectional Estimation Initial time t

Sampling and estimation process



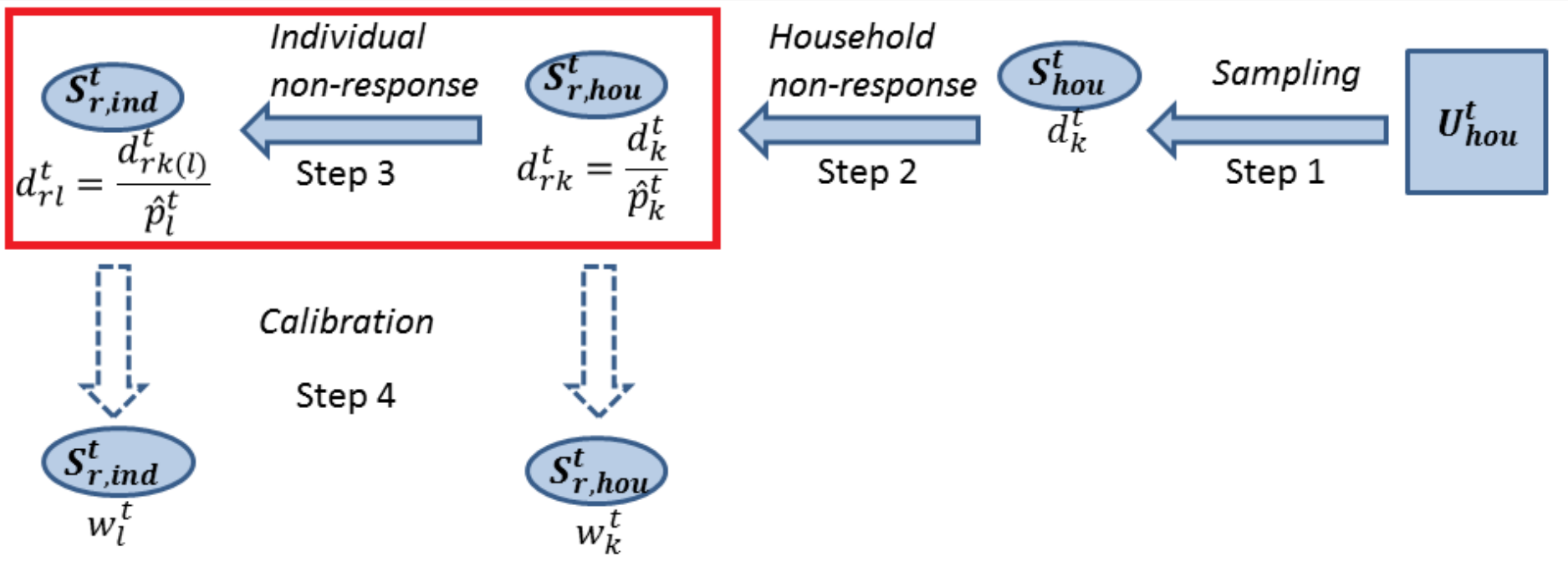
Sample of households selected by multistage sampling : sample of n_I PSUs (e.g., municipalities) and subsampling of households within.
We obtain a sample of households S^t_{hou} with sampling weights d^t_k.

Sampling and estimation process



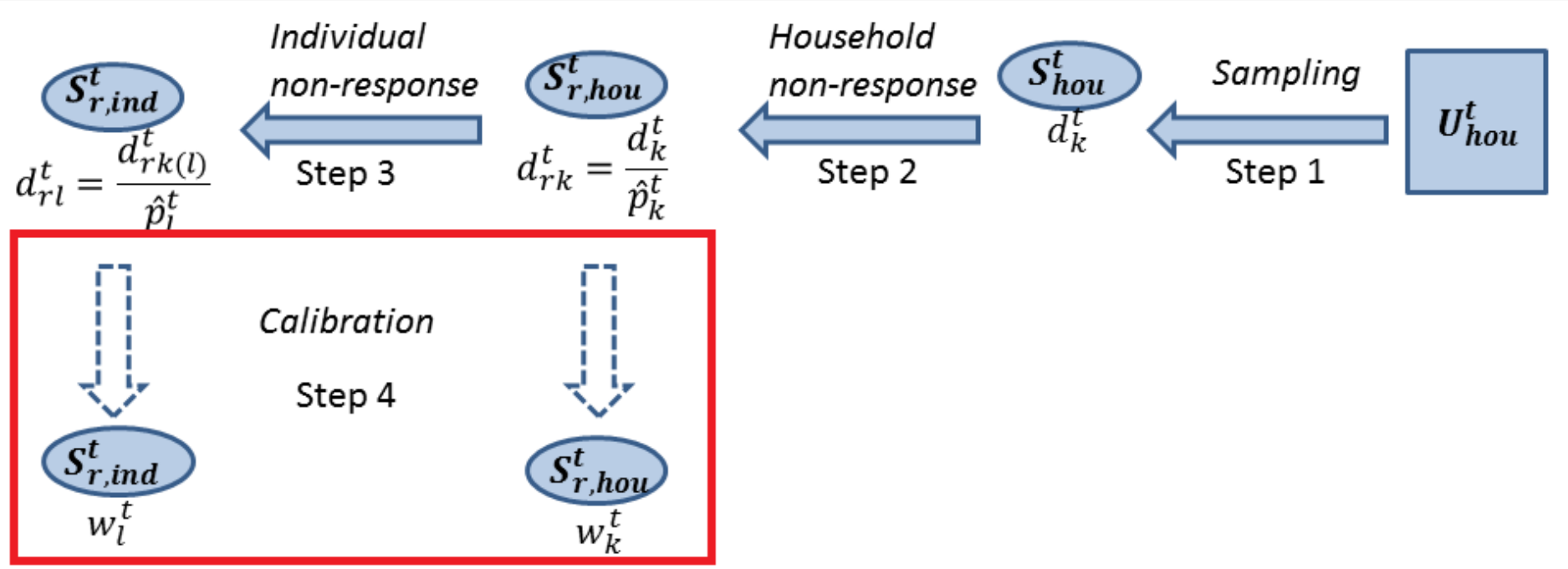
We account for the non-response of households, e.g. via Response Homogeneity Groups (RHGs).
Estimation of the response probability $\hat{p}^t_k \Rightarrow$ household weights d^t_{rk}.

Sampling and estimation process



We account for the non-response of individuals, e.g. via Response Homogeneity Groups (RHGs).
Estimation of the response probability $\hat{p}^t_l$ $\Rightarrow$ individual weights d^t_{rl}.

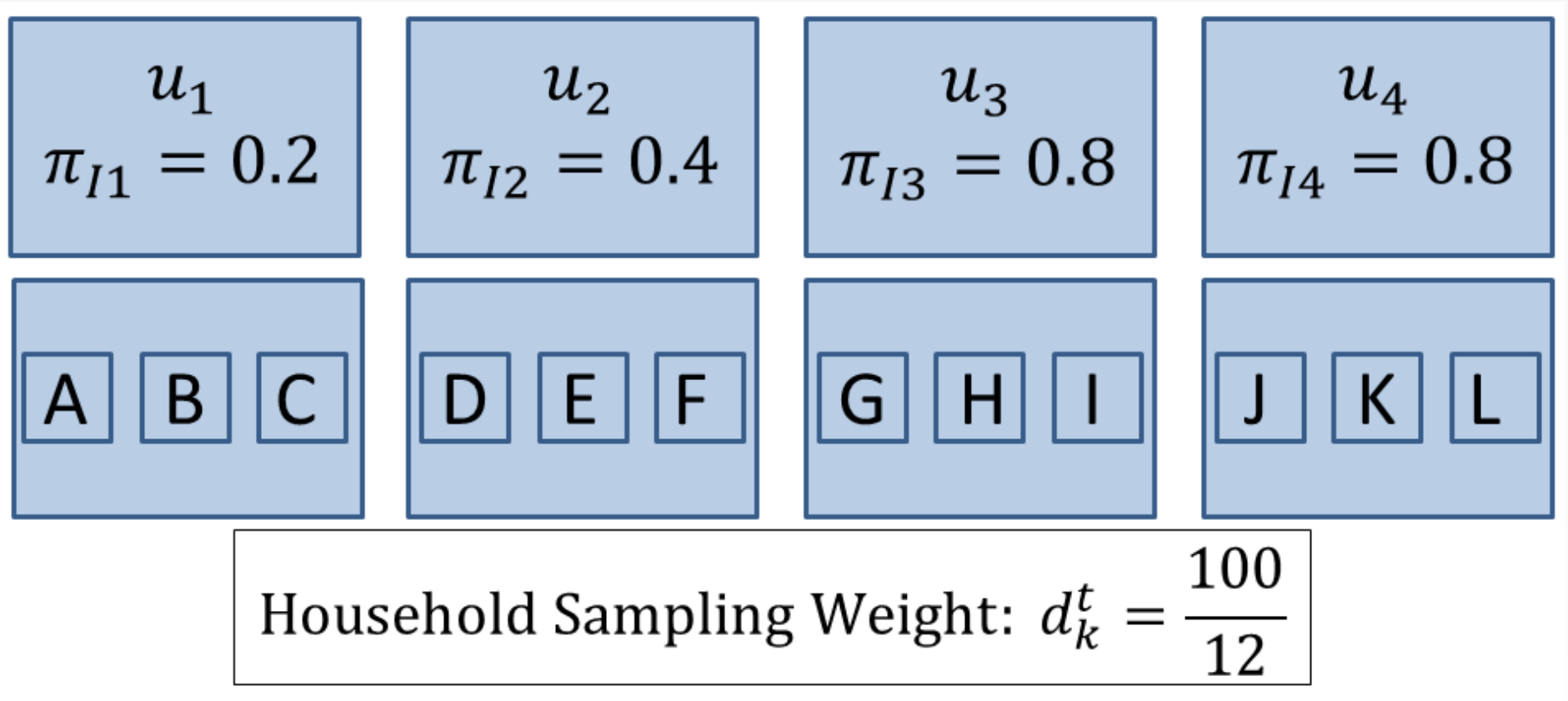
Sampling and estimation process



The weights adjusted for non-response are calibrated on auxiliary information on households and individuals.
Leads to calibrated weights w_k^t for households and w_l^t for individuals.

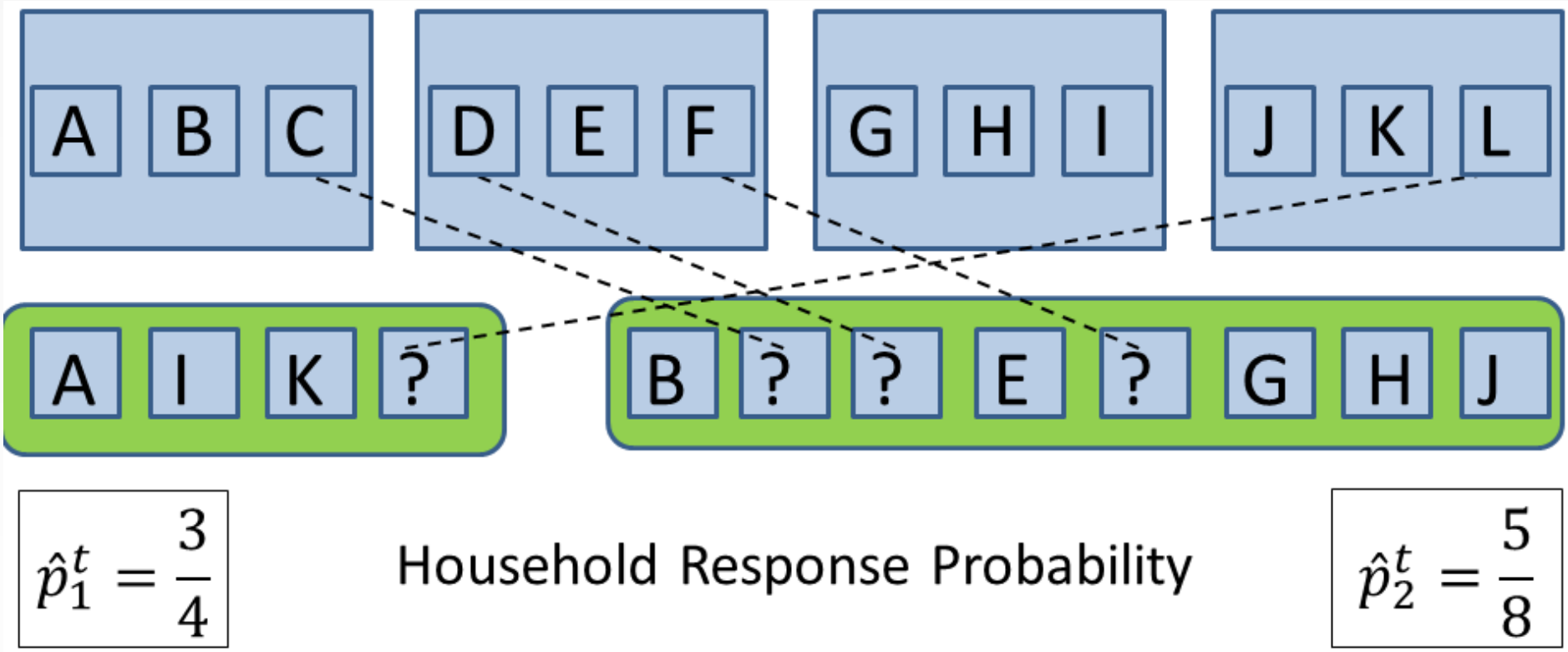
An example with two-stage sampling

Selection of the households



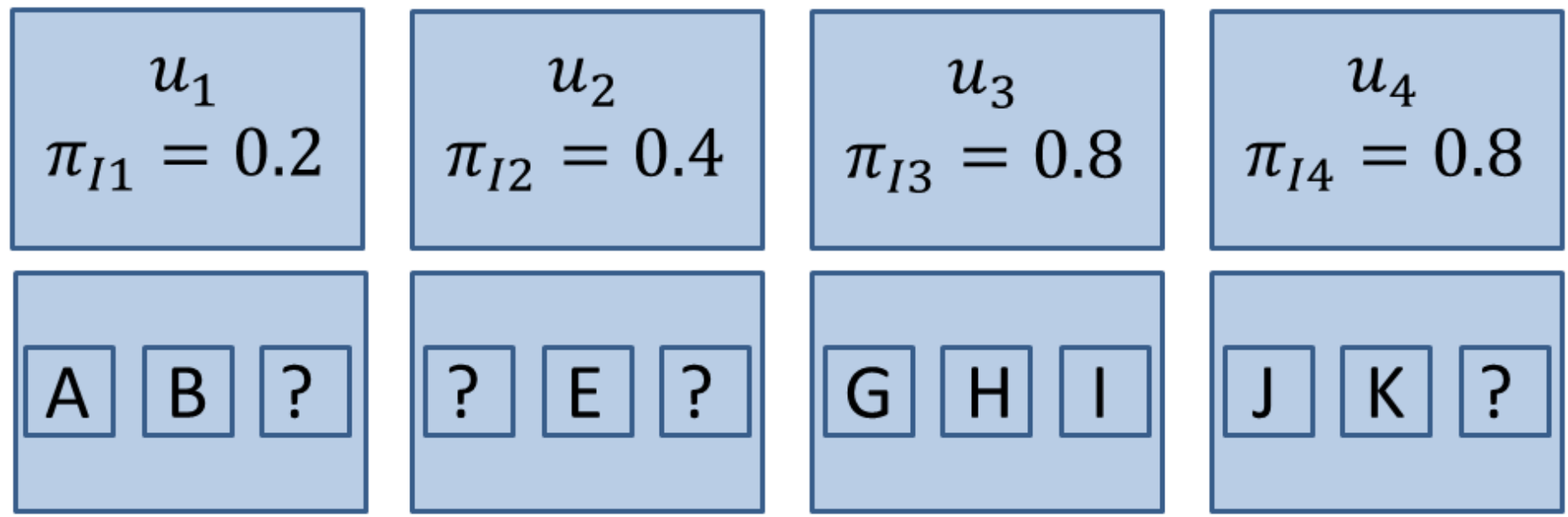
Sample of $n_I = 4$ PSUs selected by a self-weighted two-stage sampling.
We obtain equal sampling weights d_k for the households.

Correction of household non-response



Non-response of households corrected via 2 response Homogeneity Groups. Response probabilities estimated by the weighted response frequency inside a RHG.

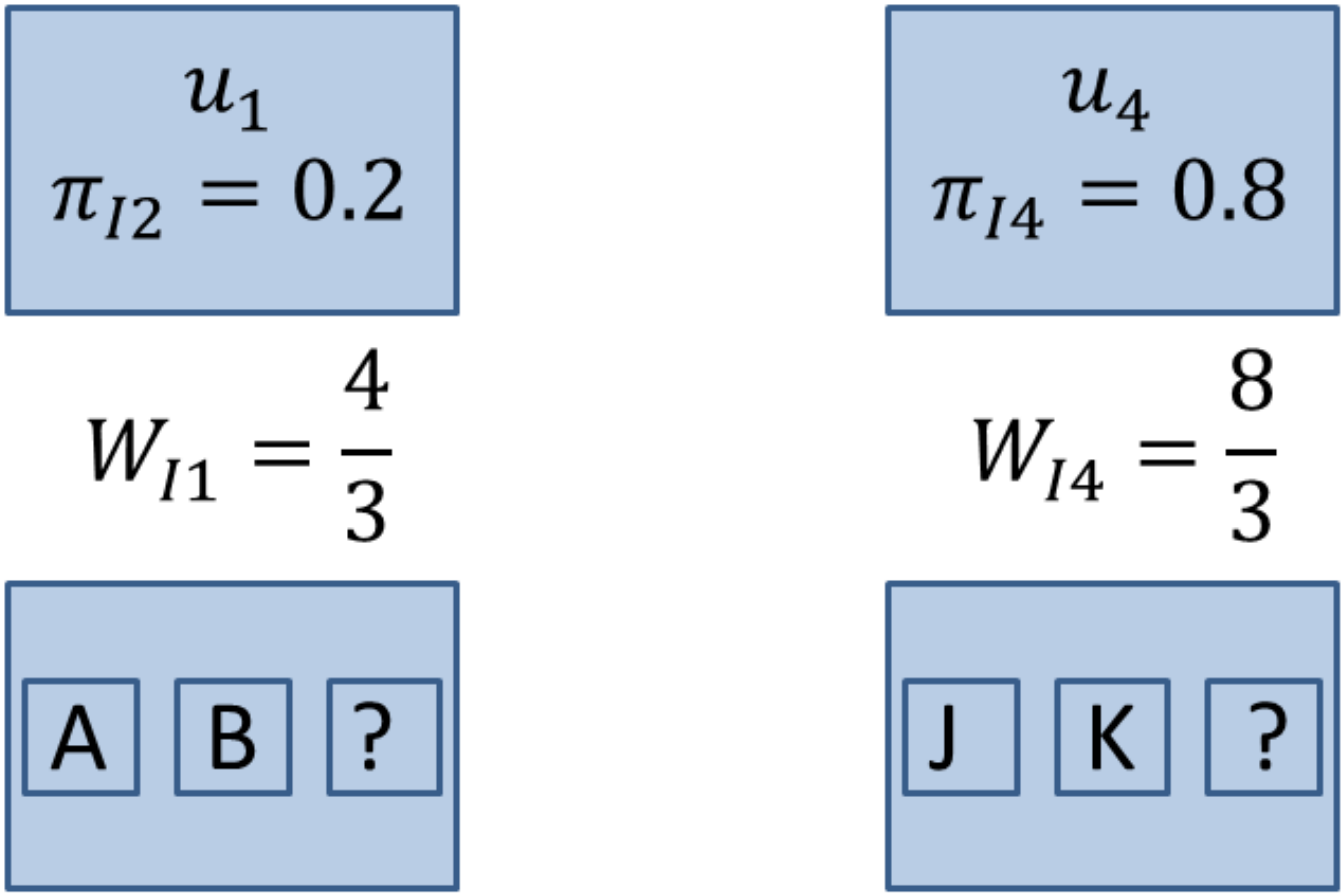
Calibration



$$\text{Household Adjusted weights } d_{rk}^t = \frac{d_k^t}{\hat{p}_k^t} \Rightarrow \text{Calibrated weights } w_k^t$$

The sampling weights are corrected to account for non-response.
The weights are finally calibrated.

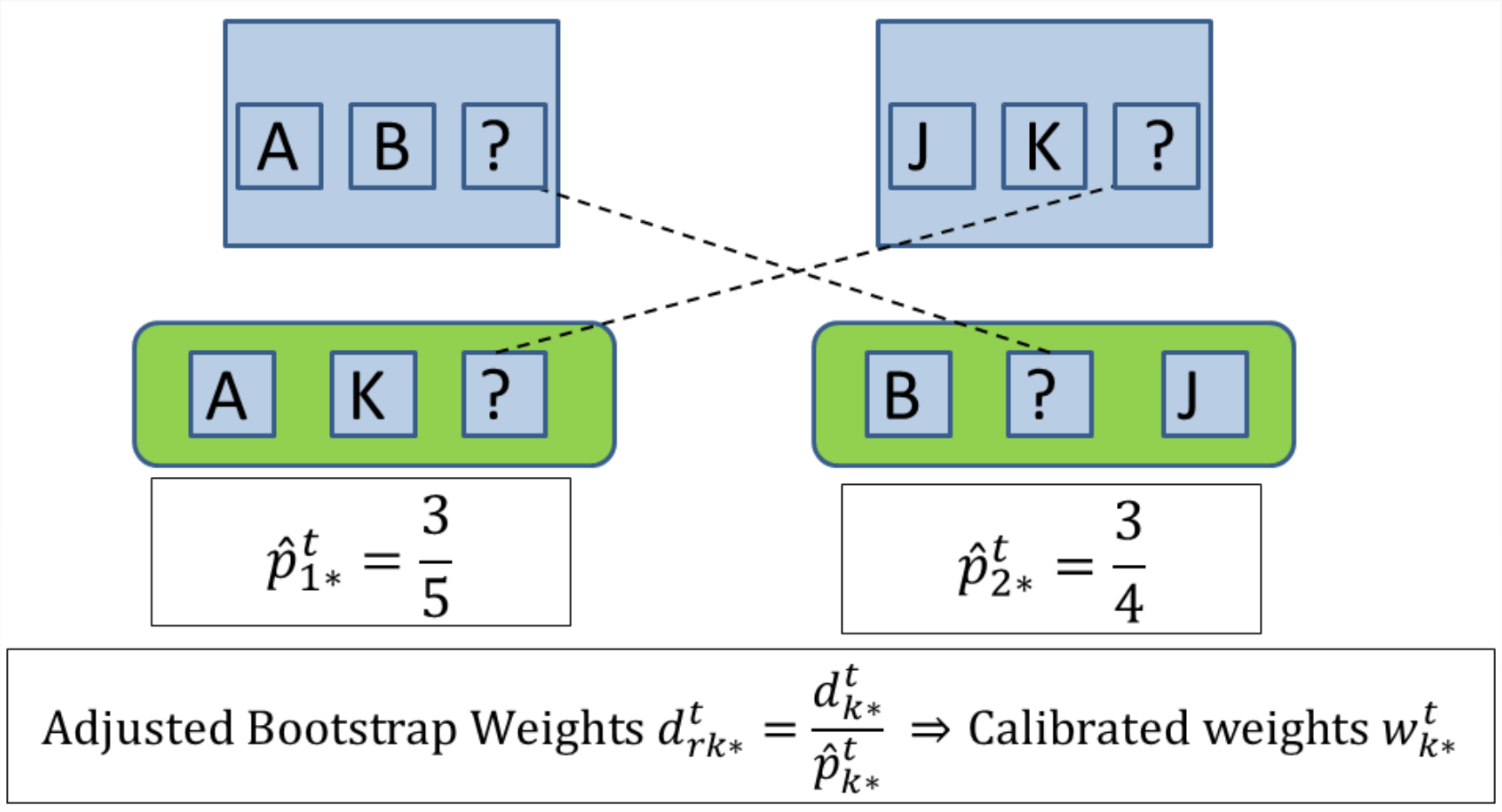
Variance estimation : bootstrapping of the PSUs



Bootstrap Household Sampling Weights $d^t_{k*} = W_{Ii} \times d^t_k$

A resample of $n_I - 1 = 3$ PSUs is selected with equal probabilities.
The resampling weight of a PSU is $4/3$ times its number of selections.

Variance estimation : bootstrapping of non-response and calibration

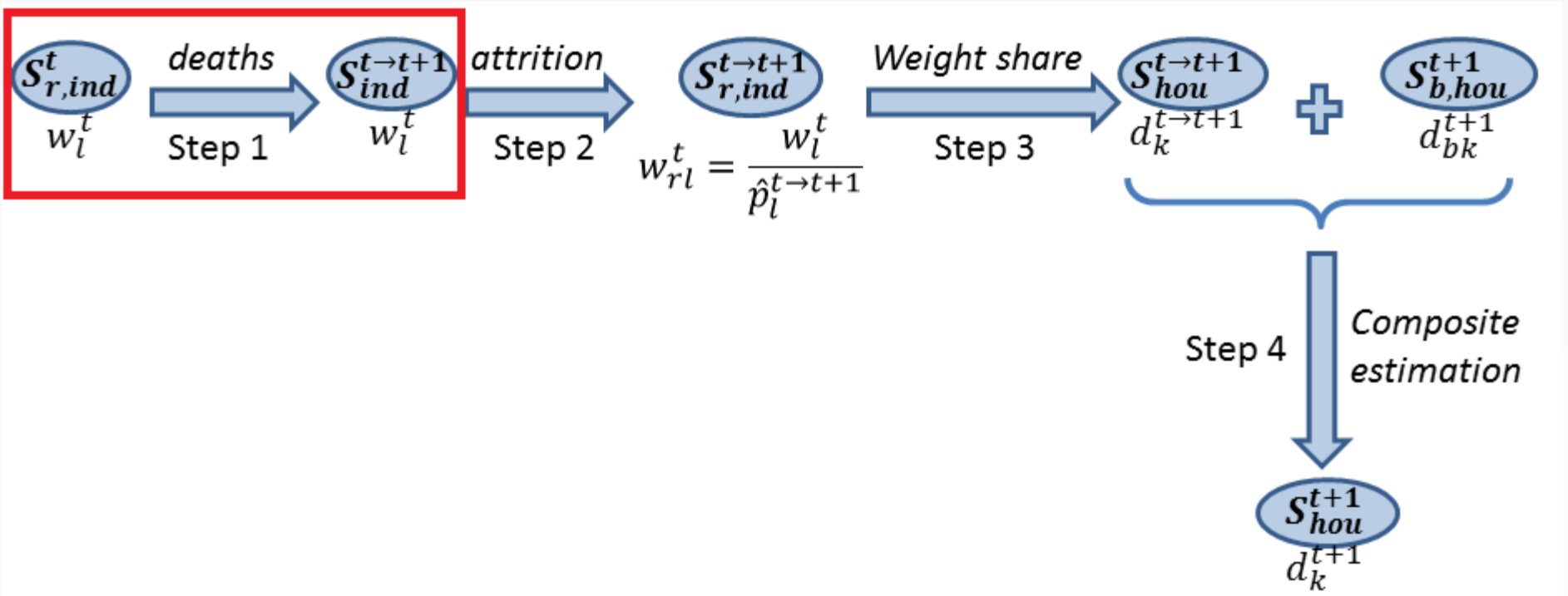


The bootstrap sampling weights are corrected for non-response as originally.
The weights are finally calibrated as originally.

Cross-sectional estimation at time $t+1$ Panel and refreshment sample

Cross-sectional Estimation : panel + refreshment sample

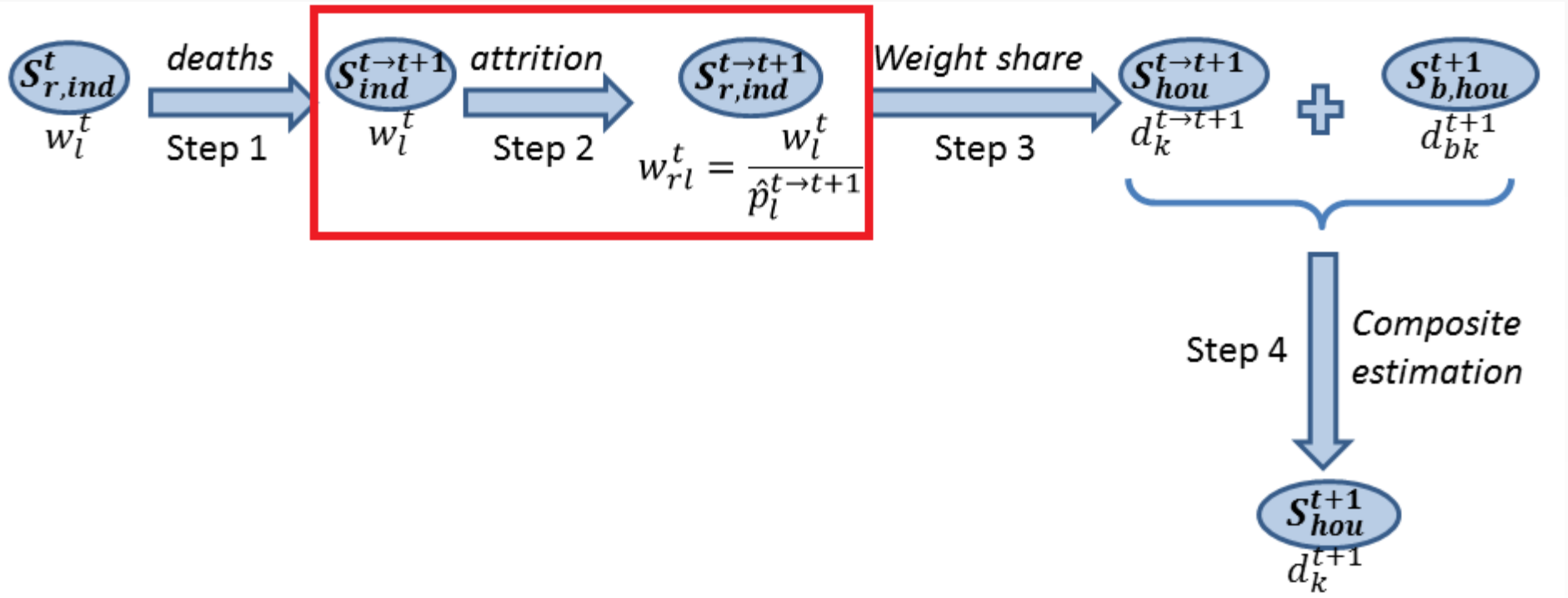
Tracking individuals : elimination of the leaving dead



We suppose that individuals are followed over time, with weights w_l^t.
We account for deaths between time t and $t+1$. The weights remain unchanged

Cross-sectional Estimation : panel + refreshment sample

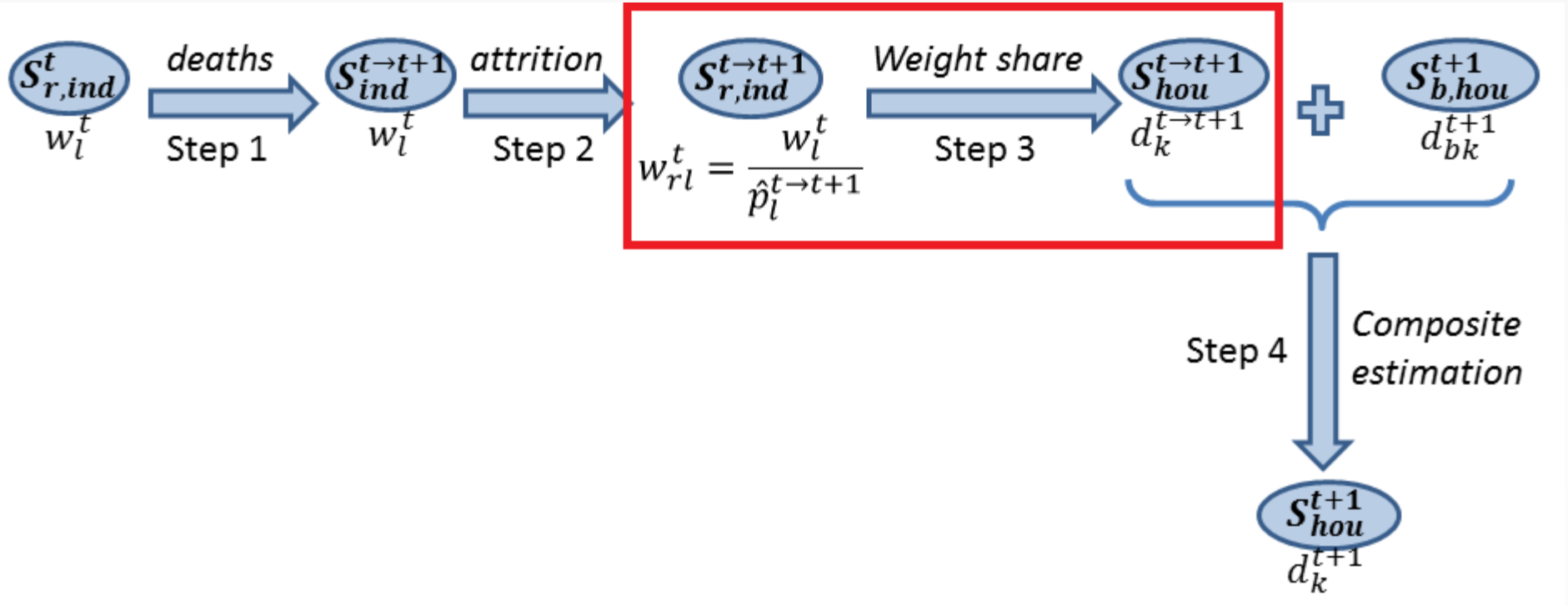
Tracking individuals : correction of attrition



We account for attrition, e.g. via Response Homogeneity Groups (RHGs).
Estimation of the response probability $\hat{p}_k^{t\to t+1}$ $\Rightarrow$ individual weights w_{rl}^t.

Cross-sectional Estimation : panel + refreshment sample

From individual to household : Weight Share Method

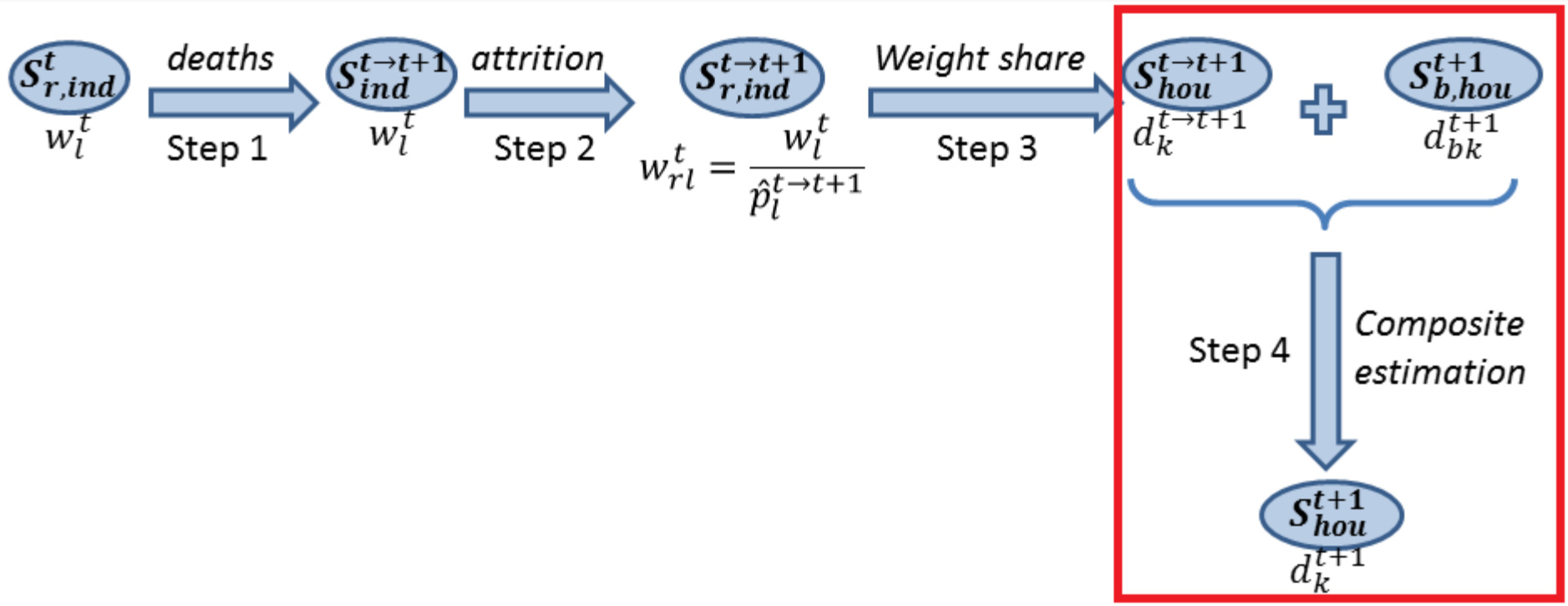


The individual weights are distributed on their household via the Weight Share Method.

This leads to the household weights $d_k^{t\to t+1}$.

Cross-sectional Estimation : panel + refreshment sample

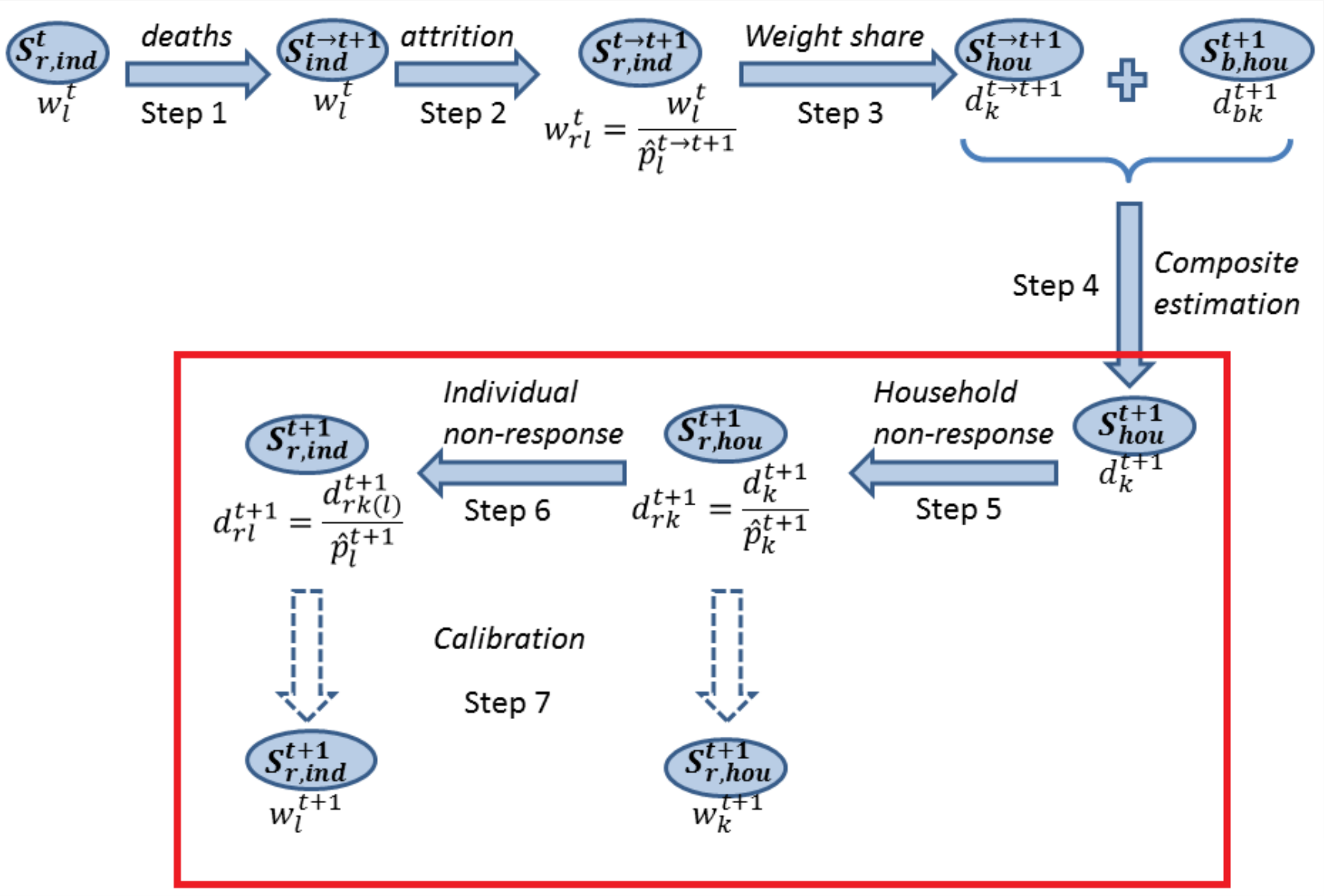
Merging the panel and the refreshment sample



The panel sample and the refreshment sample are joined, by using composite estimation.

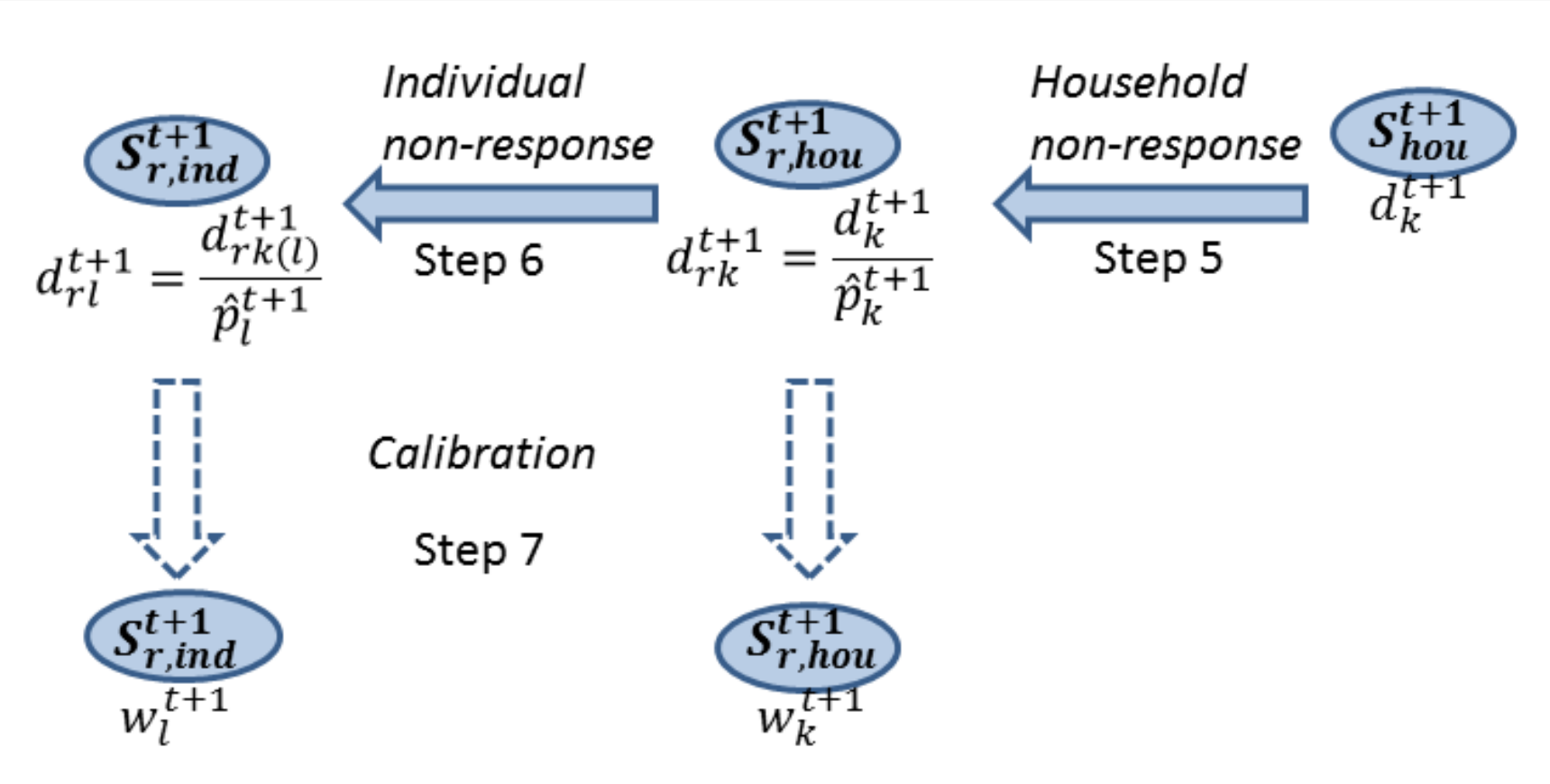
This leads to the household weights d_k^{t+1}.

Cross-sectional Estimation with a panel component



The estimations steps are then identical to that at previous time t.
Accounting for non-response of households + individuals and calibration.

Cross-sectional Estimation with a panel component



Bootstrap weights computed as if S_{hou}^{t+1} was directly selected at time $t+1$.
Expected to overestimate slightly the variance due to attrition and the variance due to household non-response at time t.
Other variance components correctly accounted for.

Related/further work

Related/further work

Longitudinal estimation can be handled similarly (slightly easier).

Need for a software : macro SAS in progress.
Interested for **applications to other surveys**.

Theoretical justification : works (fairly) well in simulations, but the properties of inverse probability weighted estimators used in case of unit non-response are more complicated to handle.


Des questions?

A small simulation study

Simulated data

Original sampling design of the PPV survey :

- stratified sample of districts (PSUs),
- inside districts, sample of households (SSUs).

Responding households and districts duplicated by inverse sampling rates $\Rightarrow$ artificial but realistic population :

- 433 districts, ranging from 600 to $23\ 000$ households,
- population of 1.1 Mo households.

Variables of interest :

- `reputation` of district (good or very good/other),
- witness of `trafficking` (never or rarely/other),
- urban `works` during last year (yes/no),
- `move` planned during year to come (yes/no).

Simulated sampling design

Unstratified self-weighted two-stage design :

- Inclusion probabilities π_{Ii} proportional to size ($n_I = 20$ or $n_I = 40$),
  $\Rightarrow$ with $n_I = 20$, π_{Ii} ranging from 0.01 to 0.42 (mean 0.05).
- Sample of $n_0 = 50$ households inside each selected district.

6 Response Homogeneity groups obtained by crossing :

- low-income housing (yes/no),
- region (Ile-de-France/north/south).

Average response probability : 74%.

Sampling design + non-response repeated $B = 5\ 000$ times to obtain a Monte Carlo variance approximation.
Simplified variance estimators computed $C = 500$ times.

Relative bias of the simplified variance estimators

	$\hat{V}_{SIMP1}$		$\hat{V}_{SIMP2}$	
	$n_I = 20$	$n_I = 40$	$n_I = 20$	$n_I = 40$
reputation	-5.7 %	-3.2 %	+3.8 %	+18.2 %
trafficking	-4.3 %	-0.9 %	+6.3 %	+23.5 %
works	+0.4 %	-3.6 %	+7.2 %	+10.4 %
move	-3.8 %	-9.8 %	+6.5 %	+12.3 %

Relative bias of the simplified variance estimators

	$\hat{V}_{SIMP1}$		$\hat{V}_{SIMP2}$	
	$n_I = 20$	$n_I = 40$	$n_I = 20$	$n_I = 40$
reputation	-5.7 %	-3.2 %	+3.8 %	+18.2 %
trafficking	-4.3 %	-0.9 %	+6.3 %	+23.5 %
works	+0.4 %	-3.6 %	+7.2 %	+10.4 %
move	-3.8 %	-9.8 %	+6.5 %	+12.3 %